

Desarrollo de una estrategia para promover

la plataforma de IA en el gobierno público

**Estándares Éticos de la IA para el Gobierno Digital**

**de Perú**

2024.12

**Índice**

[1. Generalidades 5](#_Toc191313189)

[1.1. Antecedentes 5](#_Toc191313190)

[1.2. ¿Qué es la ética de la inteligencia artificial? 6](#_Toc191313191)

[1.3. Objetivo de los estándares éticos 7](#_Toc191313192)

[1.4. Aplicación de estándares éticos 8](#_Toc191313193)

[2. Principios básicos 10](#_Toc191313194)

[2.1. Principios de la dignidad humana 10](#_Toc191313195)

[2.2. Principio de equidad 11](#_Toc191313196)

[2.3. Principio de transparencia 12](#_Toc191313197)

[2.4. Principio de robustez 13](#_Toc191313198)

[2.5. Principio de responsabilidad 14](#_Toc191313199)

[2.6. Conflicto entre los principios 15](#_Toc191313200)

[3. Requisitos clave para el cumplimiento de los principios 18](#_Toc191313201)

[3.1. Protección de los derechos humanos 18](#_Toc191313202)

[3.2. Protección de la información personal y los datos 20](#_Toc191313203)

[3.3. Promoción de la publicidad y el valor social 21](#_Toc191313204)

[3.4. Fortalecimiento de la accesibilidad y la inclusión social 23](#_Toc191313205)

[3.5. Garantía de la explicabilidad 25](#_Toc191313206)

[3.6. Garantía de la confianza y la seguridad técnica 27](#_Toc191313207)

[3.7. Desarrollo y operaciones de la IA sostenible 29](#_Toc191313208)

[3.8. Cooperación internacional 30](#_Toc191313209)

[4. Ítem de verificación por requisitos básicos 33](#_Toc191313210)

[4.1. Ítem de verificación para la protección de los derechos humanos 33](#_Toc191313211)

[4.2. Ítem de verificación para la protección de información personal y datos 35](#_Toc191313212)

[4.3. Ítem de verificación para mejorar la publicidad y el valor social 37](#_Toc191313213)

[4.4. Ítem de verificación de accesibilidad e inclusión social 40](#_Toc191313214)

[4.5. Ítem de verificación para garantizar la explicabilidad 43](#_Toc191313215)

[4.6. Ítem de verificación para el aseguramiento de la fiabilidad y seguridad técnica 45](#_Toc191313216)

[4.7. Ítem de verificación para el desarrollo y funcionamiento de la IA sostenible 48](#_Toc191313217)

[4.8. Ítem de verificación para la cooperación internacional 51](#_Toc191313218)

[5. Anexo 54](#_Toc191313219)

[5.1. Tabla de ítem de verificación 54](#_Toc191313220)

[5.2. Glosario de IA 56](#_Toc191313221)

**Índice de tablas**

[<Tabla 1 – Ítem de verificación para la protección de los derechos humanos> 33](#_Toc191313263)

[<Tabla 2 - Cuestionario de evaluación del impacto sobre los derechos fundamentales (FRIA) de ‘ALTAI’ de la UE> 34](#_Toc191313264)

[<Tabla 3 – Directrices de la OMS para el uso ético de la IA> 35](#_Toc191313265)

[<Tabla 4 – Ítem de verificación para la protección de información personal y datos> 35](#_Toc191313266)

[<Tabla 5 – Cuestionario de evaluación de IA de ‘ALTAI’ de la UE> 37](#_Toc191313267)

[<Tabla 6 – Directrices de gestión de la calidad de los datos para el aprendizaje de la IA> 37](#_Toc191313268)

[<Tabla 7 – Ítem de verificación para mejorar la publicidad y el valor social> 38](#_Toc191313269)

[<Tabla 8 – Cuestionario similar en herramientas de autoevaluación en el extranjero que representa la publicidad> 39](#_Toc191313270)

[<Tabla 9 – “Directrices éticas para una inteligencia artificial fiable” de la Unión Europea (UE)> 39](#_Toc191313271)

[<Tabla 10 – “Principios de la ética de la IA” de Australia> 40](#_Toc191313272)

[<Tabla 11 – Ítem de verificación de accesibilidad e inclusión social> 41](#_Toc191313273)

[<Tabla 12 – Principios de IA de Google> 42](#_Toc191313274)

[<Tabla 13 – Directrices sobre el uso de IA y algoritmos de la Comisión Federal de Comercio de EEUU (FTC por sus siglas en inglés)> 43](#_Toc191313275)

[<Tabla 14 – Caso de inclusión social > 43](#_Toc191313276)

[<Tabla 15 – Ítem de verificación para garantizar la explicabilidad> 44](#_Toc191313277)

[<Tabla 16 – Algoritmo de “caja negra”> 45](#_Toc191313278)

[<Tabla 17 – La IA explicable> 45](#_Toc191313279)

[<Tabla 18 – Cuestiones relacionadas con el principio de transparencia> 45](#_Toc191313280)

[<Tabla 19 – Ítem de verificación para el aseguramiento de la fiabilidad y seguridad técnica> 46](#_Toc191313281)

[<Tabla 20 – Establecimiento de vulnerabilidades de seguridad y contramedidas para sistemas de IA> 47](#_Toc191313282)

[<Tabla 21 – Casos relacionados con la ‘seguridad’> 48](#_Toc191313283)

[<Tabla 22 – Ítem de verificación para el desarrollo y funcionamiento de la IA sostenible> 48](#_Toc191313284)

[<Tabla 23 – Cuestionario de la evaluación de IA de ‘ALTAI’ de UE> 50](#_Toc191313285)

[<Tabla 24 – Diseño Éticamente Alineado del IEEE> 50](#_Toc191313286)

[<Tabla 25 – Cuestiones de dónde recae la responsabilidad > 51](#_Toc191313287)

[<Tabla 26 – Cuestiones ambientales> 51](#_Toc191313288)

[<Tabla 27 – Ítem de verificación para la cooperación internacional> 52](#_Toc191313289)

[<Tabla 28 – Ítem de verificación> 56](#_Toc191313290)

**Documento de presentación**

Estimado presidente de la Agencia Nacional de la Sociedad de la Información:

Le presentamos este informe como “Estándares Éticos de la IA para el Gobierno Digital del Perú” del “Proyecto de Cooperación Conjunta DGCC Corea-Perú de 2024”.

Diciembre de 2024

Director de investigación conjunta:

Jae-ho Lee (Director General de Cooperación Global de la Agencia Nacional de la Sociedad de la Información Inteligente de Corea)

Encargado de la tarea:

Jung-won Yoo (Jefe del Equipo de Cooperación en Gobierno Digital de la Agencia Nacional de la Sociedad de la Información Inteligente de Corea)

Daul Shin (Director del Centro Corea-Perú DGCC de la Agencia Nacional de la Sociedad de la Información Inteligente de Corea)

Director de investigación conjunta:

Jiyoon Choi (Presidente de Korea IT Consulting Co., Ltd.)

Investigadores participantes:

Kyu-hyuk Chae (Director Ejecutivo de Korea IT Consulting Co., Ltd.)

Jung-wan Park (Consultor Principal de Korea IT Consulting Co., Ltd.)

Seung-heon Lee (Consultor Principal de Korea IT Consulting Co., Ltd.)

Hyo-jae Choi (Consultor Responsable de Korea IT Consulting Co., Ltd.)

1. Generalidades
	1. Antecedentes

La tecnología de inteligencia artificial (IA) se está desarrollando rápidamente y está liderando la innovación en diversos campos. Perú también está adoptando activamente la IA para promover la transformación digital y la innovación en los servicios públicos, y está realizando esfuerzos para expandir el uso de la IA en diversas áreas, como aumentar la eficiencia administrativa, mejorar los servicios médicos, innovar en la educación y optimizar el transporte y la gestión urbana. Se espera que el avance de la IA contribuya a aumentar la productividad y la competitividad del país y a mejorar la calidad de vida de los ciudadanos.

Sin embargo, junto con las oportunidades que trae la tecnología de IA, también están surgiendo desafíos éticos y sociales como sesgo de datos, discriminación, cuestión de protección de información personal, cambios de puestos de trabajo derivados de la automatización, mal uso de la tecnología y amenazas a la seguridad. En particular, en los países donde coexisten diversas clases sociales y económicas, es esencial que las tecnologías de IA se diseñen y operen de manera que no exacerben la desigualdad social. Es importante garantizar la equidad, la transparencia, la seguridad y la explicabilidad en el proceso de toma de decisiones de los sistemas de IA, y establecer una gobernanza de la IA en la que los ciudadanos puedan confiar.

En consecuencia, los estándares éticos de la IA para el gobierno digital del Perú presentan los principios éticos básicos y los estándares de práctica que las instituciones públicas, el sector privado, las instituciones de investigación y la sociedad civil deben respetar durante todo el proceso de desarrollo y utilización de la IA. Estos estándares éticos proporcionan orientación para garantizar que la IA proteja los derechos humanos, promueva los valores sociales y opere de manera sostenible y responsable. Con ello pretendemos establecer una base normativa equilibrada para que las tecnologías de IA puedan promover la innovación respetando los derechos fundamentales de los ciudadanos y utilizarse como una herramienta para el bien público.

Los estándares éticos de la IA para el gobierno digital del Perú establecen la responsabilidad social y los estándares éticos para la IA, mediando así en los conflictos éticos que puedan surgir durante el desarrollo y la operación de la IA y generando confianza entre las partes interesadas. Asimismo, proporciona orientación práctica que refleja el entorno legal e institucional y el contexto social del Perú, manteniendo al mismo tiempo la coherencia con los principios éticos internacionales de la IA.

Los estándares éticos de la IA no deberían ser normas fijas, sino que deberían evolucionar continuamente en respuesta a los cambios tecnológicos y sociales. Para ello, estos estándares éticos se complementarán y revisarán continuamente en cooperación con diversas partes interesadas para garantizar el uso ético y público de la IA.

Las políticas y la visión del Perú se centran en cómo el uso de las tecnologías digitales genera confianza social, cómo las innovaciones tecnológicas brindan beneficios reales a todos los ciudadanos y cómo logran el desarrollo sostenible. Los esfuerzos para fortalecer la participación y la responsabilidad de los ciudadanos; y ampliar la accesibilidad a los servicios públicos reduciendo la brecha digital a través del uso ético de la tecnología de IA, permitirán la IA que se convierta en un elemento clave en la creación de valor social más allá de una mera herramienta técnica.

* 1. ¿Qué es la ética de la inteligencia artificial?

La ética de la inteligencia artificial (IA) se refiere a un sistema de principios y valores que garantizan que los sistemas de IA se desarrollen, implementen y operen de manera centrada en el ser humano y socialmente responsable. La IA puede afectar una variedad de actividades sociales y económicas a través de la recopilación de datos, el aprendizaje algorítmico y la toma de decisiones automatizada, por lo que es necesario utilizar la tecnología de IA de manera confiable y manteniendo al mismo tiempo la dignidad humana, los derechos fundamentales, la equidad y la transparencia.

El Perú está considerando introducir activamente la IA en el proceso de promoción de la innovación digital y modernización de la gestión pública, con el objetivo de aumentar la eficiencia de los servicios públicos y mejorar la calidad de vida de los ciudadanos.

La ética de la IA va más allá del simple cumplimiento de las regulaciones legales y tiene como objetivo presentar orientación a la tecnología de IA para generar confianza social y lograr un desarrollo inclusivo y sostenible. Los principios éticos de la IA deben centrarse en valores como la dignidad humana, la equidad y la responsabilidad, y estos principios deben ser aplicables de manera práctica y viable.

Para lograrlo, los estándares éticos de la IA para el gobierno digital del Perú presentan directrices éticas para garantizar que la IA se utilice en el sector público para promover el valor social, proteger los derechos de los ciudadanos y contribuir al interés público. En particular, las agencias gubernamentales y las instituciones públicas que utilizan IA deben introducir la tecnología de IA de una manera justa y responsable, manteniendo al mismo tiempo la confianza de los ciudadanos y, en el proceso, es necesario un enfoque abierto e inclusivo que refleje las opiniones de las distintas partes interesadas.

La ética de la IA es un elemento clave para mantener un equilibrio entre el avance tecnológico y la responsabilidad social. La IA ética debe ir más allá del simple cumplimiento de las regulaciones para garantizar que las tecnologías de IA promuevan el bienestar humano, alivien las desigualdades sociales y logren un desarrollo sostenible.

* 1. Objetivo de los estándares éticos

Los estándares éticos de la IA para el gobierno digital del Perú tienen como meta implementar una IA centrada en el ser humano, y siguientes objetivos para garantizar que las tecnologías de IA se utilicen de una manera que promueva el desarrollo nacional y el interés público:

* Permitir que la IA proteja y promueva la dignidad humana y los derechos fundamentales
* Garantizar que los sistemas de IA estén diseñados y operados de conformidad con la Constitución peruana y los estándares internacionales de derechos humanos y no infrinjan los derechos de los ciudadanos.
* Garantizar la equidad y la no discriminación para evitar que la IA agrave las desigualdades sociales.
* Proporcionar principios y pautas para desarrollar una IA confiable.
* Fortalecer la transparencia, la responsabilidad, la equidad y la robustez en el proceso de desarrollo y operación de la IA para establecer un sistema en el que los ciudadanos puedan confiar.
* Apoyar la realización de una IA ética al exigir que el proceso de toma de decisiones de los sistemas de IA sea explicable y verificable.
* Proporcionar orientaciones políticas para que la tecnología de IA pueda utilizarse en el sector público de manera confiable basada en los valores democráticos y el estado de derecho.
* Promover el desarrollo sostenible y el uso ético de la IA
* Establecer una dirección para armonizar la innovación tecnológica y la responsabilidad ética para que el desarrollo de la IA y la implementación ética no entren en conflicto.
* Establecer una base política para que la tecnología de IA pueda apoyar la creación de valor social, la mejora de los servicios públicos y la expansión de la inclusión digital.
* Fomentar que la tecnología de IA contribuya a los Objetivos de Desarrollo Sostenible (ODS), incluida la protección del medio ambiente, la respuesta al cambio climático y la protección de los ecosistemas.
* Establecer una gobernanza ética de la IA basada en la cooperación entre diversas partes interesadas
* Establecer una gobernanza en la que diversas partes interesadas, como el gobierno, las instituciones públicas, la academia, las empresas privadas y la sociedad civil, puedan cooperar en el proceso de establecer políticas y regulaciones relacionadas con la IA.
* Promover la innovación responsable y la cooperación internacional para el uso ético y la mejora continua de la IA.

Estos estándares éticos funcionan como normas voluntarias que complementan las normas legales, y se discutirán y mejorarán continuamente para responder de manera flexible a los cambios tecnológicos y sociales.

El objetivo de los estándares éticos de la IA para el gobierno digital del Perú es garantizar que las tecnologías de IA respeten los derechos y la dignidad humana de todos los ciudadanos y protejan las libertades fundamentales. A través de esto se permitirá que las tecnologías de IA agreguen confianza a los servicios públicos, y promuevan la equidad social y la transparencia. Asimismo, proporcionará la dirección para que se diseñen y utilicen de manera responsable. Los estándares éticos tienen como objetivo fortalecer la sostenibilidad ambiental y la inclusión social y, permiten establecer una base institucional en que los avances tecnológicos brindan beneficios a todos los ciudadanos al armonizar las políticas nacionales de transformación digital y la ley marco de IA.

* 1. Aplicación de estándares éticos
		1. Alcance de aplicación de los estándares éticos

El alcance de los estándares éticos de la IA del Perú para el sector público abarca las consideraciones éticas a lo largo de todo el ciclo de vida de los sistemas de IA: desarrollo, despliegue, utilización, eliminación, etc. Estos lineamientos analizan comprensivamente los impactos sociales, económicos y legales que pueden surgir cuando las tecnologías de IA se utilizan en el sector público y sugieren direcciones para proteger los derechos humanos, generar confianza social y lograr el desarrollo sostenible.

La ética de la IA no se trata simplemente del cumplimiento legal; es un concepto integral que incluye la confiabilidad, la responsabilidad social y el diseño ético de la tecnología. Por lo tanto, los estándares éticos proporcionan principios para garantizar que los sistemas de IA respeten la dignidad y los derechos fundamentales de las personas, sean transparentes y confiables, y apoyen el interés público y el desarrollo sostenible.

Además, presentan requisitos éticos adecuados al contexto social, económico y cultural del Perú, tomando en cuenta la coherencia con la legislación peruana y los estándares éticos de IA de organizaciones internacionales. Estas normas se revisarán y complementarán continuamente de acuerdo con los avances tecnológicos y los cambios sociales, y tienen como objetivo establecer una base para que las instituciones públicas y el sector privado practiquen la ética de la IA a través de la cooperación.

* + 1. Objetos de aplicación de los estándares éticos

Los estándares éticos se aplican a todas las partes interesadas que desarrollan, operan y utilizan sistemas de IA en el gobierno y las instituciones públicas peruanas.

* Objetos a los que se aplican estos estándares éticos:
* Gobierno e instituciones públicas: agencias gubernamentales centrales y locales que planifican y operan servicios públicos basados ​​en IA, institutos de investigación del sector público, etc.
* Desarrolladores de IA y proveedores de tecnología: empresas nacionales y extranjeras, institutos de investigación, universidades, etc. que estudian y desarrollan sistemas de IA del sector público.
* Proveedores de servicios públicos: entidades y empresas que operan servicios y sistemas basados en IA utilizando datos públicos.
* Usuarios de sistemas de IA: ciudadanos y partes interesadas que utilizan directamente los sistemas de IA del sector público

No son regulaciones legales obligatorias, sino más bien directrices éticas que fomentan la práctica voluntaria de la ética de la IA, ayudando a todas las partes interesadas a respetar los estándares éticos de la IA y cumplir con sus responsabilidades sociales. Los estándares éticos de IA del Perú se basan en principios éticos internacionales, pero buscan reflejar el entorno legal y social del Perú para aumentar la aplicabilidad práctica. En consecuencia, los lineamientos pueden utilizarse como directrices para establecer un sistema de gobernanza de la IA en el sector público, garantizar la fiabilidad de la tecnología de la IA y promover el valor público.

1. Principios básicos

Las tecnologías de la IA deben operarse según los siguientes cinco principios básicos para que se utilicen de forma ética.

* 1. Principios de la dignidad humana
* **La IA debería desempeñar un papel de apoyo y mejora para los humanos, y respetar la autonomía y la toma de decisiones humanas.**
* **Los sistemas de IA no deben violar los derechos fundamentales garantizados por la Constitución del Perú.**
* **La IA en el sector público debe diseñarse y operarse de manera que garantice los derechos de los ciudadanos.**
* Descripción de los principios de la dignidad humana

La inteligencia artificial (IA) puede utilizarse como herramienta para mejorar la vida humana, pero también tiene el potencial de socavar la dignidad humana. Los estándares éticos de la IA del Perú se basan en garantizar que la IA se desarrolle y se utilice de una manera que respete, proteja y promueva la dignidad y los derechos inherentes de todas las personas.

Todas las personas tienen la libertad y el derecho de decidir su propia vida, y los sistemas de IA no deben infringirlos. Se debe garantizar que el uso de IA no resulte en vigilancia indebida, engaño, manipulación mental o discriminación injusta, asimismo que la tecnología de IA desempeñe un papel de apoyo para permitir que los humanos tomen sus propias decisiones de forma independiente. En particular, se debe garantizar que los socialmente vulnerables y marginados no queden excluidos o desfavorecidos debido a la IA, y también se requiere un papel activo del gobierno y las instituciones públicas que pueden garantizar un acceso justo a la IA.

La IA debería estar diseñada para respetar la integridad física y mental humana y la identidad individual, en lugar de tratar a los humanos como meros datos u objetos de análisis mecánico. Los sistemas de IA deben desarrollarse y operarse de manera que no comprometan la salud y el bienestar humanos y deben cumplir los siguientes principios de diseño:

* Los sistemas de IA deberían ser diseñados con la premisa de que los humanos son seres con valor que no pueden intercambiarse con sistemas de IA.
* Las tecnologías de IA deben desarrollarse y operarse de una manera que no dañen la salud física y mental humana.
* Se debe evitar el riesgo de deño a los humanos al garantizar la seguridad y la robustez de los sistemas de IA.
* Se debe llevar a cabo una vigilancia y evaluación continuas para garantizar que el uso de la IA no atente contra los derechos y las libertades humanas.

El diseño y la operación de los sistemas de IA deben tener como premisa que todos los humanos son seres de igual valor y que la IA no debe tomar decisiones desventajosas o sesgadas hacia determinados grupos. Los estándares éticos de la IA del Perú reconocen que todos los individuos tienen igual dignidad, independientemente de su género, edad, raza, origen, situación económica o discapacidad, y proporcionan una base para establecer medidas técnicas e institucionales para asegurar que los sistemas de IA puedan garantizar esto.

* 1. Principio de equidad
* **La IA debe diseñarse y utilizarse de manera que todos los miembros de la sociedad puedan beneficiarse de manera justa y no debe agravar la desigualdad social.**
* **Los sistemas de IA deben evitar la discriminación teniendo en cuenta diversos factores sociales y culturales, como el género, la edad, el contexto económico, la región, la discapacidad y los pueblos indígenas.**
* **Se debe garantizar que los grupos vulnerables tengan igualdad de oportunidades al aumentar la accesibilidad a los servicios públicos de IA.**
* Explicación del principio de equidad

La equidad significa que los sistemas de inteligencia artificial (IA) no discriminan a individuos o grupos específicos y brindan igualdad de oportunidades en toda la sociedad. En otras palabras, los sistemas de IA deben diseñarse y operarse de manera que no afecten injustamente a individuos o grupos específicos, y deben desarrollarse de manera que logren la inclusión y la diversidad. Esto significa garantizar que los sistemas de IA brinden igualdad de oportunidades en toda la sociedad y eviten que las desigualdades sociales existentes se agraven.

Para garantizar la equidad en los sistemas de IA, se debe tener cuidado de que no se produzcan sesgos en el proceso de recopilación, procesamiento y utilización de datos; y, asimismo, hacer consideración para que ciertas clases o grupos no se vean en desventaja durante los procesos de diseño del algoritmo y de toma de decisiones. Diseñar una IA que incluya, en lugar de excluir, a los socialmente desfavorecidos y vulnerables es un elemento clave del principio de equidad.

Además, se necesita revisar continuamente si las decisiones de los sistemas de IA se toman de manera justa y racional, teniendo en cuenta el impacto en la vida y los derechos humanos. Para ello, se debe establecer criterios de evaluación transparentes y objetivos y asegurar la explicabilidad del proceso de toma de decisiones para que la tecnología de IA pueda operarse sobre la base de la confianza. El gobierno peruano adoptará un enfoque inclusivo para evitar sesgos injustos y permitir que los grupos socialmente desfavorecidos y vulnerables no queden excluidos de los desarrollos tecnológicos en el proceso de introducción y utilización de las tecnologías de IA en el sector público. Para tal fin, se debe estipular normas y procedimientos para revisar la equidad en todo el proceso de desarrollo y operación de los sistemas de IA, y, asimismo, realizar evaluaciones continuas a fin de que ningún grupo o clase específica sea discriminada.

* 1. Principio de transparencia
* **Los sistemas de IA deben tener suficiente responsabilidad de explicación para que las partes interesadas puedan comprender la manera de funcionamiento y los procesos de toma de decisiones de estos sistemas.**
* **Los procesos clave de toma de decisiones para los servicios públicos basados ​​en IA deben hacerse públicos para que los ciudadanos puedan acceder a ellos**
* Descripción del principio de transparencia

La transparencia es un elemento esencial de cualquier sistema de IA confiable y tiene como objetivo garantizar que todas las partes interesadas tengan una comprensión clara de cómo funciona la IA y el proceso de toma de decisiones. La transparencia en los sistemas de IA es un elemento clave para generar confianza pública y permitir que la IA se utilice con el fin de mejorar el valor social. Para que los sistemas de IA sean confiables y se utilicen en diversos contextos sociales, se debe divulgar claramente cómo funciona el sistema, qué datos se utiliza para aprender y tomar decisiones, y qué limitaciones e incertidumbres existen en el proceso.

La transparencia en los sistemas de IA consta de los siguientes elementos:

* **Trazabilidad**: Los procesos de recopilación de datos, entrenamiento de algoritmos, construcción de modelos y toma de decisiones de los sistemas de IA deben documentarse, de modo que sea posible rastrear el proceso mediante el cual el sistema de IA tomó una decisión particular.
* **Explicabilidad**: Se debe proporcionar una explicabilidad técnica para que se pueda entender la manera de funcionamiento y la lógica de toma de decisiones de los sistemas de IA. Esto supone que la IA no se limita a tomar decisiones, sino que proporciona una justificación lógica de por qué se tomaron esas decisiones.
* **Divulgación responsable**: Los sistemas de IA utilizados en los servicios públicos deben divulgar qué datos y algoritmos se utilizaron para tomar decisiones. En particular, cuando las decisiones tomadas por los sistemas de IA se aplican con sensibilidad en áreas clave para la vida y los derechos humanos, como la administración pública, los procedimientos judiciales, la salud y las finanzas, es necesario contar con mecanismos de supervisión y retroalimentación para garantizar que las decisiones cumplan con los valores democráticos y los principios legales.

Cuando los sistemas de IA se utilicen en el sector público, los ciudadanos deben poder confiar en que las decisiones basadas en IA son justas y confiables y que el proceso de toma de decisiones debe ser verificable. Por lo tanto, la información proporcionada por los sistemas de IA debe ser comprensible, clara y accesible para todos.

La transparencia de los sistemas de IA es un principio para asegurar la confiabilidad técnica y la responsabilidad ética. En particular, cuando se utiliza IA en el sector público, el proceso de toma de decisiones de la IA debe estar documentado, ser comprensible y monitoreable para garantizar que las decisiones tomadas por los sistemas se lleven a cabo en forma confiable. Las tecnologías de IA pueden contribuir a mejorar el valor social y fortalecer la gobernanza democrática cumpliendo el principio de transparencia.

* 1. Principio de robustez
* **Los sistemas de IA deben funcionar de manera fiable, segura y predecible.**
* **Los sistemas de IA deben minimizar los errores y las amenazas a la seguridad.**

* Explicación del principio de robustez

Para que los sistemas de inteligencia artificial (IA) se operen de manera confiable en el sector público del Perú, es necesario asegurar su robustez técnica y social. Para ello, los sistemas de IA deben ser robustos frente a errores inesperados y ataques externos, y deben funcionar de forma fiable incluso en situaciones impredecibles durante el proceso de prestación de servicios públicos.

El principio de robustez garantiza que los sistemas de IA se diseñen, desarrollen, implementen y operen de manera confiable. Los sistemas de IA deben minimizar los errores inherentes, prevenir amenazas a la seguridad y funcionar correctamente en entornos inesperados. Además, el diseño transparente y los procedimientos documentados son esenciales para que los sistemas de IA mantengan su robustez, y se debe mantener la calidad de los datos y la equidad algorítmica. Los procesos de aprendizaje y toma de decisiones de los modelos de IA deben revisarse exhaustivamente y mejorarse continuamente para minimizar la incertidumbre y el sesgo.

Los sistemas de IA robustos no solo deben ser confiables en entornos específicos, sino también funcionar de manera consistente en una variedad de contextos sociales y económicos. En particular, cuando se utilizan sistemas de IA en procesos de toma de decisiones en el sector público, se debe tener cuidado para que los errores en el sistema no exacerben la desigualdad social o tengan un efecto perjudicial sobre grupos específicos.

* 1. Principio de responsabilidad
* **Se debe establecer un sistema de gestión para que las tecnologías de IA promuevan el desarrollo ambiental, económico y social al tiempo que cumple con las responsabilidades éticas.**
* **Los sistemas de IA deben utilizarse para crear valor social y deben apuntar a mejorar la calidad de los servicios públicos y promover beneficios sociales.**
* **Las decisiones automatizadas de la IA deben diseñarse de modo que no socaven la seguridad ciudadana ni la confianza social.**
* Explicación del principio de responsabilidad

El principio de responsabilidad tiene como objetivo garantizar que los sistemas de IA funcionen de manera ética y confiable, lo que puede vincularse con la responsabilidad legal y administrativa, y social, la garantía de sostenibilidad e incluso la cooperación internacional.

Dado el impacto de las decisiones de los sistemas de IA en los ciudadanos y la sociedad, los gobiernos y los organismos pertinentes deben establecer sistemas de responsabilidad transparentes y claros. Las organizaciones e instituciones que diseñan y desarrollan sistemas de IA deben ser responsables de los resultados producidos por esos sistemas y deben contar con procedimientos claros para corregir y minimizar los daños cuando ocurren errores o efectos secundarios inesperados en los sistemas de IA.

En particular, cuando se utilizan sistemas de IA en el sector público para tomar decisiones políticas, realizar procedimientos administrativos y prestar servicios públicos, se debe considerar cuidadosamente el impacto en los ciudadanos. El proceso de toma de decisiones de los sistemas de IA debe ser verificable y debe haber procedimientos claros de responsabilidad para corregir y complementar decisiones incorrectas.

Además, el desarrollo y uso de la tecnología de IA debe tener en cuenta la sostenibilidad económica y social, y los sistemas de IA deben estar diseñados para contribuir al bienestar social a largo plazo, en lugar de buscar beneficios a corto plazo. En particular, las tecnologías de IA deben desarrollarse y operarse teniendo en cuenta la sostenibilidad ambiental y la responsabilidad social. Los sistemas de IA deben diseñarse de manera que se aumenta la eficiencia energética y se minimiza el impacto ambiental, y, asimismo, estar alineados con políticas que tengan en cuenta los valores sociales y éticos.

Además, la utilización de la IA debe realizarse en conjunción con marcos legales e institucionales para apoyar los objetivos de desarrollo sostenible para las comunidades y el gobierno. Los sistemas de IA pueden operar a través de fronteras entre los países, y se alienta al gobierno peruano a consultar los estándares internacionales de tecnología de IA, las pautas éticas y las políticas de protección de datos.

* 1. Conflicto entre los principios

Los cinco principios de dignidad humana, equidad, transparencia, robustez y responsabilidad presentados en los estándares éticos de IA del Perú desempeñan funciones complementarias en el desarrollo, el despliegue y la utilización de los sistemas de IA, pero pueden entrar en conflicto en determinadas situaciones. Reconocer y equilibrar las tensiones entre estos principios es esencial para el funcionamiento ético de la IA.

* Ejemplos de conflictos entre los principios
* **La tensión entre la dignidad humana y la robustez**: En el proceso del diseño de los sistemas de IA con un alto nivel de robustez, existe el riesgo de que la introducción de sistemas de vigilancia excesivos o de toma de decisiones automatizada limiten los derechos y las libertades de las personas. Esto supone el riesgo de violar la dignidad humana y la autodeterminación, lo que requiere un enfoque de equilibrar la seguridad y el diseño centrado en las personas.
* **Conflicto entre la equidad y la transparencia**: Eliminar el sesgo de datos y garantizar la equidad son esenciales para que los sistemas de IA funcionen de manera justa para todos los usuarios. Sin embargo, modificar los algoritmos para aumentar la equidad corre el riesgo de complicar el funcionamiento de la IA, reduciendo la explicabilidad y la transparencia. Por lo tanto, se necesitan medidas complementarias para mantener un equilibrio tecnológico y de políticas que pueda garantizar simultáneamente la equidad y la transparencia, y para proporcionar una explicación suficiente del proceso de toma de decisiones de la IA.
* **Contradicción entre la transparencia y la robustez**: Para mejorar la robustez y seguridad de la IA, algunos algoritmos pueden mantenerse confidenciales o estar sujetos a protecciones reforzadas. Sin embargo, si estas medidas son excesivas, el proceso de toma de decisiones de la IA puede volverse opaco, lo que puede reducir la confianza. Por lo tanto, se necesitan políticas y mecanismos técnicos para equilibrar la seguridad y la transparencia.
* **Equilibrio entre la responsabilidad y la equidad**: Asignar una responsabilidad excesiva a organizaciones o individuos específicos para garantizar la responsabilidad por los sistemas de IA podría socavar la equidad. Por el contrario, si la responsabilidad se distribuye entre múltiples partes interesadas, puede surgir el problema de la evasión de responsabilidad. Por lo tanto, para garantizar la responsabilidad de los sistemas de IA, se debe establecer una distribución clara de roles y un sistema de supervisión de la implementación.

A fin de resolver los conflictos entre estos principios, los estándares éticos de la IA del Perú enfatizan la colaboración y el debate continuos entre las partes interesadas. Diversas partes interesadas, incluidas las instituciones públicas, los desarrolladores de IA, la sociedad civil y las empresas, deberían seguir debatiendo el funcionamiento ético de los sistemas de IA y establecer un sistema de coordinación flexible para encontrar el punto de equilibrio óptimo en situaciones específicas. Además, cuando surgen conflictos éticos durante el desarrollo y funcionamiento de sistemas de IA, es importante establecer procedimientos y directrices que puedan resolverlos basándose en el consenso social.

Los conflictos entre los principios éticos de la IA son inevitables y no existe una única solución para ellos. Por lo tanto, se deben establecer procedimientos deliberativos responsables para resolver estas tensiones, basados ​​en los principios de participación democrática, debido proceso y participación política abierta. Por ejemplo, la “policía predictiva” basada en IA podría ayudar a prevenir el delito, pero también podría vulnerar la libertad y la privacidad de las personas. Es necesario sopesar los beneficios esperados y los riesgos individuales de los sistemas de IA y emitir juicios éticos que permiten mantener la confianza social para resolver estas tensiones entre los principios.

Los profesionales de la IA deberían abordar los dilemas éticos y la solución de compromiso mediante una reflexión basada en evidencia, en lugar de simplemente encontrar la respuesta correcta basándose en principios. Sin embargo, hay casos en los que no se puede encontrar un compromiso éticamente aceptable. Por ejemplo, ciertos derechos fundamentales y principios éticos, como la dignidad humana, no son negociables y deben protegerse absolutamente.

Los sistemas de IA deben reconocer los posibles conflictos entre estos principios y diseñar soluciones para resolverlos. En particular, se necesita un enfoque más cauteloso en grupos vulnerables como los niños y los discapacitados que corren el riesgo de verse desfavorecidos o excluidos, y en áreas como las relaciones laborales o la protección del consumidor donde la asimetría de poder e información son prominentes.

Los principios y requisitos establecidos en estos estándares éticos pueden entrar en conflicto según el contexto y la situación, y los métodos de resolución también pueden variar según el caso individual. Por lo tanto, estos estándares éticos no presentan una prioridad fija entre los principios, sino que más bien alientan a las diversas partes interesadas a buscar soluciones óptimas a través de una discusión y deliberación continuas.

1. Requisitos clave para el cumplimiento de los principios

Los ocho requisitos fundamentales para el cumplimiento de los 5 principios de los estándares éticos del Perú son los siguientes:

* **Protección de los derechos humanos**
* **Protección de la información personal y los datos**
* **Promoción de la publicidad y el valor social**
* **Fortalecimiento de la accesibilidad y la inclusión social**
* **Garantía de la explicabilidad**
* **Garantía de la fiabilidad y la seguridad técnica**
* **Desarrollo y operación de IA sostenible**
* **Cooperación internacional**
	1. Protección de los derechos humanos
* **Los sistemas de IA deben proteger los derechos fundamentales de los ciudadanos (libertad, privacidad, igualdad, etc.) y deben tener cuidado de no provocar resultados discriminatorios.**
* **Se debe mantener la imparcialidad de los datos y los algoritmos para evitar sesgos en el proceso de toma de decisiones de la IA.**
* Descripción de la protección de los derechos humanos

Los sistemas de IA deben apoyar la autonomía humana y la toma de decisiones de acuerdo con el principio de respeto a la autonomía humana. Para lograrlo, los sistemas de IA deben actuar como facilitadores de sociedades democráticas, prósperas y equitativas apoyando las elecciones de los usuarios, promoviendo los derechos fundamentales y permitiendo la supervisión humana.

Como muchas otras tecnologías, los sistemas de IA pueden facilitar o socavar derechos fundamentales. Por ejemplo, puede beneficiar a las personas ayudándolas a rastrear sus datos personales o apoyando su derecho a la educación aumentando el acceso a la educación. Sin embargo, dado el alcance y las capacidades de los sistemas de IA, pueden tener implicaciones negativas para los derechos fundamentales. En situaciones en que existan tales riesgos, se debe realizar una evaluación del impacto sobre los derechos fundamentales. Esto debe hacerse antes de desarrollar los sistemas y debería incluir una evaluación de si estos riesgos pueden reducirse o justificarse en una sociedad democrática, respetando al mismo tiempo los derechos y las libertades de los demás. Además, debería establecer mecanismos para recibir retroalimentación externa sobre los sistemas de IA que puedan vulnerar derechos fundamentales.

Los usuarios deberían poder tomar decisiones informadas y autónomas sobre los sistemas de IA. Se necesita que los usuarios reciban el conocimiento y las herramientas que permiten comprender los sistemas de IA en un grado satisfactorio e, interactuar con ellos. Además, los usuarios deberían poder autoevaluar o cuestionar razonablemente los sistemas. Los sistemas de IA deberían ayudar a las personas a tomar decisiones sobre la base de la mejor información en función de sus objetivos. Los sistemas de IA pueden moldear e influir en el comportamiento humano a través de mecanismos que son difíciles de detectar, ya que pueden utilizar procesos inconscientes como diversas formas de manipulación injusta, engaño, control y condicionamiento, todos los cuales pueden amenazar la autonomía individual, por lo que a veces pueden oponerse. El principio general de la autonomía del usuario debe ser central para el funcionamiento del sistema. En esencia, el derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado cuando éste produzca efectos jurídicos o afecte significativamente al usuario de modo similar.

La supervisión humana ayuda a garantizar que los sistemas de IA no socaven la autonomía humana ni provoquen otros efectos adversos. La supervisión se puede llevar a cabo a través de mecanismos de gobernanza como los enfoques de *HITL (Human-in-the-loop)*, *HOTL (Human-on-the-loop)* o *HIC(Human-in-command)*. *HITL* se refiere a un método en que los humanos pueden intervenir en todo el ciclo de toma de decisiones de un sistema, lo que en muchos casos no es ni posible ni deseable. *HOTL* es un método en que los humanos pueden intervenir y monitorear en el ciclo de diseño del sistema. *HIC* se refiere a un método en que los humanos supervisan las actividades generales de un sistema de IA (incluidas sus amplios impactos económicos, sociales, legales y éticos) y deciden cuándo y cómo utilizar el sistema en situaciones específicas. Esto podría incluir decidir no utilizar un sistema de IA en determinadas situaciones, establecer un nivel de discreción humana al utilizar el sistema o garantizar la función de anular las decisiones tomadas por el sistema. Además, se debe garantizar que los ejecutantes públicos puedan ejercer sus poderes de supervisión de conformidad con su autoridad. Dependiendo de la aplicación y los riesgos potenciales del sistema de IA, pueden requerirse diferentes niveles de mecanismos de supervisión para respaldar diferentes medidas de seguridad y control. En igualdad de condiciones, cuanto menos supervisión tengan los humanos sobre los sistemas de IA, más pruebas exhaustivas y gobernanza rigurosa requerirán.

* 1. Protección de la información personal y los datos
* **Se debe cumplir con las medidas de protección de la información personal y evitar el mal uso y abuso de los datos durante el proceso de recopilación, procesamiento, almacenamiento y compartición de los datos de IA.**
* **La información sensible (la de salud, financiera, biométrica, etc.) procesada automáticamente por sistemas de IA debe estar sujeta a fuertes medidas de protección.**
* **Las instituciones públicas deben garantizar la transparencia en el uso de los datos y tomar medidas para garantizar que los ciudadanos sepan cómo se utilizan sus datos.**
* Descripción de protección de la información personal y los datos

El desarrollo y la difusión de los sistemas de IA se basan en el uso de datos masivos y, como resultado, las cuestiones éticas de la protección de la información personal y la gestión de datos están adquiriendo cada vez más importancia. Por lo tanto, cuando se utilizan sistemas de IA en el sector público, se debe establecer un sistema de protección de datos exhaustivo para garantizar que no se violen los derechos fundamentales de los ciudadanos.

Cuando los sistemas de IA procesan automáticamente la información personal, es necesario reforzar las medidas de protección de la información sensible (la de salud, financiera, biométrica, etc.) y establecer un sistema de gestión para evitar el uso indebido de los datos. Para ello, los sistemas de IA deben cumplir con los principios de protección de la información personal y aplicar medidas de protección tecnológica como la encriptación, la anonimización y la seudonimización. Además, se debería introducir procesos de autoevaluación para confirmar que los sistemas de IA cumplan con los derechos de protección de datos, asegurando que se mantenga la integridad y la calidad de los datos.

Se debe implementar mecanismos para verificar si hay sesgos en los conjuntos de datos y ajustarlos para que los sistemas de IA no afecten negativamente a grupos específicos. Los datos con los que se entrenan los modelos de IA deben ser de alta calidad e integridad, y deben gestionarse para evitar resultados discriminatorios resultantes del uso de datos incompletos o sesgados. Para ello, los sistemas de IA del sector público deben introducir un proceso de revisión de la calidad de los datos y evaluar y mejorar continuamente los datos que utiliza la IA para tomar decisiones.

Las instituciones públicas deben tomar medidas para garantizar la transparencia sobre los datos que los sistemas de IA utilizan, y tomar medidas a fin de que los ciudadanos sepan cómo se utilizan sus datos. Los ciudadanos deberían poder confirmar cómo se está utilizando su información personal en los sistemas de IA y decidir libremente si consienten la recopilación y el uso de sus datos. Los sistemas de IA deben garantizar el derecho de acceso a los datos, el derecho de borrado (el derecho de ser olvidado), el derecho de oposición y el derecho a retirar el consentimiento, y deben contar con procedimientos claros que permitan ejercerlos.

Es necesario que la protección de la información personal se convierta en un principio ético más allá del simple cumplimiento de regulaciones a través de las políticas de gobernanza de los datos de IA Para ello, el sector público debe aumentar la transparencia en la gestión de datos de IA y brindar oportunidades de educación y participación para que los ciudadanos protejan sus propios datos. Además, se deben establecer fuertes medidas de protección jurídica para evitar que los datos de IA se exploten comercialmente.

* 1. Promoción de la publicidad y el valor social
* **La tecnología de IA debe utilizarse en el sector público para maximizar los beneficios sociales.**
* **Los servicios públicos basados ​​en IA deberían desempeñar un papel en la promoción no solo de los individuos sino también de la comunidad social, y deberían funcionar considerando el interés público**
* Descripción de la promoción de la publicidad y el valor social

La tecnología de IA debe utilizarse en el sector público para maximizar los beneficios sociales y debe posicionarse como una tecnología central que genere interés público y valor social más allá de una simple herramienta de desarrollo económico. Para tal fin, los servicios públicos basados ​​en IA deben operar teniendo en cuenta valores sociales integrales como la promoción de la publicidad social, la resolución de la brecha digital, la protección de los socialmente desfavorecidos y la ampliación de la participación ciudadana.

Los sistemas de IA deben utilizarse para mejorar la transparencia y la eficiencia en la administración pública y los servicios públicos. Al introducir la IA en la administración pública, se debe dar máxima prioridad a la protección de los derechos de los ciudadanos y al cumplimiento de las normas éticas, y se debe garantizar la transparencia de la toma de decisiones basada en IA. Además, la IA debe mantener la equidad en el proceso de prestación de servicios públicos y permitir que todos los ciudadanos disfruten de beneficios iguales.

Eliminación de la brecha digital y garantía de la accesibilidad son elementos clave para hacer realidad la naturaleza pública de la IA. La tecnología de IA debe desarrollarse y distribuirse de manera que sea igualmente accesible para todos los ciudadanos, y debe diseñarse de manera que los grupos socialmente desfavorecidos y vulnerables no se vean perjudicados por ella. Se necesita establecer un sistema de monitoreo continuo para que los sistemas de IA no excluyan ni discriminen a grupos específicos y, a través de esto, las tecnologías de IA deberían desempeñar un papel en la promoción de la inclusión social.

Los procesos de toma de decisiones políticas basados ​​en IA deben diseñarse para promover la participación ciudadana. Para ello, se debe permitir que los ciudadanos y las partes interesadas puedan presentar activamente sus opiniones en el proceso de establecer y operar las políticas de IA al utilizar las audiencias públicas, las plataformas de recopilación de opiniones y los sistemas de divulgación de información. Asimismo, se debe cumplir con los procedimientos democráticos en proceso del uso público de la IA, y crear un entorno en que los ciudadanos puedan proporcionar retroalimentación directa en el proceso de establecimiento e implementación de las políticas de IA.

Las tecnologías de IA debe utilizarse para satisfacer el interés público y contribuir a resolver diversos problemas sociales, como la reducción de la pobreza, la ampliación de las oportunidades educativas y la mejora de los servicios de salud y bienestar. Para tal fin, el gobierno debe fortalecer el apoyo político para que la tecnología de IA pueda contribuir a la realización de valor social, y las instituciones públicas y el sector privado deben cooperar para promover la innovación social utilizando IA.

La garantía de la equidad en el ecosistema de IA también es un factor importante. En el proceso de desarrollo y despliegue de las tecnologías de IA, las instituciones públicas y el sector privado deben cumplir con sus responsabilidades éticas para crear un entorno competitivo justo y minimizar los efectos secundarios de los sistemas de IA. Los proyectos de desarrollo de IA en el sector público deben priorizar el interés público y operar de una manera que permita a los ciudadanos experimentar directamente los beneficios.

Las políticas de IA deberían diseñarse no para obtener resultados a corto plazo, sino desarrollarse de forma sostenible. Para ello, las instituciones públicas y el sector privado deben establecer las estrategias para utilizar la IA que tengan en cuenta los intereses públicos a largo plazo y cumplan con los estándares éticos. Además, es necesario establecer un sistema de evaluación y supervisión, y asimismo mejorarlo continuamente reflejando las opiniones de los ciudadanos y los expertos para garantizar el funcionamiento sostenible de la IA en el sector público.

En el proceso de establecer y operar políticas de IA, también puede valer la pena considerar el establecimiento de un sistema de cooperación multilateral. Se fortalece los estándares éticos de la IA y se desarrolla las políticas de IA para crear valor social a través de la cooperación entre los gobiernos, las empresas privadas, el mundo académico y la sociedad civil. Para ello es necesario crear una organización de cooperación multilateral y poner en funcionamiento un sistema de recopilación y seguimiento continuo de opiniones sobre el uso público de la IA.

* 1. Fortalecimiento de la accesibilidad y la inclusión social
* **Se debe implementar políticas para cerrar la brecha digital y aumentar la accesibilidad a las tecnologías y servicios de IA.**
* **Los sistemas de IA en el sector público deben diseñarse para ser accesibles a los grupos vulnerables (pueblos indígenas, personas con discapacidad, personas de bajos ingresos, etc.).**
* Descripción del fortalecimiento de la accesibilidad y la Inclusión social

La tecnología de IA debe desarrollarse y operarse de manera que todos los ciudadanos puedan acceder a ella por igual y, en particular, se necesitan esfuerzos políticos para que los grupos socialmente desfavorecidos y vulnerables no queden excluidos del uso de la tecnología. Los sistemas de IA en el sector público deben tener como objetivos centrales el establecimiento de políticas que garanticen la accesibilidad digital, amplíen la inclusión social y tengan en cuenta el impacto de la adopción de IA en el mercado laboral.

Los sistemas de IA utilizados en el sector público deben estar diseñados para ser igualmente accesibles para todos los usuarios, independientemente de su edad, género, discapacidad, situación económica o región. Los productos y servicios de IA deben cumplir los principios del Diseño Universal para que diversos usuarios puedan utilizar la tecnología de IA sin discriminación. Además, debe garantizarse la compatibilidad con las tecnologías de asistencia para que las personas con discapacidad y las personas socialmente desfavorecidas puedan utilizar sin problemas los sistemas de IA, y deben proporcionarse medios alternativos cuando sea necesario.

Cuando se utilizan sistemas de IA para prestar servicios públicos, es necesario realizar un seguimiento y una evaluación continuos para garantizar que no se conviertan en factores que profundicen las brechas sociales. Se deben establecer procedimientos de evaluación de equidad y auditoría de accesibilidad a fin de que los sistemas de IA no excluyan o afecten negativamente a determinadas clases, y a través de esto debería permitir que los servicios públicos basados ​​en IA funcionen de manera inclusiva y justa.

Para que la tecnología de IA no esté sesgada hacia una clase específica, se deben establecer procedimientos para revisar los procesos de recopilación de datos y de aprendizaje de IA con el fin de que no se refleje ese sesgo. En el proceso de entrenamiento de modelos de IA se debe introducir un mecanismo que garantice la representatividad y diversidad de los datos para evitar la discriminación por género, edad, raza, nivel económico, etc.

Además, se necesita apoyo político para que la IA pueda contribuir a reducir las brechas sociales en áreas como la educación, el empleo y los servicios públicos. Al diseñar servicios públicos basados ​​en IA, las instituciones públicas deben ofrecer programas de educación y apoyo a fin de que los grupos vulnerables no tengan dificultades para utilizar las tecnologías digitales. Los sistemas de IA deben desarrollarse de forma inclusiva, incluido un apoyo personalizado para los excluidos digitalmente, con el objetivo de que la brecha tecnológica no conduzca a la desigualdad social.

Es necesario evaluar de antemano el impacto de la introducción de la IA en el mercado laboral y la estructura de puestos de trabajo. De acuerdo con el resultado debería establecer políticas de protección de los trabajadores. Las tecnologías de IA pueden brindar oportunidades para aumentar la productividad y crear nuevos empleos, pero también puede traer desafíos como reemplazar empleos existentes, cambiar el entorno de trabajo y ampliar la brecha digital.

Para abordar esto, se debe promover activamente programas de recapacitación (*reskilling*) y mejora de las habilidades (*upskilling*) de los trabajadores que puedan verse afectados por la automatización basada en IA. Los sectores público y privado deben trabajar juntos para establecer estrategias para crear nuevos empleos y entornos laborales favorables a la IA, y desarrollar políticas personalizadas para proteger a los grupos vulnerables, incluidos los trabajadores poco cualificados, los trabajadores de industrias tradicionales, los jóvenes y los trabajadores de mayor edad.

Además, debe fortalecerse el sistema de protección legal para que la introducción de la IA no afecte los derechos de los trabajadores, y debe establecerse un sistema de monitoreo y evaluación exhaustiva a fin de que la toma de decisiones basada en IA sea justa, transparente y no discriminatoria. Se debe ampliar el diálogo social y los sistemas de cooperación para que la introducción de la IA no tenga impactos negativos en los trabajadores, garantizando la participación de los sindicatos y las organizaciones de la sociedad civil.

Los sectores público y privado deben cumplir con los principios de protección social y transición justa, y establecer estructuras de gobernanza colaborativa para que las opiniones de los trabajadores se reflejen en el proceso de introducción de la IA. A través de esto se debería lograr al mismo tiempo el crecimiento económico y la equidad social con el uso de IA en el mercado laboral.

* 1. Garantía de la explicabilidad
* **Los servicios públicos basados ​​en IA deben poder explicar a los ciudadanos cómo se utilizan los algoritmos y los datos.**
* **Se debería proporcionar las explicaciones e información comprensibles en el proceso de toma de decisiones de la IA, asimismo incluir un proceso para apelar las decisiones automatizadas.**
* Descripción de la garantía de la explicabilidad

Cuando se utilizan sistemas de IA en el sector público, los ciudadanos deberían poder comprender el proceso de toma de decisiones basado en IA y sus resultados, y tener derecho a que se les proporcione información suficiente sobre las maneras de funcionamiento de la IA. La explicabilidad de los sistemas de IA garantiza la transparencia en la toma de decisiones, genera confianza ciudadana y permite que las decisiones automatizadas se tomen de manera justa. Para este fin, los sistemas de IA deberían poder explicar cómo funcionan y qué decisiones toman de una manera que los usuarios puedan entender, e incluir procedimientos de apelación y revisión cuando sea necesario.

Cuando los sistemas de IA se aplican a los servicios y la administración públicos, deben diseñarse para explicar claramente a los ciudadanos cómo se utilizan los algoritmos y los datos. En particular, debe establecerse un sistema para explicar públicamente cómo se utilizan los modelos de IA en las decisiones políticas basadas en IA y en la prestación de servicios públicos, el proceso de ingreso de datos y cómo se obtienen los resultados.

Los datos y algoritmos utilizados en el proceso de toma de decisiones de los sistemas de IA deben registrarse y gestionarse a través de un sistema de registro que garantice la trazabilidad, y los registros deben estar disponibles para la revisión cuando sea necesario. Si un sistema de IA funciona como una caja negra, es difícil que los ciudadanos y las instituciones confíen en él, por lo que se deben utilizar técnicas de IA explicable (XAI) para que el proceso de toma de decisiones de la IA sea interpretable. Esto permitirá la verificación de decisiones tomadas automáticamente por la IA, asegurando tanto la confiabilidad como la responsabilidad.

Cuando se utilizan sistemas de IA en servicios públicos y procesos de toma de decisiones políticas, la información presentada por la IA debe proporcionarse de una manera fácil de usar para que los ciudadanos comunes puedan entenderla fácilmente. Esto se puede lograr mediante el uso de elementos visuales intuitivos, explicaciones concisas y paneles de control fáciles de entender para ayudar a los ciudadanos a comprender claramente la lógica detrás de las decisiones de IA y las maneras de funcionamiento.

Además, cuando los sistemas de IA toman decisiones automatizadas, deben existir procedimientos que permitan a los ciudadanos solicitar explicaciones y apelaciones. Los sistemas de IA del sector público deberían proteger los derechos de los ciudadanos introduciendo sistemas de solicitud de explicaciones y proporcionando procedimientos de reexamen y revisión de decisiones específicas. Durante este proceso, se debe revisar la legitimidad de la toma de decisiones de la IA y tomar medidas para corregir errores o fortalecer la equidad cuando sea necesario.

La explicabilidad de los sistemas de IA debe equilibrarse con la protección de información personal y la seguridad de los datos. Se debe establecer las medidas de seguridad exhaustivas para evitar que la información personal sensible se exponga innecesariamente en el proceso de mejora de la transparencia de los sistemas de IA. Asimismo, en el proceso de divulgación de datos y algoritmos utilizados por la IA, se debe establecer un nivel adecuado de estrategia de divulgación de información, teniendo en cuenta la confidencialidad comercial, el mantenimiento de la seguridad y las restricciones legales.

* 1. Garantía de la confianza y la seguridad técnica
* **Los sistemas de IA deben funcionar de forma estable y minimizar los riesgos impredecibles.**
* **El rendimiento y la estabilidad de la tecnología de IA deben revisarse y mantenerse continuamente para garantizar que no haya vulnerabilidades de seguridad.**
* **Se debe establecer un mecanismo para corregir las decisiones de la IA que sean incorrectas.**
* Descripción de garantía de la confianza y la seguridad técnica

Para que los sistemas de IA funcionen de manera confiable y segura en el sector público, es esencial garantizar la confianza técnica, la seguridad, la previsibilidad y la reproducibilidad. A medida que la tecnología de IA se aplica a los servicios públicos, los procedimientos administrativos y la infraestructura social, es necesario minimizar los daños causados ​​por errores inesperados o ciberataques externos. Para ello, es necesario establecer sistemáticamente la seguridad y la gestión de riesgos de los sistemas de IA, y realizar una verificación y evaluación continuas para mantener la integridad de los datos y la robustez de los modelos.

Los sistemas de IA deben estar diseñados para funcionar de manera confiable en entornos cambiantes y contar con medidas de seguridad para protegerse contra riesgos impredecibles. Los servicios públicos basados ​​en IA deben ir más allá del simple desempeño funcional para ser tolerantes a fallas y resilientes, y deben poder operar de manera continua incluso en situaciones inesperadas. En particular, cuando un sistema de IA está expuesto a acciones hostiles externas o errores del sistema, se debe evitar la interrupción del servicio y minimizar los daños mediante procedimientos de recuperación rápida.

Los sistemas de IA deben tener un sistema de seguridad sólido para protegerse contra diversas amenazas de seguridad, como piratería, envenenamiento de datos, manipulación de algoritmos e inversión de modelo. Cuando se utilizan sistemas de IA en el sector público, es necesario revisar continuamente los protocolos de seguridad y cumplir periódicamente con las revisiones de seguridad como pruebas de equipo rojo y pruebas de penetración, para detectar amenazas con anticipación.

Para mantener la integridad de los sistemas de IA, se deben fortalecer las medidas de protección de datos y establecer un sistema de gestión de actualizaciones de seguridad y parches de vulnerabilidad para garantizar capacidades de respuesta continua. A través de esto deberían hacer que los sistemas de IA tengan una base confiable en el funcionamiento de los servicios públicos.

También podemos considerar establecer procedimientos para detectar y corregir de forma proactiva errores que puedan ocurrir en el proceso de toma de decisiones de los sistemas de IA. Para garantizar que la toma de decisiones basada en IA siga siendo justa y confiable, el rendimiento del modelo debe verificarse continuamente y minimizar los sesgos. Para ello, las instituciones públicas deben realizar periódicamente procesos de evaluación del desempeño de los modelos de IA y operar sistemas que eviten errores inesperados en la toma de decisiones. Además, se deben identificar los factores que degradan el rendimiento e implementar medidas correctivas para garantizar que los sistemas de IA sigan siendo confiables.

En particular, se deben establecer procedimientos de evaluación de la calidad de los datos y de verificación de modelos para evitar sesgos que puedan producirse durante el proceso de aprendizaje de nuevos datos por parte de los modelos de IA. Cuando un sistema de IA produce resultados no deseados, es necesario introducir un mecanismo de compensación para identificarlos y corregirlos para mantener la imparcialidad del sistema.

Para garantizar la confiabilidad de los sistemas de IA, se debe garantizar la reproducibilidad y la previsibilidad. Se debería mantener la coherencia del algoritmo para que los mismos valores de entrada puedan producir los mismos resultados, y para ello se requiere procedimientos de prueba repetida y validación. Los sistemas de IA del sector público deben adoptar las estrategias de aprendizaje continuo para que los modelos de IA puedan reflejar nuevos datos y, al mismo tiempo, mantener la toma de decisiones anterior incluso en los casos que los modelos de IA continúan aprendiendo.

Si un sistema de IA produce resultados inesperados, también es necesario contar con un proceso para analizarlos y determinar la causa. Para ello, es necesario registrar sistemáticamente el historial operativo del sistema de IA y fortalecer el sistema de registro y monitoreo del proceso de toma de decisiones de la IA para permitir una respuesta rápida cuando ocurre un problema.

* 1. Desarrollo y operaciones de la IA sostenible
* **Las instituciones públicas que introducen y operan los sistemas de IA deben establecer un sistema de responsabilidad claro y establecer una gobernanza para supervisar la operación ética.**
* **Se debe establecer políticas para que la IA pueda contribuir a los Objetivos de Desarrollo Sostenible (ODS), incluida la respuesta al cambio climático, la protección del medio ambiente y la promoción del bienestar público.**
* Descripción de desarrollo y operaciones de la IA sostenible

Para garantizar la operación ética de la tecnología de IA, es necesario fortalecer la gobernanza pública y establecer claramente sistemas de responsabilidad. El gobierno debería establecer políticas y marcos regulatorios para una IA confiable y obligar las evaluaciones del impacto ético de los sistemas de IA. Con estas medidas se debería operar un organismo supervisor independiente que supervisa y evalúa el uso de la IA en las instituciones públicas y el sector privado, y establecer un sistema que permita prevenir con antelación los riesgos de los sistemas de IA y minimizar los efectos secundarios inesperados.

Además, cuando se utilizan los sistemas de IA en el sector público, se debe garantizar la gestión transparente de los datos y la divulgación de los procesos de toma de decisiones sobre IA, y se deben establecer procedimientos de diálogo social y de debate público para que los ciudadanos y las partes interesadas puedan participar en el proceso de establecimiento de las políticas de IA. Para ello, es necesario formar una estructura cooperativa en la que la gobernanza de la IA se funciona de forma democrática y diversas partes interesadas incluidas instituciones públicas, empresas privadas, expertos en tecnología y organizaciones de la sociedad civil.

Las tecnologías de IA deben diseñarse y operarse de manera justa e inclusiva, teniendo cuidado de no exacerbar las desigualdades sociales. Al prestar servicios públicos, los sistemas de IA deben garantizar la igualdad de acceso para todos los ciudadanos, incluidos los socialmente desfavorecidos y vulnerables. Se debe establecer un sistema para realizar evaluaciones periódicas del impacto social a fin de minimizar el sesgo social en el diseño y el funcionamiento de la IA y evitar afectar negativamente a determinadas clases.

Se necesita apoyo político para que los sistemas de IA puedan contribuir a promover el interés público y crear valor social, y para este fin, se debe establecer un sistema de cooperación entre el gobierno, el sector privado, la sociedad civil y la academia para fomentar un ecosistema de IA sostenible. Son necesarios los esfuerzos de políticas a fin de evaluar el impacto de la tecnología de IA en el mercado laboral y apoyar los programas de transición laboral y *reskilling* ante los cambios de puestos del trabajo debidos a la IA.

El proceso de transformación digital del Perú requiere una gestión sistemática y una gobernanza ética para garantizar que las tecnologías de IA se desarrollen de una manera que genere valor social sin comprometer la sostenibilidad ambiental. El desarrollo y el funcionamiento de la IA deben estar en consonancia con los Objetivos de Desarrollo Sostenible (ODS), teniendo en cuenta la protección del medio ambiente, la inclusión social, el fortalecimiento de la gobernanza pública y la cooperación internacional.

Los sistemas de IA deben considerar la sostenibilidad ambiental durante todo su desarrollo y funcionamiento y, en particular, deben reflejar principios como la optimización del consumo de energía, la reducción de las emisiones de carbono y la aplicación de tecnologías verdes. El alto consumo de energía durante el entrenamiento y la distribución de modelos de IA puede tener un impacto negativo en el medio ambiente, por lo que se debe mejorar la eficiencia energética mediante el desarrollo de algoritmos eficientes y la utilización de hardware de bajo consumo desde la etapa de diseño de los sistemas de IA.

* 1. Cooperación internacional
* **Si la IA tiene un impacto más allá de las fronteras internacionales, se debe cumplir con la gobernanza global y las responsabilidades internacionales.**
* **El Perú debe cooperar con organizaciones internacionales y países vecinos para fortalecer la ética de la IA.**
* Descripción de la cooperación internacional

Para fortalecer la ética de la IA, es importante coordinar políticas con la comunidad internacional y establecer estrategias de IA que cumplan con los estándares globales.

La IA es una tecnología que opera a través de las fronteras y, como tal, es necesario garantizar la coherencia de las políticas y la interoperabilidad entre países. Al adherirse a los estándares éticos globales de IA, Perú puede fomentar un ecosistema de IA confiable y trabajar con la comunidad internacional para abordar cuestiones éticas y, al mismo tiempo, promover el avance tecnológico.

El Perú debe cumplir con las normas internacionales de IA y reflejarlas en las políticas nacionales. Para garantizar que la tecnología de IA avance de manera ética y responsable, el Perú debe participar activamente en el sistema de gobernanza global de IA a través de la cooperación con la comunidad internacional. Es necesario establecer una estrategia que refleje las tendencias internacionales de políticas de IA a través de un sistema de cooperación global para que la tecnología de IA pueda contribuir a objetivos sostenibles como responder al cambio climático, promover el bienestar público y proteger el medio ambiente.

Adoptar los principios éticos y los estándares de IA aceptados internacionalmente es esencial para que el Perú construya sistemas de IA confiables y asegure la competitividad en el mercado global de IA. Para tal fin, el Perú debe participar activamente en los debates internacionales sobre políticas de IA y mantenerse alineado con los estándares éticos globales en materia de IA.

El Perú debe trabajar estrechamente con organizaciones internacionales y organismos multilaterales de cooperación para promover la investigación y el desarrollo en IA y contribuir a reducir la brecha tecnológica. Se debe adoptar las siguientes medidas para avanzar y utilizar éticamente las tecnologías de IA:

* Participación en el foro colaborativo global para la ética y la formulación de políticas de IA
* Establecimiento de una plataforma de cooperación internacional para la investigación y la innovación en IA
* Promoción del intercambio de conocimientos y tecnologías de IA con países en desarrollo y economías emergentes
* Fortalecer la cooperación entre los sectores público y privado y activar los intercambios globales de tecnología

En un entorno donde los sistemas de IA se utilizan a nivel internacional, se debe establecer estándares legales y éticos relacionados con la transferencia transfronteriza de datos. Se debe establecer políticas que protejan la soberanía digital nacional manteniendo al mismo tiempo la coherencia con las regulaciones internacionales de protección de información personal.

Además, es necesario introducir sistemas de transparencia, explicabilidad y auditoría que cumplan con los estándares internacionales para garantizar la responsabilidad de la toma de decisiones basada en IA. Deberían aceptar los estándares jurídicos y técnicos internacionales y mejorar continuamente las políticas nacionales para que las tecnologías de IA puedan funcionar éticamente en un entorno global.

1. Ítem de verificación por requisitos básicos
	1. Ítem de verificación para la protección de los derechos humanos

|  |  |
| --- | --- |
| Número | Ítem de verificación |
| 1.1 | ¿Los sistemas de IA están diseñados y operados de manera que no violen la dignidad humana y los derechos fundamentales (por ejemplo, el derecho a la vida, la seguridad física y mental, la autonomía, etc.)? |
| 1.2 | ¿Existen procedimientos establecidos para garantizar que los sistemas de IA no discriminen injustamente por motivos de género, raza, edad, discapacidad, religión, origen socioeconómico, etc.? |
| 1.3 | ¿Se garantiza que los sistemas de IA no infrinjan la libertad de expresión, el acceso a la información y los derechos fundamentales de los ciudadanos en el proceso de prestación de servicios públicos y toma de decisiones de políticas? |
| 1.4 | ¿Existen procedimientos para realizar evaluaciones de impacto previas para determinar si es probable que los sistemas de IA infrinjan los derechos humanos y para corregir y remediar los problemas si surgen? |

<Tabla 1 – Ítem de verificación para la protección de los derechos humanos>

* Explicación de ítem de verificación

El desarrollo y el uso de sistemas de IA pueden dar lugar a violaciones de las libertades y derechos universales, inviolables y absolutos de los seres humanos. Por lo tanto, es importante que los desarrolladores y operadores de sistemas de IA evalúen y comprueben si no socavan la dignidad humana ni infringen derechos básicos como el derecho a la libertad y la igualdad, protegidos por las leyes y las normas internacionales de derechos humanos, como la Carta de las Naciones Unidas y la Declaración Universal de Derechos Humanos. En esta sección se presenta los valores que tienen que tener en cuenta desde la etapa de planificación y diseño de los sistemas de inteligencia artificial para respetar al mismo tiempo los derechos humanos y la autonomía, y se hace un llamamiento a realizar esfuerzos previos y posteriores en tal sentido.

|  |
| --- |
| Referencia-Cuestionario de evaluación del impacto sobre los derechos fundamentales (FRIA por sus siglas en inglés) de ‘ALTAI’ de la UE |
| La Unión Europea propone la Lista de evaluación de la IA confiable (ALTAI por sus siglas en inglés) como directriz para una ética de la IA confiable. ALTAI recomienda realizar una autoevaluación basada en las siguientes preguntas para evaluar de forma preventiva el impacto de los sistemas de IA en derechos fundamentales como el derecho a la igualdad.1. ¿El sistema de IA puede discriminar potencialmente de forma negativa a las personas por alguno de los siguientes motivos: sexo, raza, color, origen étnico o social, características genéticas, idioma, religión o creencias, opiniones políticas o de cualquier otro tipo, pertenencia a una minoría nacional, propiedad, nacimiento, discapacidad, edad, orientación sexual, etc.?* ¿Ha implementado procesos para evaluar y monitorear posibles discriminaciones negativas (sesgos) durante las fases de desarrollo, despliegue y utilización del sistema de IA?
* ¿Ha implementado procesos para abordar y corregir posibles discriminaciones negativas (sesgos) en el sistema de IA?

2. ¿El sistema de IA respeta los derechos del niño, por ejemplo en lo que respecta a su protección y teniendo en cuenta sus intereses superiores?* ¿Ha implementado procesos para abordar y corregir posibles daños a los niños causados ​​por el sistema de IA?
* ¿Ha implementado procesos para probar y monitorear posibles daños a los niños durante las fases de desarrollo, despliegue y utilización del sistema de IA?

3. ¿El sistema de IA protege los datos personales de las personas de acuerdo con el Reglamento General de Protección de Datos (GDPR por sus siglas en inglés)?* ¿Ha establecido procesos para evaluar en detalle la necesidad de una evaluación de impacto de la protección de datos, incluida una evaluación de la necesidad y proporcionalidad de las operaciones de tratamiento en relación con su finalidad, con respecto a las fases de desarrollo, despliegue y utilización del sistema de IA?
* ¿Ha establecido medidas previstas para abordar los riesgos, incluidas salvaguardas, medidas de seguridad y mecanismos para garantizar la protección de los datos personales con respecto a las fases de desarrollo, despliegue y utilización del sistema de IA?

4. ¿El sistema de IA respeta la libertad de expresión e información y/o la libertad de reunión y asociación?* ¿Ha puesto en marcha procesos para evaluar y supervisar posibles infracciones a la libertad de expresión e información, o a la libertad de reunión y asociación, durante las fases de desarrollo, despliegue y utilización del sistema de IA?
* ¿Ha puesto en marcha procesos para abordar y rectificar posibles infracciones a la libertad de expresión e información, o a la libertad de reunión y asociación, en el sistema de IA?
 |
| Fuente: EU, Assessment List for Trustworthy Artificial Intelligence (2020) |

<Tabla 2 - Cuestionario de evaluación del impacto sobre los derechos fundamentales (FRIA) de ‘ALTAI’ de la UE>

|  |
| --- |
| Referencia: Directrices de la OMS para el uso ético de la IA |
| La OMS ha publicado directrices éticas para el uso de la IA en la atención de la salud y sugiere los siguientes principios y consideraciones para los desarrolladores de tecnología de IA en la atención de la salud con el fin de garantizar la autonomía humana (preservación y mejora de la autonomía humana). La tecnología de IA en la atención de la salud no debe reemplazar a los profesionales médicos, sino que debe mejorar y empoderar la toma de decisiones humana.- El juicio humano debe participar a la hora de recomendar tratamientos o predecir enfermedades utilizando tecnología de IA.- El Departamento de Salud debería especificar los tipos de información que utilizan los médicos para emitir juicios independientes sobre los resultados de la IA.- Se debe informar a los pacientes que las recomendaciones se basan en tecnología de IA y se debe brindar información significativa y clara para la toma de decisiones del paciente. |
| Fuente: WHO, Ethics and governance of artificial intelligence for health: WHO guidance(2021), p143 |

<Tabla 3 – Directrices de la OMS para el uso ético de la IA>

* 1. Ítem de verificación para la protección de información personal y datos

|  |  |
| --- | --- |
| Número | Ítem de verificación |
| 2.1 | ¿El sistema de IA cumple con las leyes de privacidad peruanas y los estándares internacionales? |
| 2.2 | ¿Tiene el sistema de IA una base legal clara (por ejemplo, consentimiento del usuario, contrato, obligación legal, etc.) para recopilar, procesar y almacenar información personal? |
| 2.3 | ¿Existen procedimientos establecidos para permitir que los ciudadanos que utilizan servicios de IA ejerzan su derecho a acceder, corregir o eliminar su información personal (incluido el derecho al olvido)? |
| 2.4 | ¿El sistema de IA tiene un proceso de gestión de seguridad para evitar fugas de datos, mal uso, etc.? |
| 2.5 | ¿Los datos recopilados y procesados ​​por sistemas de IA están protegidos contra el acceso no autorizado o la falsificación? |

<Tabla 4 – Ítem de verificación para la protección de información personal y datos>

* Explicación de ítem de verificación

Las grandes cantidades de datos que se utilizan para desarrollar y utilizar la IA pueden contener diversos tipos de información personal y de privacidad sensible. Los datos pueden incluir información personal recopilada en línea, junto con información personal estructurada. En general, es difícil para los proveedores de información y los usuarios saber cómo se procesa su información personal, y si esta se vulnera o filtra, puede tener un impacto negativo en la vida de las personas y la sociedad, por lo que se están realizando esfuerzos para proteger la privacidad en el campo de la IA.

Los sistemas de IA se desarrollan utilizando datos, y los datos recopilados durante el proceso de operación se utilizan nuevamente para desarrollar o mejorar el sistema de IA. Por lo tanto, para asegurar la confianza en los sistemas de IA, es importante utilizar los datos para el propósito previsto y gestionar la calidad de los datos.

Esta sección verifica el cumplimiento de la Ley de Protección de Información Personal y los esfuerzos para proteger la privacidad de las personas por parte de entidades que desarrollan u operan sistemas de IA. Además, al controlar tareas como la recopilación y el procesamiento de datos necesarios en las etapas de desarrollo y operación de los sistemas de IA, se gestiona la calidad y el riesgo de los datos y se verifica los esfuerzos internos para evitar que los datos se utilicen para fines distintos a los previstos.

|  |
| --- |
| Referencia - Cuestionario de evaluación de IA de ‘ALTAI’ de la UE |
| ALTAI de la UE presenta las siguientes evaluaciones necesarias para garantizar el cumplimiento de los principios de protección de la privacidad y gobernanza de datos:1. Evaluación de la privacidad→ ¿Se ha tenido en cuenta el impacto del sistema de IA en el derecho a la privacidad, el derecho a la integridad física, mental y/o moral y el derecho a la protección de datos?→ Dependiendo del caso de uso, ¿se han establecido mecanismos que permitan señalar cuestiones relacionadas con la privacidad que afecten al sistema de IA?2. Evaluación de la gobernanza de datos→ ¿Su sistema de IA se desarrolló utilizando o procesando datos personales (incluidas categorías especiales de datos personales)?→ ¿Han cumplido con las contramedidas obligatorias de acuerdo con el Reglamento General de Protección de Datos (RGPD) o un equivalente de un país no europeo?• Las medidas abarcan la inclusión de Evaluación de Impacto de la Protección de Datos (EIPD); la designación de un delegado de protección de datos (DPD) y su inclusión en las primeras etapas de desarrollo, adquisición o uso de sistemas de IA; mecanismos de supervisión del procesamiento de datos (incluida la limitación del acceso a personal calificado, mecanismos para registrar el acceso a los datos y realizar modificaciones); medidas para lograr la privacidad por diseño y la encriptación, la seudonimización y la anonimización; minimización de datos, en particular datos personales; si han implementado el derecho a retirar el consentimiento, el derecho a objetar y el derecho a ser olvidado; si han considerado las implicaciones para la privacidad y la protección de datos personales recopilados, generados y procesados a lo largo del ciclo de vida del sistema de IA.→ ¿Los datos de entrenamiento no personales u otros datos no personales procesados del sistema de IA han considerado las implicaciones de los datos personales y la protección de datos?→ ¿Ha adaptado el sistema de IA a los estándares relevantes (por ejemplo, la Organización Internacional de Normalización (ISO por sus siglas en inglés) e Instituto de Ingenieros Eléctricos y Electrónicos (IEEE por sus siglas en inglés) o a protocolos ampliamente adoptados para la gestión y gobernanza de datos (diarios)? |
| Fuente: EU, Assessment List for Trustworthy Artificial Intelligence (2020) |

<Tabla 5 – Cuestionario de evaluación de IA de ‘ALTAI’ de la UE>

|  |
| --- |
| Referencia – “Corea” Directrices de gestión de la calidad de los datos para el aprendizaje de la IA |
| Esta directriz V1.0 define la organización, los procedimientos, los estándares de calidad, los métodos de control de calidad y las actividades necesarias para asegurar datos de aprendizaje de alta calidad durante el proceso de desarrollo de IA, y respalda una serie de actividades para inspeccionar y tomar medidas, y presenta los siguientes principios de control de calidad de datos.- La gestión de la calidad de los datos para el aprendizaje de la IA debe garantizar la calidad durante todo el ciclo de vida de los datos. - La gestión de la calidad de los datos para el aprendizaje de la IA debe ser capaz de mejorar continuamente la calidad. - Para gestionar la calidad de los datos para el aprendizaje de IA, se debe establecer una organización y llevarla a cabo de acuerdo con los roles y responsabilidades establecidos.- Para gestionar la calidad de los datos para el aprendizaje de IA, se debe asegurar un sistema de soporte como la capacitación en gestión de calidad para que la organización pueda asegurar capacidades de gestión de calidad. |
| Fuente: “Directrices de gestión de la calidad de datos de aprendizaje de la IA” (2021) de NIA |

<Tabla 6 – Directrices de gestión de la calidad de los datos para el aprendizaje de la IA>

* 1. Ítem de verificación para mejorar la publicidad y el valor social

|  |  |
| --- | --- |
| Número | Ítem de verificación |
| 3.1 | ¿Los sistemas de IA están diseñados y operados para promover el interés público y mejorar la calidad de vida de los ciudadanos? |
| 3.2 | ¿Los sistemas de IA desempeñan un papel de reflejar las opiniones de los ciudadanos en la administración pública y promover su participación en los procesos de formulación de políticas? |
| 3.3 | ¿Se realizan periódicamente evaluaciones de impacto ético para garantizar que los sistemas de IA mantengan su valor social en la administración y los servicios públicos? |

<Tabla 7 – Ítem de verificación para mejorar la publicidad y el valor social>

* Explicación de ítem de verificación

Existe un creciente consenso internacional respecto de que el desarrollo de la IA debe orientarse a contribuir a la seguridad y el bienestar de toda la humanidad. Las normas nacionales e internacionales, incluida la Constitución del Perú, la Carta de las Naciones Unidas y los Convenios Fundamentales de la Organización Internacional del Trabajo, buscan asegurar la prosperidad humana y el bienestar económico desde la perspectiva de la publicidad social. La publicidad se puede garantizar respetando y cumpliendo las normas y valores perseguidos por la comunidad internacional en el campo de lA.

La IA debe utilizarse en el sector público para mejorar la calidad de vida de los ciudadanos y contribuir a la creación de valor social. Es esencial garantizar el interés público para que la tecnología de IA no se utilice de forma abusiva en beneficio de individuos o grupos específicos y, para ello, es necesario comprobar si los sistemas de IA se desarrollan y operan teniendo como prioridad el interés público.

Se necesita evaluar si los sistemas de IA están proporcionando beneficios reales a los ciudadanos en la administración y los servicios públicos. El tema principal de la revisión es si la IA se está utilizando de una manera que mantenga la equidad en el proceso de establecimiento e implementación de políticas públicas y mejore la eficiencia de la administración pública. Pueden surgir problemas de publicidad cuando los sistemas de IA no están diseñados para mejorar los servicios públicos o cuando los beneficios se concentran sólo en ciertos grupos.

La IA debería desempeñar un papel de reflejar las opiniones de los ciudadanos en la administración pública y promover procedimientos democráticos en la toma de decisiones políticas. La toma de decisiones basada en IA debe estructurarse para comunicarse con los ciudadanos, y el proceso de toma de decisiones debe ser transparente y abierto al público para que los ciudadanos puedan confiar en él.

Es necesario realizar evaluaciones de impacto periódicas para garantizar que la IA opere en el sector público de manera ética y sostenible. Debería comprobar si los sistemas de IA funcionan de forma que mantengan la equidad social en la administración y los servicios públicos y garanticen que todos los ciudadanos, incluidos los desfavorecidos, reciban beneficios iguales.

Cuando se aplican sistemas de IA en el sector público, se debe divulgar información para garantizar la transparencia en el proceso de toma de decisiones políticas y permitir que los ciudadanos comprendan plenamente el papel y el impacto de la IA. Es importante verificar si la IA se está utilizando de una manera que refuerce los procesos democráticos en los servicios públicos, y deben existir sistemas que permitan a los ciudadanos acceder y proporcionar retroalimentación sobre el proceso de toma de decisiones de la IA.

|  |
| --- |
| Referencia: Cuestionario similar en herramientas de autoevaluación en el extranjero que representa la publicidad |
| (Caso del cuestionario de autoevaluación de ética de la IA de *Digital Dubai*)→ Los desarrolladores de sistemas de IA que respaldan la toma de decisiones críticas deben involucrar a expertos con experiencia en ciencias sociales, políticas u otros campos en el proceso de desarrollo para preparar una amplia evaluación del impacto social de su trabajo. |
| Fuente: Digital Dubai, ‘AI ETHICS GUIDELINES’(2020) |

<Tabla 8 – Cuestionario similar en herramientas de autoevaluación en el extranjero que representa la publicidad>

|  |
| --- |
| Referencia: “Directrices éticas para una inteligencia artificial fiable” de la Unión Europea (UE) |
| El Grupo de Expertos sobre Inteligencia Artificial de la Unión Europea, un grupo de expertos independiente creado por el Parlamento Europeo, comenzó a redactar un conjunto de normas éticas de IA confiable en 2018 y desde entonces ha celebrado audiencias públicas con más de 500 participantes y docenas de talleres y conferencias para proponer directrices para normas éticas de IA que puedan aplicarse en todos los estados miembros de la UE.1.6 Bienestar social y ambiental- La exposición ubicua a los sistemas sociales de IA en todas las esferas de nuestra vida (sea en ámbitos como la educación, el trabajo, el cuidado o el entretenimiento) pueden alterar nuestra concepción de la acción social o afectar a nuestras relaciones y vínculos sociales.→ Por lo tanto, será necesario tener en cuenta y llevar a cabo un seguimiento exhaustivo de los efectos de esos sistemas- Además, se debería evaluar el impacto del sistema de IA en la etapa del desarrollo y la utilización teniendo en cuenta sus repercusiones amplias sobre la democracia → El uso de sistemas de IA debería ser objeto de un estudio pormenorizado, sobre todo en situaciones relacionadas con el proceso democrático, no solo en el terreno de la adopción de decisiones políticas sino también en contextos electorales. |
| Fuente: EU, ‘Ethics guidelines for trustworthy AI’(2019) |

<Tabla 9 – “Directrices éticas para una inteligencia artificial fiable” de la Unión Europea (UE)>

|  |
| --- |
| Referencia: “Principios de la ética de la IA” de Australia |
| El Departamento de Industria, Ciencia, Energía y Recursos de Australia presenta un marco integral de la ética de IA para ayudar a las empresas privadas y a los gobiernos a diseñar, desarrollar y utilizar la IA de manera responsable, proporcionando principios como orientación.En el marco de la estrategia nacional del gobierno australiano para el desarrollo responsable e inclusivo de la IA, desarrollaron principios éticos relacionados con la IA y presentaron procedimientos para su aplicación, pero los principios no son de aplicación obligatoria sino voluntaria.El principio de contestabilidad. Cuando los sistemas de IA tienen un impacto significativo en las personas, las comunidades, los grupos sociales o el medio ambiente, debería ser posible cuestionar los resultados de los sistemas de IA y el uso de la IA mediante procedimientos adecuados.→ ‘Impacto significativo’ se refiere a todos los impactos que pueden ocurrir en los resultados de IA, y en el proceso de uso de IA, y asimismo, su definición puede variar dependiendo del proceso de aplicación, el contexto de aplicaciones y el impacto general de la IA.· Este principio es un elemento importante para garantizar la confianza pública en la IA al informar que se puede corregir (o compensar) el daño cuando los sistemas de IA tienen un impacto negativo en diversas áreas de la sociedad.· En particular, es necesario hacer más hincapié en estos principios en relación con los grupos vulnerables dentro de la sociedad.· Mientras tanto, el acceso a la información de algoritmos de IA o un sistema de monitoreo efectivo pueden utilizarse como alternativas para fortalecer el principio de objeción. |
| Fuente: DISER, ‘Australia's AI Ethics Principles' (2019) |

<Tabla 10 – “Principios de la ética de la IA” de Australia>

* 1. Ítem de verificación de accesibilidad e inclusión social

|  |  |
| --- | --- |
| Número | Ítem de verificación |
| 4.1 | ¿Están los sistemas de IA diseñados para que los grupos socialmente vulnerables, como los discapacitados, los ancianos y las personas de bajos ingresos puedan tener el acceso igualitario a los servicios públicos? |
| 4.2 | ¿Se aplican los sistemas de IA de manera justa a todos los ciudadanos sin discriminación en el proceso de prestación de servicios públicos? |
| 4.3 | ¿El sistema de IA tiene una interfaz fácil de usar y un diseño intuitivo que tiene en cuenta los antecedentes técnicos y el nivel de capacidad digital del usuario? |
| 4.4 | Al expandir los sistemas de IA en el sector público, ¿están implementando programas de educación y capacitación dirigidos a la clase de baja accesiblidad? |
| 4.5 | ¿El sistema de IA refleja políticas de administración pública inclusivas sin excluir a ciertas clases sociales o regiones? |

<Tabla 11 – Ítem de verificación de accesibilidad e inclusión social>

* Explicación de ítem de verificación

Los sistemas de IA pueden repetir y reproducir estereotipos y prejuicios inherentes a la sociedad existente, lo que puede causar problemas en términos de sesgo, idoneidad y equidad. Esto se debe a que el sesgo humano puede reflejarse en los datos necesarios para desarrollar y operar sistemas de IA. En un esfuerzo por minimizar la discriminación resultante del uso de sistemas de IA, es importante no sólo reducir el sesgo reflejado en los datos o garantizar la accesibilidad a la tecnología o los servicios para los socialmente desfavorecidos, sino también escuchar, revisar, evaluar y reflejar diversas opiniones desde las etapas de desarrollo y operación. En esta sección se evalúa los esfuerzos para establecer evaluaciones y procedimientos que consideren a los diversos miembros de la sociedad y minimicen los sesgos y la discriminación.

Los avances en la tecnología de IA deben diseñarse para beneficiar a todos los ciudadanos por igual, y se debe garantizar la accesibilidad e inclusión integrales para evitar marginar o discriminar a determinados grupos. Se necesita comprobar si los sistemas de IA funcionan de forma que ofrezcan igualdad de oportunidades a las personas socialmente desfavorecidas, como los discapacitados, los ancianos y las personas con bajos ingresos, y reduzcan la brecha tecnológica.

Los sistemas de IA deben diseñarse para garantizar la igualdad de acceso a los servicios públicos y a la administración digital para las personas en riesgo de ser marginadas en entornos tecnológicos, como las personas con discapacidad, los ancianos y las personas de bajos ingresos. Para tal fin, se deben aplicar principios de Diseño Universal con el objetivo de que distintos usuarios puedan utilizar fácilmente los sistemas de IA.

Los sistemas de IA tienen el potencial de perjudicar a determinados grupos debido a sesgos inherentes en sus datos de entrenamiento. Cuando se utiliza IA en servicios públicos, se debe analizar el impacto en los grupos socialmente desfavorecidos y vulnerables y es necesario establecer procedimientos para evitar resultados discriminatorios no deseados.

En el proceso de utilización de la tecnología de IA, se debe tener cuidado para que los usuarios que carecen de conocimientos técnicos o capacidades digitales no queden excluidos. Al operar sistemas de IA en el sector público, es necesario aplicar una interfaz de usuario (IU) intuitiva a fin de que los ciudadanos puedan usarla fácilmente y se deben proporcionar materiales de orientación fáciles de entender. Especialmente para regiones o poblaciones con baja accesibilidad digital, es importante diseñar para que sea compatible con tecnología de asistencia o brindar enfoques alternativos que puedan vincularse con servicios fuera de línea. Los sistemas públicos de IA deben ir más allá de simplemente aumentar la eficiencia, y funcionar como tecnologías inclusivas que todos los ciudadanos puedan utilizar por igual.

Para evitar que la brecha digital se amplíe en el proceso de la difusión de la tecnología de IA, es necesario implementar programas de educación sobre el uso de IA y alfabetización digital dirigidos a las clases de bajo acceso. Cuando las instituciones públicas introducen sistemas de IA, deben brindar educación a los ciudadanos para que puedan usarlos de manera efectiva, y, asimismo, establecer un sistema de apoyo activo a fin de que los grupos socialmente desfavorecidos y vulnerables puedan beneficiarse de la tecnología de IA.

Además, en el proceso de establecer políticas y regulaciones relacionadas con la IA, se debe establecer una estructura para recopilar opiniones de diversas partes interesadas y se debe realizar una revisión continua para garantizar que ningún grupo quede excluido durante el proceso de desarrollo y operación de la tecnología.

|  |
| --- |
| Referencia: “Principios de IA” de Google |
| Es una orientación sobre consideraciones éticas para quienes desarrollan aplicaciones de IA en Google- Presenta el principio central de no crear ni promover prejuicios injustos y prohíbe el trato injusto basado en raza, género, nacionalidad, ingresos, orientación sexual, capacidad, creencias políticas o religiosas. |
| Fuente: Google, ‘Artificial Intelligence at Google: Our Principles’ (2018) |

<Tabla 12 – Principios de IA de Google>

|  |
| --- |
| Referencia: Directrices sobre el uso de IA y algoritmos de la Comisión Federal de Comercio de EEUU (FTC por sus siglas en inglés) |
| En abril de 2020, la Comisión Federal de Comercio de Estados Unidos (FTC) emitió una directriz que tiene cinco propuestas clave para orientar la dirección de regulaciones de la IA y los algoritmos y alentar a las empresas a responder en forma adecuada.- 1) Garantizar la transparencia, 2) Explicar la toma de decisiones con la utilización de IA y algoritmos, 3) Garantizar la equidad en la toma de decisiones, 4) Garantizar la robustez y la validez empírica de los modelos de datos, 5) Mantener la responsabilidad por el cumplimiento de regulaciones, la ética, la imparcialidad y el mantenimiento de la responsabilidad de lucha contra la discriminación.El tercer punto es garantizar la equidad en la toma de decisiones de la siguiente manera:- Garantizar la equidad en la toma de decisiones1) No se debe discriminar a determinados grupos o clases y realizar pruebas rigurosas antes y después de utilizar el algoritmo.2) Se debe revisar la equidad de los resultados.3) A los consumidores se les debe garantizar el derecho a acceder a la información y la oportunidad de corregirla. |
| Fuente: FTC, ‘Using Artificial Intelligence and Algorithms’ (2020) |

<Tabla 13 – Directrices sobre el uso de IA y algoritmos de la Comisión Federal de Comercio de EEUU (FTC por sus siglas en inglés)>

|  |
| --- |
| Referencia – Caso de inclusión social |
| Puede que determinados grupos sociales se enfrenten a la desigualdad de oportunidades cuando la IA se utiliza para fines de contratación o calificación crediticia, proporcionando resultados sesgados para ciertas razas, géneros y grupos de edad Existe la preocupación de que ciertos grupos puedan experimentar desigualdad de oportunidades cuando se utiliza inteligencia artificial.(Investigación relacionada) Más del 90% de las personas expuestas a anuncios de empleo para trabajadores madereros son hombres y más del 70% son blancos. Más del 65% de las vacantes de empleo para limpiadores son para mujeres y más del 75% son para personas de raza negra. Dado que los investigadores publicaron los anuncios en las mismas condiciones, dedujeron que los resultados se debían a la focalización de anuncios del algoritmo de la plataforma. Además, señalaron que las ofertas de trabajo publicadas a través de anuncios de Facebook están más expuestas a ciertas razas o géneros según el tipo de trabajo (Ali et al., 2019).(Caso 1) En enero de 2021, el chatbot de IA Iruda fue suspendido del servicio debido a problemas como discursos de odio contra ciertos grupos como discapacitados, minorías sexuales y razas, y expresiones de prejuicio de género debido al sesgo en los datos que aprendió.(Caso 2) El gobierno holandés desarrolló y utilizó un sistema de detección de riesgos basado en IA (SyRi) para acabar con el fraude en las prestaciones sociales y el fraude fiscal, pero fue retirado en febrero de 2020 tras la sentencia judicial que planteó inquietudes sobre su uso discriminatorio contra minorías y grupos vulnerables. |
| Fuente: Ali, M., et al., ‘Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes’, Proceedings of the ACM on Human-Computer Interaction, 3(CSCW) (2019) |

<Tabla 14 – Caso de inclusión social >

* 1. Ítem de verificación para garantizar la explicabilidad

|  |  |
| --- | --- |
| Número | Ítem de verificación |
| 5.1 | ¿Los datos, algoritmos y criterios de toma de decisiones que utilizan los sistemas de IA son transparentes y fácilmente comprensibles para los ciudadanos y las partes interesadas? |
| 5.2 | Cuando un sistema de IA toma una decisión, ¿existen procedimientos establecidos para proporcionar a los humanos información suficiente para comprender y revisar esa decisión? |
| 5.3 | ¿Están los sistemas de IA tomando medidas técnicas e institucionales para abordar el problema de la “caja negra” (el problema de la opaca estructura interna del algoritmo)? |
| 5.4 | Si la decisión de un sistema de IA se basa en datos específicos, ¿se registra ese proceso y se pone a disposición para su revisión? |

<Tabla 15 – Ítem de verificación para garantizar la explicabilidad>

* Explicación de ítem de verificación

La explicabilidad de los sistemas de IA es un principio importante para garantizar la confiabilidad de la IA. A diferencia de otras tecnologías, no es técnicamente fácil rastrear el proceso operativo y los principios de los sistemas de IA, como los algoritmos de aprendizaje profundo. Sin embargo, los principales principios éticos nacionales e internacionales sobre IA suelen enfatizar la explicabilidad puesto que la IA reemplaza las decisiones humanas o influye en ellas.

A medida que los sistemas de IA se vuelven más frecuentes en los servicios públicos y las decisiones políticas, la explicabilidad se convierte en un elemento esencial para permitir que los ciudadanos y las partes interesadas comprendan cómo funciona la IA y el proceso de toma de decisiones. Es muy importante verificar la transparencia y explicabilidad de los sistemas de IA, ya que la toma de decisiones opaca de la IA puede generar problemas como menor confianza de los ciudadanos, mayor injusticia, sesgo automatizado y errores.

En esta sección se verifica si se están realizando esfuerzos por proporcionar la información necesaria para garantizar el derecho del usuario a saber.

|  |
| --- |
| Referencia - Algoritmo de “caja negra” |
| Debido al avance de la tecnología de IA, la demanda de transparencia en los sistemas de IA está aumentando. Esto se refiere a que no se puede averiguar qué está haciendo el algoritmo internamente a parte de los datos de entrada y salida.- Muchas empresas nacionales e internacionales están utilizando entrevistas de IA en el proceso de contratación, pero surgen dudas sobre si el sistema de IA respalda una evaluación justa.- Los “Principios éticos para una IA fiable” de la UE destacan que las decisiones tomadas por los sistemas de IA deben ser comprensibles y rastreables para los humanos.- Las directrices de ética de la IA del Ministerio de Empleo y Economía de Finlandia entienden la transparencia como la divulgación del propósito para el cual se recopilan los datos y qué objetivos apoya y determina el algoritmo. |

<Tabla 16 – Algoritmo de “caja negra”>

|  |
| --- |
| Referencia – La IA explicable |
| Existe una demanda cada vez mayor de que cuando las partes interesadas especificadas en la ley solicitan una explicación de la IA, los valores de entrada, los procesos internos, los tipos y los estados de la IA deben explicarse de una manera que el solicitante pueda comprender, y se están realizando esfuerzos para garantizar la explicabilidad de la IA.- El "Código de Ética de Inteligencia Artificial" del Ministerio de Salud y Asistencia Social del Reino Unido establece estándares estrictos que requieren que el desarrollo de algoritmos demuestre cómo aprenden, revele qué tipos de algoritmos se desarrollaron y utilizaron, a qué revisión ética se sometieron los datos utilizados y cómo los resultados pueden integrarse en los servicios de salud. |

<Tabla 17 – La IA explicable>

|  |
| --- |
| Referencia – Cuestiones relacionadas con el principio de transparencia |
| Centro de información sobre privacidad electrónica (EPIC por sus siglas en inglés), una organización estadounidense sin fines de lucro dedicada a la privacidad, presentó una petición ante la Comisión Federal de Comercio (FTC) alegando que HireVue, una solución de reclutamiento basada en IA, no cumple con los estándares nacionales e internacionales de transparencia, equidad y responsabilidad (Forbes, febrero de 2020).- En particular, plantea preocupaciones sobre la tecnología de reconocimiento facial y los algoritmos no probados utilizados en los procesos de entrevistas de IA.- HireVue contrató a una empresa de auditoría externa para realizar una auditoría del algoritmo\* y anunció que el algoritmo no estaba sesgado, pero luego eliminó la función principal del software, que era analizar las expresiones faciales de los entrevistados basándose en entrevistas grabadas.\* La auditoría del algoritmo fue realizada por una organización externa (O’Neil Risk Consulting & Algorithmic Auditing), pero los resultados de la auditoría no fueron publicados. |

<Tabla 18 – Cuestiones relacionadas con el principio de transparencia>

* 1. Ítem de verificación para el aseguramiento de la fiabilidad y seguridad técnica

|  |  |
| --- | --- |
| Número | Ítem de verificación |
| 6.1 | ¿Son los sistemas de IA resilientes y seguros frente a ciberataques como piratería, envenenamiento de datos y evasión de modelos? |
| 6.2 | ¿El sistema de IA cumple con los procesos de gestión de la calidad de los datos (frecuencia, precisión, integridad, etc.) y tiene procedimientos establecidos para evaluar y mejorar continuamente el rendimiento del modelo? |
| 6.3 | ¿El sistema de IA tiene procesos establecidos para detectar y abordar falsos positivos y predicciones inexactas? |
| 6.4 | ¿Tiene sistemas de respaldo o procesos alternativos implementados para minimizar el impacto de fallas o mal funcionamiento del sistema de IA? |
| 6.5 | ¿Existen procedimientos regulares de inspección y actualización para mantener la fiabilidad y seguridad de los sistemas de IA? |

<Tabla 19 – Ítem de verificación para el aseguramiento de la fiabilidad y seguridad técnica>

* Explicación de ítem de verificación

La seguridad de un sistema de IA incluye no solo el concepto estrecho de seguridad que requiere que funcione de manera estable durante todo su uso, sino también el concepto amplio de tener la capacidad de hacer inferencias y juicios válidos incluso en datos, entornos y situaciones problemáticas que difieren de la experiencia de aprendizaje. Por lo tanto, debe funcionar de forma estable incluso en situaciones anormales, como ataques externos intencionales o errores importantes del sistema.

Los sistemas de IA pueden ser vulnerables a ciberataques como piratería, envenenamiento de datos y evasión de modelos, y si los sistemas de IA funcionan en la nube o están vinculados a una infraestructura de Internet de las cosas (IoT), corren un mayor riesgo de quedar fácilmente expuestos a amenazas de seguridad externas. En particular, en el caso de los sistemas de IA directamente relacionados con la vida humana, como los vehículos autónomos, los sistemas de transporte de IA y los sistemas de IA médica, existe una alta posibilidad de que se produzcan accidentes fatales debido a vulnerabilidades de seguridad.

Si se produce una brecha entre los datos con los que un modelo de IA fue entrenado y el entorno operativo real, esto puede provocar una degradación del rendimiento y errores en la toma de decisiones. Para evitar esto, se necesita evaluar continuamente el rendimiento de los modelos de IA y establecer procesos de detección y corrección de errores.

Además, cuando los sistemas de IA toman decisiones incorrectas o brindan predicciones inexactas, se requiere un proceso de respuesta, y cuando los sistemas de IA realizan una toma de decisiones automatizada, se requieren revisiones y acciones correctivas para detectar y corregir falsos positivos.

Por ejemplo, se necesita establecer un sistema de respuesta para evitar que juicios incorrectos en las calificaciones crediticias, la selección de personal y los sistemas de automatización del servicio público basadas en la IA se conduzcan a violaciones de los derechos individuales.

Esta sección se identifica los tipos de peligros potenciales o su probabilidad de ocurrencia y se verifica los esfuerzos para garantizar la fiabilidad técnica y las medidas de seguridad.

|  |
| --- |
| Referencia - Establecimiento de vulnerabilidades de seguridad y contramedidas para sistemas de IA |
| Los sistemas de IA utilizan activamente computadoras y redes de comunicación junto con la infraestructura de Internet de las Cosas, lo que los hace vulnerables a ataques cibernéticos externos y fácilmente expuestos a riesgos.- Los vehículos autónomos puedan verse afectados en el control remoto y la conducción autónoma por ataques cibernéticos externos.- El secuestro y el terrorismo de personalidades importantes de un país podrían ser posibles con el uso de vehículos autónomos- Los sistemas de transporte con IA deben proporcionar funciones de seguridad ya que, si fallan debido a piratería, pueden provocar accidentes graves.- Chrysler retira 1,4 millones de vehículos tras detectar vulnerabilidades en los sistemas de a bordo que podrían permitir ataques para controlar el funcionamiento y frenado del vehículo (’15) |

<Tabla 20 – Establecimiento de vulnerabilidades de seguridad y contramedidas para sistemas de IA>

|  |
| --- |
| Referencia – Casos relacionados con la ‘seguridad’ |
| A medida que se expanden los campos en los que se aplica la tecnología de IA, surgen problemas relacionados con la seguridad.(Caso 1) Accidente: Los fabricantes mundiales de automóviles, incluidos Tesla, Uber, Google, GM y Audi, están investigando e invirtiendo en el desarrollo de automóviles autónomos con funciones de piloto automático, pero los accidentes causados ​​por el piloto automático continúan ocurriendo.- Existe controversia en cada caso sobre si la causa del accidente fue la negligencia del vehículo autónomo o la negligencia del otro vehículo, del conductor o del peatón, pero son necesarias medidas legales, institucionales y éticas para garantizar la seguridad.- También son necesarios esfuerzos para mejorar la seguridad considerando no sólo la robustez de la tecnología sino también los patrones de comportamiento de los usuarios que la utilizan.(Caso 2) Lesiones físicas: Un caso en el que un vehículo autónomo de Uber, durante una prueba en Arizona, EE.UU., atropelló y mató a un peatón que cruzaba la calle de forma imprudente fuera de un cruce peatonal (marzo de 2018) |

<Tabla 21 – Casos relacionados con la ‘seguridad’>

* 1. Ítem de verificación para el desarrollo y funcionamiento de la IA sostenible

|  |  |
| --- | --- |
| Número | Ítem de verificación |
| 7.1 | ¿Las instituciones públicas que operan sistemas de IA han establecido un sistema claro de responsabilidad y, una gobernanza para cumplir con las responsabilidades legales y éticas en el proceso de utilización de la IA? |
| 7.2 | ¿Se ha establecido una institución de gobernanza que puede abordar problemas éticos inesperados en el proceso de la operación y el desarrollo de sistemas de IA (por ejemplo, sesgo, injusticia, violaciones de la privacidad, etc.)? |
| 7.3 | ¿Los sistemas de IA están diseñados para promover el bienestar público y la inclusión social y funcionan de manera que fomenten la participación de la comunidad y los ciudadanos? |
| 7.4 | ¿Se están diseñando los sistemas de IA de manera sostenible para generar ahorros de costos y mejoras en la eficiencia operativa a largo plazo en el sector público? |
| 7.5 | ¿Los sistemas de IA están diseñados y respaldados por políticas para contribuir a la respuesta al cambio climático, la protección del medio ambiente y la conservación de los ecosistemas? |

<Tabla 22 – Ítem de verificación para el desarrollo y funcionamiento de la IA sostenible>

* Explicación de ítem de verificación

Cuando las instituciones públicas operan sistemas de IA, en términos de la sostenibilidad, es importante establecer un sistema de responsabilidad claro y una gobernanza sistemática para implementar las responsabilidades legales y éticas.

Para que la tecnología de IA asegure la fiabilidad en el proceso de los servicios públicos y de la toma de decisiones políticas, es necesario garantizar la responsabilidad y la transparencia de la entidad operadora. Para ello, es necesario definir claramente los roles y responsabilidades de cada organización y establecer procedimientos para gestionar los riesgos que puedan surgir durante el proceso de operación. Además, se debe realizar un seguimiento y una supervisión continuos para que el uso de los sistemas de IA no viole los derechos de los ciudadanos y el interés público. Las instituciones públicas necesitan aplicar normas y estándares éticos que fortalezcan la responsabilidad de las operaciones de IA y evaluarlas y mejorarlas continuamente.

Existe la posibilidad de que surjan problemas éticos inesperados, como sesgos, injusticias y violaciones de la privacidad, durante el desarrollo y el funcionamiento de los sistemas de IA. Cuando se utiliza la IA en el sector público, es esencial establecer una gobernanza ética que pueda detectar estos problemas de forma temprana y responder de manera eficaz. Es necesario realizar revisiones éticas previas y operar un proceso de monitoreo y mejora continua para garantizar que los modelos de IA no tengan un impacto perjudicial en determinadas clases o grupos. Además, se requiere esfuerzos para minimizar los problemas éticos reflejando diversos antecedentes sociales y estableciendo procedimientos para recopilar opiniones de las partes interesadas desde la etapa de diseño de los sistemas de IA.

La tecnología de IA debe utilizarse para promover el bienestar público y mejorar la inclusión social. Cuando las instituciones públicas utilizan IA, deben aplicar un enfoque integral para que determinadas clases sociales o regiones no queden excluidas de los beneficios tecnológicos, y es importante que los sistemas de IA en la administración pública se operen de una manera que promueva la participación ciudadana y garantice la equidad. Para ello, es necesario diseñar servicios públicos basados ​​en IA que tengan en cuenta la accesibilidad de las distintas clases con el fin de que puedan contribuir a resolver la brecha digital, y se debe diseñar un sistema de apoyo para que los grupos socialmente desfavorecidos y vulnerables también puedan utilizar la tecnología de IA. Además, los proyectos de IA del sector público deben operar reflejando las opiniones de diversas partes interesadas, como el gobierno, la sociedad civil, el mundo académico y las empresas.

Para que la tecnología de IA funcione de manera estable en el sector público a largo plazo, la reducción de costos y la mejora de la eficiencia que tengan en cuenta la sostenibilidad son esenciales. Los sistemas de IA deben diseñarse y respaldarse con políticas que consideren el RdI (Rendimiento de la Inversión) a largo plazo, en lugar de centrarse únicamente en el rendimiento a corto plazo. Esto requiere esfuerzos para la mejora de gestión de datos, la optimización de algoritmos y la eficiencia de la infraestructura para que los sistemas de IA puedan mejorar la calidad de los servicios públicos y al mismo tiempo optimizar los costos de mantenimiento y operación. Además, es importante establecer una estrategia de inversión pública que promueva la investigación, el desarrollo y la innovación para que los servicios públicos basados ​​en IA puedan desarrollarse continuamente, y construir un modelo operativo considerando la sostenibilidad económica.

Las tecnologías de IA pueden realizar cálculos a gran escala y potencialmente generar un alto consumo de energía y emisiones de carbono, que son consideraciones importantes en términos de respuesta al cambio climático y protección del medio ambiente. Por lo tanto, se deben estipular políticas para utilizar centros de datos ecológicos y modelos de IA de bajo consumo eléctrico para minimizar el impacto de los sistemas de IA en el medio ambiente. Se necesitan esfuerzos para monitorear el consumo de energía y las emisiones de carbono de los sistemas de IA e implementar algoritmos de ahorro de energía. Además, es importante establecer políticas de IA respetuosas con el medio ambiente y aplicar estrategias operativas de IA que tengan en cuenta la sostenibilidad en el sector público para que la tecnología de IA pueda contribuir a responder al cambio climático y preservar el ecosistema.

|  |
| --- |
| Referencia – Cuestionario de la evaluación de IA de ‘ALTAI’ de UE |
| La ALTAI de la Unión Europea obliga a realizar una evaluación de la responsabilidad, que incluye:→ La responsabilidad significa que deben existir mecanismos para garantizar la responsabilidad por el desarrollo, el despliegue o la utilización de los sistemas de IA.→ Esto está estrechamente relacionado con la gestión de riesgos, que identifica y mitiga los riesgos de una manera transparente que pueda ser explicada y monitoreada por terceros, y debe establecerse un mecanismo de responsabilidad para corregir adecuadamente cualquier impacto indebido o adverso que pueda surgir. |
| Fuente: EU, ‘Assessment List for Trustworthy Artificial Intelligence’ (2020) |

<Tabla 23 – Cuestionario de la evaluación de IA de ‘ALTAI’ de UE>

|  |
| --- |
| Referencia – Diseño Éticamente Alineado del Instituto de Ingenieros Eléctricos y Electrónicos (IEEE por sus siglas en inglés) |
| Diseño Éticamente Alineado de IEEE surge recomendaciones prácticas para la IA y proporciona directriz específica sobre estándares, certificaciones, regulaciones o leyes para la fabricación y el uso.- En términos de responsabilidad, se enfatiza que la responsabilidad debe establecerse claramente permitiendo que siempre se sepa quién es legalmente responsable de un sistema de IA específico mediante un sistema de inscripción y mantenimiento de registros. |
| Fuente: IEEE, ‘Ethically Aligned Design’ (2019) |

<Tabla 24 – Diseño Éticamente Alineado del IEEE>

|  |
| --- |
| Referencia – Cuestiones de dónde recae la responsabilidad |
| Debido a la naturaleza de “caja negra” del algoritmo, es difícil determinar si se produjo un error, dónde se produjo, etc., lo que dificulta determinar dónde recae la responsabilidad incluso si se ha producido un daño.(Descripción general) Los robo-advisors utilizados en el sector financiero son servicios automatizados que utilizan IA y algoritmos para gestionar y operar los activos financieros de los clientes. Han surgido problemas relacionados con la determinación de un responsable y la ausencia de un sistema para la supervisión cuando se produce un daño al usuario debido a errores o defectos en el algoritmo.(Contramedida) Se enfatiza que debe haber un departamento o persona a cargo que pueda plantear objeciones cuando se produce un daño, ya que las empresas pueden evitar la responsabilidad enfatizando el efecto de “caja negra” de los algoritmos. |

<Tabla 25 – Cuestiones de dónde recae la responsabilidad >

|  |
| --- |
| Referencia – Cuestiones ambientales |
| A medida que la tecnología de IA se entrena y se desarrolla con base en cantidades masivas de datos, existe la preocupación de que la cantidad de electricidad que consume y el carbono que emite tengan un impacto negativo en el medio ambiente. Como resultado, los valores medioambientales se enfatizan cada vez más en el desarrollo de la IA.(Ejemplo) Se estima que las emisiones de carbono emitidas durante el proceso de entrenamiento de GPT-3, un modelo de lenguaje de Open AI en EE.UU., ascienden a 522 toneladas métricas, lo que equivale a la cantidad generada al conducir 120 automóviles de pasajeros durante un año.(Contramedida) La Unión Europea planea establecer un sistema para medir la eficiencia energética de los centros de datos de las empresas europeas como parte del ‘Índice de Economía y Sociedad Digitales’, con el objetivo de lograr la neutralidad de carbono en los centros de datos y la infraestructura en la nube para 2030. |
| Fuente: Fortune, 'A.I.’s carbon footprint is big, but easy to reduce, Google researchers say' (2021.4.22.) |

<Tabla 26 – Cuestiones ambientales>

* 1. Ítem de verificación para la cooperación internacional

|  |  |
| --- | --- |
| Número | Ítem de verificación |
| 8.1 | ¿La política operativa del sistema de IA cumple con los estándares internacionales de derechos humanos (como la Declaración Universal de Derechos Humanos de las Naciones Unidas) y aplica estándares éticos y legales reconocidos internacionalmente? |
| 8.2 | ¿Se están participando activamente en proyectos internacionales de investigación sobre IA y en debates sobre el desarrollo de estándares relacionados con los sistemas de IA? |
| 8.3 | ¿Tiene un sistema de cooperación internacional para abordar cuestiones éticas y legales globales que puedan surgir al operar sistemas de IA? |
| 8.4 | ¿Cumplen los sistemas de IA con las regulaciones internacionales de privacidad respecto del movimiento transfronterizo de datos? |

<Tabla 27 – Ítem de verificación para la cooperación internacional>

* Explicación de ítem de verificación

A medida que las tecnologías de IA se utilizan a través de las fronteras, es importante que las políticas nacionales y las normas jurídicas estén armonizadas internacionalmente. Garantizar que la política de IA del Perú cumpla con los estándares de derechos humanos reconocidos internacionalmente es un requisito fundamental para que las tecnologías de IA puedan operar de manera confiable en el sector público. Es necesario establecer políticas que reflejen los estándares globales de derechos humanos, como la Declaración Universal de Derechos Humanos de las Naciones Unidas, y revisar si los sistemas de IA utilizados en el sector público están diseñados de acuerdo con estos estándares. Además, es una tarea clave aplicar fielmente los estándares legales y éticos para que la IA pueda operar como una tecnología centrada en el ser humano sin violar los derechos de los ciudadanos en los servicios públicos.

Los debates relacionados con el desarrollo de la tecnología de IA están avanzando rápidamente a nivel internacional, y los gobiernos y los institutos de investigación de cada país cooperan para establecer la dirección y los estándares para el desarrollo de la IA. La participación activa del gobierno peruano y de las instituciones públicas en proyectos internacionales de investigación en IA y en las discusiones sobre el desarrollo de estándares es un elemento clave para garantizar el desarrollo ético y confiable de las tecnologías de IA. Se necesita una colaboración mundial en materia de investigación para garantizar que las tecnologías de IA sean interoperables entre países y abordar conjuntamente los sesgos técnicos y las cuestiones éticas.

Es probable que la tecnología de IA plantee cuestiones éticas y jurídicas no sólo a nivel nacional sino también internacional. En particular, debido a que la equidad y la responsabilidad de los sistemas de IA pueden interpretarse de manera diferente en los distintos países, es importante establecer normas y medidas de respuesta consistentes a través de la cooperación global. El gobierno peruano debería verificar si está estableciendo un sistema de cooperación internacional y operando una estructura de cooperación que pueda resolver los dilemas éticos y las disputas legales que pueda causar la IA.

A medida que aumenta el movimiento transfronterizo de datos, el cumplimiento de las regulaciones internacionales sobre privacidad es esencial. La confirmación de si los sistemas de IA que operan en Perú aplican las políticas coherentes con las regulaciones globales de protección de datos (por ejemplo, el RGPD de la UE, las Directrices de Privacidad de la OCDE, etc.) es fundamental para aumentar la fiabilidad de las tecnologías de IA. Se requiere un enfoque equilibrado para que las tecnologías de IA puedan utilizarse sin problemas y al mismo tiempo cumpliendo con los principios de privacidad, y se necesitan medidas políticas que tengan en cuenta la soberanía de los datos y la interoperabilidad global.

1. Anexo
	1. Tabla de ítem de verificación

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Núm. | Ítem de verificación | Sí | No | N/A |
| 1. Protección de los derechos humanos
 |
| 1.1 | ¿Los sistemas de IA están diseñados y operados de manera que no violen la dignidad humana y los derechos fundamentales (por ejemplo, el derecho a la vida, la seguridad física y mental, la autonomía, etc.)? |  |  |  |
| 1.2 | ¿Existen procedimientos establecidos para garantizar que los sistemas de IA no discriminen injustamente por motivos de género, raza, edad, discapacidad, religión, origen socioeconómico, etc.? |  |  |  |
| 1.3 | ¿Se garantiza que los sistemas de IA no infrinjan la libertad de expresión, el acceso a la información y los derechos fundamentales de los ciudadanos en el proceso de prestación de servicios públicos y toma de decisiones de políticas? |  |  |  |
| 1.4 | ¿Existen procedimientos para realizar evaluaciones de impacto previas para determinar si es probable que los sistemas de IA infrinjan los derechos humanos y para corregir y remediar los problemas si surgen? |  |  |  |
| 1. Protección de información personal y datos
 |
| 2.1 | ¿El sistema de IA cumple con las leyes de privacidad peruanas y los estándares internacionales? |  |  |  |
| 2.2 | ¿Tiene el sistema de IA una base legal clara (por ejemplo, consentimiento del usuario, contrato, obligación legal, etc.) para recopilar, procesar y almacenar información personal? |  |  |  |
| 2.3 | ¿Existen procedimientos establecidos para permitir que los ciudadanos que utilizan servicios de IA ejerzan su derecho a acceder, corregir o eliminar su información personal (incluido el derecho al olvido)? |  |  |  |
| 2.4 | ¿El sistema de IA tiene un proceso de gestión de seguridad para evitar fugas de datos, mal uso, etc.? |  |  |  |
| 2.5 | ¿Los datos recopilados y procesados ​​por sistemas de IA están protegidos contra el acceso no autorizado o la falsificación? |  |  |  |
| 1. Mejorar la publicidad y el valor social
 |
| 3.1 | ¿Los sistemas de IA están diseñados y operados para promover el interés público y mejorar la calidad de vida de los ciudadanos? |  |  |  |
| 3.2 | ¿Los sistemas de IA desempeñan un papel de reflejar las opiniones de los ciudadanos en la administración pública y promover su participación en los procesos de formulación de políticas? |  |  |  |
| 3.3 | ¿Se realizan periódicamente evaluaciones de impacto ético para garantizar que los sistemas de IA mantengan su valor social en la administración y los servicios públicos? |  |  |  |
| 1. Fortalecimiento de la accesibilidad y la inclusión social
 |
| 4.1 | ¿Están los sistemas de IA diseñados para que los grupos socialmente vulnerables, como los discapacitados, los ancianos y las personas de bajos ingresos puedan tener el acceso igualitario a los servicios públicos? |  |  |  |
| 4.2 | ¿Se aplican los sistemas de IA de manera justa a todos los ciudadanos sin discriminación en el proceso de prestación de servicios públicos? |  |  |  |
| 4.3 | ¿El sistema de IA tiene una interfaz fácil de usar y un diseño intuitivo que tiene en cuenta los antecedentes técnicos y el nivel de capacidad digital del usuario? |  |  |  |
| 4.4 | Al expandir los sistemas de IA en el sector público, ¿están implementando programas de educación y capacitación dirigidos a la clase de baja accesibilidad? |  |  |  |
| 4.5 | ¿El sistema de IA refleja políticas de administración pública inclusivas sin excluir a ciertas clases sociales o regiones? |  |  |  |
| 1. Garantizar la explicabilidad
 |
| 5.1 | ¿Los datos, algoritmos y criterios de toma de decisiones que utilizan los sistemas de IA son transparentes y fácilmente comprensibles para los ciudadanos y las partes interesadas? |  |  |  |
| 5.2 | Cuando un sistema de IA toma una decisión, ¿existen procedimientos establecidos para proporcionar a los humanos información suficiente para comprender y revisar esa decisión? |  |  |  |
| 5.3 | ¿Están los sistemas de IA tomando medidas técnicas e institucionales para abordar el problema de la “caja negra” (el problema de la opaca estructura interna del algoritmo)? |  |  |  |
| 5.4 | Si la decisión de un sistema de IA se basa en datos específicos, ¿se registra ese proceso y se pone a disposición para su revisión? |  |  |  |
| 1. Aseguramiento de la fiabilidad y la seguridad técnica
 |
| 6.1 | ¿Son los sistemas de IA resilientes y seguros frente a ciberataques como piratería, envenenamiento de datos y evasión de modelos? |  |  |  |
| 6.2 | ¿El sistema de IA cumple con los procesos de gestión de la calidad de los datos (frecuencia, precisión, integridad, etc.) y tiene procedimientos establecidos para evaluar y mejorar continuamente el rendimiento del modelo? |  |  |  |
| 6.3 | ¿El sistema de IA tiene procesos establecidos para detectar y abordar falsos positivos y predicciones inexactas? |  |  |  |
| 6.4 | ¿Tiene sistemas de respaldo o procesos alternativos implementados para minimizar el impacto de fallas o mal funcionamiento del sistema de IA? |  |  |  |
| 6.5 | ¿Existen procedimientos regulares de inspección y actualización para mantener la fiabilidad y seguridad de los sistemas de IA? |  |  |  |
| 1. Desarrollo y funcionamiento de la IA sostenible
 |
| 7.1 | ¿Las instituciones públicas que operan sistemas de IA han establecido un sistema claro de responsabilidad y, una gobernanza para cumplir con las responsabilidades legales y éticas en el proceso de utilización de la IA? |  |  |  |
| 7.2 | ¿Se ha establecido una institución de gobernanza que puede abordar problemas éticos inesperados en el proceso de la operación y el desarrollo de sistemas de IA (por ejemplo, sesgo, injusticia, violaciones de la privacidad, etc.)? |  |  |  |
| 7.3 | ¿Los sistemas de IA están diseñados para promover el bienestar público y la inclusión social y funcionan de manera que fomenten la participación de la comunidad y los ciudadanos? |  |  |  |
| 7.4 | ¿Se están diseñando los sistemas de IA de manera sostenible para generar ahorros de costos y mejoras en la eficiencia operativa a largo plazo en el sector público? |  |  |  |
| 7.5 | ¿Los sistemas de IA están diseñados y respaldados por políticas para contribuir a la respuesta al cambio climático, la protección del medio ambiente y la conservación de los ecosistemas? |  |  |  |
| 1. Cooperación internacional
 |
| 8.1 | ¿La política operativa del sistema de IA cumple con los estándares internacionales de derechos humanos (como la Declaración Universal de Derechos Humanos de las Naciones Unidas) y aplica estándares éticos y legales reconocidos internacionalmente? |  |  |  |
| 8.2 | ¿Se están participando activamente en proyectos internacionales de investigación sobre IA y en debates sobre el desarrollo de estándares relacionados con los sistemas de IA? |  |  |  |
| 8.3 | ¿Tiene un sistema de cooperación internacional para abordar cuestiones éticas y legales globales que puedan surgir al operar sistemas de IA? |  |  |  |
| 8.4 | ¿Cumplen los sistemas de IA con las regulaciones internacionales de privacidad respecto del movimiento transfronterizo de datos? |  |  |  |

<Tabla 28 – Ítem de verificación>

* 1. Glosario de IA

Este glosario es relevante para estos estándares éticos, y tiene como objetivo ayudar a comprender los términos utilizados en este documento o la terminología de IA relacionada.

* Accessibility / Accesibilidad

La accesibilidad se refiere al grado de conveniencia en uso en que los productos, los sistemas, los servicios, los entornos e instalaciones están diseñados en consideración de las diversas necesidades, características y capacidades de los usuarios. También, incluye el grado en que un usuario puede usarlo directamente o con la ayuda de tecnología de asistencia para lograr un objetivo en un contexto específico.

* Accuracy / Precisión

La precisión se refiere a la capacidad de un modelo de IA para mantener un alto rendimiento de predicción incluso con datos nuevos con los que no ha sido entrenado, y se mide a través de varias métricas de evaluación.

* Adversarial Evasion / Evasión de modelos

La evasión de modelos es uno de los ataques más comunes a los modelos de aprendizaje automático (ML por sus siglas en inglés), una técnica de ataque que hace que el modelo clasifique erróneamente los datos de entrada.

Los atacantes pueden manipular intencionalmente los datos de entrada para provocar un mal funcionamiento del sistema de IA, alterando así los resultados de predicción del modelo. Por ejemplo, en un sistema de reconocimiento de imágenes, simplemente transformar un píxel específico puede hacer que el modelo reconozca el objetivo como un objeto completamente diferente. Este es un método de ataque representativo que utiliza Ejemplos Adversarios.

Para fortalecer la seguridad de los modelos de IA, se necesitan métodos como el entrenamiento adversario, algoritmos de detección y el fortalecimiento de la robustez de los modelos.

* AI Bias / Sesgo de IA

El sesgo de la IA se refiere a errores sistemáticos que ocurren en el diseño, la recopilación de datos y los procesos de aprendizaje de los algoritmos, que pueden afectar injustamente a determinados grupos.

* AI Designer / Diseñador de IA

Los diseñadores de IA son responsables de cerrar la brecha entre las capacidades de IA y las necesidades de los usuarios. Por ejemplo, antes de desarrollar un producto de IA, puede crear un prototipo que muestre sus características y cómo podría usarse el producto una vez implementado. También trabajan con el equipo de desarrollo para comprender mejor las necesidades de los usuarios y encontrar formas de crear tecnologías para abordarlas. Además, diseñan una plataforma que admite la recopilación y anotación de datos para apoyar a los desarrolladores de IA y administran la seguridad y la equidad de los datos durante este proceso.

* AI Developer / Desarrollador de IA

Los desarrolladores de IA son expertos que realizan diversas tareas relacionadas con el desarrollo de IA. El desarrollo de IA es el proceso general de creación y mantenimiento de aplicaciones, marcos u otros componentes de IA, incluidas etapas como la concepción, el diseño, la recopilación de datos, el aprendizaje, la programación, las pruebas y la corrección de errores. Los desarrolladores de IA participan ampliamente en todo, desde la concepción inicial del software hasta la implementación y operación final, y desempeñan un papel fundamental a la hora de garantizar la estabilidad y fiabilidad de los sistemas de IA.

* AI Ethics Review Committee / Comité de Revisión Ética de IA

El Comité de Revisión Ética de IA es un grupo de partes interesadas y expertos que tienen en cuenta diversos factores como el género, la edad y los antecedentes. El comité evaluará las consideraciones éticas en el desarrollo y el funcionamiento de la IA y proporcionará supervisión para que la IA se utilice de forma justa y responsable. Al formar un comité, se debe garantizar que funcione de forma independiente y libre de intereses económicos o de otro tipo, y que se minimicen los sesgos y los conflictos de intereses. El tamaño del comité puede ajustarse de acuerdo al alcance del trabajo, y es necesario asegurar su eficacia garantizando suficientemente los roles y la autoridad a desempeñar y la accesibilidad a la información.

* AI Practitioner / Encargado de IA

Un encargado de IA se entiende cualquier individuo u organización involucrada en la investigación, el diseño, la provisión de datos, el desarrollo y la distribución (incluida la implementación) de sistemas de IA. Sin embargo, los usuarios finales o consumidores no están incluidos en los encargados de IA.

* AI System Environment / Entorno del sistema de IA

El entorno del sistema de IA incluye elementos externos que rodean al sistema de IA y se refiere a datos externos que no forman parte del sistema en sí.

* AI System Lifecycle / Ciclo de vida de los sistemas de IA

El ciclo de vida de un sistema de IA consta de la fase de desarrollo que incluye investigación, diseño, provisión de datos, pruebas y desarrollo, la fase de distribución (incluida la implementación) y, finalmente, la fase de operación y mantenimiento.

* Artificial Intelligence (AI) System / Sistema de Inteligencia Artificial (IA)

Los sistemas de IA son sistemas de software y hardware diseñados por humanos para percibir el entorno, recopilar e interpretar datos estructurados y no estructurados, inferir conocimientos o procesar información para determinar acciones óptimas para lograr objetivos complejos determinados. Los sistemas de IA pueden basarse en reglas simbólicas o aprender modelos numéricos, e incluso pueden tener la capacidad de ajustarse analizando el impacto de acciones anteriores en el entorno. Las áreas de tecnología de IA incluyen el aprendizaje automático (incluido el aprendizaje profundo y el aprendizaje de refuerzo), el razonamiento de máquina (planificación, programación, representación e inferencia de conocimientos, optimización, etc.) y la robótica (control, sensores, actuadores e integración de sistemas ciberfísicos).

* Assistive Technology / Tecnología de asistencia

La tecnología de asistencia se refiere al software y hardware que se agregan o se integran a los sistemas de tecnología de la información y las comunicaciones (TIC) para aumentar la accesibilidad. En particular, suelen estar diseñados para ayudar a las personas con discapacidad a realizar actividades cotidianas. Las tecnologías de asistencia típicas incluyen sillas de ruedas, máquinas de lectura y dispositivos de asistencia, y las tecnologías de asistencia basadas en software, como lectores de pantalla, amplificadores de pantalla, sintetizadores de voz y software de entrada de voz, también se utilizan en relación con la accesibilidad web. Las tecnologías de asistencia de hardware también incluyen teclados alternativos y dispositivos señaladores especiales.

* Audit / Auditoría

Una auditoría es un proceso de investigación y evaluación independiente de los atributos esenciales de una empresa, producto, software, etc. Esto ayudará a verificar si hay representaciones erróneas sobre temas específicos y garantizar la fiabilidad para las distintas partes interesadas. Aunque las auditorías se realizan más comúnmente sobre la información financiera de las corporaciones, pueden aplicarse en muchas áreas, incluidas las auditorías que evalúan la ética y la seguridad de los sistemas de IA.

* Auditability / Auditabilidad

La auditabilidad se refiere a la capacidad de los algoritmos, datos y procesos de diseño de un sistema de IA para ser evaluados. Esto no significa necesariamente revelar modelos de negocios o información de propiedad intelectual, pero puede fortalecerse garantizando mecanismos de trazabilidad y registro desde las primeras etapas de diseño del sistema.

* Autonomous AI System / Sistema de IA autónomo

Los sistemas de IA autónomos son sistemas de IA que actúan o realizan tareas con un alto grado de autonomía sin intervención externa directa. Estos sistemas están diseñados para tomar decisiones por sí solos y operar de forma independiente en entornos específicos, y se utilizan en una variedad de campos, incluidos los automóviles autónomos, los sistemas de toma de decisiones automatizadas y la robótica.

* Bias / Sesgo

El sesgo se refiere al fenómeno en el cual las nociones preconcebidas sobre un individuo, grupo o concepto específico son inherentes a un sistema de IA. En los sistemas de IA basados ​​en aprendizaje automático, el sesgo puede producirse en el proceso de recopilación y aprendizaje de datos, y en los sistemas basados ​​en reglas, el sesgo puede formarse dependiendo de las medidas de configuración de los ingenieros de conocimiento. Además, existe la posibilidad de que los sistemas de IA adquieran sesgos a través del aprendizaje y la interacción en línea. Algunos sesgos pueden dar lugar a injusticias, lo que se define como “sesgo injusto”.

* Confidence Score / Puntuación de confianza

Puntuación de confianza es un concepto que expresa numéricamente la probabilidad de que la salida generada por un sistema de IA para una entrada específica sea correcta. Esto se utiliza para evaluar cuantitativamente la fiabilidad de los resultados generados por el modelo de IA, y una puntuación de confianza baja significa que la precisión de los resultados es baja o la incertidumbre es alta. Cuando se utilizan sistemas de IA para la toma de decisiones prácticas, se necesitan medidas para complementar la fiabilidad y precisión del modelo considerando la puntuación de confianza.

* Data Governance / Gobernanza de datos

La gobernanza de datos es un concepto que se aplica tanto a nivel macro como micro. A nivel macro, la gobernanza de datos se refiere a la gestión de los flujos de datos a través de las fronteras internacionales y, más específicamente, a la gobernanza de datos internacional. Por otro lado, a nivel micro, se refiere a un sistema que gestiona eficazmente los datos para que una organización pueda mantener una alta calidad durante todo el ciclo de vida de los datos y respaldar los objetivos comerciales. Los elementos centrales de la gobernanza de datos incluyen la disponibilidad, usabilidad, consistencia, integridad y compartición de datos, y los procesos de gestión que permiten a las empresas y las instituciones evitar los impactos negativos de la mala calidad de los datos y aprovecharlos de manera efectiva en toda la organización.

* Data Poisoning / Envenenamiento de datos

El envenenamiento de datos se refiere a una técnica de ataque en la que un atacante inyecta intencionalmente información falsa en los datos de entrenamiento de un modelo de IA explotando un sistema de IA. Esto puede provocar que los sistemas de IA aprendan incorrectamente y, en casos graves, la precisión de los modelos de IA puede disminuir significativamente. Los ataques de envenenamiento de datos se presentan de muchas formas, algunas de las cuales tienen como objetivo cambiar el comportamiento de los sistemas de IA. En particular, existe un método en el que los atacantes insertan datos de “puerta trasera” para enseñar a los modelos de IA patrones específicos o inducir un comportamiento no deseado. Para prevenir este tipo de ataques, se necesitan medidas para comprobar continuamente la integridad de los datos de los sistemas de IA y reforzar los sistemas de detección y verificación de anomalías.

* Data Protection Impact Assessment (DPIA) / Evaluación de Impacto de la Protección de Datos (EIPD)

Evaluación de Impacto de la Protección de Datos es un proceso de evaluación de los riesgos que pueden surgir cuando un sistema de IA procesa información personal. Esto debería realizarse en casos en los que se prevén riesgos elevados, como la toma de decisiones automatizada, la vigilancia masiva y el procesamiento de datos sensibles.

* Data Protection Officer (DPO) / Delegado de Protección de Datos (DPD)

Un Delegado de Protección de Datos (DPD) es un profesional especializado en legislación de protección de datos y normativa de privacidad, y desempeña la función de monitorear y supervisar internamente a las organizaciones públicas y privadas para garantizar que operen en cumplimiento de las leyes de privacidad. El DPD debe tener una posición independiente dentro de la organización y ser capaz de revisar y tomar medidas sobre cuestiones legales y éticas relacionadas con la protección de datos.

* Designed Remediation / Remediación diseñada

Remediación diseñada se refiere al concepto de la detección, la auditoría, y la corrección de decisiones incorrectas cuando los sistemas de IA toman decisiones incorrectas, asimismo incorporar dispositivos complementarios y procedimientos de respuesta desde la etapa de diseño para que se mejoren los sistemas si es necesario.

Los sistemas de IA pueden encontrar errores inesperados, resultados sesgados y fallos del funcionamiento al tomar decisiones automatizadas. Por lo tanto, es necesario diseñar con antelación mecanismos como redundancia, sistemas alternativos y procedimientos alternativos para prevenir y resolver estos problemas.

Por ejemplo, un sistema de diagnóstico médico basado en IA podría incluir la capacidad de detectar automáticamente diagnósticos incorrectos y permitir que un profesional médico intervenga para revisar los errores. Además, si un sistema de IA legal hace una interpretación incorrecta de la ley, es necesario establecer un proceso de objeciones para que un humano pueda revisarlo.

* Encryption, Pseudonymization, Aggregation, and Anonymization / Encriptación, seudonimización, agregación y anonimización

Se utilizan diversos métodos técnicos para proteger los datos y mejorar la privacidad, siendo las tecnologías de protección representativas la encriptación, la seudonimización, la agregación y la anonimización.

Encriptación: Un método para transformar datos de forma que usuarios no autorizados no puedan leer el contenido original. Normalmente, los datos se protegen mediante un clave hash o una clave de cifrado, y solo el usuario que posee la clave puede restaurar los datos originales.

Seudonimización: Un método para transformar datos personales de modo que no puedan vincularse directamente a un individuo específico sin información adicional. Por ejemplo, un interesado puede ser identificado utilizando un código o clave específica en lugar de un nombre, pero sólo la entidad que posee esa clave puede restaurar los datos originales.

Agregación: Un método para generar datos estadísticos grupales mediante la combinación de puntos de datos individuales. Por ejemplo, evita que se identifiquen personas individuales calculando la edad promedio de un grupo particular.

Anonimización: Un método que elimina por completo todos los elementos de identificación personal. Una vez anonimizados, los datos no pueden vincularse a un individuo específico. Estos no se consideran datos personales en una normativa de protección de datos (por ejemplo, el RGPD).

Estas tecnologías se utilizan ampliamente en sistemas de IA y procesos de gestión de datos para mejorar la privacidad y la seguridad.

* End User / Usuario final

Los usuarios finales se refieren a personas u organizaciones que utilizan directamente sistemas de IA o, en última instancia, utilizan los servicios de sistemas de IA. Esto puede abarcar una variedad de grupos de usuarios, incluidos consumidores, profesionales de instituciones públicas y empleados de organizaciones privadas.

Los roles como administradores de sistemas, administradores de bases de datos, profesionales de TI e ingenieros de software que son responsables de la gestión técnica directa y el mantenimiento de los sistemas de IA no están incluidos como usuarios finales. En cambio, los usuarios finales se refieren a usuarios habituales, clientes o no expertos que utilizan servicios o sistemas basados ​​en IA, y pueden abarcar a cualquier individuo u organización que aproveche las decisiones o recomendaciones proporcionadas por IA.

* Ethical AI / IA ética

IA ética se utiliza para referirse al desarrollo, el despliegue y la utilización de IA que garantiza el cumplimiento de las normas éticas, incluidos los derechos fundamentales y los principios éticos como derechos morales especiales.

* Ethics / Ética

La ética es una disciplina filosófica que estudia los valores morales humanos y los estándares de comportamiento, mientras que la ética de la IA es una rama de la ética aplicada que aborda cuestiones normativas relacionadas con el diseño, desarrollo, implementación y utilización de la IA. La ética de la IA incluye principios como la protección de los derechos fundamentales, la equidad, la transparencia y la responsabilidad, y sirven como elementos clave para una IA confiable.

* Explainability / Explicabilidad

La explicabilidad se refiere a la capacidad de un sistema de IA de proporcionar una comprensión de cómo y por qué tomó una decisión particular. Cuando los sistemas de IA cuentan con la explicabilidad, los usuarios y las partes interesadas pueden confiar en las maneras de funcionamiento de la IA y las conclusiones.

* Fairness / Equidad

La equidad significa que los sistemas de IA están diseñados y operados de manera que no discriminen a individuos o grupos específicos. Para garantizar la equidad de los sistemas de IA, es necesario comprobar y ajustar continuamente los sesgos de los algoritmos y los datos de aprendizaje.

* Fault Tolerance / Tolerancia a fallos

La tolerancia a fallas se refiere a una característica de diseño que permite que todo el sistema funcione normalmente incluso si algunos componentes del sistema fallan. Este es un factor importante para aumentar la estabilidad y confiabilidad del sistema, garantizando que las operaciones puedan continuar incluso si ocurre una falla.

Los sistemas con alta tolerancia a fallos están diseñados para que pequeños errores no provoquen la falla de todo el sistema y, aunque el rendimiento puede reducirse gradualmente dependiendo de la gravedad del error, se mantiene la funcionalidad principal. Por el contrario, los sistemas que carecen de tolerancia a fallos corren un mayor riesgo de experimentar un punto único de falla (SPOF por sus siglas en inglés) que puede provocar que todo el sistema suspenda.

* Human-Centric AI / IA centrada en el ser humano

La IA centrada en el ser humano se refiere a un enfoque que garantiza y protege los valores humanos fundamentales en los procesos del desarrollo, la implementación y la operación de los sistemas de IA, asegurando que los sistemas de IA se diseñen y se operen centrándose en la dignidad humana.

* Human Oversight / Supervisión humana

La supervisión humana se refiere a los mecanismos de gobernanza diseñados para permitir la intervención humana incluso cuando los sistemas de IA funcionan de forma autónoma. Esto sirve para permitir una intervención inmediata si el sistema de IA produce resultados inesperados.

Los métodos de supervisión humana se clasifican de la siguiente manera:

Human-in-the-Loop (HITL): Un método en el que los humanos intervienen directamente en los procesos de toma de decisiones de la IA

Human-on-the-Loop (HOTL): Un método en el que un humano monitorea continuamente un sistema de IA mientras está en funcionamiento.

Human-in-Command (HIC): Un método en el que un humano tiene control general sobre un sistema de IA y decide la utilización de IA cuando sea necesario.

* Model Inversion / Inversión de modelo

La inversión de modelo es un tipo de ataque a los modelos de IA, donde un atacante accede a un sistema de IA e infiere de forma inversa los datos aprendidos por el modelo.

Este ataque permite que un modelo de IA reconstruya parte o la totalidad de los datos que ha aprendido, lo que puede provocar graves violaciones de la privacidad, especialmente si los datos de entrenamiento contienen información personal sensible. Por ejemplo, si se realiza un ataque de inversión de modelo en un modelo de IA de reconocimiento facial, existe la posibilidad de que se puedan restaurar los datos de imagen facial aprendidos por el modelo.

Las contramedidas contra la inversión de modelo requieren medidas de seguridad como privacidad diferencial, entrenamiento encriptado y técnicas de protección de datos estructurados.

* Penetration Testing (Pen Test) / Examen de penetración (pentest)

Examen de penetración (pentest) o hacking ético son un método para identificar y mejorar las vulnerabilidades de seguridad mediante la realización de ataques de penetración simulados en sistemas informáticos y sistemas de IA.

Examen de penetración tienen como objetivo evaluar el estado de seguridad de los sistemas de IA y prevenir con antelación riesgos de piratería o violación de datos. Este proceso analiza las vulnerabilidades de seguridad que los atacantes pueden realmente explotar y prepara medidas de seguridad para complementarlas.

Para fortalecer la seguridad de los sistemas de IA, es importante realizar examen de penetración periódicamente y, en particular, evaluar amenazas como la evasión de modelos, el envenenamiento de datos y la inversión de modelo.

Los sistemas de seguridad basados ​​en IA pueden utilizar examen de penetración automatizado, simulaciones de ataques simulados y entrenamiento de respuesta de IA adversaria.

* Reproducibility / Reproducibilidad

La reproducibilidad se refiere a si se obtienen los mismos resultados cuando un experimento de IA se realiza repetidamente en las mismas condiciones.

* Responsible AI / IA responsable

Un enfoque para desarrollar, evaluar e implementar sistemas de IA de manera segura, confiable y ética. Los sistemas de IA son el producto de una variedad de decisiones tomadas por los usuarios que desarrollan e implementan estos sistemas. La IA responsable puede ayudar a orientar estas decisiones hacia resultados más beneficiosos y equitativos, desde cómo se utilizan los sistemas con antelación hasta cómo interactúan los usuarios con los sistemas de IA. Esto significa poner a las personas y los objetivos en el centro de las decisiones de diseño del sistema y respetar valores duraderos como la equidad, la confianza y la transparencia.

* Red Teaming / Trabajo en equipo rojo

El trabajo en equipo rojo es un método en el que un grupo de expertos independientes realiza ataques a un sistema de IA para evaluar su seguridad y estabilidad. Esta técnica se utiliza para identificar y complementar vulnerabilidades en los sistemas de IA.

* Robustness / Robustez

La robustez de los sistemas de IA es un concepto que incluye la robustez técnica (precisión y confiabilidad) y la robustez social (responsabilidad ética y equidad). Este es un elemento fundamental para garantizar que los sistemas de IA sean seguros y confiables incluso en situaciones inesperadas.

* Stakeholder / Partes interesadas

Las partes interesadas incluyen a todas las personas y organizaciones afectadas o involucradas en el diseño, el desarrollo, el despliegue y la operación de la IA. Esto abarca instituciones públicas, institutos de investigación, empresas, organizaciones de la sociedad civil, trabajadores y consumidores.

* Sustainability / Sostenibilidad

La sostenibilidad es el concepto de garantizar que los sistemas de IA funcionen teniendo en cuenta sus impactos ambientales, sociales y económicos a largo plazo. Los sistemas de IA deben diseñarse para ser energéticamente eficientes, socialmente inclusivos y evolucionar mediante investigación y desarrollo continuos.

* Traceability / Trazabilidad

La trazabilidad de un sistema de IA se refiere a la capacidad de documentar y revisar la recopilación de datos, el desarrollo de algoritmos, el despliegue y los procesos operativos del sistema.

* Trust / Confianza

La confianza es el concepto que permite a los usuarios creer que los sistemas de IA cumplen con las leyes, respetan los principios éticos y están diseñados de manera sólida. La garantía de confianza de un sistema de IA requiere la explicabilidad, la transparencia, la equidad, la responsabilidad y la robustez.

* Trustworthy AI / IA confiable

IA confiable se refiere a una IA que cumple con la ley, respeta los principios éticos y que cuenta con la robustez tecnológica, e incluye la confianza no solo del sistema de IA en sí, sino también la confianza de los procesos operativos y de gestión.

* Vulnerable Individuals and Groups / Individuos y grupos vulnerables

Individuos y grupos vulnerables son aquellos que probablemente se verán relativamente más afectados por los sistemas de IA debido a sus características sociales, económicas, legales o físicas. Esto puede definirse en un momento específico o en un contexto específico, y puede variar dependiendo de factores sociales discriminatorios (género, raza, edad, nivel económico, etc.).

**Referencias**

Recommendation of the Council on Artificial Intelligence (OECD, 2025)

Recommendation on the ethics of artificial intelligence (UNESCO, 2024)

AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense Supporting Document (Defense Innovation Board)

Criterios de ética para la inteligencia artificial (IA) (Ministerio de Ciencia y TIC de Corea del Sur, 2020)

A pro-innovation approach to AI regulation (UK Department of Science, Innovation & Technology, 2023)

Social Principles of Human-Centric AI (Japan)

ETHICS GUIDELINES FOR TRUSTWORTHY AI (EU, 2019)

THE ASSESSMENT LIST FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE (ALTAI) (EU, 2020)

Lista de verificación autónoma para la implementación de los criterios de ética en IA (Ministerio de Ciencia y TIC de Corea del Sur, 2022)

Position of the European Parliament adopted at first reading on 13 March 2024 with a view to the adoption of Regulation (EU, 2024)

WHO, Ethics and governance of artificial intelligence for health: WHO guidance (2021)

Guía para la gestión de calidad de los datos de entrenamiento de inteligencia artificial (NIA de Corea del Sur, 2021)

AI ETHICS GUIDELINES (Digital Dubai, 2020)

Australia's AI Ethics Principles (DISER, 2019)

Artificial Intelligence at Google: Our Principles (Google, 2018)

Using Artificial Intelligence and Algorithms (FTC, 2020)

Ethically Aligned Design (IEEE, 2019)