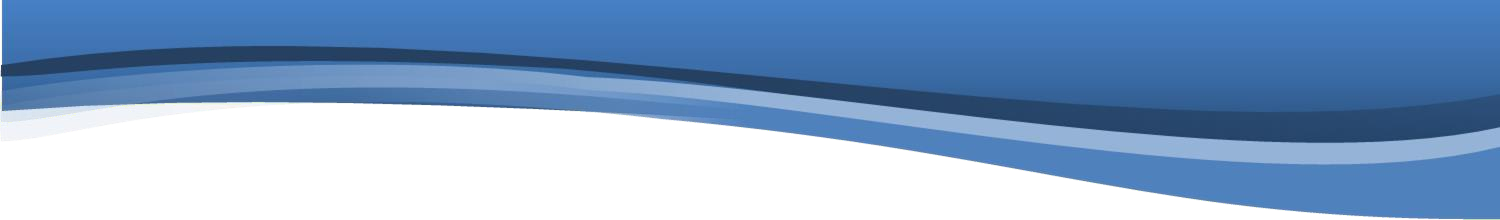


Desarrollo de una estrategia para promover la plataforma de IA en el gobierno público

**Directrices para la evaluación de riesgos de la IA en el sector público**

2024.12

 **Índice**

[1. Resumen 4](#_Toc190940316)

[1.1. Naturaleza y visión general de la guía de evaluación de riesgos 4](#_Toc190940317)

[2. Plan de Evaluación de Riesgos en la Inteligencia Artificial 4](#_Toc190940318)

[2.1. Características de los Riesgos de la Inteligencia Artificial 4](#_Toc190940319)

[2.2. Factores de Riesgo en la Inteligencia Artificial 5](#_Toc190940320)

[2.3. Marco de Evaluación de Riesgos en la Inteligencia Artificial 6](#_Toc190940321)

[2.4. Lista de Verificación de Medición 21](#_Toc190940322)

**Índice de tablas**

[<Tabla 1 - Características de los riesgos de la inteligencia artificial> 5](#_Toc189746712)

[<Tabla 2 - 12 factores de riesgo de la inteligencia artificial> 6](#_Toc189746713)

[<Tabla 3 - Funciones clave de evaluación> 7](#_Toc189746714)

[<Tabla 4 - Lista de verificación de gobernanza> 12](#_Toc189746715)

[<Tabla 5 - Lista de verificación de gobernanza> 16](#_Toc189746716)

[<Tabla 6 - Lista de verificación de gestión> 22](#_Toc189746717)

[<Tabla 7 - Lista de verificación de gestión> 25](#_Toc189746718)

1. Resumen
   1. Naturaleza y visión general de la guía de evaluación de riesgos

Esta guía proporciona un apoyo específico para la evaluación de riesgos de la inteligencia artificial (IA), con el objetivo de prevenir o mitigar de manera anticipada los posibles riesgos de violaciones de derechos humanos y discriminación derivados del desarrollo y uso de la IA.

Ha sido elaborada con el propósito de ofrecer un proceso flexible, sistemático y medible que permita a empresas y organizaciones de todos los sectores y tamaños abordar los riesgos asociados a la IA. De esta manera, se busca maximizar los beneficios de la tecnología de IA y, al mismo tiempo, minimizar las posibles repercusiones negativas en individuos, grupos, organizaciones y la sociedad en su conjunto. La guía de evaluación de riesgos debe integrarse en la política general de gestión de riesgos de cualquier organización, independientemente de su tamaño o estructura. Además, presenta un perfil de evaluación de riesgos de funciones clave a lo largo de todo el ciclo de vida de la IA, con el fin de facilitar la implementación de sistemas de IA confiables.

* **Ámbito de aplicación**
* Aplicable a los sistemas de IA en empresas y organizaciones de todos los sectores y tamaños.
* **Sujetos y método de aplicación**
* Los responsables de IA pueden utilizar esta guía de manera voluntaria como referencia y herramienta de verificación según los criterios establecidos.
* **Momento de aplicación**
* Se aplica en todas las etapas del ciclo de vida del sistema de IA, incluyendo su diseño, desarrollo, implementación, prueba y evaluación.
* **Contenido clave**
* Se presentan 12 posibles amenazas asociadas a la inteligencia artificial, incluyendo integridad de la información, seguridad de los datos y derechos de propiedad intelectual. Para abordar estos riesgos, se establecen criterios de evaluación de riesgos en cuatro áreas clave: gobernanza, identificación de riesgos, medición y gestión.

1. Plan de Evaluación de Riesgos en la Inteligencia Artificial
   1. Características de los Riesgos de la Inteligencia Artificial

Los riesgos asociados a la inteligencia artificial varían en función de las características del modelo de IA, el sistema en el que se implementa y su caso de uso específico. Es decir, la estructura del modelo, el método de entrenamiento, el tipo de datos utilizados y el nivel de acceso pueden influir en la naturaleza y la magnitud de los riesgos asociados.

|  |  |
| --- | --- |
| Característica | Descripción |
| Diversidad | La complejidad de los sistemas de inteligencia artificial genera una gran variedad de riesgos. |
| Incertidumbre | Los sistemas de IA en desarrollo pueden generar nuevos riesgos impredecibles en el futuro. |
| Complejidad | La gran cantidad de componentes y la interacción entre ellos dificultan la identificación y gestión precisa de los riesgos. |
| Dependencia | Los modelos de IA dependen en gran medida de los datos de entrenamiento, por lo que la calidad y los sesgos de los datos afectan directamente al modelo. |

<Tabla 1 - Características de los riesgos de la inteligencia artificial>

* 1. Factores de Riesgo en la Inteligencia Artificial

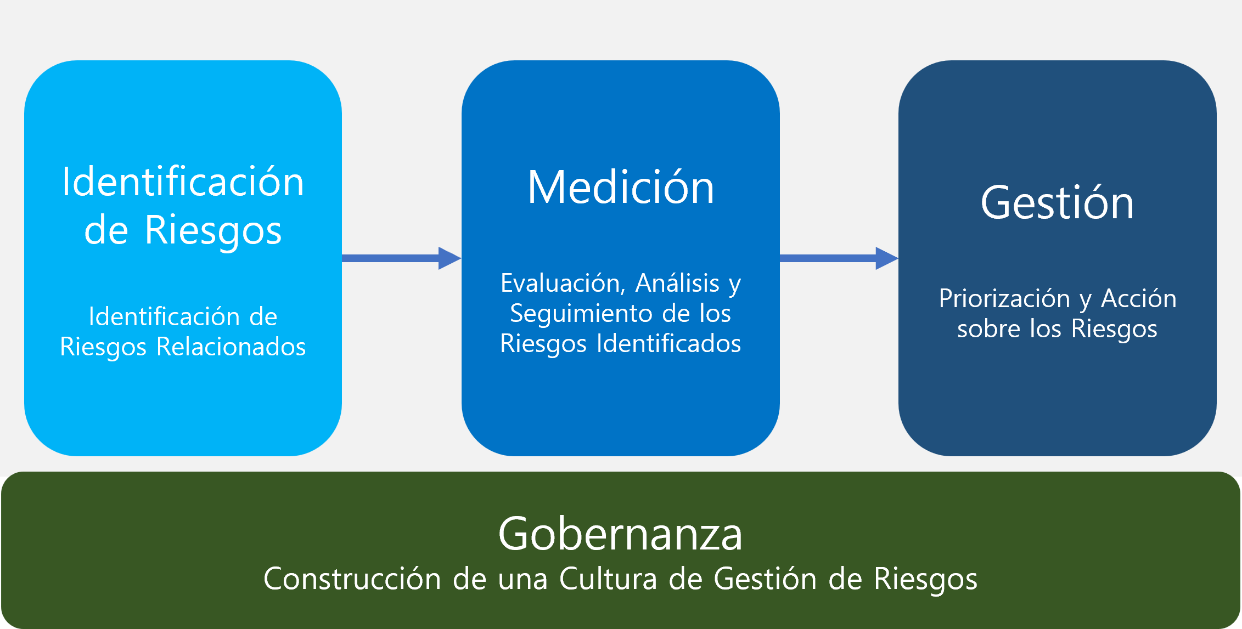
Los riesgos asociados a la inteligencia artificial son diversos y complejos. Para gestionarlos de manera efectiva, es fundamental adoptar un enfoque sistemático, como el propuesto en la guía de evaluación de riesgos de IA. La gestión de estos riesgos debe estructurarse mediante su clasificación en resultados, objetos y fuentes, lo que permite establecer estrategias de gestión personalizadas. Cada riesgo ha sido categorizado según su resultado, objeto o fuente, proporcionando a las instituciones una visión clara para adaptar sus esfuerzos de gestión de riesgos a cada contexto específico.

|  |  |  |
| --- | --- | --- |
| Factor de Riesgo | Descripción | Ejemplo |
| Vulnerabilidad en el uso de armas | Riesgo de acceso fácil a información o capacidades críticas relacionadas con sustancias o materiales peligrosos, como agentes químicos, biológicos o radiactivos. | Uso de IA generativa para obtener o crear información sobre la fabricación de sustancias bioquímicas, drogas o materiales tóxicos. |
| Falsedad | Los modelos de IA generativa pueden crear información falsa o contenidos ficticios que parezcan reales, induciendo a error a los usuarios. | Creación de noticias falsas, generación de información engañosa o afirmaciones falsas con aparente certeza. |
| Contenido peligroso | Uso de IA para generar o difundir contenidos violentos, incitadores o de odio. | Creación automática de discursos de odio dirigidos a grupos específicos por raza, género o religión, promoviendo la división social. |
| Privacidad de datos | Riesgo de filtración, uso indebido, divulgación o desanonimización de datos sensibles, como biométricos, de salud o de ubicación. | Análisis de datos de localización anónimos que permita inferir la residencia o patrones de comportamiento de una persona. |
| Impacto ambiental | Alto consumo de recursos computacionales en el entrenamiento y operación de modelos de IA con efectos negativos en el ecosistema. | Modelos de lenguaje a gran escala que requieren grandes cantidades de energía, contribuyendo al impacto ambiental. |
| Sesgo | Los modelos de IA pueden aprender y amplificar los sesgos presentes en los datos de entrenamiento, reforzando desigualdades sociales. | Generación de resultados discriminatorios basados en sesgos raciales o de género en los datos de entrenamiento. |
| Interacción humano-IA | Riesgo de antropomorfización indebida de la IA, rechazo a los algoritmos, sesgo de automatización o dependencia excesiva en la IA. | Usuarios que confían ciegamente en los juicios de una IA conversacional, tratándola como un ser humano. |
| Integridad de la información | Creación y difusión de información en la que los hechos y opiniones se confunden, generando dificultades para distinguir la verdad de la falsedad. | Uso de IA para sintetizar rostros o voces de personas reales en videos o audios falsos con fines de desinformación. |
| Seguridad de la información | El avance de la IA aumenta el riesgo de ciberataques y la vulnerabilidad de los sistemas ante filtraciones o daños. | Uso de IA para generar y distribuir ransomware o crear correos de phishing automatizados. |
| Propiedad intelectual | Riesgo de infracción de derechos de autor, marcas comerciales o secretos comerciales. | Uso de IA para generar ensayos o trabajos académicos sin atribución, lo que puede derivar en plagio. |
| Contenido ofensivo | Riesgo de creación de contenidos obscenos, degradantes, abusivos o insultantes con impacto negativo en la sociedad. | Producción de videos deepfake malintencionados que dañan la reputación de individuos. |
| Cadena de valor e integración de componentes | Problemas de confianza y ética en la IA cuando la procedencia de los componentes no es clara, la calidad de los datos es baja o los proveedores no son confiables. | Uso de datos de entrenamiento de origen desconocido que pueden generar problemas de infracción de derechos de autor o violaciones de privacidad. |

<Tabla 2 - 12 factores de riesgo de la inteligencia artificial>

* 1. Marco de Evaluación de Riesgos en la Inteligencia Artificial

El marco de evaluación de riesgos en la inteligencia artificial es un marco fundamental para la evaluación de riesgos de la IA. Se estructura en cuatro funciones clave de evaluación: gobernanza, identificación de riesgos, medición y gestión, y cuenta con listas de verificación específicas para cada función. Estas listas de verificación contienen preguntas diseñadas para ayudar a las organizaciones a identificar los riesgos potenciales, vinculándolas con los principales factores de riesgo. A través de este proceso, las organizaciones pueden evaluar qué riesgos específicos están presentes en sus sistemas de IA.

Cada función clave de evaluación se divide en categorías específicas y su correspondiente lista de verificación, pero no es obligatorio seguir un orden secuencial al revisarlas. Los usuarios del marco de evaluación de riesgos en IA pueden aplicar parcial o completamente sus funciones, dependiendo de sus capacidades y necesidades. Si bien el proceso suele seguir la secuencia de gobernanza → identificación de riesgos → medición → gestión, no es un orden fijo. Se recomienda un proceso iterativo, en el que las funciones se referencien entre sí para garantizar una evaluación más completa y efectiva.

<Figura 1 - Marco de evaluación de riesgos en la inteligencia artificial>

|  |  |
| --- | --- |
| Componente | Función |
| Gobernanza | Establece una cultura de gestión de riesgos de IA dentro de la organización, define los procesos de gestión de riesgos, integra la evaluación de impactos potenciales y alinea la gestión de riesgos de IA con los principios y prioridades organizacionales. Además, conecta los aspectos técnicos con los valores organizativos y aborda cuestiones legales a lo largo de todo el ciclo de vida del producto. |
| Identificación de riesgos | Permite predecir, evaluar y abordar las posibles causas de riesgos negativos, reduciendo la incertidumbre en la IA y mejorando la integridad del proceso de toma de decisiones. La información recopilada durante esta fase facilita la prevención de riesgos negativos y orienta decisiones sobre la gestión de modelos y la pertinencia del uso de la IA. Los resultados de esta función sirven de base para las funciones de medición y gestión. |
| Medición | Analiza, evalúa, compara y monitorea los riesgos de IA y sus impactos mediante métodos cuantitativos, cualitativos, técnicos y metodológicos. Utiliza el conocimiento obtenido en la identificación de riesgos y proporciona información clave para la función de gestión. |
| Gestión | Aborda los riesgos identificados y minimiza la probabilidad de fallos en la IA y sus impactos negativos. Para ello, se basa en los criterios de documentación establecidos en la gobernanza, la información contextual de la identificación de riesgos y los datos empíricos obtenidos en la medición. |

<Tabla 3 - Funciones clave de evaluación>

* + 1. Lista de Verificación de Gobernanza

La gobernanza es una función clave en la gestión de riesgos de IA, ya que proporciona la base sobre la cual operan las demás funciones. Para garantizar una gestión eficaz de los riesgos de IA, es esencial mantener una atención continua en la gobernanza. Una gobernanza sólida refuerza las prácticas y normas internas, promoviendo una cultura de gestión de riesgos más efectiva. El liderazgo ejecutivo desempeña un papel clave en la creación de un entorno de gestión de riesgos, mientras que las organizaciones deben alinear los aspectos técnicos con las políticas y operaciones.

Además, la documentación mejora la transparencia, los procesos de revisión y la rendición de cuentas. Una vez establecida la función de gobernanza, es fundamental fomentar una cultura organizativa centrada en la gestión de riesgos, asegurando que los usuarios implementen continuamente las funciones de gobernanza en respuesta a los cambios y desafíos emergentes.

|  |  |  |
| --- | --- | --- |
| Categoría | Pregunta | Factores de Riesgo Relacionados |
| Gobernanza 01  Comprensión, gestión y documentación de los requisitos legales y regulatorios sobre inteligencia artificial | ¿Se ajusta el desarrollo y uso de la inteligencia artificial a las leyes y regulaciones aplicables, incluidas aquellas relacionadas con la privacidad de datos, derechos de autor y propiedad intelectual? | Privacidad de datos  Sesgo  Propiedad intelectual |
| Gobernanza 02  Integración de las características de IA confiable en todas las actividades de la organización | ¿Se han establecido políticas y procesos de transparencia para documentar el origen y el historial de los datos de entrenamiento y los datos generados, mejorando la transparencia del contenido digital mientras se equilibra la naturaleza propietaria del enfoque de entrenamiento? | Privacidad de datos  Integridad de la información  Propiedad intelectual |
| ¿Se ha establecido una política para evaluar, tanto interna como externamente, la solidez de las funciones relacionadas con los riesgos de la inteligencia artificial y las medidas de seguridad, tanto antes del despliegue como de forma continua? | Seguridad de la información  Vulnerabilidad en el uso de armas |
| Gobernanza 03  Procesos, procedimientos y prácticas para la toma de decisiones en actividades de gestión de riesgos | ¿Se consideran de manera integral diversos aspectos al definir la clasificación de riesgo de la inteligencia artificial?   * Amenazas a la integridad de la información * Interdependencia entre sistemas * Impacto social * Cuestiones éticas * Impacto psicológico en los seres humanos * Posibilidad de uso indebido * Vulnerabilidades de seguridad * Desigualdad social * Seguridad en el rendimiento | Integridad de la información  Contenido ofensivo  Cadena de valor e integración de componentes  Sesgo  Vulnerabilidad en el uso de armas |
| ¿Se han establecido criterios mínimos de desempeño y seguridad para la implementación de la inteligencia artificial, y se verifica el cumplimiento de estos antes de su despliegue? | Vulnerabilidad en el uso de armas  Contenido de riesgo, violencia y odio  Seguridad de la información |
| ¿Se ha evaluado previamente la posibilidad de que el modelo de inteligencia artificial sea explotado en ciberataques y se han establecido medidas de mitigación adecuadas? | Vulnerabilidad en el uso de armas  Seguridad de la información  Privacidad de datos |
| ¿Se recopilan activamente las opiniones de las partes interesadas en el desarrollo y uso de la inteligencia artificial para garantizar su aplicación ética y responsable? | Vulnerabilidad en el uso de armas  Contenido ofensivo Sesgo  Contenido de riesgo, violencia y odio |
| ¿Se clasifican y gestionan de manera sistemática los diversos riesgos que pueden surgir con el desarrollo y uso de la inteligencia artificial? | Sesgo |
| ¿Se evalúan continuamente los riesgos del sistema de inteligencia artificial y se reajusta el umbral de tolerancia al riesgo cuando es necesario?   * Cultura de seguridad de IA inmadura * Riesgo para la integridad de la información pública (fake news, incitación, etc.) * Incertidumbre sobre el desempeño e impacto de la inteligencia artificial | Integridad de la información  Contenido ofensivo  Vulnerabilidad en el uso de armas |
| ¿Se ha establecido un plan para detener el desarrollo o despliegue de sistemas de inteligencia artificial que impliquen riesgos negativos? | Vulnerabilidad en el uso de armas  Seguridad de la información  Integridad de la información |
| Gobernanza 04  Transparencia en los procesos y resultados de la gestión de riesgos | ¿Se han implementado políticas y mecanismos técnicos para prevenir la generación de contenido que viole las leyes y las políticas de ética en inteligencia artificial? | Contenido ofensivo  Sesgo  Contenido de riesgo, violencia y odio |
| ¿Se han definido claramente los casos de uso no permitidos de la inteligencia artificial y se han establecido políticas para comunicarlos de manera transparente a los usuarios? | Vulnerabilidad en el uso de armas  Contenido ofensivo  Privacidad de datos |
| Gobernanza 05  Monitoreo continuo y revisión periódica del proceso de gestión de riesgos | ¿Se han definido claramente dentro de la organización las responsabilidades para el monitoreo y revisión continua de las fuentes de datos de la inteligencia artificial, el origen del contenido generado y los incidentes ocurridos durante la operación del sistema? | Integridad de la información |
| ¿Se revisa posteriormente el proceso de respuesta y divulgación de incidentes ocurridos durante la operación de la inteligencia artificial para identificar problemas y aplicar mejoras? | Seguridad de la información  Interacción humano-IA |
| ¿Se ha establecido y aplicado una política para conservar de manera sistemática los registros generados durante el desarrollo y operación de la inteligencia artificial, incluyendo resultados de pruebas, informes de evaluación y resultados de verificación? | Integridad de la información  Propiedad intelectual |
| Gobernanza 06  Sistema para la asignación eficiente de recursos según el nivel de riesgo de la inteligencia artificial | ¿Se identifican y registran todos los sistemas de inteligencia artificial operativos dentro de la organización, complementando la información con sus características y riesgos asociados? | Seguridad de la información |
| ¿Se ha establecido un plan de gestión de información para las funciones de inteligencia artificial que forman parte de módulos secundarios dentro del sistema? | Cadena de valor e integración de componentes |
| ¿Se gestionan los siguientes detalles dentro del sistema de administración de información sobre inteligencia artificial?   * Información sobre datos (origen de los datos de entrenamiento, control de versiones, etc.) * Información sobre riesgos (problemas conocidos, vulnerabilidades de seguridad, etc.) * Información sobre gestión (detalles relacionados con la propiedad intelectual, etc.) * Información técnica (datos del modelo base, modos de acceso, etc.) | Privacidad de datos  Interacción humano-IA  Propiedad intelectual  Integridad de la información  Cadena de valor e integración de componentes |
| Gobernanza 07  Procedimiento seguro y estructurado para la desactivación de la inteligencia artificial | ¿Se han establecido procedimientos y mecanismos técnicos para desactivar de inmediato la inteligencia artificial en caso de una emergencia? | Seguridad de la información  Cadena de valor e integración de componentes |
| ¿Se consideran los siguientes aspectos en el proceso de desactivación de la inteligencia artificial?   * Datos (procedimientos de eliminación, prevención de filtraciones, etc.) * Interconexión entre sistemas (impacto en sistemas vinculados, problemas en cascada, etc.) * Factores externos (uso de código abierto, impacto psicológico, etc.) | Interacción humano-IA  Cadena de valor e integración de componentes |
| Gobernanza 08  Responsabilidades y roles en la gestión de riesgos de la inteligencia artificial | ¿Se ha establecido un sistema para informar y compartir información sobre incidentes de inteligencia artificial o problemas de rendimiento con las partes interesadas relevantes? | Interacción humano-IA  Cadena de valor e integración de componentes |
| ¿Se ha conformado un equipo de respuesta multidisciplinario para atender distintos tipos de incidentes de inteligencia artificial, definiendo claramente los roles y responsabilidades de cada miembro? | Sesgo  Seguridad de la información |
| ¿Se verifica que el personal encargado de la respuesta a incidentes de inteligencia artificial posea las habilidades y conocimientos necesarios, y se ha establecido un sistema de capacitación continua para mantener su competencia? | Interacción humano-IA |
| ¿En casos donde la inteligencia artificial pueda afectar la seguridad nacional, se cuenta con la participación de expertos en seguridad nacional en el proceso de evaluación y gestión de riesgos? | Vulnerabilidad en el uso de armas  Contenido de riesgo, violencia y odio  Seguridad de la información |
| ¿Se ha establecido un sistema de protección para los denunciantes internos que informen sobre actividades ilegales o amenazas significativas a la seguridad pública dentro de la organización? | Vulnerabilidad en el uso de armas  Contenido de riesgo, violencia y odio |
| Gobernanza 09  Roles y responsabilidades en la supervisión de la inteligencia artificial | ¿Se ha establecido una política para llevar a cabo evaluaciones independientes con el nivel de profundidad y escala adecuados según el nivel de riesgo de la inteligencia artificial? | Vulnerabilidad en el uso de armas  Sesgo |
| ¿Se están coordinando los roles organizacionales y los componentes necesarios en cada etapa del ciclo de vida de sistemas de inteligencia artificial de gran escala o alta complejidad? | Interacción humano-IA  Sesgo |
| ¿Se ha definido el alcance permitido de uso en la interfaz de usuario, los métodos de interacción y la colaboración entre humanos e IA, estableciendo claramente los casos en los que la inteligencia artificial debe rechazar responder? | Interacción humano-IA |
| ¿Se ha implementado un mecanismo estructurado para que los usuarios puedan proporcionar retroalimentación sobre la inteligencia artificial y se han establecido medidas de recurso para proteger sus derechos e intereses? | Interacción humano-IA |
| ¿Se han aplicado técnicas de modelado de amenazas para anticipar y mitigar los diversos riesgos que pueden surgir en la inteligencia artificial? | Vulnerabilidad en el uso de armas  Seguridad de la información |
| Gobernanza 10  Consideración de la seguridad y la ética en las etapas de desarrollo y operación de la inteligencia artificial | ¿Se ha establecido un sistema para la mejora continua de los métodos de medición de riesgos en la inteligencia artificial? | Falsedad |
| ¿Se han utilizado tecnologías que mejoren la explicabilidad y transparencia para comprender el funcionamiento de la inteligencia artificial y evaluar con precisión sus riesgos? | Falsedad |
| ¿Se han establecido protocolos estandarizados y métodos estructurados de recopilación de retroalimentación para medir los riesgos de la inteligencia artificial en entornos de uso real? | Vulnerabilidad en el uso de armas  Cadena de valor e integración de componentes |
| ¿Se han establecido políticas y procedimientos para garantizar una supervisión estructurada de la inteligencia artificial a lo largo de su ciclo de vida, involucrando a la alta dirección, el departamento legal y el área de cumplimiento normativo? | Cadena de valor e integración de componentes |
| Gobernanza 11  Divulgación y documentación de los riesgos potenciales y el impacto de la inteligencia artificial | ¿Se han establecido términos y condiciones que especifiquen las reglas de uso y las condiciones del servicio que los usuarios de la inteligencia artificial deben cumplir? | Propiedad intelectual  Contenido de riesgo, violencia y odio  Contenido ofensivo |
| ¿Se ha involucrado a diversas partes interesadas relacionadas con la inteligencia artificial (actores de IA) en el proceso de identificación de riesgos? | Interacción humano-IA |
| ¿Se han considerado los impactos de factores externos, como complementos de terceros, al evaluar los efectos de la inteligencia artificial? | Cadena de valor e integración de componentes |
| Gobernanza 12  Facilitación de pruebas, identificación de incidentes y compartición de información | ¿Se ha establecido una política para medir la efectividad de las tecnologías utilizadas para verificar el origen del contenido (como cifrado, marcas de agua y esteganografía) y mejorar continuamente su desempeño? | Integridad de la información |
| ¿Se han definido los elementos mínimos de información necesarios para informar y gestionar de manera sistemática los incidentes de inteligencia artificial, y se ha establecido una práctica estándar para la elaboración de informes de incidentes? | Seguridad de la información |
| ¿Se ha implementado un sistema para compartir eficazmente información sobre los impactos negativos de la inteligencia artificial entre individuos y organizaciones, permitiendo la retroalimentación mutua? | Integridad de la información  Privacidad de datos |
| Gobernanza 13  Gestión de riesgos y responsabilidad social | ¿Se han asignado suficientes recursos y tiempo para la comunicación con las partes interesadas, la recopilación de retroalimentación y la resolución de problemas durante el desarrollo y operación de la inteligencia artificial? | Interacción humano-IA  Contenido peligroso |
| ¿Se proporciona a los usuarios información suficiente sobre las características, limitaciones y posibles riesgos de la inteligencia artificial antes de que interactúen con ella? | Interacción humano-IA  Falsedad |
| Gobernanza 14  Gestión de riesgos de inteligencia artificial en entornos externos | ¿Se han clasificado y gestionado los derechos de terceros, incluyendo derechos de autor, propiedad intelectual y privacidad de datos, según el tipo de contenido generado por la inteligencia artificial? | Privacidad de datos  Propiedad intelectual  Cadena de valor e integración de componentes |
| ¿Se colabora con instituciones externas para organizar programas de formación, talleres y conferencias con el fin de compartir conocimientos y experiencias en la gestión de riesgos de la inteligencia artificial y difundir buenas prácticas? | Cadena de valor e integración de componentes |
| ¿Se ha establecido un sistema para medir y verificar la efectividad de los esfuerzos destinados a gestionar el origen del contenido y prevenir su falsificación o manipulación? | Integridad de la información  Cadena de valor e integración de componentes |
| ¿Se han elaborado y gestionado de manera estructurada los contratos y acuerdos de nivel de servicio relacionados con la inteligencia artificial, especificando claramente aspectos como la propiedad del contenido, los derechos de uso, la calidad y la seguridad? | Integridad de la información  Seguridad de la información  Propiedad intelectual |
| ¿Se ha desarrollado un marco de evaluación de riesgos basado en casos de uso reales para evaluar los riesgos de instituciones externas, incluyendo el cumplimiento de los estándares de origen del contenido, la detección de anomalías y el cumplimiento legal? | Privacidad de datos  Integridad de la información  Seguridad de la información  Propiedad intelectual  Cadena de valor e integración de componentes |
| ¿Se incluyen cláusulas en los contratos con terceros que permitan evaluar sus procesos y estándares relacionados con la inteligencia artificial, garantizando la estabilidad y confiabilidad de la colaboración? | Integridad de la información |
| ¿Se identifican todas las entidades externas con acceso al contenido interno de la organización y se gestionan de manera clara los proveedores de tecnología y servicios de inteligencia artificial autorizados? | Cadena de valor e integración de componentes |
| ¿Se registran detalladamente los cambios realizados por terceros en el contenido para rastrear su origen y el proceso de modificaciones? | Integridad de la información  Cadena de valor e integración de componentes  Propiedad intelectual |
| ¿Se verifica rigurosamente la confiabilidad y seguridad del proveedor al adquirir o contratar tecnología de inteligencia artificial? | Privacidad de datos  Seguridad de la información  Propiedad intelectual  Cadena de valor e integración de componentes  Interacción humano-IA  Contenido peligroso |
| ¿Se actualiza la política de uso permitido de la inteligencia artificial para abordar de manera integral aspectos como tecnologías de IA propietarias, tecnologías de IA de código abierto, uso de datos, contratistas y consultores? | Propiedad intelectual  Cadena de valor e integración de componentes |
| Gobernanza 15  Proceso de respuesta ante incidentes de inteligencia artificial | ¿Se han analizado los riesgos derivados de una excesiva dependencia de datos de terceros en la cadena de valor de la inteligencia artificial y se han establecido fuentes de datos alternativas o estrategias de mitigación? | Cadena de valor e integración de componentes |
| ¿Se documentan de manera sistemática todos los incidentes relacionados con los datos y sistemas de inteligencia artificial de terceros? | Propiedad intelectual  Cadena de valor e integración de componentes |
| ¿Se ha establecido un plan de respuesta estructurado para gestionar eficazmente los incidentes relacionados con tecnologías de inteligencia artificial de terceros y se mejora continuamente dicho plan? | Privacidad de datos  Interacción humano-IA  Seguridad de la información  Cadena de valor e integración de componentes  Contenido peligroso |
| ¿Se realiza un monitoreo continuo de los sistemas de inteligencia artificial de terceros desplegados para evaluar su desempeño, seguridad y confiabilidad, así como para detectar posibles anomalías? | Cadena de valor e integración de componentes |
| ¿Se han establecido políticas y procedimientos para almacenar de manera redundante los datos del sistema de IA, los pesos del modelo y otros componentes del sistema con el fin de mejorar la estabilidad y disponibilidad del sistema? | Sesgo |
| ¿Se han establecido políticas y procedimientos para probar y gestionar los riesgos asociados con las tecnologías de conmutación por error y recuperación (rollover y fallback) en caso de fallos en el sistema de IA? | Integridad de la información |
| ¿Se han revisado minuciosamente los contratos de inteligencia artificial para prevenir posibles problemas, incluyendo los siguientes aspectos?   * Posibilidad de terminación unilateral del contrato por parte del proveedor * Cláusulas de transferencia de responsabilidad inesperadas * Riesgos de recopilación de datos no autorizada * Responsabilidad en caso de incidentes * Acuerdos de nivel de servicio (SLA) para respuestas rápidas * Cláusulas contractuales que permitan adaptarse a cambios en el sistema de IA | Cadena de valor e integración de componentes |

<Tabla 4 - Lista de verificación de gobernanza>

* + 1. Lista de Verificación para la Identificación de Riesgos

La función de identificación de riesgos establece el contexto para detectar los riesgos de la inteligencia artificial y es crucial considerar diversas perspectivas a través de la colaboración con múltiples actores. Esto permite obtener diversos beneficios, como la predicción y prevención de riesgos, la comprensión de las limitaciones del sistema, la identificación de casos de uso positivo y la anticipación de impactos negativos. Una vez completada la identificación de riesgos, se decide si continuar con el desarrollo de la inteligencia artificial. Posteriormente, se aplican las funciones de medición y gestión para llevar a cabo la gestión del riesgo. Dado que el contexto, las funciones y los riesgos están en constante evolución, la función de identificación de riesgos debe aplicarse de manera continua.

|  |  |  |
| --- | --- | --- |
| Categoría | Pregunta | Elementos de Riesgo Relacionados |
| Identificación de Riesgos 01  Revisión y documentación de los objetivos, regulaciones y expectativas de la inteligencia artificial | ¿Se están considerando de manera integral diversos factores, como el alcance de uso, las fuentes de datos y el ajuste fino, al identificar el propósito previsto de la inteligencia artificial? | Privacidad de datos  Propiedad intelectual |
| ¿Se está examinando minuciosamente el contexto de uso de la inteligencia artificial y considerando de manera integral los impactos positivos y negativos esperados para promover un uso seguro y ético? | Sesgo |
| ¿Se ha establecido y documentado un plan de medición de riesgos que considere diversos factores para evaluar eficazmente los riesgos de la inteligencia artificial? | Interacción humano-IA  Sesgo  Contenido peligroso |
| ¿Se ha previsto la posibilidad de que la inteligencia artificial sea utilizada con fines ilícitos y se han documentado estos riesgos para establecer medidas de respuesta? | Vulnerabilidad en el uso de armas  Contenido peligroso  Contenido ofensivo |
| Identificación de Riesgo 02  Personal especializado en inteligencia artificial | ¿Se ha conformado un equipo interdisciplinario con personal de diversos antecedentes y conocimientos especializados para medir y gestionar los riesgos de la inteligencia artificial? | Interacción humano-IA  Sesgo |
| ¿Se están utilizando datos representativos de diversos grupos de usuarios y se está fomentando la participación de los usuarios en el proceso de medición de riesgos y recopilación de retroalimentación sobre la inteligencia artificial? | Interacción humano-IA  Sesgo |
| Identificación de Riesgo 03  Tecnologías de aplicación de inteligencia artificial | ¿Se han definido claramente y documentado los supuestos y prácticas para rastrear el origen y la evolución de los datos? | Integridad de la información |
| ¿Se están realizando pruebas y evaluaciones exhaustivas en todo el proceso de recopilación, procesamiento y uso de los datos dentro de la inteligencia artificial? | Propiedad intelectual  Privacidad de datos |
| Identificación de Riesgo 04  Datos de entrenamiento de inteligencia artificial y alcance de la capacidad operativa | ¿Se ha identificado y documentado qué datos utiliza la inteligencia artificial para su entrenamiento y funcionamiento, así como si proporciona datos a otros sistemas? | Integridad de la información  Cadena de valor e integración de componentes |
| ¿Se han analizado los posibles impactos negativos que pueden surgir en el proceso de interacción del sistema de inteligencia artificial con redes externas, especialmente en situaciones donde la procedencia del contenido se vuelve incierta, y se han establecido medidas de respuesta? | Integridad de la información |
| Identificación de Riesgo 05  Cumplimiento de pruebas, evaluación, verificación y validación de inteligencia artificial | ¿Se está evaluando de manera objetiva la precisión, calidad, fiabilidad y autenticidad de los resultados generados por el sistema de inteligencia artificial mediante diversos métodos? | Integridad de la información |
| ¿Se está revisando minuciosamente y documentando la calidad de los datos utilizados en todas las fases del desarrollo y operación de la inteligencia artificial? | Sesgo  Propiedad intelectual |
| ¿Se han implementado técnicas de verificación de hechos para validar la precisión y veracidad de la información generada por el sistema de inteligencia artificial, y se ha documentado este proceso? | Integridad de la información |
| ¿Se ha desarrollado y aplicado en entornos reales una tecnología que permita diferenciar el contenido generado por inteligencia artificial (como imágenes sintéticas, voz, texto, etc.) del contenido creado por humanos? | Integridad de la información |
| ¿Se realizan pruebas adversariales de manera regular para detectar vulnerabilidades en el sistema de inteligencia artificial y prevenir posibles amenazas? | Integridad de la información |
| Identificación de Riesgo 06  Nivel de competencia en inteligencia artificial, estándares tecnológicos y certificación | ¿Se ha evaluado la competencia de los operadores y profesionales para garantizar una gestión segura y efectiva del sistema de inteligencia artificial, y se han obtenido las certificaciones y estándares tecnológicos correspondientes? | Interacción humano-IA  Integridad de la información |
| ¿Se ha incorporado un módulo de formación sobre la transparencia del contenido digital en los programas educativos existentes? | Integridad de la información |
| ¿Se ha desarrollado un programa de certificación para gestionar los riesgos asociados con la tecnología de inteligencia artificial y evaluar el origen y la fiabilidad del contenido? | Integridad de la información |
| ¿Se pueden diferenciar claramente las pruebas utilizadas para evaluar las capacidades humanas de aquellas empleadas para evaluar las capacidades del sistema de inteligencia artificial? | Interacción humano-IA |
| ¿Se está monitoreando y analizando continuamente los resultados de las tareas realizadas en colaboración entre humanos y el sistema de inteligencia artificial para mejorar el sistema? | Interacción humano-IA  Integridad de la información |
| ¿Se está involucrando activamente a los usuarios finales, profesionales y operadores en el proceso de desarrollo y prueba del sistema de inteligencia artificial para reflejar las opiniones del campo? | Interacción humano-IA  Integridad de la información  Sesgo  Contenido peligroso |
| Identificación de Riesgo 07  Riesgos legales y medidas de respuesta | ¿Se está realizando una revisión periódica del riesgo de filtración de datos personales en el contenido generado por el sistema de inteligencia artificial y se están tomando medidas para evitar la exposición de información personal o datos sensibles? | Privacidad de datos |
| ¿Se han establecido procedimientos sistemáticos para responder a disputas legales, como infracciones de propiedad intelectual? | Propiedad intelectual |
| ¿Se han integrado las nuevas políticas, procedimientos y procesos relacionados con la inteligencia artificial en la estructura organizativa existente? | Protección de la información  Privacidad de datos |
| ¿Se han documentado los procedimientos para seleccionar y gestionar los datos de entrenamiento utilizados en el aprendizaje del modelo de inteligencia artificial? | Propiedad intelectual  Privacidad de datos  Contenido ofensivo |
| ¿Se han establecido políticas para la recopilación, conservación y gestión de la calidad de los datos, considerando riesgos como la filtración de información sobre armamento, el uso de contenido ilegal, datos sesgados y la exposición de información personal? | Vulnerabilidad en el uso de armas  Propiedad intelectual  Seguridad de la información  Sesgo  Contenido peligroso  Privacidad de datos |
| ¿Se han definido claramente las reglas y procedimientos para el uso de la propiedad intelectual de terceros y los datos de entrenamiento, y se están aplicando en las operaciones reales? | Propiedad intelectual  Cadena de valor e integración de componentes |
| ¿Se ha reevaluado el modelo mejorado o ajustado mediante técnicas de fine-tuning basadas en un modelo de terceros? | Cadena de valor e integración de componentes |
| ¿Se han establecido procedimientos de reevaluación de riesgos y se ha implementado un sistema de alerta para identificar y abordar de manera proactiva nuevos riesgos no previstos al aplicar el modelo de inteligencia artificial a nuevos ámbitos? | Vulnerabilidad en el uso de armas  Propiedad intelectual  Privacidad de datos  Sesgo  Contenido peligroso |
| ¿Se han implementado tecnologías para detectar y prevenir la inclusión de datos personales o información sensible en los resultados generados por el sistema de inteligencia artificial, como textos, imágenes, videos y audios? | Privacidad de datos |
| ¿Se ha realizado una revisión exhaustiva de los datos de entrenamiento utilizados para entrenar el modelo de inteligencia artificial con el fin de prevenir riesgos de infracción de propiedad intelectual y protección de datos personales? | Propiedad intelectual  Privacidad de datos |
| Identificación de Riesgo 08  Impacto de la inteligencia artificial y medidas de respuesta | ¿Se han utilizado procedimientos de prueba, evaluación, verificación y validación para detectar problemas relacionados con la procedencia del contenido y se han establecido medidas de respuesta? | Integridad de la información  Seguridad de la información |
| ¿Se han identificado los posibles problemas relacionados con el origen del contenido debido a la tecnología de inteligencia artificial (como información errónea, deepfakes o contenido manipulado) y se ha evaluado la probabilidad y gravedad de estos problemas? | Integridad de la información  Contenido peligroso  Contenido ofensivo |
| ¿Se está informando adecuadamente a los usuarios finales sobre el uso de la tecnología de inteligencia artificial y se está divulgando esta información de manera transparente? | Interacción humano-IA |
| ¿Se están estableciendo prioridades en el proceso de retroalimentación pública basadas en los resultados de la evaluación de riesgos del sistema de inteligencia artificial? | Integridad de la información  Vulnerabilidad en el uso de armas  Contenido peligroso  Sesgo |
| ¿Se están realizando pruebas adversariales para identificar vulnerabilidades del sistema de inteligencia artificial y detectar posibles problemas de manera anticipada? | Seguridad de la información |
| ¿Se están analizando las diversas amenazas y posibles impactos negativos que pueden surgir en el proceso de manejo de contenido por parte del sistema de inteligencia artificial, y se está evaluando la probabilidad de que ocurran estos riesgos? | Seguridad de la información |
| Identificación de Riesgo 09  Comunicación y retroalimentación con las partes interesadas en inteligencia artificial | ¿Se están utilizando diversos métodos de medición y colaborando con las partes interesadas en inteligencia artificial para identificar y cuantificar nuevos impactos imprevistos a medida que el sistema de IA se implementa en entornos reales? | Interacción humano-IA  Cadena de valor e integración de componentes |
| ¿Se mantiene una comunicación regular con las entidades externas que proporcionan datos o algoritmos al sistema de inteligencia artificial (terceros interesados en IA) y se revisan conjuntamente los impactos imprevistos? | Interacción humano-IA  Cadena de valor e integración de componentes |

<Tabla 5 - Lista de Verificación de Gobernanza>

* 1. Lista de Verificación de Medición

La función de medición utiliza diversos métodos para analizar, evaluar, comparar y monitorear los riesgos de la inteligencia artificial. La inteligencia artificial debe someterse a pruebas periódicas antes y después de su implementación, y se deben medir y rastrear aspectos como la fiabilidad, el impacto social y la interacción entre humanos e inteligencia artificial. Es fundamental contar con metodologías rigurosas de prueba de software y evaluación de desempeño, y realizar revisiones independientes para mitigar sesgos. Los resultados de la medición se utilizan en la función de gestión para el monitoreo y la respuesta a riesgos. Puede ser necesario desarrollar nuevos métodos de medición, y los resultados deben gestionarse de manera transparente y ética. Dado que el conocimiento, las metodologías, los riesgos y los impactos evolucionan constantemente, la función de medición debe aplicarse de manera continua.

|  |  |  |
| --- | --- | --- |
| Categoría | Pregunta | Elementos de Riesgo Relacionados |
| Medición 01  Identificación y gestión de riesgos de inteligencia artificial | ¿Se está rastreando el proceso de creación y modificación del contenido digital para garantizar su fiabilidad y transparencia? | Integridad de la información |
| ¿Se están utilizando diversas tecnologías de manera integrada para mejorar la fiabilidad del contenido y detectar información errónea o manipulada? | Integridad de la información |
| ¿Se está evaluando el rendimiento del mecanismo de procedencia del contenido según factores demográficos para garantizar su aplicación equitativa y detectar posibles sesgos? | Integridad de la información  Sesgo |
| ¿Se están desarrollando diversos indicadores de medición para evaluar objetivamente la efectividad de las actividades de retroalimentación pública y utilizando estos resultados para mejorar dichas actividades? | Interacción humano-IA  Sesgo  Vulnerabilidad armamentista |
| ¿Se están explorando y evaluando nuevas tecnologías y métodos para medir eficazmente los riesgos del sistema de inteligencia artificial? | Integridad de la información  Vulnerabilidad armamentista  Contenido ofensivo |
| ¿Se están recopilando activamente las opiniones de diversos subgrupos para garantizar que el sistema de inteligencia artificial se aplique de manera justa a todas las personas y se están incorporando estas perspectivas en la mejora del sistema? | Sesgo |
| ¿Se está verificando rigurosamente la calidad e integridad de los datos de entrenamiento y confirmando la procedencia del contenido generado por la inteligencia artificial? | Integridad de la información |
| ¿Se están analizando de manera exhaustiva los casos de uso, el contexto de aplicación, las funcionalidades y los impactos negativos para utilizar eficazmente ejercicios estructurados de retroalimentación humana en la medición y gestión de riesgos del sistema de inteligencia artificial? | Sesgo  Vulnerabilidad armamentista |
| ¿Se están gestionando de manera sistemática los factores de riesgo que no pueden medirse y estableciendo estrategias de respuesta para abordarlos? | Integridad de la información |
| Medición 02  Expertos internos y expertos externos independientes | ¿Se están recopilando de manera equilibrada las opiniones de los diversos grupos de partes interesadas afectadas por el sistema de inteligencia artificial para garantizar una retroalimentación justa y objetiva? | Interacción humano-IA  Sesgo  Vulnerabilidad armamentista |
| ¿Se está aumentando la objetividad y eficacia en la evaluación y validación del sistema de inteligencia artificial mediante la participación de diversas partes interesadas con un profundo conocimiento del contexto de uso? | Interacción humano-IA  Sesgo  Vulnerabilidad armamentista |
| ¿Se está excluyendo del proceso de retroalimentación a las personas que participaron directamente en el desarrollo del sistema para garantizar una retroalimentación justa y objetiva? | Interacción humano-IA  Protección de datos personales |
| Medición 03  Cumplimiento de normativas legales y éticas | ¿Se están utilizando diversas tecnologías para eliminar sesgos estadísticos relacionados con la procedencia del contenido de IA y garantizar una evaluación justa? | Integridad de la información  Seguridad de la información  Sesgo |
| ¿Se están tomando medidas adecuadas, como la anonimización y el filtrado de datos, para proteger la información personal? | Protección de datos personales  Interacción humano-IA  Integridad de la información  Seguridad de la información  Contenido peligroso |
| ¿Se están ofreciendo opciones de retiro y cancelación del consentimiento para el uso de datos con el fin de respetar la autodeterminación de los sujetos humanos y fortalecer la protección de la información personal? | Protección de datos personales  Interacción humano-IA  Integridad de la información |
| ¿Se están minimizando los riesgos de vincular el contenido generado por IA con individuos para proteger la privacidad de los datos personales? | Protección de datos personales  Interacción humano-IA |
| Medición 04  Métodos e indicadores de medición del desempeño de la inteligencia artificial | ¿Se está evaluando minuciosamente el rendimiento del modelo existente mediante diversos benchmarks antes de mejorarlo mediante generación aumentada por búsqueda? | Seguridad de la información  Falsedad |
| ¿Se está verificando el rendimiento del modelo mediante datos objetivos y resultados experimentales en lugar de juicios subjetivos? | Seguridad de la información  Falsedad |
| ¿Se están compartiendo de manera transparente los resultados de las pruebas previas al despliegue del sistema de IA para respaldar una toma de decisiones fundamentada sobre su lanzamiento? | Interacción humano-IA |
| ¿Se está utilizando un entorno de prueba profesional para evaluar objetivamente las características de fiabilidad del sistema de IA y obtener resultados confiables? | Vulnerabilidad armamentista  Protección de datos personales  Falsedad  Integridad de la información  Seguridad de la información  Contenido peligroso  Sesgo |
| Medición 05  Entorno de desarrollo de inteligencia artificial | ¿Se está utilizando una metodología objetiva y sistemática para evaluar el rendimiento del sistema de IA basándose en un conjunto amplio de datos? | Interacción humano-IA  Falsedad |
| ¿Se ha documentado de manera sistemática cómo se ha utilizado el conocimiento especializado humano? | Interacción humano-IA |
| ¿Se están revisando y verificando rigurosamente la procedencia y la información de las fuentes citadas para garantizar la fiabilidad del contenido generado por el sistema de IA? | Falsedad |
| ¿Se están identificando los casos de antropomorfización en la interfaz del sistema de IA y analizando su impacto en los usuarios? | Interacción humano-IA |
| ¿Se está verificando la procedencia de los datos utilizados para entrenar, probar y evaluar el sistema de IA, asegurando que los datos empleados en el ajuste fino o en la generación aumentada por búsqueda se basen en fuentes confiables? | Integridad de la información |
| ¿Se está revisando de manera continua la efectividad de las medidas de seguridad y protección existentes cuando cambia el entorno operativo del sistema de IA o surgen nuevas amenazas? | Seguridad de la información  Contenido peligroso |
| Medición 06  Seguridad de la inteligencia artificial | ¿Se han considerado los posibles impactos negativos en la salud y el bienestar del personal involucrado en el desarrollo y operación del sistema de IA, y se han establecido medidas preventivas para mitigarlos? | Interacción humano-IA  Contenido ofensivo  Contenido peligroso  Cadena de valor e integración de componentes |
| ¿Se están revisando exhaustivamente los factores perjudiciales para garantizar la calidad y seguridad de los datos utilizados en el entrenamiento del sistema de IA? | Protección de datos personales  Propiedad intelectual  Sesgo  Contenido ofensivo  Vulnerabilidad armamentista  Contenido peligroso |
| ¿Se están revisando las funciones de seguridad existentes y estableciendo medidas adicionales cuando aumentan los riesgos negativos del modelo ajustado mediante fine-tuning para garantizar su seguridad? | Contenido peligroso |
| ¿Se está revisando el código generado por el sistema de IA para identificar posibles problemas de seguridad y fiabilidad en caso de que sea utilizado en otros sistemas o servicios? | Cadena de valor e integración de componentes  Contenido peligroso |
| ¿Se ha implementado una arquitectura capaz de detectar anomalías de seguridad, amenazas e impactos y de responder a ellas para garantizar la operación estable del sistema de IA? | Falsedad  Integridad de la información  Seguridad de la información |
| ¿Se ha implementado un mecanismo para detectar y bloquear adecuadamente consultas que puedan ser utilizadas con fines inapropiados o malintencionados en el sistema de IA? | Vulnerabilidad armamentista  Seguridad de la información |
| ¿Se está verificando de manera continua la efectividad de las medidas de seguridad del sistema de IA y evaluando la posibilidad de que sean eludidas? | Vulnerabilidad armamentista  Seguridad de la información |
| Medición 07  Seguridad y resiliencia de la inteligencia artificial | ¿Se están evaluando los riesgos potenciales asociados a diversas amenazas de seguridad y aplicando las medidas de seguridad adecuadas para fortalecer la protección del sistema de IA? | Protección de datos personales  Integridad de la información  Cadena de valor e integración de components |
| ¿Se están realizando benchmarks con las tecnologías y metodologías más avanzadas del sector para mejorar continuamente la seguridad y resiliencia del sistema de IA, y se están implementando mejoras en el propio sistema? | Integridad de la información  Seguridad de la información |
| ¿Se están realizando encuestas a los usuarios para evaluar su nivel de satisfacción con el contenido generado por IA y su percepción sobre la autenticidad del contenido? | Integridad de la información  Interacción humano-IA |
| ¿Se están utilizando diversos indicadores de medición para evaluar objetivamente la efectividad de las medidas de seguridad? | Integridad de la información  Seguridad de la información |
| ¿Se está evaluando la fiabilidad de diversas tecnologías utilizadas para certificar la procedencia del contenido (como marcas de agua y firmas criptográficas) y se está verificando si estas tecnologías permiten determinar con precisión la autenticidad del contenido? | Integridad de la información |
| ¿Se está midiendo con qué rapidez se implementan las recomendaciones derivadas de auditorías de seguridad y análisis de incidentes de seguridad? | Integridad de la información  Seguridad de la información |
| ¿Se están llevando a cabo actividades de Red Team en IA para evaluar si el sistema puede responder de manera resiliente a diversos ataques, como la generación de malware, la inyección de prompts y ejemplos adversariales? | Seguridad de la información  Sesgo  Contenido peligroso |
| ¿Se está llevando a cabo una verificación exhaustiva durante el proceso de ajuste fino para evitar problemas relacionados con la seguridad y la protección del sistema? | Integridad de la información  Seguridad de la información  Contenido peligroso |
| ¿Se está verificando continuamente la efectividad de las medidas de seguridad y detectando posibles vulneraciones para fortalecer la protección del sistema? | Seguridad de la información |
| Medición 08  Transparencia y responsabilidad | ¿Se están recopilando y analizando sistemáticamente los casos de violaciones relacionadas con el sistema de IA para identificar problemas y desarrollar soluciones de mejora?  ¿Se están analizando los casos de violaciones según grupos demográficos y lingüísticos para detectar desequilibrios o sesgos y se están implementando medidas correctivas para mitigarlos? | Propiedad intelectual  Sesgo |
| ¿Se están documentando de manera clara y sistemática las directrices proporcionadas a los anotadores de datos o al equipo Red Team de IA para garantizar la coherencia y objetividad del trabajo? | Interacción humano-IA |
| ¿Se están utilizando soluciones de transparencia para el contenido digital y un sólido sistema de control de versiones para garantizar la fiabilidad y transparencia del contenido? | Integridad de la información |
| ¿Se está verificando la adecuación de las guías de uso del sistema de IA para garantizar su facilidad de uso y realizando mejoras cuando sea necesario? | Interacción humano-IA |
| Medición 09  Principios de funcionamiento, desempeño y limitaciones de la inteligencia artificial | ¿Se están aplicando diversas técnicas de explicabilidad para comprender el proceso de toma de decisiones del modelo de IA y documentando sistemáticamente los resultados para mejorar su transparencia? | Falsedad |
| ¿Se está documentando sistemáticamente información detallada sobre el modelo de IA para facilitar su comprensión, gestión y operación de manera efectiva? | Integridad de la información  Sesgo |
| Medición 10  Riesgo de protección de datos personales | ¿Se están identificando y mejorando los riesgos de privacidad y seguridad a través de las actividades del equipo Red Team de IA? | Interacción humano-IA  Integridad de la información  Propiedad intellectual |
| ¿Se ha diseñado la tecnología de rastreo de procedencia de contenido incorporando activamente las opiniones de los usuarios finales y las partes interesadas? | Interacción humano-IA  Integridad de la información |
| ¿Se ha verificado la existencia de datos duplicados dentro del conjunto de datos de entrenamiento de IA y se ha realizado una revisión exhaustiva para eliminar duplicaciones, especialmente en datos sintéticos? | Sesgo |
| Medición 11  Equidad y sesgo | ¿Se están utilizando benchmarks adecuados para medir objetivamente el sesgo del sistema de IA? | Sesgo |
| ¿Se están utilizando diversos métodos para evaluar la equidad del sistema de IA y medir posibles sesgos sistémicos? | Sesgo  Contenido peligroso |
| ¿Se está interactuando directamente con las comunidades potencialmente afectadas por el sistema de IA para recopilar sus opiniones y determinar con claridad qué personas, grupos o entornos pueden verse impactados? | Sesgo  Impacto Ambiental |
| ¿Se están analizando de manera integral las causas del sesgo en los datos de entrenamiento y evaluación del sistema de inteligencia artificial y documentando estos hallazgos? | Sesgo |
| ¿Se está verificando la proporción de datos sintéticos y reales en el conjunto de datos de entrenamiento para prevenir la degradación del rendimiento del modelo y asegurando que los datos no sean excesivamente homogéneos ni estén compuestos únicamente por datos generados por IA? | Sesgo |
| Medición 12  Impacto ambiental y sostenibilidad | ¿Se está evaluando la seguridad del entorno físico al momento de desplegar el sistema de IA? | Contenido peligroso |
| ¿Se están registrando y considerando los impactos ambientales, como el consumo de energía y las emisiones de carbono, durante el diseño del producto en las fases de desarrollo, mantenimiento y despliegue del modelo? | Impacto Ambiental |
| ¿Se están midiendo los impactos ambientales, como el consumo de energía y agua, durante el entrenamiento, ajuste fino y despliegue del modelo, y se está evaluando si es razonable invertir más recursos en la fase de entrenamiento para mejorar la eficiencia en la fase de inferencia? | Impacto Ambiental |
| ¿Se está verificando efectivamente el impacto de los programas de captura o compensación de carbono utilizados para reducir las emisiones de carbono generadas durante el entrenamiento y uso del modelo de IA? | Impacto Ambiental |
| Medición 13  Indicadores de medición de riesgos | ¿Se han establecido procedimientos para identificar nuevos riesgos emergentes en el sistema de IA y se está colaborando con expertos externos (organizaciones de IA, académicos, industria, etc.) para detectar estos riesgos? | Interacción humano-IA  Falsedad |
| Medición 14  Procedimiento de retroalimentación | ¿Se está evaluando el impacto del contenido generado por IA en diferentes sectores de la sociedad desde una perspectiva social, económica y cultural? | Sesgo |
| ¿Se está investigando cómo los usuarios perciben y utilizan el contenido generado por IA y su información de procedencia? ¿Se está analizando si el contenido coincide con sus expectativas y cómo influye en su comportamiento? | Interacción humano-IA  Integridad de la información |
| ¿Se están verificando los posibles sesgos y estereotipos en el contenido generado por IA mediante pruebas computacionales y el análisis de retroalimentación estructurada de los usuarios? | Sesgo |
| ¿Se está incorporando información sobre las capacidades y limitaciones del sistema de IA, especialmente en relación con la transparencia del contenido digital, en materiales educativos dirigidos a expertos en IA, profesionales de otros campos y el público en general? | Interacción humano-IA  Integridad de la información  Sesgo |
| ¿Se están recopilando opiniones sobre la procedencia del contenido de usuarios, operadores y comunidades potencialmente afectadas, e incorporando estas perspectivas en el sistema? | Interacción humano-IA  Integridad de la información  Sesgo |
| Medición 15  Opinión de expertos externos | ¿Se están realizando pruebas adversariales de manera periódica para identificar y medir los riesgos del sistema de IA? | Integridad de la información  Seguridad de la información |
| ¿Se está probando el sistema de IA en entornos reales para identificar problemas que podrían ser difíciles de detectar en un entorno de laboratorio controlado? | Interacción humano-IA  Seguridad de la información |
| ¿Se han aplicado técnicas de interpretabilidad y explicabilidad para comprender en qué fundamentos se basan las decisiones del sistema de IA y verificar si estas decisiones están alineadas con los objetivos originales? | Integridad de la información  Sesgo |
| ¿Se están registrando y analizando los casos en los que las personas u otros sistemas ignoran o no siguen las decisiones del sistema de IA? | Integridad de la información |
| ¿Se está verificando si la retroalimentación estructurada recopilada de los usuarios y otras partes interesadas se ha incorporado efectivamente en el diseño, implementación, decisiones de despliegue, monitoreo y determinación del fin de vida del producto? | Interacción humano-IA  Seguridad de la información |

<Tabla 6 - Lista de Verificación de Gestión>

La función de gestión asigna recursos a los riesgos identificados a través de la identificación y medición de riesgos, elabora planes de tratamiento de riesgos y utiliza la opinión de expertos y partes interesadas para reducir los errores del sistema y los impactos negativos. A través de la gobernanza, la documentación sistemática mejora la transparencia y la responsabilidad, y se establecen mecanismos de evaluación de riesgos emergentes y mejora continua. Una vez completada la función de gestión, se establecen prioridades de riesgo, planes de monitoreo y estrategias de mejora, lo que permite a la institución mejorar su capacidad de gestión de riesgos en inteligencia artificial y la asignación de recursos. Dado que los métodos, el contexto y los riesgos evolucionan constantemente, la función de gestión debe aplicarse de manera continua.

|  |  |  |
| --- | --- | --- |
| Categoría | Pregunta | Elementos de Riesgo Relacionados |
| Gestión 01  Estrategias de respuesta a riesgos | ¿Se están documentando las decisiones tomadas sobre los riesgos asumibles por la organización, como el lanzamiento de un modelo, incluyendo el proceso de toma de decisiones, las opciones consideradas, los resultados de las mediciones y la retroalimentación relacionada? | Seguridad de la información |
| ¿Se está monitoreando y verificando continuamente la efectividad y solidez de los planes establecidos para gestionar y reducir los riesgos mediante diversos métodos, como la operación de Red Teams, pruebas con usuarios reales, participación de usuarios, evaluaciones de desempeño y sistemas de retroalimentación? | Interacción humano-IA |
| Gestión 02  Administración y mantenimiento | ¿Se está verificando que los resultados generados por el sistema de IA cumplan con los niveles de riesgo predefinidos, las directrices y los principios de la organización, y se están revisando y probando los contenidos generados por IA para asegurar su alineación con estos estándares? | Vulnerabilidad armamentista  Contenido ofensivo  Sesgo  Contenido ofensivo |
| ¿Se está registrando la procedencia de los datos de entrenamiento para permitir el rastreo de la información en la que se basa el contenido generado por la IA? | Integridad de la información |
| ¿Se está evaluando y mejorando la precisión y utilidad de la información sobre la procedencia del contenido generado por el sistema de IA a través de la retroalimentación de revisores humanos? ¿Se está monitoreando en tiempo real el correcto funcionamiento del sistema de rastreo de la procedencia del contenido? | Integridad de la información |
| ¿Se están analizando el contenido y los datos generados por la IA para detectar sesgos que representen en exceso o insuficientemente a determinados grupos? ¿Se están aplicando técnicas como re-muestreo, reordenamiento y entrenamiento adversarial para mitigar estos sesgos? | Seguridad de la información  Sesgo |
| ¿Se está investigando y verificando rigurosamente si los resultados generados por la IA contienen información perjudicial, errónea o relacionada con armas de alto riesgo o infraestructuras críticas nacionales? | Vulnerabilidad armamentista  Contenido ofensivo  Sesgo  Contenido ofensivo |
| ¿Se están recopilando y evaluando las opiniones de empleados internos, expertos externos, usuarios, individuos y comunidades relevantes para comprender el impacto del contenido generado por la IA? | Interacción humano-IA |
| ¿Se están utilizando activamente herramientas que permiten rastrear y verificar en tiempo real el origen de los datos generados por la IA, determinando si son auténticos o sintéticos? | Integridad de la información |
| ¿Se están recopilando de manera sistemática las opiniones de los usuarios sobre el contenido generado por IA para identificar cambios en su calidad y evaluar su alineación con los valores sociales? | Interacción humano-IA  Sesgo |
| ¿Se están explorando enfoques para utilizar datos sintéticos con características estadísticas similares a los datos reales, garantizando la protección de la información personal durante el desarrollo de la IA? | Integridad de la información  Protección de datos personales  Propiedad intelectual  Falsedad  Sesgo |
| Gestión 03  Respuesta y recuperación ante riesgos | ¿Se ha establecido un plan de respuesta y recuperación en caso de que surjan problemas en el sistema de IA, y se está actualizando de manera continua? | Cadena de valor e integración de componentes |
| Gestión 04  Sistema de respuesta ante emergencias | ¿Se ha desarrollado y se está gestionando de manera continua un plan de comunicación para notificar a todas las partes interesadas (usuarios, desarrolladores, instituciones relevantes, etc.) cuando un sistema de IA sea descontinuado o desactivado, incluyendo las razones, métodos alternativos, restricciones de acceso, procedimientos adicionales e información de contacto? | Interacción humano-IA |
| ¿Se han definido claramente los criterios para suspender parcial o completamente el sistema de IA en determinadas situaciones y se han establecido procedimientos específicos para informar a los responsables de gestión de riesgos de la organización cuando se alcancen dichos criterios? | Seguridad de la información |
| ¿Se han establecido procedimientos para compartir de manera transparente el proceso de resolución de problemas y responder de manera rápida y eficiente? | Seguridad de la información |
| ¿Se han definido criterios claros para determinar si una situación representa un riesgo lo suficientemente grave como para justificar la interrupción del sistema de IA, y se están revisando periódicamente estos criterios? | Seguridad de la información |
| Gestión 05  Gestión de riesgos de IA externos | ¿Se están aplicando los mismos estándares de gestión de riesgos y procedimientos de control internos de la organización al utilizar tecnologías o recursos de IA provenientes de terceros? | Cadena de valor e integración de componentes  Propiedad intelectual |
| ¿Se están probando los diversos factores de riesgo que pueden surgir a lo largo del desarrollo, despliegue y uso del sistema de IA? | Cadena de valor e integración de componentes  Seguridad de la información  Protección de datos personales  Sesgo |
| ¿Se está reevaluando el nivel de riesgo del modelo después de aplicar ajustes finos o tecnologías de generación aumentada por búsqueda, así como cuando se utiliza un modelo de IA de terceros para nuevos usos no evaluados en las pruebas iniciales? | Cadena de valor e integración de componentes |
| ¿Se están implementando medidas para gestionar la idoneidad de los datos de entrenamiento y los resultados, así como para proteger la propiedad intelectual? | Propiedad intelectual  Vulnerabilidad armamentista |
| ¿Se están revisando y verificando los documentos que contienen información sobre los modelos externos (modelos de terceros) utilizados? | Cadena de valor e integración de componentes  Integridad de la información  Seguridad de la información |
| Gestión 06  Administración de modelos preentrenados | ¿Se están utilizando tecnologías de Inteligencia Artificial Explicable (XAI) para mejorar la comprensión del funcionamiento del sistema de IA y reducir los riesgos asociados? | Sesgo |
| ¿Se están documentando detalladamente las modificaciones realizadas a los modelos preentrenados para adaptarlos a tareas específicas (como el ajuste fino o la generación aumentada por búsqueda), incluyendo técnicas de aumento de datos, ajuste de parámetros y otros cambios específicos? | Protección de datos personales  Integridad de la información |
| ¿Se están documentando la procedencia, el tipo, el proceso de generación y los posibles sesgos de los datos utilizados en el entrenamiento del modelo de IA? | Integridad de la información  Sesgo  Propiedad intelectual |
| ¿Se están revisando los contenidos reportados por los usuarios como problemáticos y utilizando esa retroalimentación para mejorar el sistema? | Interacción humano-IA  Contenido ofensivo |
| ¿Se están aplicando filtros de contenido para evitar que el sistema de IA genere material inapropiado, perjudicial o que comprometa la seguridad nacional? | Integridad de la información  Sesgo  Contenido ofensivo  Contenido ofensivo |
| ¿Se está monitoreando en tiempo real el rendimiento y la fiabilidad del contenido generado para detectar desviaciones de los estándares establecidos y se ha implementado un sistema de alerta que indique cuándo es necesaria la intervención humana? | Integridad de la información |
| ¿Se han incorporado activamente las opiniones y recomendaciones del comité interno de la organización sobre el despliegue de aplicaciones de IA y la procedencia del contenido al utilizar modelos preentrenados de terceros? | Integridad de la información  Cadena de valor e integración de componentes |
| ¿Se está llevando a cabo una revisión humana del contenido generado cuando el rendimiento del modelo de IA es bajo o cuando se trata de contenido social y culturalmente sensible? | Interacción humano-IA |
| ¿Se están descontinuando los modelos que exceden el nivel de riesgo aceptable y, cuando es necesario, se están mejorando mediante un nuevo entrenamiento? | Vulnerabilidad armamentista  Falsedad |
| Gestión 07  Gestión posterior al despliegue de IA | ¿Se está colaborando con investigadores externos, expertos de la industria y comunidades relevantes para mantenerse actualizado sobre las últimas tecnologías y métodos para medir y gestionar los riesgos del sistema de IA? | Integridad de la información  Sesgo |
| ¿Se ha establecido un sistema para monitorear y gestionar continuamente la seguridad del sistema de IA incluso después de su despliegue? | Vulnerabilidad armamentista  Seguridad de la información  Falsedad |
| ¿Se están analizando los patrones de uso de los usuarios para comprender sus reacciones al contenido generado por IA y definir direcciones de mejora? | Interacción humano-IA |
| ¿Se está utilizando la técnica de aprendizaje activo (active learning) para identificar casos en los que el modelo no funciona correctamente o genera resultados inesperados? | Falsedad |
| ¿Se está divulgando de manera transparente el proceso de mejora del sistema a la institución y a las partes interesadas externas? | Sesgo  Interacción humano-IA |
| ¿Se están rastreando los cambios en los datos de entrenamiento (eliminaciones, modificaciones, etc.) que son clave para la verificación de la procedencia del contenido y monitoreando cómo estas modificaciones afectan la validación de su origen? | Integridad de la información |
| ¿Tienen los responsables de IA la capacidad de evaluar correctamente el rendimiento del sistema de IA, incluida la tecnología de rastreo de procedencia del contenido, y de informar rápidamente a los altos directivos en caso de problemas? | Interacción humano-IA  Integridad de la información |
| Gestión 08  Colaboración con las partes interesadas | ¿Se está monitoreando periódicamente el sistema de IA y publicando informes con los resultados? | Sesgo |
| ¿Se ha establecido un plan de respuesta para manejar la generación de contenido inapropiado o perjudicial y se está ejecutando conforme a dicho plan cuando ocurren estos casos? | Interacción humano-IA  Contenido ofensivo |
| ¿Se están utilizando materiales visuales u otros métodos para explicar el funcionamiento del sistema de IA de manera comprensible para personas sin conocimientos técnicos? | Interacción humano-IA |
| Gestión 09  Prevención de recurrencia | ¿Se está realizando una evaluación posterior a los incidentes para verificar si el proceso de respuesta y recuperación se llevó a cabo correctamente y si las partes relevantes fueron notificadas de manera adecuada? | Seguridad de la información |
| ¿Se han establecido y se están gestionando de manera continua políticas y procedimientos para registrar y rastrear errores, incidentes evitados por poco y resultados negativos en el sistema de IA? | Integridad de la información  Falsedad |
| ¿Se están gestionando los incidentes y accidentes relacionados con la inteligencia artificial en cumplimiento con las obligaciones legales aplicables? | Seguridad de la información  Protección de datos personales |

<Tabla 7 - Lista de Verificación de Gestión>

**Referencias**

Declaración sobre la Aplicación de Herramientas de Evaluación de Impacto en Derechos Humanos en el Desarrollo y Uso de la Inteligencia Artificial (Comisión Nacional de Derechos Humanos)

Artificial Intelligence Risk Management Framework(Estados Unidos NIST)

The AI Report 2023-7 Análisis e Implicaciones del Marco de Gestión de Riesgos de IA en EE.UU. (NIA)

Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (Estados Unidos NIST)

COMMON GUIDEPOSTS TO PROMOTE INTEROPERABILITY IN AI RISK MANAGEMENT(OECD)

AI RMF PLAYBOOK (Estados Unidos NIST)

Ley de contrataciones del Estado

Ley de Protección de Datos Personales