TIC

Lineamientos de Seguridad y Privacidad de la Información para Sistemas de Inteligencia Artificial

Ministerio de tecnologías de la información y las comunicaciones

MSPI

Carina Murcia Yela – Ministra de Tecnologías de la Información y las Comunicaciones
Giovanny Andrés López Cabezas - Viceministro de Transformación Digital
Lucy Elena Urón Rincón - Directora de Gobierno Digital
Lucy Estella Palacios Valoyes - Subdirectora de Estándares y Arquitectura de TI
Johanna Marcela Forero Varela - Funcionaria de la Subdirección de Estándares y Arquitectura de TI
Tairo Elias Mendoza Piedrahita – Funcionario de la Dirección de Gobierno Digital
Jose Luis Alejandro Carrillo Valderrama – Funcionario de la Subdirección de Estándares y Arquitectura de TI
Germán García Filoth - Contratista de la Subdirección de Estándares y Arquitectura de TI
Danny Alejandro Garzón Aristizábal – Contratista de la Subdirección de Estándares y Arquitectura de TI
Lourdes María Acuña Acuña - Contratista de la Subdirección de Estándares y Arquitectura de TI.

Ministerio de Tecnologías de la Información y las Comunicaciones
Viceministerio de Transformación Digital
Dirección de Gobierno Digital

Versión	**Observaciones**
Versión 1 19/12/2025	Lineamientos de Seguridad y Privacidad de la Información para Sistemas de Inteligencia Artificial Dirigida a las entidades del Estado

Comentarios, sugerencias o correcciones pueden ser enviadas al correo electrónico: gobiernodigital@mintic.gov.co

Lineamientos de Seguridad y Privacidad de la Información para Sistemas de Inteligencia Artificial
V 1.0

Tabla de contenido

Listado de tablas

Listado de ilustraciones

Lineamientos de Seguridad y Privacidad de la Información para Sistemas de Inteligencia Artificial

1. Derechos de autor

Para el desarrollo de este lineamiento, se recogieron aspectos importantes de mejores prácticas y documentos de uso libre por parte del NIST (National Institute of Standards and Technology – (Artificial Intelligence Risk Management Framework (AI RMF 1.0)) y ((International Organization for Standardization) / IEC (International Electrotechnical Commission). (2023). ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management System).

2. Audiencia

Entidades públicas de orden nacional y entidades públicas del orden territorial, así como proveedores de servicios de Gobierno Digital, y terceros que deseen adoptar el Modelo de Seguridad y Privacidad de TI en el marco de la estrategia de seguridad digital.

3. Definiciones

- **Ajuste fino (Fine-tuning):** Proceso mediante el cual un modelo de inteligencia artificial preentrenado (como un Modelo de Lenguaje Grande) se somete a una fase adicional de entrenamiento utilizando un conjunto de datos más pequeño, específico y especializado. Su objetivo es adaptar el modelo para que realice tareas particulares o domine un dominio de conocimiento concreto con mayor precisión. Desde la perspectiva de la seguridad y privacidad, este proceso requiere controles estrictos, ya que la utilización de

datos no verificados, contaminados o confidenciales durante el ajuste fino puede introducir vulnerabilidades, sesgos, o derivar en la memorización y posterior fuga de información sensible.

- **Alucinación (Hallucination):** Fenómeno asociado principalmente a los modelos de inteligencia artificial generativa (como los LLM), en el cual el sistema produce respuestas, datos o afirmaciones que, aunque se presentan con un alto grado de confianza y coherencia gramatical, son fácticamente incorrectas, carecen de sentido lógico o no están respaldadas por sus datos de entrenamiento ni por el contexto proporcionado. En entornos operativos o de toma de decisiones, las alucinaciones representan un riesgo crítico para la fiabilidad e integridad de la información.

- **Ataque Adversarial (Adversarial Attack):** Técnica maliciosa diseñada para engañar y manipular modelos de aprendizaje automático mediante la alteración sutil y calculada de los datos de entrada. Su propósito es provocar que el sistema cometa errores de clasificación o genere resultados incorrectos sin que dicha manipulación sea evidente para un observador humano.

- **Ataque por latencia energética:** Un ataque que explota la dependencia del rendimiento respecto a optimizaciones de hardware y modelos para anular los efectos de las optimizaciones de hardware, aumentar la latencia computacional, incrementar la temperatura del hardware y aumentar masivamente la cantidad de energía consumida.

- **Ataque Rowhammer:** Un ataque de inyección de fallos basado en software que explota errores dinámicos de perturbación de memoria de acceso aleatorio a través de aplicaciones en espacio de usuario y permite al atacante inferir información sobre ciertos secretos de víctimas almacenados en celdas de memoria. Llevar a cabo este ataque requiere que el atacante controle un proceso no privilegiado en espacio de usuario que se ejecute en la misma máquina que el modelo de aprendizaje automático de la víctima.

- **Ataques a la privacidad de datos:** Ataques contra modelos de aprendizaje automático que extraen información sensible sobre datos de entrenamiento.

- **Ataques de envenenamiento (Poisoning attacks):** Ataques adversariales en los que un adversario interfiere con un modelo durante su etapa de entrenamiento, como insertando datos de entrenamiento maliciosos (envenenamiento de datos) o modificando el propio proceso de entrenamiento (envenenamiento de modelo). Algunos ataques conocidos son:
  - **Extracción de prompts (Prompt Extraction)**: Un ataque que intenta divulgar el prompt del sistema u otra información en el contexto de un modelo de lenguaje grande que normalmente estaría oculto a un usuario.
  - **Inyección de prompt (Prompt injection)**: Un ataque que explota la concatenación de entradas no confiables con un prompt construido por una parte de mayor confianza, como el diseñador de la aplicación.

    - **Inferencia de propiedades:** Un ataque de privacidad de datos que infiere una propiedad global sobre los datos de entrenamiento de un modelo de aprendizaje automático.
    - **Acceso a consultas (query access):** Una capacidad con la que un atacante puede enviar consultas a un modelo entrenado de aprendizaje automático y obtener predicciones o generaciones.

- **Ataques de inferencia de atributos:** Un ataque contra modelos de aprendizaje automático que infiere atributos sensibles de un registro de datos de entrenamiento, dado conocimiento parcial sobre el registro.

- **Ataque de inyección directa de *prompt*:** En el contexto de la IA generativa, es un ataque realizado por un usuario a través del acceso a la interfaz de consultas (*query access*), mediante el cual introduce instrucciones maliciosas diseñadas para manipular el modelo, evadir sus restricciones o alterar su comportamiento previsto.

- **Confidencialidad de datos:** Un concepto bien establecido en ciberseguridad que se refiere a la protección de información sensible frente a accesos y divulgaciones no autorizadas.

- **Control del código fuente (Source code control):** Nivel de acceso indebido o vulnerabilidad mediante el cual un atacante logra alterar, manipular o comprometer la integridad del código base subyacente de un algoritmo o sistema de aprendizaje automático. Esta capacidad le permite al atacante introducir comportamientos maliciosos, puertas traseras (backdoors) o fallos estructurales directamente desde la fase de diseño, desarrollo o implementación del modelo.

- **Envenenamiento de datos (Data poisoning):** Tipo de ataque en el cual un adversario manipula, altera o controla intencionalmente una parte del conjunto de datos utilizado para entrenar o ajustar un modelo de inteligencia artificial. Su objetivo principal es comprometer la integridad, precisión o seguridad del modelo resultante, induciendo comportamientos erróneos, sesgos o creando puertas traseras (backdoors).

- **Expectativa sobre transformación (Expectation of transformation):** Un método para fortalecer ejemplos adversariales y mantenerse adversariales bajo transformaciones de imagen que ocurren en el mundo real, como cambios de ángulo y punto de vista. EOT modela estas perturbaciones dentro del procedimiento de optimización. En lugar de optimizar la verosimilitud logarítmica de un solo ejemplo, EOT utiliza una distribución elegida de funciones de transformación que llevan una entrada controlada por el adversario a la entrada "verdadera" percibida por el clasificador.

- **Extracción de datos de entrenamiento:** La capacidad de un atacante para extraer los datos de entrenamiento de un modelo generativo solicitando al modelo entradas específicas.

- **Geovallado (Geofencing):** Tecnología o mecanismo de control basado en la ubicación que establece perímetros o barreras virtuales alrededor de un área geográfica física real (utilizando señales como GPS, RFID, Wi-Fi o datos celulares). En el ámbito de la ciberseguridad, se implementa para restringir, condicionar o monitorear el acceso a sistemas, infraestructuras críticas o conjuntos de datos, garantizando que los usuarios o dispositivos solo puedan interactuar con los recursos si se encuentran dentro de las zonas geográficas o instalaciones previamente autorizadas.

- **Guardarraíls (Guardrails):** Son mecanismos de control, políticas y salvaguardas técnicas diseñadas para asegurar que los modelos de Inteligencia Artificial, especialmente los de IA Generativa y LLMs, operen dentro de límites seguros, éticos y legales.

- **Inteligencia Artificial Generativa (Generative AI):** Rama de la inteligencia artificial centrada en la creación de contenido nuevo y original (como texto, imágenes, audio o código), a partir de los patrones y estructuras aprendidas de los datos de entrenamiento, distinguiéndose de la IA tradicional enfocada en clasificación o predicción.

- **Inyección indirecta de prompt (Indirect prompt injection):** Un tipo de inyección de prompt ejecutada mediante control de recursos en lugar de mediante entrada proporcionada por el usuario como en una inyección directa por prompt.
- **Jailbreak:** Un ataque directo de prompting destinado a eludir restricciones impuestas a los resultados del modelo, como eludir comportamientos de rechazo para permitir un mal uso.

- **Modelo discriminativo:** Un tipo de método de aprendizaje automático diseñado para aprender a diferenciar o clasificar entre distintas clases o categorías de datos.

- **Modelos de difusión:** Una clase de modelos generativos de variables latentes que consta de tres componentes principales: un proceso directo, un proceso inverso y un procedimiento de muestreo. Su objetivo es aprender un proceso de difusión que genere la distribución de probabilidad de un conjunto de datos dado. Se utiliza ampliamente en visión por ordenador en tareas que incluyen la reducción de ruido, la restauración (inpainting), la superresolución y la generación de imágenes.

- **Modelo de Lenguaje Grande (Large Language Model - LLM):** Sistema de inteligencia artificial basado en arquitecturas de aprendizaje profundo que ha sido entrenado con cantidades masivas de datos textuales. Está diseñado para comprender, generar y procesar lenguaje natural a gran escala, permitiendo interacciones complejas.

- **Patrón Puerta Trasera (Backdoor):** Una transformación o inserción aplicada a una muestra de datos que desencadena un comportamiento especificado por el adversario en un modelo que ha sido objeto de un ataque de envenenamiento por puerta trasera. Por ejemplo, en visión por ordenador, un adversario podría envenenar un modelo de tal manera que la inserción de un cuadrado de píxeles blancos induce una etiqueta objetivo deseada.

- **Prompting:** Es la técnica de formular instrucciones efectivas para obtener los mejores resultados de un sistema de inteligencia artificial.

- **Generación Aumentada por Recuperación (Retrieval-Augmented Generation - RAG):** Patrón de arquitectura de inteligencia artificial que mejora la precisión y confiabilidad de los Modelos de Lenguaje Grandes (LLM) al conectarlos con bases de datos o repositorios de conocimiento externos. Este enfoque permite que el sistema recupere información específica, verificable y actualizada antes de generar una respuesta, mitigando riesgos de seguridad como las “alucinaciones” y garantizando que el contenido se base estrictamente en fuentes de datos autorizadas.

- **Red neuronal de gráficos (Graph Neural Networks - GNN):** Una red neuronal diseñada para procesar datos estructurados en grafos. Los GNN realizan transformaciones optimizables sobre atributos de grafos (por ejemplo, nodos, aristas, contexto global) mientras preservan simetrías de grafos como la invariancia por permutación. Las GNN utilizan una arquitectura de "graph-in, graph-out" que toma un grafo de entrada con información y lo transforma progresivamente en un grafo de salida con la misma conectividad que el grafo de entrada.

- **Restauración de imágenes (Inpainting):** Técnica de visión por computadora en la que un modelo generativo reconstruye, restaura o rellena de forma autónoma áreas faltantes, corruptas o enmascaradas de una imagen, manteniendo la coherencia semántica y visual con el contexto original.

4. Introducción

Este anexo entrega los lineamientos de seguridad y privacidad de la información para sistemas de Inteligencia Artificial (IA), el cual está concebido para que se puedan integrar y fortalecer los lineamientos del modelo de seguridad y privacidad de la información (MSPI), del Ministerio de Tecnologías de la Información y las Comunicaciones – MinTIC.

La seguridad de la información se constituye como un pilar estratégico para la protección de la información clasificada y reservada, de acuerdo como lo establece la ley de transparencia (Ley 1712 de 2014), para las entidades públicas.

Frente a accesos no autorizados, alteraciones indebidas o destrucción accidental o malintencionada. Con la incorporación acelerada de nuevas tecnologías, también se ha incrementado la sofisticación y la frecuencia de las amenazas cibernéticas. Por ello, garantizar la confidencialidad, integridad y disponibilidad de la información no solo preserva la reputación y continuidad operacional de las entidades y del Gobierno Nacional, sino que también consolida la confianza de la ciudadanía en la prestación segura y eficiente de los servicios públicos digitales.

La seguridad de la información y la privacidad adquieren entonces una relevancia aún mayor en el uso de tecnologías de IA. En un escenario donde grandes volúmenes de información

sensible, clasificada o reservada son procesados por sistemas automatizados, la protección de los datos se vuelve indispensable. La IA depende del acceso a datos para entrenar modelos, optimizar su desempeño y tomar decisiones; sin embargo, este beneficio viene acompañado de riesgos que deben gestionarse de manera rigurosa. La ausencia de controles adecuados puede derivar en fugas de información, sesgos, accesos indebidos o usos no autorizados. En este sentido, garantizar la confidencialidad, integridad y disponibilidad de los datos no solo mitiga la materialización de riesgos, sino que también fortalece la confianza institucional, protege los derechos de los ciudadanos y asegura el cumplimiento de estándares éticos, legales y normativos.

5. Justificación

Teniendo en cuenta la tendencia de hacer un uso rápido y despliegues basados en sistemas de inteligencia artificial (IA), generando oportunidades para mejorar la infraestructura tecnológica, la prestación de los servicios y la gestión pública. Los modelos y plataformas de IA procesan grandes volúmenes de información, incluidas categorías sensibles y datos personales, y participan en decisiones que afectan derechos, servicios y la continuidad operativa. Por ello, es imprescindible incorporar salvaguardas específicas de seguridad y privacidad que reduzcan la probabilidad de incidentes y mitiguen sus impactos.

En coherencia con el CONPES 4144 de 2025 y el Modelo de Seguridad y Privacidad de la Información (MSPI), se justifica que todas las entidades públicas adopten prácticas robustas y obligatorias de mitigación de riesgo que incluyan, como mínimo, capacidades de detección de amenazas adaptadas a entornos con IA. Estas capacidades deben contemplar monitoreo continuo, análisis de telemetría y anomalías, correlación inteligente de eventos, pruebas de robustez de modelos, y procesos automatizados para la contención y respuesta a incidentes.

Garantizar estas capacidades no solo reduce el riesgo operativo y jurídico (incumplimiento normativo, pérdida de datos, interrupción de servicios), sino que protege derechos ciudadanos, preserva la confianza pública y facilita la armonización del MSPI con futuros cambios normativos. Por tanto, la adopción de herramientas y procesos especializados para IA debe considerarse una exigencia mínima para la planificación, preparación y gestión de la seguridad y privacidad de la información en las entidades públicas.

Los lineamientos expuestos en este documento son un complemento del MSPI para la planificación y preparación de la seguridad y privacidad de la información, en los sistemas de IA de las entidades públicas y sirve como referente para entidades privadas.

6. Objetivos

El objetivo principal es establecer un conjunto de directrices técnicas, organizativas y normativas que orienten para el diseño, desarrollo, implementación, operación, funcionamiento y mantenimiento de los sistemas de Inteligencia Artificial, asegurando la protección, garantizando la seguridad de la información, la preservación de la privacidad de

la información, la adecuada gestión de riesgos y amenazas digitales y el estricto cumplimiento del marco regulatorio y la normativa vigente, alineado con los principios, lineamientos y capacidades definidas en el Modelo de Seguridad y Privacidad de la Información.

Los objetivos específicos del presente documento son:

- Establecer roles y responsabilidades que aseguren la supervisión, rendición de cuentas y toma de decisiones sobre el diseño, despliegue y operación de sistemas basados en IA.
- Desarrollar y formalizar metodologías de evaluación del impacto de la Inteligencia Artificial en materia de seguridad y privacidad de la información, permitiendo valorar de manera sistemática los riesgos asociados al ciclo de vida de los sistemas de IA y su interacción con datos sensibles o clasificados.
- Garantizar que las infraestructuras, plataformas y servicios que incorporan IA cumplan con estándares y buenas prácticas de seguridad reconocidos a nacional e internacionalmente, asegurando su capacidad para enfrentar amenazas emergentes y escenarios de ataque cada vez más sofisticados.
- Verificar que los sistemas basados en Inteligencia Artificial aseguren la confidencialidad, integridad y disponibilidad, mediante controles técnicos y diseñar capacidades de resiliencia ante fallos y ataques.
- Asegurar el diseño, desarrollo y uso de aplicaciones de Inteligencia Artificial se ajusten estrictamente a la normativa vigente en materia de protección de datos personales, incluyendo los lineamientos establecidos en la Ley 1581 de 2012 y sus disposiciones reglamentarias.
- Incorporar mecanismos robustos de anonimización, ofuscamiento de datos y protección de datos personales en los sistemas y modelos de Inteligencia Artificial, con el propósito de mitigar riesgos asociados al tratamiento de información sensible y garantizar el respeto de los principios de privacidad, proporcionalidad y minimización de datos.

7. Alcance

Estos lineamientos aplican a todos los sujetos obligados señalados en el artículo 2.2.9.1.1.2. del Decreto 1078 de 2015 (DUR-TIC), "Por medio del cual se expide el Decreto Único Reglamentario del sector de Tecnologías de la Información y las Comunicaciones", los cuales deben adoptar medidas técnicas, administrativas y de talento humano para garantizar que la seguridad digital se incorpore al plan de seguridad y privacidad de la información y así mitigar riesgos relacionados con la protección y la privacidad de la información e incidentes de seguridad digital, que puedan afectar a la confidencialidad, integridad o disponibilidad de la información de todos los procesos, trámites, sistemas de información, infraestructura tecnológica e infraestructura crítica de los sujetos obligados, adoptando las medidas de seguridad alineadas al MSPI, para prestar servicios de confianza, generando protección de la información de los ciudadanos, gestionando los riesgos y los incidentes de seguridad digital. Además, debe comprender el relacionamiento de un efectivo tratamiento de la evidencia digital y la referenciación basada en la Resolución 500 del 2021 del MinTIC, la norma ISO/IEC 27001:2022 y la política de Gobierno Digital con el Modelo de Seguridad y Privacidad de la Información – MSPI.

8. Marco normativo

La protección de los datos personales es un principio esencial para garantizar la integridad, dignidad y derechos fundamentales de los ciudadanos. En el contexto del uso creciente de tecnologías de Inteligencia Artificial (IA) por parte de entidades públicas, dicha protección adquiere un carácter estratégico y obligatorio. El fundamento jurídico de estas obligaciones se encuentra consagrado en diversos artículos de la Constitución Política de Colombia, los cuales orientan la forma en que deben diseñarse, implementarse y gestionarse los sistemas de IA en el Estado.

A continuación, se destacan los artículos constitucionales más relevantes:

Artículo 13 – Derecho a la igualdad. El artículo 13 establece que todas las personas nacen libres e iguales ante la ley y tienen derecho a recibir la misma protección y trato de las autoridades sin discriminación alguna. Este principio implica que las soluciones de Inteligencia Artificial deben evitar sesgos algorítmicos, discriminación automatizada o decisiones injustas basadas en variables sensibles. El Estado tiene la obligación de adoptar medidas para asegurar que el uso de tecnologías emergentes, como la IA, no profundice brechas sociales ni afecte a poblaciones en situación de vulnerabilidad.

Artículo 15 – Derecho a la intimidad y protección de datos personales. Este artículo reconoce el derecho fundamental a la intimidad personal y familiar, así como el derecho a conocer, actualizar y rectificar la información contenida en bases de datos públicas o privadas. Establece además que la recolección, tratamiento y circulación de datos debe respetar las libertades y garantías constitucionales.

Este artículo constituye el pilar fundamental para la regulación del tratamiento de datos personales en sistemas de Inteligencia Artificial. Cualquier plataforma o modelo de IA utilizado por entidades públicas debe garantizar, entre otros aspectos:

- la inviolabilidad de la información privada,
- el cumplimiento de principios de necesidad, finalidad y proporcionalidad,
- la trazabilidad y transparencia en el procesamiento de datos,
- la existencia de mecanismos para ejercer derechos de habeas data.

Asimismo, las comunicaciones privadas solo pueden ser intervenidas mediante orden judicial, reforzando la necesidad de controles estrictos cuando modelos de IA procesan información sensible.

Artículo 20 – Libertad de expresión e información. Este artículo garantiza la libertad de expresar y difundir opiniones, así como, el derecho a recibir información veraz e imparcial. Este principio es relevante para sistemas de IA que procesan, clasifican o generan información, ya que deben promover transparencia algorítmica, evitar la difusión de contenidos engañosos y respetar el derecho a la rectificación en condiciones de equidad. Las entidades públicas que usen IA deben procurar que las decisiones automatizadas no afecten el pluralismo informativo ni la libertad de expresión.

Artículo 25 – Derecho al trabajo. El artículo 25 establece la protección del trabajo en condiciones dignas y justas. En el contexto de la IA, este derecho se relaciona con:

- la adopción de tecnologías que complementen, y no vulneren, los derechos laborales;
- la obligación estatal de garantizar condiciones seguras, responsables y éticas en la interacción entre personas y sistemas automatizados;
- la necesidad de capacitación y adaptación laboral ante el cambio tecnológico.
- El diseño de políticas públicas en IA debe promover la inclusión, la capacitación y la transición laboral justa.

Además del marco constitucional, existen leyes y regulaciones que establecen principios, obligaciones y responsabilidades para el tratamiento de datos personales y el uso de tecnologías basadas en Inteligencia Artificial:

- **Ley 1266 de 2008 – Habeas Data.** Regula el manejo de información financiera, crediticia y comercial, y consagra el derecho a actualizar y rectificar datos. Es aplicable cuando sistemas de IA procesan información de esta naturaleza.
- **Ley 1273 de 2009 - Protección de la información y de los datos.** Por medio de la cual se modifica el Código Penal, se crea un nuevo bien jurídico tutelado y se preservan integralmente los sistemas que utilicen las tecnologías de la información y las comunicaciones.
- **Ley 1581 de 2012 – Protección de Datos Personales.** Establece el régimen general de protección de datos personales en Colombia, definiendo principios, derechos de los titulares y obligaciones de responsables y encargados. Es la base normativa para todo sistema de IA que procese datos personales.
- **Decreto 1377 de 2013.** Reglamenta aspectos operativos de la Ley 1581 de 2012, incluyendo autorizaciones, políticas de tratamiento, avisos de privacidad y transferencias internacionales de datos.
- **Ley 1712 de 2014 – Transparencia y Acceso a la Información Pública.** Establece obligaciones para garantizar el acceso a la información pública, así como, límites y excepciones en caso de información clasificada o reservada. La IA implementada por entidades públicas debe cumplir con los principios de transparencia activa y pasiva, sin comprometer derechos fundamentales ni información protegida.
- **CONPES 3995 de 2020 – Política Nacional de Confianza y Seguridad Digital.**
- Decreto 767 de 2022 - Establece los lineamientos generales de la Política de Gobierno Digital, entendida como el uso y aprovechamiento de las Tecnologías de la Información y las Comunicaciones, con el objetivo de impactar positivamente la calidad de vida de los ciudadanos y, en general, los habitantes del territorio nacional y la competitividad del país, promoviendo la generación de valor público a través de la transformación digital del Estado, de manera proactiva, confiable, articulada y colaborativa entre los Grupos de Interés y permitir el ejercicio de los derechos de los usuarios del ciberespacio.
- **CONPES 4144 de 2025 - Política Nacional de Inteligencia Artificial -** Su objetivo es impulsar el desarrollo, adopción y uso ético, seguro y sostenible de la IA, promoviendo la investigación, el talento digital y la infraestructura necesaria para la transformación socioeconómica.

Colombia toma como referente los principios y buenas prácticas basados en marcos internacionales que orientan el uso responsable, ético y seguro de la Inteligencia Artificial.

9. Estado del arte del desarrollo de modelos de inteligencia artificial a nivel internacional y en Colombia

A nivel global, las recomendaciones, normativas y regulaciones sobre inteligencia artificial (IA) están en un estado constante de cambio y adaptándose continuamente a medida que la tecnología avanza y se integran en más aspectos de la vida cotidiana. Dentro de los enfoques y desarrollos más significativos, se encuentran los siguientes:

- **Legislación en EE. UU.:** Estados Unidos está en una fase activa de desarrollo de una legislación centrada exclusivamente en la IA. Además de las leyes federales que ya existen, se contempla la creación de una entidad reguladora dedicada a proporcionar supervisión y directrices sobre el uso de la Inteligencia Artificial.

- **Principios de la OCDE, UNESCO y UNICEF:** La Organización para la Cooperación y el Desarrollo Económicos (OCDE) ha definido varios principios que guían la implementación y uso ético de la IA. Por su parte, UNESCO y UNICEF han ofrecido recomendaciones para asegurar que la IA sea desarrollada y utilizada de manera ética, complementando otros marcos internacionales como el AI RMF (National Institute of Standards and Technology, 2023), ISO 42001 ((International Organization for Standardization) / IEC (International Electrotechnical Commission), 2023) (UNICEF, 2021) y IEEE 7000, que refuerzan buenas prácticas en la materia.

- **Acta de Inteligencia Artificial (IA) de la Unión Europea (UE):** La Unión Europea ha introducido el Acta de Inteligencia Artificial (IA) de la Unión Europea (UE), que adopta un esquema de gobernanza que prioriza la gestión de riesgos. Este marco tiene potencial para influir en regulaciones en otras naciones, estableciendo un estándar que otros podrían seguir.

- **Fragmentación del mercado global:** Con la diversidad en las regulaciones sobre IA en distintas partes del mundo, hay un riesgo creciente de fragmentación del mercado[1]. Esto crea dificultades para las empresas que operan internacionalmente, ya que deben navegar un paisaje regulatorio cada vez más complejo.

- **Desarrollo de un marco global:** Se están explorando iniciativas para crear un marco de gobernanza a nivel mundial que ayude a mitigar los riesgos asociados y fomente un desarrollo responsable de la Inteligencia Artificial. Este marco busca facilitar la

[1] Gobernanza de la Inteligencia Artificial en beneficio de la Humanidad: Informe final, 2024, Naciones Unidas. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_es.pdf

cooperación internacional, asegurando que los beneficios de la IA se compartan equitativamente.

- **Comparación de enfoques:** Las diferentes regiones del mundo están adoptando estrategias diversas para regular la IA. Esto subraya la importancia de realizar un análisis comparativo constante que permita identificar las mejores prácticas que se implementen de acuerdo con las tendencias y cambios tecnológicos y de esta manera entender las implicaciones que tienen estas distintas regulaciones.

En definitiva, la regulación de la Inteligencia Artificial es un desafiante campo de estudio que continúa evolucionando, con un enfoque creciente en la ética, la responsabilidad y la mitigación de riesgos en un contexto global[2].

Tema	Descripción
Enfoque basado en riesgos	Regulación adaptada al nivel de riesgo de cada aplicación de IA, promoviendo prácticas éticas y sostenibles.
Legislación en EE. UU.	Desarrollo activo de legislación centrada en IA, con la creación de una entidad reguladora dedicada.
Principios de la OCDE y UNESCO	Definición de principios por la OCDE y recomendaciones de UNESCO para un uso ético de la IA.
Acta de inteligencia Artificial (IA) de la Unión Europea (UE)	Marco desarrollado por la Unión Europea (UE) que prioriza la gestión de riesgos en IA, influyendo potencialmente en otras naciones.
Fragmentación del mercado global[3]	Diversidad en regulaciones que puede dificultar operaciones de empresas internacionales.
Desarrollo de un marco global	Iniciativas para crear un marco de gobernanza mundial que mitigue riesgos y fomente un desarrollo responsable de la IA.
Comparación de enfoques	Importancia de análisis comparativos para identificar mejores prácticas en la regulación de la IA.

Tabla 1. Regulación de riesgos de Inteligencia Artificial (IA)

País	Avances de marcos normativos de IA basado en riesgos
Estados Unidos	En desarrollo activo de legislación centrada en IA; se prevé la creación de una entidad reguladora. Enfocada en establecer estándares para la seguridad y la confiabilidad de los sistemas de IA. Se contempla la creación de entidades reguladoras que supervisen la implementación y el uso de la IA.
Unión Europea	Introducción del Acta de IA de la UE, que clasifica los sistemas según su nivel de riesgo y establece requisitos estrictos, con un enfoque en la gestión de riesgos en IA de alto riesgo. Los sistemas de alto riesgo deben cumplir con estrictos requisitos de transparencia, supervisión humana y gestión de riesgos a lo largo de todo su ciclo de vida. Esto incluye evaluación de impacto y documentación técnica detallada.
Argentina	Proyecto de ley para fortalecer la industria del software, sin clasificación específica de riesgo, pero con beneficios y regulaciones existentes que abogan por un crecimiento responsable.

[2] Gutiérrez, Juan David (2024). «Regulación sobre IA». Blog Foro Administración, Gestión y Política Pública. Disponible en: https://forogpp.com/inteligencia-artificial-y-sector-publico/regulacion-sobre-ia/

[3] Informe (Navegando por la fragmentación del sistema financiero mundial) del Foro Económico Mundial, elaborado en colaboración con Oliver Wyman. https://es.weforum.org/stories/2025/12/como-entender-la-fragmentacion-financiera-global-para-mitigar-sus-riesgos/

Brasil	Proyecto de ley que establece categorías de riesgo para sistemas de IA, con evaluaciones preliminares y responsabilidades por daños causados; prohíbe sistemas de alto riesgo que induzcan comportamientos perjudiciales.
Chile	Clasificación de sistemas de IA en cuatro categorías de riesgo y principios éticos como supervisión humana y seguridad técnica; enfoque en la intervención humana y protección de derechos.
Ecuador	Establecimiento de un marco regulatorio con categorías de riesgo y un registro de sistemas de IA, que requiere evaluaciones de impacto para sistemas de alta complejidad; se enfoca en la transparencia y la protección de datos.
Perú	Ley que promueve el uso de IA en sectores clave, promoviendo la colaboración y el desarrollo responsable; se observa un enfoque en la ética y la implementación de medidas de supervisión.
Uruguay	Marco regulatorio que exige etiquetado digital y supervisión de IA, además de un sistema de seguimiento para garantizar el cumplimiento de la ley; se espera garantizar transparencia y ética en el uso de IA.
Costa Rica	Proyecto de ley que regula el uso de IA, contempla evaluaciones de impacto para sistemas de alto riesgo, y asegura la protección de derechos fundamentales y la dignidad humana.
México	Propuesta de ley que regula la ética en IA y la robótica, promoviendo la protección de derechos humanos y la equidad, sin una clasificación detallada de riesgos, pero establece supervisión y un consejo regulador.
Panamá	Proyecto de ley que regula el desarrollo de IA, enfocado en derechos, seguridad y privacidad; establece prohibiciones sobre el uso malintencionado de IA y proporciona derechos a las personas afectadas por decisiones automatizadas.

Tabla 2. Comparación de avances de regulación de riesgos de Inteligencia Artificial (IA)

La identificación y gestión de riesgos es esencial para el desarrollo sostenible y ético de la Inteligencia Artificial. Este análisis refleja la necesidad de establecer regulaciones y marcos que protejan tanto a los individuos como a la sociedad en su conjunto.

Resumen de riesgos de seguridad identificados en el análisis internacional de IA[4]

- Los sistemas de IA pueden perpetuar o amplificar sesgos existentes, resultando en decisiones discriminatorias en áreas como empleo, justicia y acceso a servicios.
- Las aplicaciones de IA son vulnerables a ataques cibernéticos, que pueden comprometer la integridad y la confidencialidad de los datos, así como, llevar a malinterpretaciones o manipulación.
- Muchos sistemas de IA son vistos como "cajas negras", donde las decisiones automatizadas son difíciles de entender, lo que genera desconfianza entre los usuarios y reguladores.
- La falta de claridad sobre quién es responsable por decisiones tomadas por sistemas de IA puede dificultar la rendición de cuentas y resultar en conflictos legales complejos.

[4] Gutiérrez, Juan David (2024). «Regulación sobre IA». Blog Foro Administración, Gestión y Política Pública. Disponible en: https://forogpp.com/inteligencia-artificial-y-sector-publico/regulacion-sobre-ia/

- La IA puede ser utilizada con fines dañinos, como la creación de contenido falso (Deepfake) o la manipulación de la opinión pública, que puede comprometer la seguridad social.
- La recopilación y el análisis de datos personales mediante IA pueden violar la privacidad de los individuos, especialmente, si se manejan datos sensibles sin consentimiento adecuado.
- Las tecnologías avanzadas pueden exacerbar la brecha digital, donde ciertos grupos no tienen acceso equitativo a beneficios derivados de la IA, resultando en desigualdades económicas y sociales.
- La automatización de trabajos mediante IA puede llevar a la pérdida de empleos en sectores vulnerables, generando preocupaciones económicas y sociales.

En conclusión, la Inteligencia Artificial representa un conjunto de riesgos de seguridad estratégicos, identificados de forma consistente en el análisis internacional. Estos riesgos se derivan principalmente de su capacidad para escalar usos maliciosos, incluyendo ciberataques, desinformación y fraude, así como, la de su integración en sistemas críticos de defensa e infraestructuras esenciales, donde fallos técnicos o decisiones automatizadas pueden generar impactos de alto alcance.

Asimismo, se destacan riesgos sistémicos vinculados a la pérdida de control efectivo y a problemas de alineación de sistemas complejos, la concentración de capacidades tecnológicas en un número limitado de actores, y la insuficiencia de marcos de gobernanza, regulación y rendición de cuentas a nivel internacional. El consenso experto concluye que la amenaza principal no reside en la tecnología en sí, sino en su despliegue sin salvaguardas adecuadas, supervisión humana significativa y mecanismos de cooperación y control internacionales robustos.

10. Principios rectores para la implementación de IA segura

El gobierno nacional mediante la guía para la implementación, desarrollo y uso de sistemas de Inteligencia Artificial (IA) en entidades públicas en Colombia, ha establecido los principios y valores que sirvan como marco orientador que garantice que los beneficios de la IA lleguen, de manera equitativa, inclusiva y ética en todo el territorio nacional.

Por lo expuesto anteriormente, a continuación, se presentan los principios y valores que se consideran fundamentales para contar con sistemas de Inteligencia Artificial (IA), éticos y con responsabilidad adaptados al contexto colombiano.

Ilustración 1. Principios rectores para implementación de IA segura.

Elaboración propia basada en la Guía Ética para la Implementación, Desarrollo y Uso de Sistemas de Inteligencia Artificial en Entidades Públicas de Colombia (MinTIC, 2025)

10.1. Centralidad humana y bien público

Los sistemas de IA en el sector público deben ser diseñados, desarrollados y desplegados con el objetivo primordial de promover el bienestar de las personas, proteger los derechos humanos y fortalecer el interés general. Su propósito fundamental es aumentar las capacidades humanas y no dejar la toma de decisiones críticas exclusivamente en manos de la IA. Este enfoque garantiza que el despliegue de las tecnologías en el estado colombiano se realice bajo principios de soberanía digital, equidad, inclusión y rendición de cuentas, asegurando que la innovación tecnológica sea una herramienta al servicio de las personas, mejorando su calidad de vida, fortaleciendo los valores democráticos.

La aplicación de este principio conlleva la priorización de las necesidades ciudadanas, la garantía de un acceso equitativo a los servicios y el fomento de sociedades inclusivas. La finalidad última es que la IA impulse un gobierno más eficiente, justo y sensible a las necesidades de su población.

10.2. Transparencia, explicabilidad y rendición de cuentas

Los sistemas de IA en el sector público deben operar bajo principios de transparencia, garantizando que su propósito, las fuentes de datos utilizadas, los mecanismos de procesamiento y toma de decisiones sean comprensibles para todos los actores involucrados. Sus resultados deben ser explicables, es decir, debe permitir a los usuarios humanos y a las personas afectadas comprender el razonamiento detrás de las decisiones

impulsadas por la IA, incorporando, por ejemplo, registros técnicos obligatorios como control de versiones del modelo, bitácoras de entrenamiento, metadatos, fuentes de datos y logs de auditoría. Además, deben establecerse líneas claras de rendición de cuentas para cada paso del proceso desde el diseño, despliegue, desarrollo, implementación hasta los resultados de la IA.

Este principio es crucial para fomentar la confianza pública, permitir la supervisión democrática y facilitar mecanismos de reparación, de apelación y revisión para la identificación de oportunidades de mejora. Adicionalmente, se debe asegurar la auditabilidad (Auditability) y la responsabilidad continua (Answerability) a lo largo del ciclo de vida de la IA, con el fin de fortalecer la justicia ciudadana y la supervisión democrática de los sistemas implementados para su permanente actualización. La transparencia y la explicabilidad son vitales para construir la confianza pública, mientras que los mecanismos de rendición de cuentas deben garantizar la supervisión humana y la responsabilidad por los resultados impulsados por la IA.

10.3. Equidad, igualdad y No discriminación

Los sistemas de IA en el sector público deben ser diseñados e implementados para garantizar un trato justo y equitativo para todos los ciudadanos, trabajando activamente para prevenir y mitigar el sesgo algorítmico que podría conducir a la discriminación o exacerbar las desigualdades sociales existentes.

La aplicación de este principio implica asegurar la igualdad de acceso a los servicios públicos, prevenir impactos desproporcionados en grupos vulnerables y defender los principios de justicia. Esto requiere medidas proactivas para identificar y mitigar el sesgo en los datos, y los algoritmos en el despliegue de sistemas de IA.

10.4. Privacidad, gobernanza de datos y seguridad

Los sistemas de IA en el sector público deben garantizar el respeto y la protección efectiva de la privacidad de los ciudadanos en consonancia con los derechos fundamentales y principios de soberanía digital. Esto exige marcos de gobernanza de datos robustos que regulen de manera rigurosa la recolección, almacenamiento, procesamiento y uso de datos conforme a estrictos estándares éticos y legales, incluyendo la minimización de datos, la anonimización y su uso seguro y responsable. La implementación de medidas de ciberseguridad debe ser prioritaria con el fin de prevenir accesos no autorizados, vulneraciones de datos y riesgos asociados a la integridad de los sistemas. Así mismo, se requiere establecer mecanismos claros, accesibles y verificables para la obtención del consentimiento informado por parte de los titulares de los datos, garantizando que las personas comprendan cómo, por qué y para qué se utiliza su información en entornos de Inteligencia Artificial Pública.

La observancia de este principio es vital para mantener la confianza ciudadana y prevenir el uso indebido de información sensible con especial atención en los casos de niños, niñas y adolescentes.

10.5. Robustez, fiabilidad y seguridad

Los sistemas de IA implementados en el sector público deben cumplir con altos estándares de robustez técnica, fiabilidad operativa y seguridad en su funcionamiento. Esto significa que deben funcionar de manera consistente y precisa en diferentes situaciones y entornos diversos, resistir errores y posibles ataques y minimizar el riesgo de producir resultados inesperados o dañinos.

La aplicación rigurosa de este principio fortalece la integridad de los servicios ofrecidos al público, previene interrupciones que pueden afectar derechos fundamentales y contribuye a la sostenibilidad y eficiencia de las operaciones institucionales. Para ello se requiere un enfoque integral que incluya pruebas exhaustivas, validación técnica continua y mecanismos de monitoreo permanente, orientados a anticipar fallos, mitigar vulnerabilidades y garantizar la seguridad del sistema en todo su ciclo de vida.

10.6. Sostenibilidad ambiental y bienestar

El diseño, desarrollo y despliegue de la IA en el sector público deben incorporar una evaluación integral de su huella ambiental (incluyendo consumo energético, hídrico y de otros recursos) así como sus impactos sociales a largo plazo como la transformación del mercado laboral, el riesgo de exclusión digital.

Es imperativo que estos procesos se orienten a minimizar las consecuencias negativas y potenciar los beneficios sociales, económicos y ambientales en coherencia con los Objetivos de Desarrollo Sostenible (ODS).

Este principio implica la promoción activa de prácticas de IA ecológicas y ambientalmente sostenibles, que reduzcan su impacto ecológico a lo largo de todo el ciclo de vida tecnológico. Así mismo, exige preparar a la fuerza laboral para los cambios estructurales que la IA genera en los entornos productivos, mediante estrategias de capacitación, reconversión y fortalecimiento de habilidades digitales para lograr un acceso equitativo a la tecnología.

11. Recomendaciones para la identificación y gestión de riesgos asociados a la IA.

Las entidades, deben incluir dentro del proceso de gestión de riesgos de seguridad de la información, la identificación y gestión de los riesgos asociados con Inteligencia Artificial (IA), como parte del Modelo de Seguridad y Privacidad de la Información (MSPI), se debe realizar la identificación de activos de información, riesgos, amenazas y vulnerabilidades, para llevar a cabo un análisis de los riesgos asociados con IA, posteriormente se deben implementar los controles diseñados para mitigar estos riesgos y el proceso de reporte de estos, con el fin de ofrecer una vía para minimizar los posibles impactos que se puedan presentar si se materializa un riesgo relacionado con IA, los riesgos deben estar documentados y gestionarse eficazmente.

Lineamiento: Incluir dentro del proceso de gestión de riesgos de seguridad de la información la valoración de riesgos de IA, identificación y clasificación de los activos de información que tengan información clasificada y reservada y que cuenten con sistemas de información de IA, que permita:

- Clasificar los activos de información de acuerdo con los tres principios de seguridad de la información: Integridad, confidencialidad y disponibilidad para garantizar que la información recibe los niveles de protección adecuados, como lo establece el MSPI.
- Actualizar el inventario y la clasificación de los activos por los propietarios y custodios de los activos de forma periódica.

Entradas recomendadas	***Salidas***
• Inventario de activos de información. • Matriz de riesgos.	• Inventario actualizado incluyendo sistemas de IA. • Matriz de Riesgos actualizada con los riesgos relacionados con IA y riesgos emergentes. • Plan de tratamiento de riesgos de sistemas basados en IA.

11.1. Comprendiendo y abordando riesgos, impactos y perjuicios

En el contexto de la Inteligencia Artificial, el riesgo se entiende como la combinación entre la probabilidad de que ocurra un evento y la magnitud de las consecuencias asociadas, conforme a lo establecido en la Guía para la Gestión Integral del Riesgo en Entidades Públicas[5]. Los impactos derivados del uso de sistemas de IA pueden ser tanto positivos como negativos, generando oportunidades o amenazas para las personas, las entidades públicas, la sociedad y el entorno.

La evaluación del riesgo considera principalmente el nivel de daño potencial y la probabilidad de que este se materialice. Estos impactos pueden afectar a individuos, comunidades, instituciones y al medio ambiente, por lo que su adecuada gestión resulta fundamental para un uso responsable de la IA.

La gestión de riesgos, entendida como el conjunto de actividades orientadas a dirigir y controlar una organización frente al riesgo, permite no solo mitigar efectos negativos, sino también identificar oportunidades para maximizar los beneficios de la Inteligencia Artificial. Una gestión adecuada contribuye al desarrollo de sistemas más confiables, seguros y

[5] Guía para la Gestión Integral del Riesgo en Entidades Públicas, Versión 7, del año 2025, del Departamento Administrativo de la Función Pública: https://www.funcionpublica.gov.co/detalle-publicacion?entryId=963161

aceptados, al considerar las limitaciones, incertidumbres y posibles efectos no previstos de los modelos.

Dado que la IA evoluciona de manera constante, la gestión de riesgos debe ser flexible y adaptativa, capaz de responder a nuevos escenarios y amenazas emergentes. Asimismo, es importante evitar una confianza excesiva en estos sistemas, reconociendo que pueden presentar errores o sesgos y que requieren supervisión humana continua. Los sistemas de IA fiables y su uso responsable pueden mitigar riesgos negativos y contribuir a beneficios para las personas, las entidades y los ecosistemas.

Ilustración 2. Ejemplos de posibles daños relacionados con sistemas de IA.

11.2. Desafíos para la gestión de riesgos en IA

A continuación, se describen varios desafíos que deben tenerse en cuenta al gestionar riesgos en busca de la fiabilidad de la IA.

11.2.1. Medición de riesgos

La medición de riesgos asociados a la IA puede resultar compleja cuando los posibles fallos o impactos no están claramente definidos o no se comprenden adecuadamente. La dificultad para medir un riesgo ya sea de forma cuantitativa o cualitativa, no implica necesariamente que un sistema de IA sea de alto o bajo riesgo, sino que requiere enfoques de análisis más rigurosos y adaptativos.

Uno de los principales desafíos se relaciona con el uso de software, hardware o datos de terceros. Si bien estos componentes pueden acelerar el desarrollo y la adopción de soluciones de IA, también pueden dificultar la medición y gestión del riesgo. Las métricas y metodologías empleadas por los desarrolladores pueden diferir de las utilizadas por quienes despliegan u operan los sistemas, y no siempre existe transparencia sobre los criterios de evaluación aplicados. En todos los casos, los distintos actores involucrados comparten la responsabilidad de gestionar los riesgos de los sistemas de IA que desarrollan, integran o utilizan.

Otro reto relevante es el seguimiento de riesgos emergentes, dado que los sistemas de IA evolucionan con el tiempo y pueden generar impactos no previstos inicialmente. En este contexto, los enfoques de evaluación de impacto resultan útiles para identificar posibles perjuicios en escenarios específicos y fortalecer la toma de decisiones.

La disponibilidad de métricas confiables constituye también un desafío, debido a la falta de consenso sobre métodos estandarizados y verificables para medir riesgos y confiabilidad en distintos casos de uso. Adicionalmente, algunas métricas pueden simplificar en exceso la realidad, no considerar diferencias contextuales o afectar de manera desigual a distintos grupos poblacionales.

La medición del riesgo varía según la etapa del ciclo de vida del sistema de IA, ya que ciertos riesgos pueden permanecer latentes y manifestarse posteriormente durante su operación. Asimismo, los distintos actores pueden tener percepciones diferentes del riesgo, dependiendo de su rol en el desarrollo, despliegue o uso del sistema, lo que refuerza la necesidad de una gestión compartida y coordinada.

Finalmente, es importante considerar que los riesgos observados en entornos controlados o de laboratorio pueden diferir de aquellos que surgen en escenarios reales de operación. A ello se suma la limitada interpretabilidad o transparencia de algunos sistemas de IA, lo que dificulta la identificación y medición de riesgos. En los casos en que la IA complementa o reemplaza actividades humanas, la ausencia de una línea base comparable con el desempeño humano representa un desafío adicional para evaluar su impacto.

11.2.2. Tolerancia al riesgo

Aunque se pueden priorizar riesgos, no se prescribe tolerancia al riesgo. La tolerancia al riesgo se refiere a la disposición de la entidad o del actor de IA, para asumir el riesgo y alcanzar sus objetivos. La tolerancia al riesgo puede verse influida por requisitos legales o regulatorios. La tolerancia al riesgo y el nivel de riesgo aceptable para las entidades o la sociedad son altamente contextuales y específicos de la aplicación y del caso de uso. La tolerancia al riesgo puede verse influenciada por políticas y normas establecidas por propietarios de sistemas de IA, entidades, industrias, comunidades o responsables políticos. Es probable que las tolerancias al riesgo cambien con el tiempo a medida que evolucionen los sistemas, políticas y normas de IA. Diferentes entidades pueden tener tolerancias al riesgo variables debido a sus prioridades organizativas particulares y consideraciones de recursos.

El conocimiento y los métodos emergentes para informar mejor los compromisos entre daño y costo - beneficio seguirán siendo desarrollados y debatidos por empresas, gobiernos, el mundo académico y la sociedad civil. En la medida en que los desafíos para especificar tolerancias al riesgo de IA permanezcan sin resolver, puede haber contextos en los que un marco de gestión de riesgos aún no sea fácilmente aplicable para mitigar riesgos negativos de IA.

La gestión del riesgo puede ser flexible y complementar las prácticas de riesgo existentes que deben estar alineadas con las leyes, normativas y normas aplicables. Las entidades

deben seguir las normativas y directrices existentes sobre los criterios de riesgo, la tolerancia y la respuesta establecidos por requisitos organizacionales, de dominio, disciplina, sectorial o profesionales. Algunos sectores o industrias pueden tener definiciones establecidas de daño o requisitos de documentación, informes y divulgación. Dentro de los sectores, la gestión de riesgos puede depender de las directrices existentes para aplicaciones específicas y entornos de casos de uso. Cuando no existan directrices establecidas, las entidades deberían definir una tolerancia razonable al riesgo. Una vez definida la tolerancia, la gestión del riesgo de IA puede utilizarse para gestionar riesgos y documentar procesos de gestión de riesgos.

11.2.3. Priorización de riesgos

En la gestión de riesgos asociados a la IA, no resulta viable ni eficiente intentar eliminar todos los riesgos negativos. Las entidades deben reconocer que no todos los riesgos tienen la misma relevancia y que una priorización adecuada permite asignar los recursos de manera efectiva y proporcional al impacto potencial.

La priorización de riesgos debe basarse en el nivel de impacto y la probabilidad de ocurrencia, teniendo en cuenta los criterios establecidos en el capítulo V, de la Guía para la Gestión Integral del Riesgo en Entidades Públicas, considerando el contexto de uso del sistema de IA, su grado de adaptación y los posibles efectos sobre las personas, la seguridad o los derechos fundamentales. Los sistemas que presenten riesgos elevados o inaceptables deben ser gestionados con mayor urgencia y, de ser necesario, su desarrollo o despliegue debe limitarse hasta que dichos riesgos sean mitigados, los controles definidos para mitigar el riesgo deben estar alineados con los controles del Modelo de Seguridad y Privacidad de la Información (MSPI).

Los sistemas de IA que interactúan directamente con personas y procesan datos sensibles o influyen en decisiones relevantes requieren un nivel de priorización mayor. En contraste, aquellos orientados exclusivamente a procesos técnicos y entrenados con datos no sensibles pueden requerir una priorización inicial menor, sin perjuicio de evaluaciones periódicas.

Finalmente, es fundamental identificar y documentar los riesgos residuales, informando de manera transparente a los usuarios y partes interesadas sobre los posibles impactos negativos, como parte de un enfoque de rendición de cuentas y gestión responsable de la IA.

11.2.4. Integración organizativa y gestión del riesgo

Los riesgos asociados a la IA no deben abordarse de manera aislada ni fragmentada. A lo largo del ciclo de vida de un sistema de IA intervienen diversos actores, desarrolladores, integradores, operadores y usuarios, cada uno con responsabilidades, niveles de información y capacidades distintas. En muchos casos, las entidades que desarrollan un sistema de IA no cuentan con plena visibilidad sobre los contextos específicos en los que este será utilizado, lo que refuerza la necesidad de una gestión de riesgos coordinada y transversal.

En este sentido, la gestión de riesgos de la Inteligencia Artificial debe integrarse dentro de los marcos generales de gestión de riesgos organizacionales, y no tratarse como un ejercicio independiente. Abordar los riesgos de la IA de manera articulada con otros riesgos críticos, como la ciberseguridad, la protección de datos personales, la continuidad del negocio y la sostenibilidad ambiental permite una visión más integral, reduce duplicidades y genera eficiencias operativas y de control.

Muchos de los riesgos asociados a los sistemas de IA son compartidos con otros procesos de desarrollo y despliegue de software. Entre ellos se incluyen los riesgos relacionados con la privacidad derivados del uso de datos para entrenamiento, las implicaciones energéticas y ambientales asociadas a infraestructuras intensivas en cómputo, las amenazas a la confidencialidad, integridad y disponibilidad de los sistemas y de los datos, así como, los riesgos inherentes a la seguridad del hardware y software subyacente. Por ello, la gestión de riesgos de IA puede y debe complementarse con marcos y directrices existentes en materia de seguridad de la información, gestión tecnológica y riesgo empresarial.

Para que la gestión de riesgos de IA sea efectiva, las entidades deben establecer y mantener mecanismos claros de rendición de cuentas, definir roles y responsabilidades, y promover una cultura organizacional orientada a la gestión proactiva del riesgo. La aplicación aislada de metodologías o marcos de riesgo no resulta suficiente si no existe un compromiso explícito de la alta dirección y estructuras de incentivos que respalden estas prácticas. En muchos casos, lograr una gestión eficaz del riesgo requerirá ajustes organizacionales y cambios culturales.

Finalmente, es importante reconocer que las entidades pequeñas y medianas pueden enfrentar desafíos particulares en la implementación de la gestión de riesgos de IA, debido a limitaciones de recursos, capacidades técnicas o madurez institucional. Por ello, los enfoques de gestión deben ser proporcionales y adaptables, garantizando que todas las entidades, independientemente de su tamaño, puedan gestionar de manera responsable los riesgos asociados al uso de la inteligencia artificial.

11.3. Audiencia

La identificación y gestión de los riesgos e impactos asociados a la Inteligencia Artificial, tanto positivos como negativos, requiere la participación de un conjunto amplio y diverso de actores a lo largo de todo el ciclo de vida de los sistemas de IA. Una gestión de riesgos efectiva se fortalece cuando incorpora múltiples perspectivas, conocimientos técnicos, disciplinas y experiencias, promoviendo equipos diversos desde el punto de vista técnico, social y demográfico.

La gestión de riesgos de IA está concebida para ser aplicada por los distintos actores involucrados en el diseño, desarrollo, despliegue, operación, evaluación y uso de sistemas de Inteligencia Artificial, incluyendo el contexto de aplicación, los datos y entradas, el modelo de IA, la tarea y salida, y el entorno operativo. Los actores que participan en estas dimensiones y que asumen responsabilidades en la toma de decisiones, el desarrollo técnico,

la evaluación y el uso de los sistemas conforman la audiencia principal de la gestión de riesgos de IA.

De manera complementaria, la dimensión de personas y del planeta, ubicada en el centro del marco, representa la protección de los derechos humanos, el bienestar social y el impacto ambiental. Los actores asociados a esta dimensión constituyen una audiencia adicional que aporta insumos relevantes a la gestión de riesgos de IA. Entre ellos se incluyen organismos de estandarización, asociaciones sectoriales, investigadores, organizaciones de la sociedad civil, grupos de defensa de derechos y organizaciones ambientales, entre otros. Su participación resulta clave para anticipar impactos, evaluar riesgos emergentes y fortalecer una gobernanza de la inteligencia artificial alineada con valores éticos, sociales y ambientales.

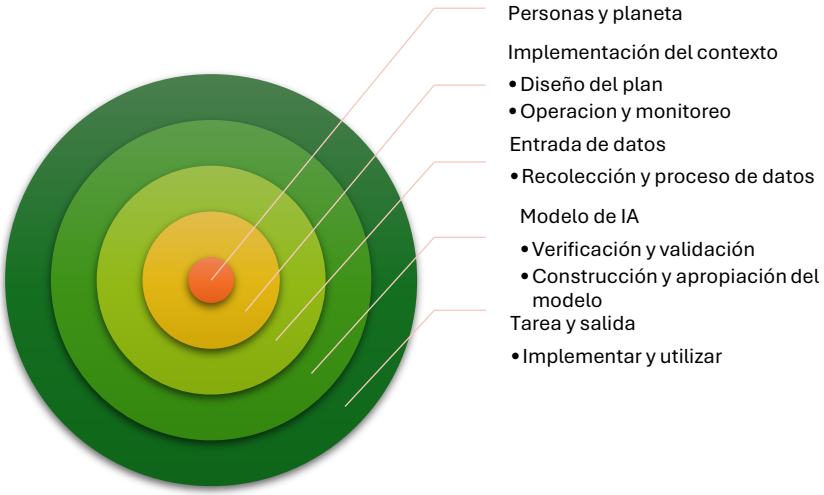

Ilustración 3. Ciclo de vida y dimensiones clave de un sistema de IA. Modificado del Marco OCDE (2022) para la Clasificación de Sistemas de IA — Documentos de la Economía Digital de la OCDE.

Los dos círculos interiores muestran las dimensiones clave de los sistemas de IA y el círculo exterior muestra las etapas del ciclo de vida de la IA. Idealmente, los esfuerzos de gestión de riesgos comienzan con la función de Planificación y Diseño en el contexto de la aplicación y se llevan a cabo a lo largo del ciclo de vida del sistema de IA.

Estos actores desempeñan un papel fundamental en la gestión de riesgos de la Inteligencia Artificial, al contribuir de manera activa en diferentes dimensiones del proceso. En particular, pueden:

- Aportar contexto y facilitar la comprensión de los impactos potenciales y reales derivados del uso de sistemas de IA.
- Servir como fuente de normas, estándares y orientaciones formales o casi formales que apoyen la gestión responsable de los riesgos asociados a la IA.
- Definir y delimitar los márgenes de operación de los sistemas de IA, considerando aspectos técnicos, sociales, legales y éticos.

- Promover espacios de diálogo informados sobre los compromisos necesarios para equilibrar valores y prioridades sociales, tales como la protección de las libertades y derechos civiles, la equidad, la sostenibilidad ambiental y el desarrollo económico.

La gestión eficaz del riesgo en Inteligencia Artificial se fundamenta en un enfoque de responsabilidad colectiva entre los distintos actores involucrados. Las funciones de gestión de riesgos requieren la participación de perspectivas, disciplinas, profesiones y experiencias diversas, lo que fortalece la identificación y evaluación de riesgos. La conformación de equipos diversos facilita un intercambio más abierto y crítico de ideas, haciendo explícitos supuestos que de otro modo podrían permanecer implícitos. Esta visión colectiva amplía la capacidad de anticipar impactos, identificar riesgos existentes y detectar riesgos emergentes, contribuyendo al desarrollo de sistemas de IA confiables y responsables.

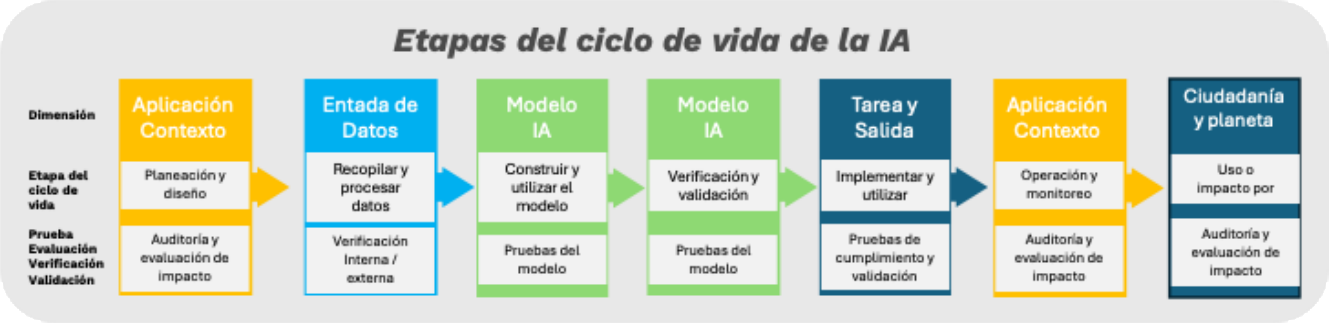

Ilustración 4. Etapas del ciclo de vida de la IA.

11.4. Riesgos y fiabilidad de la IA

Para que los sistemas de Inteligencia Artificial sean confiables, deben responder a múltiples criterios relevantes para las partes interesadas. Los enfoques orientados a fortalecer la confiabilidad de la IA contribuyen a reducir los riesgos y los impactos negativos asociados a su uso.

Este marco identifica las principales características de una IA confiable, entre las que se incluyen sistemas válidos y fiables, seguros y resilientes, responsables y transparentes, comprensibles e interpretables, respetuosos de la privacidad y justos, con mecanismos para identificar y gestionar sesgos dañinos. La aplicación de estas características debe equilibrarse de acuerdo con el contexto específico de uso del sistema.

Si bien estas características son atributos sociotécnicos, la rendición de cuentas y la transparencia también dependen de los procesos organizativos y del entorno institucional. La omisión de cualquiera de estos elementos puede aumentar la probabilidad y el impacto de consecuencias negativas.

A continuación, en la Ilustración 5, se presenta la organización de dichas características; cabe aclarar que la característica “Válido y fiable” es una condición necesaria para la confiabilidad y se muestra como base para otras características de confiabilidad. De igual

manera, la característica “Responsable y transparente” se muestra como una caja vertical porque es transversal y requerida por todas las demás características.

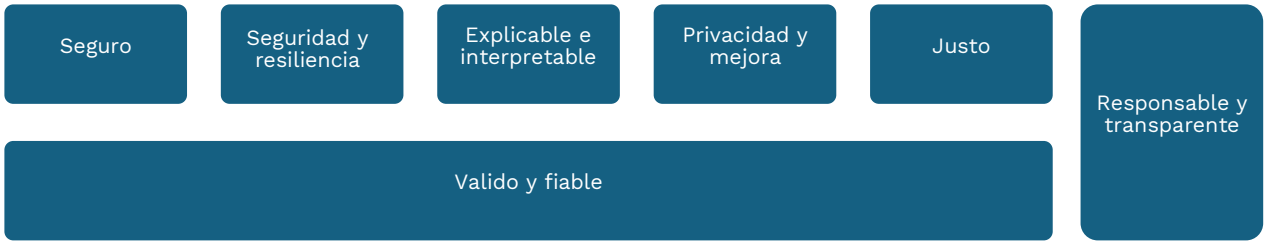

Ilustración 5. Características de sistemas de IA fiables.

Las características de fiabilidad de los sistemas de Inteligencia Artificial están estrechamente relacionadas con factores sociales y organizacionales, la calidad y naturaleza de los datos utilizados, la selección de modelos y algoritmos, y las decisiones adoptadas por quienes diseñan, desarrollan y supervisan estos sistemas. La definición de métricas y umbrales asociados a dichas características requiere necesariamente del juicio humano, considerando el contexto específico de uso.

Abordar de manera aislada las características de fiabilidad no garantiza que un sistema de IA sea confiable. En la práctica, existen compromisos y tensiones entre ellas, y no todas resultan igualmente relevantes en todos los escenarios. La confiabilidad debe entenderse como un concepto sociotécnico y contextual, cuya solidez depende del equilibrio entre sus distintos atributos.

Durante la gestión de riesgos de IA, las entidades pueden enfrentar decisiones complejas al equilibrar aspectos como interpretabilidad, privacidad, precisión, equidad y seguridad. Estas compensaciones varían según el contexto y los valores en juego, por lo que deben analizarse y resolverse de forma transparente, justificada y documentada.

Para fortalecer la comprensión del contexto a lo largo del ciclo de vida de la IA, resulta clave la participación de expertos y partes interesadas, así como, la aplicación de procesos de prueba, evaluación, verificación y validación (TEVV), alineados con las condiciones reales de despliegue. La diversidad de perspectivas contribuye a identificar riesgos, impactos y beneficios potenciales, especialmente aquellos que emergen en entornos sociales.

Finalmente, la responsabilidad de equilibrar las características de fiabilidad y determinar la idoneidad del uso de la IA es compartida entre todos los actores involucrados. La decisión de desarrollar o desplegar un sistema de IA debe basarse en una evaluación contextual integral de los riesgos, impactos, costos y beneficios, informada por un conjunto amplio y representativo de partes interesadas.

11.4.1. Válido y fiable

La validez de un sistema de inteligencia artificial se refiere a la confirmación, basada en evidencia objetiva, de que cumple los requisitos para su uso previsto, conforme a la

definición establecida en la norma ISO 9000:2015. El despliegue de sistemas de IA inexactos, poco fiables o que no generalizan adecuadamente a contextos distintos de aquellos para los que fueron entrenados incrementa los riesgos y reduce la confianza en su funcionamiento.

La fiabilidad, de acuerdo con la ISO/IEC TS 5723:2022, corresponde a la capacidad de un sistema para operar de manera consistente y sin fallos durante un período determinado y bajo condiciones específicas. En el contexto de la IA, la fiabilidad implica que el sistema funcione correctamente a lo largo de todo su ciclo de vida, dentro de los escenarios de uso esperados.

La precisión y la robustez son atributos clave que contribuyen a la validez y fiabilidad de los sistemas de IA, aunque en algunos casos pueden presentar tensiones entre sí. La precisión hace referencia a la cercanía de los resultados generados por el sistema a valores considerados verdaderos, y su medición debe realizarse mediante conjuntos de prueba representativos de las condiciones reales de uso, documentando claramente la metodología empleada. Asimismo, resulta recomendable analizar los resultados de manera desagregada para identificar posibles diferencias entre distintos segmentos de datos.

Por su parte, la robustez se define como la capacidad del sistema para mantener un nivel adecuado de desempeño frente a variaciones en las condiciones de operación, incluyendo escenarios no previstos inicialmente. Un sistema robusto no solo debe funcionar correctamente en condiciones esperadas, sino también minimizar posibles daños cuando opera en entornos imprevistos.

La evaluación de la validez y fiabilidad de los sistemas de IA debe realizarse mediante pruebas y monitoreo continuo, permitiendo identificar fallos y degradaciones en el desempeño. Estas mediciones son esenciales para la gestión de riesgos, en particular cuando ciertos errores pueden generar impactos significativos. En estos casos, puede ser necesaria la intervención humana para detectar, corregir o mitigar errores que el sistema de IA no pueda gestionar de manera autónoma.

11.4.2. Seguro

Los sistemas de inteligencia artificial deben operar de manera segura, evitando que, bajo condiciones definidas, se generen situaciones que pongan en riesgo la vida, la salud, la propiedad o el medio ambiente, conforme a lo establecido en la ISO/IEC TS 5723:2022. La seguridad de los sistemas de IA se fortalece mediante prácticas responsables a lo largo de todo su ciclo de vida.

El funcionamiento seguro de la IA requiere un diseño, desarrollo y despliegue responsables, así como, información clara y adecuada para quienes operan o utilizan estos sistemas. Asimismo, resulta fundamental que las decisiones adoptadas por los usuarios finales y operadores se basen en una comprensión adecuada de los riesgos, apoyada por documentación y explicaciones sustentadas en evidencia empírica.

La gestión de los riesgos de seguridad asociados a la IA debe adaptarse al contexto y a la gravedad de los posibles impactos. Los riesgos que puedan derivar en daños graves o pérdida de vidas humanas deben ser priorizados y gestionados mediante procesos más rigurosos y exhaustivos.

Incorporar consideraciones de seguridad desde las etapas iniciales de planificación y diseño permite prevenir fallos críticos. Adicionalmente, prácticas como la realización de pruebas y simulaciones rigurosas, el monitoreo en tiempo real y la disponibilidad de mecanismos de intervención humana o desactivación segura son esenciales para responder a desviaciones del comportamiento esperado.

Finalmente, los enfoques de seguridad en IA deben alinearse con las mejores prácticas y estándares aplicables en sectores críticos como el transporte y la salud, garantizando coherencia con las regulaciones y lineamientos específicos de cada ámbito de aplicación.

11.4.3. Seguridad y resiliencia

Los sistemas de inteligencia artificial, así como los entornos en los que se despliegan pueden considerarse resilientes cuando son capaces de soportar eventos adversos o cambios inesperados en su entorno o forma de uso, manteniendo sus funciones esenciales y degradándose de manera segura cuando sea necesario, conforme a la ISO/IEC TS 5723:2022.

Las principales preocupaciones de seguridad en los sistemas de IA incluyen amenazas como ciberataques, envenenamiento de datos y la exfiltración de modelos, datos de entrenamiento o información sensible a través de los puntos de acceso del sistema. Un sistema de IA se considera seguro cuando protege adecuadamente la confidencialidad, integridad y disponibilidad de la información mediante controles que previenen accesos o usos no autorizados, en alineación con los lineamientos del Marco de Ciberseguridad del NIST y el Marco de Gestión de Riesgos.

Si bien la seguridad y la resiliencia están estrechamente relacionadas, no son conceptos equivalentes. La resiliencia se refiere a la capacidad del sistema para recuperarse y continuar operando tras un evento adverso, mientras que la seguridad abarca un enfoque más amplio que incluye la prevención, detección, respuesta y recuperación frente a amenazas. En este sentido, la resiliencia se vincula con la robustez del sistema y considera no solo la calidad de los datos, sino también escenarios de uso indebido, abuso o comportamientos adversariales.

11.4.4. Responsable y transparente

La confianza en los sistemas de inteligencia artificial depende directamente de la rendición de cuentas, la cual se sustenta en la transparencia. La transparencia, se refiere al grado en que la información relevante sobre un sistema de IA y sus resultados es accesible para las personas que interactúan con él, incluso cuando no son plenamente conscientes de su uso. Una transparencia efectiva debe proporcionar niveles de información adecuados a cada

etapa del ciclo de vida del sistema y ajustados al rol y conocimiento de los distintos actores involucrados.

Este principio abarca aspectos como las decisiones de diseño, los datos de entrenamiento, la estructura y el entrenamiento del modelo, los casos de uso previstos y las decisiones adoptadas durante el despliegue y la operación del sistema. La transparencia resulta esencial para facilitar acciones correctivas frente a resultados incorrectos o impactos negativos, así como, para mejorar la interacción humana–IA, por ejemplo, mediante mecanismos claros de notificación cuando se detectan resultados adversos.

Si bien un sistema transparente no garantiza por sí mismo precisión, seguridad, privacidad o equidad, la falta de transparencia dificulta la evaluación de estas características, especialmente a medida que los sistemas evolucionan. Por ello, la rendición de cuentas debe considerar el rol específico de cada actor de IA y adaptarse al contexto jurídico, sectorial y social en el que se despliega el sistema, incrementando las exigencias de transparencia y responsabilidad cuando los impactos potenciales son graves.

Las entidades deben equilibrar las medidas de transparencia y rendición de cuentas con la protección de información sensible o propietaria y con la disponibilidad de recursos. En este sentido, mantener la trazabilidad y procedencia de los datos de entrenamiento, respetando los derechos de propiedad intelectual aplicables, contribuye a fortalecer tanto la transparencia como la responsabilidad. Asimismo, se recomienda que los desarrolladores y operadores evalúen y adopten de manera conjunta herramientas de transparencia y documentación que permitan asegurar un uso adecuado y responsable de los sistemas de IA.

11.4.5. Explicable e interpretable

La explicabilidad se refiere a la capacidad de describir cómo funcionan los sistemas de inteligencia artificial, mientras que la interpretabilidad se relaciona con la comprensión del significado de los resultados que estos generan en función de su propósito previsto. En conjunto, ambas características permiten a operadores, supervisores y usuarios comprender mejor el comportamiento, la fiabilidad y los impactos de los sistemas de IA.

La falta de explicabilidad o interpretabilidad puede incrementar la percepción de riesgo, especialmente cuando los usuarios no pueden entender o contextualizar adecuadamente las decisiones automatizadas. Por ello, los sistemas de IA deben ofrecer información clara y comprensible que permita explicar su funcionamiento y los resultados producidos, ajustando el nivel de detalle según el rol, los conocimientos y las capacidades de quienes interactúan con el sistema.

La explicabilidad facilita la supervisión, el monitoreo, la auditoría y la gobernanza de los sistemas de IA, mientras que la interpretabilidad permite comprender por qué el sistema realizó una predicción o recomendación específica. Ambas contribuyen a una gestión de riesgos más efectiva y a una mayor confianza en el uso de la tecnología.

Si bien son conceptos distintos, la transparencia, la explicabilidad y la interpretabilidad se refuerzan mutuamente. La transparencia permite conocer qué ocurrió en el sistema, la explicabilidad cómo se tomó una decisión y la interpretabilidad por qué esa decisión es relevante y cuál es su significado para el usuario.

11.4.6. Privacidad mejorada

La privacidad comprende el conjunto de principios, normas y prácticas orientadas a proteger la autonomía, la identidad y la dignidad de las personas, incluyendo el control sobre la recopilación, uso y divulgación de información personal. En el contexto de la inteligencia artificial, valores como el anonimato, la confidencialidad y el control de los datos deben guiar el diseño, desarrollo y despliegue de los sistemas.

Los riesgos asociados a la privacidad están estrechamente relacionados con otras características de la IA, como la seguridad, la equidad y la transparencia, y pueden generar compromisos entre estos principios. Asimismo, ciertas capacidades de los sistemas de IA pueden introducir nuevos riesgos, al permitir inferencias que revelen información sensible o identifiquen a individuos a partir de datos aparentemente anónimos.

El uso de tecnologías que mejoran la privacidad, así como prácticas de minimización de datos como la anonimización y la agregación, contribuye al desarrollo de sistemas de IA con mayor protección de la información personal. No obstante, en algunos contextos, estas técnicas pueden afectar la precisión del sistema, lo que requiere un análisis cuidadoso para equilibrar la privacidad con otros valores relevantes, como la equidad y la eficacia del sistema.

11.4.7. Justo – con sesgo dañino controlado

La equidad en los sistemas de inteligencia artificial implica abordar de manera activa los riesgos de sesgo dañino y discriminación, reconociendo que los conceptos de justicia y equidad pueden variar según el contexto cultural, social y el caso de uso específico. Por ello, la gestión de riesgos de IA debe considerar estas diferencias y adoptar enfoques proporcionales y contextualizados.

La mitigación de sesgos no garantiza, por sí sola, la justicia de un sistema de IA. Incluso cuando se logra un cierto equilibrio entre grupos demográficos, los sistemas pueden seguir generando impactos inequitativos, por ejemplo, si resultan inaccesibles para personas con discapacidad, profundizan la brecha digital o refuerzan desigualdades estructurales existentes.

El sesgo en la IA va más allá de la representatividad de los datos y puede manifestarse en distintas formas. De manera general, se identifican tres categorías principales: sesgo sistémico, asociado a normas, prácticas organizativas y condiciones sociales preexistentes; sesgo computacional o estadístico, derivado de datos no representativos o de procesos algorítmicos; y sesgo cognitivo humano, relacionado con la forma en que las personas

diseñan, interpretan y utilizan los sistemas de IA. Estos sesgos pueden surgir incluso sin intención discriminatoria.

Dado que los sistemas de IA pueden amplificar la velocidad y el alcance de los sesgos, resulta esencial identificar, evaluar y gestionar estos riesgos de manera continua. La equidad en la IA está estrechamente vinculada con la transparencia y la rendición de cuentas, y su adecuada gestión es fundamental para prevenir impactos negativos sobre individuos, comunidades, entidades y la sociedad en general.

11.5. Eficacia de la gestión de riesgos de IA

- La eficacia de la gestión de riesgos en inteligencia artificial (IA), debe evaluarse de manera sistemática considerando su capacidad para mejorar la confiabilidad, seguridad y desempeño de los sistemas de IA. Estas evaluaciones deben incluir métricas que permitan medir avances en la identificación, mitigación y control de riesgos, así como mejoras en los resultados operativos de los sistemas.
- Las entidades deben realizar evaluaciones periódicas para determinar si sus políticas, procesos, prácticas, planes de implementación, indicadores y mecanismos de seguimiento han fortalecido efectivamente su capacidad de gestionar los riesgos asociados a la IA. Este ejercicio debe promover un enfoque colaborativo con otros actores relevantes, orientado al desarrollo y adopción de métricas, metodologías y objetivos comunes, así como al intercambio de resultados y buenas prácticas en beneficio del interés público.

Una gestión de riesgos de IA eficaz se refleja en:

- Procesos fortalecidos para identificar, analizar, evaluar, tratar, supervisar y comunicar[6] los riesgos de la IA, con documentación clara y trazable de los resultados.
- Mayor comprensión de las interrelaciones y compensaciones entre las características de fiabilidad, los enfoques sociotécnicos y los riesgos asociados a los sistemas de IA.
- Procedimientos explícitos para la toma de decisiones informadas sobre la puesta en marcha, el despliegue o la suspensión de sistemas de IA.
- Políticas y mecanismos organizativos que refuercen la rendición de cuentas en relación con los riesgos y los impactos de los sistemas de IA.
- Una cultura organizacional orientada a la identificación temprana y la gestión responsable de los riesgos y posibles impactos sobre personas, comunidades, entidades y la sociedad.
- Mejores prácticas de intercambio de información, tanto al interior como entre entidades, sobre riesgos, decisiones, lecciones aprendidas, procesos de prueba, evaluación, verificación y validación (TEVV) y enfoques de mejora continua.
- Mayor conocimiento del contexto de uso para anticipar riesgos posteriores o emergentes.
- Un fortalecimiento de la participación de las partes interesadas y actores relevantes de la IA.

[6] Guía para la Gestión Integral del Riesgo en Entidades Públicas, Versión 7, del año 2025, del Departamento Administrativo de la Función Pública: https://www.funcionpublica.gov.co/detalle-publicacion?entryId=963161.

- Capacidades ampliadas para la ejecución de procesos de TEVV y la evaluación de riesgos asociados a los sistemas de IA.

11.6. El núcleo de la gestión de riesgos de la IA

El núcleo proporciona resultados y acciones que permiten el diálogo, la comprensión y las actividades para gestionar los riesgos de IA y desarrollar de forma responsable sistemas de IA fiables. Como se ilustra en la Ilustración 6, el núcleo está compuesto por cuatro funciones: Gobernar, Mapear, Medir y Gestionar. Cada una de estas funciones de alto nivel se divide en categorías y subcategorías. Las categorías y subcategorías se subdividen en acciones y resultados específicos. Las acciones no constituyen una lista de comprobación, ni necesariamente son un conjunto ordenado de pasos.

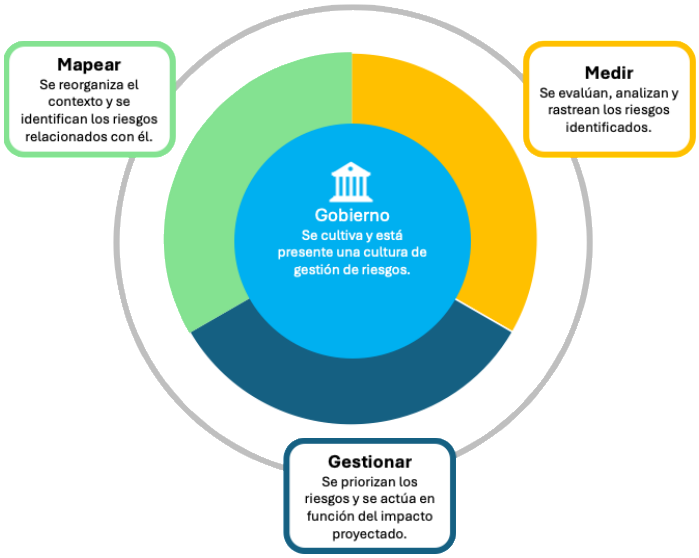

Ilustración 6. Núcleo de la gestión de riesgos de la IA.

Las funciones organizan las actividades de gestión de riesgos de IA en su nivel más alto para gobernar, mapear, medir y gestionar los riesgos de IA. La gobernanza está diseñada para ser una función transversal que informa e integra las otras tres funciones. La gestión de riesgos debe ser continua, oportuna y llevarse a cabo a lo largo de las dimensiones del ciclo de vida del sistema de IA. Las funciones principales de la gestión de riesgos de la IA deben llevarse a cabo de manera que refleje perspectivas diversas y multidisciplinares, incluyendo potencialmente las opiniones de actores de IA fuera de la entidad. Contar con un equipo diverso contribuye a un intercambio más abierto de ideas y supuestos sobre los propósitos y funciones de la tecnología que se está diseñando, desarrollando, desplegando o evaluando, lo que puede crear oportunidades para sacar a la luz problemas e identificar riesgos existentes y emergentes.

Se pueden aplicar las siguientes recomendaciones presentadas a continuación según se adapten mejor a sus necesidades para gestionar los riesgos de IA en función de sus recursos y capacidades.

11.6.1. Gobierno

La gobernanza de la inteligencia artificial tiene como objetivo consolidar una cultura organizacional de gestión de riesgos en las entidades que diseñan, desarrollan, adquieren, despliegan, evalúan o supervisan sistemas de IA. Esta gobernanza debe permitir anticipar, identificar y gestionar de manera sistemática los riesgos y posibles impactos asociados al uso de estas tecnologías.

- Un marco de gobierno eficaz establece procesos, documentos y estructuras organizativas que incorporan la evaluación de riesgos e impactos, considerando no solo a los usuarios directos, sino también a las personas y grupos que puedan verse afectados en la sociedad. Asimismo, debe definir procedimientos claros para la toma de decisiones y la rendición de cuentas.
- La gobernanza de la IA proporciona una estructura que alinea la gestión de riesgos con los principios, políticas y prioridades estratégicas de la entidad, integrando los aspectos técnicos del diseño y desarrollo de los sistemas con los valores organizacionales. Esto facilita el desarrollo de capacidades y prácticas adecuadas para los actores involucrados en la adquisición, entrenamiento, despliegue y supervisión de sistemas de IA.
- Finalmente, el enfoque de gobierno debe abarcar todo el ciclo de vida de los sistemas de IA, incluyendo consideraciones legales, contractuales y de seguridad relacionadas con el uso de software, hardware y datos de terceros, garantizando un uso responsable, coherente y sostenible de la inteligencia artificial.

11.6.2. Contexto

La gestión eficaz de los riesgos asociados a un sistema de inteligencia artificial requiere, como paso inicial, una adecuada definición del contexto en el que dicho sistema será diseñado, desarrollado, desplegado y utilizado. El ciclo de vida de la IA comprende múltiples actividades interdependientes y la participación de diversos actores, quienes, en muchos casos, no cuentan con visibilidad ni control total sobre todas las etapas o decisiones involucradas. Estas interdependencias pueden dificultar la anticipación de los impactos reales de los sistemas de IA.

- Las decisiones adoptadas en etapas tempranas como la definición de objetivos, propósitos y casos de uso pueden influir de manera significativa en el comportamiento y las capacidades del sistema. Asimismo, los cambios en el entorno de despliegue, incluyendo los usuarios finales y las personas potencialmente afectadas, pueden modificar los impactos esperados. Como resultado, decisiones bien intencionadas en una fase del ciclo de vida pueden verse afectadas por condiciones o decisiones posteriores, introduciendo incertidumbre en la gestión de riesgos.
- Definir claramente el contexto permite reducir esta incertidumbre y fortalecer la toma de decisiones, al facilitar la identificación, evaluación y mitigación de riesgos negativos. La información recopilada en esta etapa resulta clave para prevenir impactos adversos, apoyar procesos como la gestión de modelos y determinar, desde fases iniciales, la pertinencia o necesidad de una solución basada en IA.

- La contextualización se fortalece mediante la participación de equipos internos diversos y la interacción con desarrolladores, operadores, usuarios finales, comunidades potencialmente afectadas y otros actores relevantes, de acuerdo con el nivel de riesgo del sistema. Este enfoque permite a las entidades mejorar su comprensión del entorno de uso, validar supuestos, identificar desviaciones respecto al contexto previsto, reconocer beneficios potenciales, comprender limitaciones técnicas y anticipar impactos negativos tanto en usos previstos como no previstos.

11.6.3. Medición

La medición en la gestión de riesgos de la inteligencia artificial comprende el uso de herramientas, técnicas y metodologías cuantitativas, cualitativas o mixtas para analizar, evaluar, comparar y monitorear los riesgos y los impactos asociados a los sistemas de IA. Estas mediciones deben basarse en el conocimiento relevante de los riesgos identificados y aplicarse tanto antes del despliegue como de forma periódica durante la operación del sistema.

Los procesos de medición deben documentar aspectos clave de la funcionalidad, fiabilidad y desempeño de los sistemas de IA, incluyendo métricas relacionadas con características de IA confiable, impacto social y configuraciones de interacción humano–IA. Para ello, se deben aplicar metodologías rigurosas de prueba y evaluación, con métricas de incertidumbre, comparaciones con referencias de desempeño y documentación formal de resultados. La revisión independiente puede fortalecer la objetividad y reducir sesgos o conflictos de interés.

La medición proporciona una base trazable para gestionar las compensaciones entre las distintas características de fiabilidad, permitiendo decisiones informadas como la recalibración del sistema, la aplicación de controles de mitigación o, cuando sea necesario, la suspensión de su uso. Estos procesos deben integrarse en esquemas formales de prueba, evaluación, verificación y validación (TEVV), con métricas y metodologías alineadas con estándares científicos, legales y éticos, aplicadas de manera transparente.

Finalmente, dado que los sistemas de IA y sus contextos evolucionan, las actividades de medición deben mantenerse de forma continua y adaptativa, permitiendo identificar riesgos existentes y emergentes, evaluar la eficacia de las métricas utilizadas y apoyar de manera oportuna el monitoreo y la respuesta frente a riesgos.

11.6.4. Gestionar

La gestión del riesgo en sistemas de inteligencia artificial implica la asignación sistemática de recursos a los riesgos previamente identificados, medidos y priorizados. El tratamiento del riesgo debe contemplar planes claros para la respuesta, recuperación y comunicación frente a incidentes o eventos que puedan afectar el funcionamiento del sistema o generar impactos negativos, la gestión de riegos e incidentes se deben realizar teniendo en cuenta los lineamientos establecidos en el Modelo de Seguridad y Privacidad de la Información (MSPI).

La información contextual recopilada permite reducir la probabilidad de fallos y mitigar impactos adversos. Asimismo, la implementación de prácticas consistentes de documentación fortalece la gestión de riesgos de IA, incrementa la transparencia y facilita la rendición de cuentas. Estos procesos deben complementarse con mecanismos para identificar riesgos emergentes y con enfoques de mejora continua.

Las entidades deben establecer planes para la priorización del riesgo, así como esquemas de monitoreo y revisión periódica que permitan ajustar las medidas adoptadas. La aplicación continua de estas prácticas fortalece la capacidad institucional para gestionar los riesgos de los sistemas de IA en operación y para asignar recursos de manera eficiente, reconociendo que los métodos, contextos, riesgos y expectativas de los actores involucrados evolucionan con el tiempo.

11.7. Perfiles para la gestión de riesgos de IA

La definición de perfiles y responsabilidades para la gestión de riesgos de la inteligencia artificial debe basarse en las competencias y capacidades de las entidades involucradas. Estos perfiles permiten estructurar una gestión de riesgos eficaz, alineada con los requisitos legales, los principios organizacionales y las mejores prácticas internacionales. La comparación entre las capacidades actuales y los objetivos esperados facilita la identificación de brechas que deben ser abordadas de manera prioritaria.

En el ciclo de vida de la IA participan diversos actores, entre ellos desarrolladores, expertos en dominios específicos, especialistas en datos, analistas socioculturales y responsables de seguridad y riesgos. Cada uno cumple funciones diferenciadas en el diseño, desarrollo, despliegue y operación de los sistemas de IA, incluyendo la gestión de datos, la validación de modelos, la evaluación de impactos y el monitoreo continuo.

Las actividades de prueba, evaluación, verificación y validación son transversales a todo el ciclo de vida del sistema y deben involucrar a distintos perfiles para garantizar la objetividad, integridad y confiabilidad del proceso. Asimismo, la participación de usuarios finales y otras partes interesadas resulta clave para promover un enfoque centrado en las personas y mejorar la identificación de riesgos y posibles impactos.

Finalmente, la gobernanza y supervisión de la IA, así como las decisiones que sean tomadas por la IA, constituyen funciones estratégicas que deben ser asumidas por el liderazgo organizacional, con el apoyo de instancias técnicas y éticas. Este enfoque debe considerar también a los usuarios finales y a las comunidades potencialmente afectadas, reconociendo su papel en la identificación de riesgos, la retroalimentación y la construcción de sistemas de IA responsables y confiables.

Tipo de Rol	Descripción
Desarrollador de IA	Encargado del diseño y desarrollo de sistemas de IA, asegurando que se minimicen los sesgos en la programación y se integren mejores prácticas éticas.

Usuario Final	Interactúa con el sistema de IA y puede introducir sesgos a través de decisiones de uso. Su influencia puede impactar en el resultado del sistema.
Experto Humano	Puede asignarse para supervisar el sistema de IA, aportando juicio crítico y experiencia en la toma de decisiones.
Responsable de Decisiones	Evalúa y utiliza la información proporcionada por la IA para tomar decisiones informadas; puede ver la IA como una opinión adicional o una guía.
Organización/Equipo	La estructura organizativa influye en cómo se distribuyen los roles y cómo se toman las decisiones; los sesgos sistémicos pueden afectar la efectividad y la transparencia.
Responsable Político	Influye en las decisiones sobre el uso y regulación de la IA, considerando las implicancias prácticas y éticas de los sistemas de IA.

Tabla 3. Recomendaciones relacionadas con los perfiles.

12. Recomendaciones Generales

12.1. Identificación y gestión de riesgos asociados a las vulnerabilidades de los sistemas basados en Inteligencia Artificial.

La identificación y gestión de los riesgos asociados a los sistemas basados en inteligencia artificial debe contemplar, de manera prioritaria, las vulnerabilidades técnicas y operativas propias de estos modelos, incluyendo riesgos como la inyección de prompts (prompt injection), el proceso para eliminar las restricciones impuestas y saltarse los filtros de seguridad en el envenenamiento de datos y otros vectores de ataque que pueden comprometer la seguridad, integridad y confiabilidad de los sistemas.

El uso de la IA conlleva riesgos que pueden afectar tanto a las personas como a las entidades. Entre los más relevantes se encuentran el sesgo y la discriminación, la falta de transparencia, los riesgos para la privacidad y protección de datos, las amenazas a la ciberseguridad, los impactos en el empleo, la dependencia tecnológica, la manipulación de información y desinformación, los desafíos de responsabilidad legal, el impacto ambiental y la ausencia de marcos sólidos de ética y gobernanza.

Abordar estos riesgos resulta esencial para garantizar que la implementación de la IA se realice de manera ética, segura y responsable, maximizando sus beneficios y minimizando los posibles perjuicios. En este sentido, los marcos regulatorios y de gobernanza de la IA deben priorizar la protección de los derechos humanos, la mitigación de riesgos para la sociedad y, al mismo tiempo, habilitar el uso legítimo de la IA por parte del Estado y la ciudadanía.

Un riesgo relevante también surge cuando las entidades no adoptan tecnologías de IA eficaces, especialmente aquellas con capacidad para detectar y prevenir incidentes

relacionados con ciberseguridad, fraude, abuso en línea y otros daños digitales. Por ello, se recomienda un enfoque holístico basado en el riesgo, que permita aplicar medidas de protección proporcionales al nivel de riesgo y al contexto específico de cada caso de uso.

Un marco de gestión de riesgos de IA basado en el riesgo debe proporcionar criterios flexibles para evaluar la probabilidad y gravedad de los impactos, facilitar la adaptación de medidas de mitigación a riesgos reales y evitar controles innecesarios. Este enfoque debe incluir la identificación de factores y perjuicios potenciales, la orientación sobre usos de mayor o menor riesgo y la actualización continua de metodologías a partir de la experiencia acumulada.

Asimismo, dicho marco debe ser eficiente en el uso de recursos, pro-innovación y centrado en resultados, desarrollado mediante procesos abiertos y colaborativos, con un lenguaje claro que facilite la comunicación de riesgos entre equipos técnicos, directivos, entidades públicas y la sociedad. La adopción de una taxonomía común, alineada con estándares y buenas prácticas internacionales, permitirá una gestión coherente e integrada de los riesgos de la IA, incluyendo aquellos derivados de vulnerabilidades técnicas avanzadas como la manipulación de modelos y la evasión de controles de seguridad.

A continuación, se presenta el Top 10 OWASP - 2025 de riesgos y mitigaciones para LLMs y aplicaciones de IA Generativa y otros riesgos relacionados con la Inteligencia Artificial:

Vulnerabilidad	Tipo de riesgo	Impacto potencial	Medidas de mitigación
LLM01:2025 Inyección de Prompts (Prompt Injection)	Integridad del modelo	Divulgación de información no deseada o ejecución de acciones prohibidas.	Aislar la entrada del usuario usando delimitadores estrictos (ej. """ o etiquetas XML). Implementar un clasificador de intención en el backend previo al LLM para descartar heurísticas de inyección antes de la inferencia.
LLM02:2025 Divulgación de Información Sensible	Privacidad	Exposición de datos sensibles de usuarios.	Integrar un motor de sanitización en el pipeline de entrada para anonimizar PII/PHI. Aplicar filtrado basado en roles (RBAC), validando el token del usuario directamente en las consultas a la base de datos vectorial.
LLM03:2025 Cadena de Suministro Insegura	Seguridad del suministro	Fuga de información crítica.	Verificar fuentes de datos y proveedores de información. Automatizar análisis de composición (SCA) en el CI/CD para librerías de IA. Exigir firmas criptográficas y revisar el SBOM al importar modelos pre-entrenados de repositorios de terceros.

LLM04:2025 Envenenamiento de Datos y Modelo (Data Poisoning and Model)	Integridad de datos	Comportamiento del modelo sesgado o poco ético.	Monitorear hashes de los documentos base del RAG para detectar modificaciones no autorizadas. Exigir validación humana (HITL), para autorizar nuevos datos antes de introducirlos a un flujo de fine-tuning o RAG.
LLM05:2025 Manejo Inadecuado de la Salida	Disponibilidad	Ejecución de código malicioso o acceso no autorizado.	Tratar toda salida del LLM como datos sin confianza. Aplicar *Output Encoding* (ej. escape HTML/JS) en la UI. Si el agente genera código para ejecutar, forzar su ejecución en contenedores efímeros, sin acceso a red y con privilegios nulos.
LLM06:2025 Agencia Excesiva	Control	Acciones perjudiciales por exceso de autonomía del sistema.	Otorgar al LLM o agente únicamente los permisos mínimos necesarios, evitar crear "súper agentes", exigir confirmaciones manuales, restringir cantidad de acciones en un periodo de tiempo, implementar defensa en profundidad sobre los datos de entrada, monitorear registro del razonamiento y configurar alertas de comportamiento anómalo.
LLM07:2025 Filtración de prompts de sistema (System Prompt Filtering)	Confidencialidad	Compromiso del sistema por acceso a información crítica.	Instrucciones de No divulgación y Encapsulamiento. Análisis de intenciones y validación de salidas. Separación de privilegios usando un LLM para limpiar la entrada y otro para procesar, limitando el contexto histórico accesible. Implementar frameworks de Guardarraíls. No incluir información sensible en los prompts del sistema.
LLM08:2025 Debilidades de vector y representaciones vectoriales	Seguridad de consultas	Filtraciones de información entre usuarios.	Implementar aislamiento multitenencia (*Tenant Isolation*) inyectando el tenant_id en los metadatos del *embedding* y forzando el filtro de este ID en cada consulta. Exigir cifrado AES-256 en reposo para la base de datos vectorial.

LLM09:2025 Desinformación	Veracidad	Efectos reputacionales y legales.	Utilizar RAG, de preferencia semántico y no por límite de tokens o caracteres. Verificar información generada por IA. Solicitar, en los prompts, la fuente exacta de la la información, indicando reporte de desconocimiento en caso de no encontrar información válida.
LLM10:2025 Consumo Ilimitado	Disponibilidad	Caída del servicio y pérdidas financieras.	Establecer topes absolutos en el parámetro max_tokens tanto de entrada como de salida en el API. Implementar Rate Limiting por IP y usuario, junto con timeouts estrictos en el gateway para abortar inferencias atascadas.
Jailbreak de modelos (Jailbreak of models)	Seguridad / Ética	Evasión de restricciones y uso indebido del modelo	Pruebas adversariales, sobrescribir y agregar la directriz principal de seguridad al final del último prompt del usuario para que sea la última instrucción que evalúe el modelo.
Ataques adversariales	Seguridad / Disponibilidad	Manipulación de resultados y degradación del servicio	Normalizar y pre-procesar el texto de entrada (eliminando caracteres invisibles, unicodes inusuales o padding anómalo) antes de la tokenización. Desplegar un WAF configurado para detectar inyecciones orientadas a IA.
Sesgo algorítmico	Ético / Legal	Discriminación y vulneración de derechos	Implementar técnicas de re-ponderación de clases en el pre-procesamiento de datos y automatizar pruebas con métricas de equidad algorítmica en el pipeline de evaluación.
Falta de explicabilidad	Transparencia / Gobernanza	Dificultad para auditar y rendir cuentas	Integrar algoritmos de interpretabilidad y atribución de características (ej. valores de Shapley) en las respuestas. Configurar el LLM para generar y registrar internamente la cadena de razonamiento en los logs de auditoría.
Dependencia excesiva de la IA	Operacional	Pérdida de control humano y errores no detectados	Forzar la exposición de métricas de confianza en la interfaz final. Implementar firmas o marcas de agua criptográficas en el

			contenido generado y requerir revisión manual para flujos de decisión.
Fugas de datos personales	Legal / Privacidad	Incumplimiento normativo y sanciones	Aplicar técnicas de reconocimiento de entidades nombradas (NER) y expresiones regulares en el Gateway API para ejecutar un enmascaramiento y pseudoanonimización bidireccional. Aplicación periódica de DPIA.
Impacto ambiental elevado	Sostenibilidad	Aumento de huella de carbono	Aplicar técnicas de cuantización de pesos y destilación de modelos. Desplegar un sistema de caché semántico en la arquitectura para servir respuestas cacheadas a consultas vectorialmente similares y evitar la recomputación.
Fallas en actualizaciones del modelo	Disponibilidad / Confiabilidad	Interrupción del servicio o errores en producción	Implementar gestión de cambios, pruebas previas y técnicas de rollback. Automatizar pipelines de pruebas de regresión comparativa de inferencia y configurar alertas para la desviación de la distribución de datos.

Tabla 4. Top 10 OWASP - 2025 de riesgos y mitigaciones para LLMs y aplicaciones de IA Generativa y otros riesgos relacionados con la Inteligencia Artificial.[7]

12.1.1. Evaluación de Riesgos Especializada:

La evaluación especializada de riesgos en sistemas de inteligencia artificial tiene como finalidad asegurar que el sistema de gestión de la IA sea capaz de cumplir los objetivos definidos, de manera segura, confiable y alineada con la política de gestión de riesgos de la entidad.

- Este proceso incluye la realización de evaluaciones específicas para identificar los riesgos e impactos asociados al uso de la IA, así como el análisis de las posibles consecuencias para la entidad, la ciudadanía y terceros en caso de que dichos riesgos se materialicen. A partir de este análisis, se deben determinar los niveles de riesgo, evaluando tanto la probabilidad de ocurrencia como la magnitud de los impactos potenciales.

[7] Basado en Top 10 2025 de riesgos y mitigaciones para LLMs y aplicaciones de IA Generativa. OWASP Gen AI Security Project. Recuperado 16 de abril de 2026, de https://genai.owasp.org/resource/top-10-2025-de-riesgos-y-mitigaciones-para-llms-y-aplicaciones-de-ia-generativa/

- Con base en los resultados obtenidos, los riesgos deben ser priorizados para la definición e implementación de planes de tratamiento adecuados. La evaluación puede apoyarse en simulaciones y pruebas controladas que permitan comprender cómo actores maliciosos podrían explotar debilidades del sistema, incluyendo ataques como la inyección de prompts o el bypass de mecanismos de control (jailbreak).
- Adicionalmente, la evaluación debe identificar aquellos riesgos que facilitan o dificultan el cumplimiento de los objetivos de la política de gestión de riesgos, y definir las acciones necesarias para su tratamiento. Como parte de un enfoque preventivo y de mejora continua, se deben realizar análisis periódicos de vulnerabilidades sobre las herramientas de IA, con el fin de detectar oportunamente nuevos vectores de ataque y fortalecer las medidas de seguridad implementadas.

12.1.2. Fortalecimiento de Guardarraíles:

Los guardarraíles son medidas que permiten desarrollar un modelo de IA con responsabilidad, para evitar que se suponga una amenaza tecnológica, social o de seguridad, teniendo en cuenta que la IA clásica y la generativa necesitan guardarraíles diferentes, teniendo en cuenta que la generativa tiene la capacidad de interactuar directamente con usuarios finales, trae nuevos desafíos para los desarrolladores. Estos mecanismos también ayudan a reducir la degradación que experimenta el rendimiento de los modelos de IA con el tiempo, una situación que puede surgir por cambios en los datos de entrada o por una falta de actualización continua.

El nuevo Reglamento Europeo de Inteligencia Artificial, aprobado en 2024, clasifica los riesgos de los sistemas de IA en cuatro niveles:

- **Riesgo mínimo:** Sin obligaciones específicas; incluye, por ejemplo, videojuegos y herramientas de correo que usan IA para filtrar correos no deseados.
- **Riesgo limitado:** Deben cumplir con obligaciones de transparencia; por ejemplo, los chatbots deben ser identificados como tales a los usuarios.
- **Riesgo alto:** Sistemas que pueden afectar la salud, seguridad o derechos fundamentales, como robots quirúrgicos y herramientas de selección de personal. Están sujetos a requisitos estrictos de calidad, transparencia y supervisión humana.
- **Riesgo inaceptable:** Sistemas que amenazan la seguridad o derechos de las personas, prohibidos por el reglamento. Ejemplos incluyen el reconocimiento biométrico en tiempo real en espacios públicos y la manipulación de emociones.

Para sistemas de alto riesgo, se requieren salvaguardas técnicas y organizativas que protejan derechos fundamentales, garantizando la privacidad, evitando sesgos y minimizando errores. Además, se imponen obligaciones de transparencia para informar a los usuarios sobre su interacción con sistemas de IA.

Por lo anteriormente expuesto sobre IA, se proponen varios guardarraíles:

- **Tecnológicos:** Implementar modelos de control automático para prevenir sesgos, moderación de contenido que evite discurso de odio y desinformación, filtros de

seguridad para bloquear contenido ilegal, y herramientas de monitoreo constante y auditoría técnica para asegurar la explicabilidad de los sistemas.

- **De procedimiento:** Respetar normativas internas y protocolos éticos durante el desarrollo de sistemas IA, estableciendo procesos para revisar y aprobar los sistemas antes de su implementación, a fin de detectar fallos o vulnerabilidades.
- **Humanos:** Incluir procesos de supervisión directa (conocidos como "human in the loop") y la revisión por expertos y comités éticos para evaluar riesgos y decisiones sobre usos específicos de la IA.

12.1.2.1. Monitoreo y Detección Automatizada:

- Implementar sistemas de monitoreo de input/output que puedan detectar intentos de inyección de prompts o jailbreaks.
- Emplear algoritmos de detección de anomalías para identificar comportamientos fuera de lo común.
- Definir de forma clara qué aspectos del sistema serán monitoreados: seguridad, rendimiento, cumplimiento, etc.
- Clasificar los riesgos específicos que deseas mitigar con los guardarraíles.
- Utilizar herramientas adecuadas para monitorear entradas, salidas, y comportamiento del modelo.
- El monitoreo debe estar integrado en el ciclo de vida del desarrollo del modelo.
- Establecer umbrales claros para la operación normal del sistema.
- Configurar alertas para eventos anómalos o umbrales excedidos.
- Implementar detección en tiempo real para respuestas generadas por IA.
- Usar modelos de detección de anomalías para identificar comportamientos atípicos.
- Diseñar acciones automáticas cuando se detecte un evento crítico.
- Mantener un registro detallado de los incidentes para análisis posteriores.
- Realizar auditorías regulares para ajustar y mejorar los guardarraíles.
- Asegurarse de actualizar los procedimientos conforme surjan nuevas amenazas.
- Verificar periódicamente que las operaciones cumplan con las políticas internas y externas.
- Asegurarse de que el sistema respete principios éticos y derechos de los usuarios.

12.1.2.2. Pruebas de Seguridad Proactivas:

- Realizar pruebas de stress y análisis de vulnerabilidades enfocadas en inyección de prompts y jailbreak.
- Usar técnicas de fuzzing para provocar malfuncionamientos y anticipar posibles debilidades.
- Diseñar y reforzar los guardarraíles para evitar manipulaciones maliciosas de inputs, asegurando que los modelos operen dentro de los parámetros deseados.
- Los modelos se deben mantener robustos ante inputs adversarios y evitar que generen respuestas dañinas o no deseadas.

12.1.2.3. Redundancia y Aislamiento:

- Implementar redundancias en los sistemas y procesos para mitigar el impacto potencial de una vulnerabilidad explotada.
- Aislar componentes críticos para que una vulnerabilidad no afecte todo el sistema.

12.1.2.4. Capacitación y Concienciación:

- Formar a los desarrolladores y personal clave en la identificación y mitigación de inyecciones de prompts y técnicas de jailbreak.
- Promover una cultura de seguridad donde todos los involucrados estén alerta a estas amenazas.

12.1.2.5. Actualizaciones y Mejoras de Modelo:

- Mantener los modelos actualizados con las últimas técnicas de defensa contra ataques conocidos.
- Mejorar los modelos continuamente basándose en el feedback y evaluación de las amenazas emergentes.

12.1.2.6. Colaboración y Comunidad:

- Colaborar con expertos en seguridad y compartir hallazgos sobre vulnerabilidades de IA en comunidades tecnológicas.
- Participar en foros de intercambio de información para estar al tanto de nuevas técnicas de defensa.

12.2. Recomendaciones para implementar una arquitectura segura de modelos de IA.

Implementar una arquitectura segura para modelos de Inteligencia Artificial (IA) es fundamental para proteger los datos, la privacidad y la integridad de los sistemas. Aquí hay algunas recomendaciones clave:

- Aplicar técnicas de anonimización y ofuscamiento de datos sobre los datos utilizados por los sistemas de IA.
- Implementar controles de acceso estrictos para el manejo de datos sensibles.
- Garantizar que los conjuntos de datos sean representativos y estén libres de sesgos relevantes.
- Realizar auditorías periódicas de los datos y de los modelos de IA.
- Restringir el acceso a los modelos, datos y entornos de IA únicamente a personal autorizado.
- Implementar autenticación multifactor para accesos críticos a sistemas y plataformas de IA.

- Ejecutar pruebas de robustez para evaluar el comportamiento del modelo frente a entradas inesperadas o maliciosas.
- Aplicar técnicas de entrenamiento adversarial para fortalecer los modelos frente a ataques.
- Establecer mecanismos de monitoreo continuo del desempeño y la seguridad de los modelos.
- Definir planes formales de actualización y mantenimiento de modelos ante nuevas amenazas y vulnerabilidades.
- Asegurar el cumplimiento de la normativa vigente en materia de protección de datos y privacidad.
- Documentar de manera clara las decisiones clave relacionadas con el diseño, entrenamiento y despliegue de los modelos.
- Promover el uso ético y responsable de la inteligencia artificial dentro de la organización.
- Proporcionar capacitación periódica en seguridad, riesgos y buenas prácticas de IA a los equipos involucrados.

12.2.1. Separar el prompt del procesamiento de datos.

Separar el procesamiento de datos del prompt en una arquitectura de modelos de IA puede mejorar la flexibilidad y la seguridad del sistema. Aquí hay algunas recomendaciones para lograrlo:

- **Arquitectura de Microservicios:** Diseña la solución utilizando microservicios, donde cada servicio se encarga de un aspecto específico (por ejemplo, uno para el preprocesamiento de datos, otro para el modelo y otro para el post - procesamiento).
- **Interfaz de Comunicación:** Define una API clara que separe el procesamiento de datos del modelo. Esto permite que el modelo reciba datos en un formato específico sin depender de cómo se generan esos datos.
- Pipeline de Datos: Crea un pipeline de datos para preprocesar y transformar datos antes de que lleguen al modelo. Así, el prompt puede ser ajustado sin afectar el procesamiento subyacente.
- **Gestión de Configuraciones:** Utiliza archivos de configuración o sistemas de gestión para definir las reglas de procesamiento de datos. Esto permite cambios sin modificar el código del modelo mismo.
- **Modularidad:** Implementa una estructura modular donde el procesamiento de datos y la lógica del modelo pueden ser probados y desplegados por separado.
- **Almacenamiento de Datos:** Usa bases de datos o sistemas de almacenamiento que mantengan los datos procesados y brinden acceso al modelo de manera controlada.
- **Pruebas Aisladas:** Realiza pruebas independientes tanto para el procesamiento de datos como para el modelo para asegurarte de que ambos componentes funcionen correctamente sin interferencias.

12.2.2. Desarrollar defensa en profundidad.

12.2.2.1. Capa perimetral y de red

- Implementar firewalls de red y de aplicaciones para filtrar tráfico no autorizado.
- Desplegar sistemas de detección y prevención de intrusiones (IDS/IPS).
- Aplicar segmentación y microsegmentación de red para limitar el movimiento lateral.
- Restringir accesos externos mediante listas de control y redes privadas virtuales (VPN).

12.2.2.2. Capa de aplicación

- Aplicar validación y sanitización estricta de todas las entradas.
- Implementar autenticación multifactor (MFA) para accesos a sistemas críticos.
- Establecer controles de acceso basados en roles y privilegio mínimo (RBAC).
- Proteger APIs y servicios de IA con controles de seguridad específicos.
- Implementar Firewall de Aplicaciones Web (WAF), con soporte para IA.

12.2.2.3. Capa de modelos de IA

- Implementar guardarraíles para limitar comportamientos no autorizados del modelo.
- Realizar pruebas adversariales periódicas (prompt injection, jailbreak, inputs maliciosos).
- Aplicar versionamiento y control de cambios sobre modelos y configuraciones.
- Limitar el acceso a modelos y artefactos de entrenamiento.

12.2.2.4. Capa de datos

- Aplicar cifrado de datos en tránsito y en reposo.
- Implementar técnicas de anonimización y ofuscamiento de datos personales.
- Garantizar la calidad, integridad y procedencia de los datos de entrenamiento.
- Restringir el acceso a datos sensibles mediante controles técnicos y administrativos.

12.2.2.5. Capa de infraestructura

- Aplicar prácticas de endurecimiento (*hardening*) de sistemas operativos y plataformas.
- Gestionar de manera oportuna parches y actualizaciones de seguridad.
- Implementar monitoreo continuo de infraestructura y servicios.
- Registrar y auditar eventos relevantes de seguridad.

12.2.2.6. Monitoreo y detección

- Implementar sistemas de monitoreo del desempeño y comportamiento de los modelos.
- Detectar desviaciones, anomalías y comportamientos inesperados.
- Integrar alertas de seguridad con los procesos de respuesta a incidentes.

12.2.2.7. Respuesta y recuperación

- Definir y mantener planes de respuesta a incidentes específicos para IA.
- Establecer capacidades de desactivación segura (*safe shutdown*), de modelos.
- Implementar políticas de copias de seguridad y recuperación ante desastres.

12.2.2.8. Gobernanza y gestión

- Definir roles y responsabilidades claras para la seguridad de sistemas de IA.
- Integrar la gestión de riesgos de IA con la gestión de riesgos organizacional.
- Mantener documentación técnica y de decisiones clave.
- Realizar auditorías periódicas de seguridad y cumplimiento.

12.2.2.9. Personas y cultura organizacional

- Capacitar periódicamente al personal en seguridad y riesgos de IA.
- Sensibilizar sobre amenazas como ingeniería social y uso indebido de IA.
- Promover una cultura de seguridad y responsabilidad en el uso de IA.

12.2.3. Aplicar el principio de mínimo privilegio

El principio de mínimo privilegio establece que los usuarios, aplicaciones y componentes de los sistemas de inteligencia artificial deben contar únicamente con los permisos estrictamente necesarios para cumplir sus funciones. Su aplicación reduce la superficie de ataque, limita el impacto de incidentes de seguridad y fortalece la protección de los activos de información.

12.2.3.1. Definición estructurada de roles y responsabilidades

Se deben identificar y documentar formalmente los perfiles técnicos y administrativos (desarrolladores, científicos de datos, auditores, etc.), asociando a cada identidad digital únicamente los privilegios técnicos requeridos para su operación. Esto evita la asignación genérica de permisos con altos privilegios.

12.2.3.2. Auditoría y recertificación periódica de privilegios

Es imperativo establecer un ciclo de revisión de accesos para verificar que los permisos sigan siendo pertinentes. Se deben revocar de manera inmediata los privilegios de cuentas

inactivas, personal desvinculado o usuarios que hayan cambiado de rol dentro de la arquitectura del sistema de IA.

12.2.3.3. Implementación de modelos de Control de Acceso Basado en Roles (RBAC)

Se deben desplegar esquemas RBAC para automatizar la asignación de permisos según la función del usuario. Este control debe aplicarse tanto al acceso a la infraestructura de cómputo como a los conjuntos de datos de entrenamiento y a las interfaces de consulta (endpoints) del modelo.

12.2.3.4. Segregación de funciones (SoD) en tareas críticas

Se debe garantizar que ninguna persona tenga el control total sobre el ciclo de vida del sistema de IA. Por ejemplo, quien desarrolla el modelo no debe tener privilegios exclusivos para aprobar su paso a producción o para modificar los registros de auditoría (logs) del sistema.

12.2.3.5. Mecanismos de acceso condicionado y autenticación reforzada

Se requiere la implementación de Autenticación Multifactor (MFA) y políticas de acceso condicionado para interactuar con componentes críticos. El acceso debe validarse no solo mediante credenciales, sino mediante el estado de salud del dispositivo y la verificación de la identidad del solicitante.

12.2.3.6. Restricciones contextuales de tiempo y ubicación

Para mitigar riesgos operativos, se deben configurar límites de acceso basados en el contexto, restringiendo las operaciones críticas a horarios laborales predefinidos, segmentos de red específicos o zonas geográficas autorizadas mediante técnicas de geofencing.

12.2.3.7. Trazabilidad, monitoreo y auditoría continua de accesos

Se deben capturar y centralizar los registros de actividad de todos los usuarios con privilegios. El monitoreo debe estar orientado a detectar patrones anómalos, como intentos de escalamiento de privilegios o acceso inusual a volúmenes masivos de datos de entrenamiento.

12.2.3.8. Integración de soluciones de Gestión de Identidades y Accesos (IAM)

La gestión de privilegios debe centralizarse mediante herramientas de IAM que permitan aplicar políticas de seguridad de manera uniforme en toda la pila tecnológica, desde el almacenamiento en la nube hasta el despliegue de modelos en el borde de la infraestructura tecnológica.

12.2.3.9. Programas de capacitación y concienciación técnica

El personal técnico debe recibir formación específica sobre la importancia de la higiene de credenciales y los riesgos de seguridad derivados de la sobreexposición de permisos en entornos de inteligencia artificial.

La aplicación sistemática del principio de mínimo privilegio es un componente clave de una arquitectura de seguridad robusta para sistemas de inteligencia artificial, ya que limita los riesgos operativos y fortalece la confianza en el uso de estas tecnologías.

12.2.4. Minimizar la superficie de exposición.

La minimización de la superficie de exposición constituye un principio fundamental para la protección de los sistemas y aplicaciones de inteligencia artificial, ya que reduce las oportunidades de explotación por parte de actores maliciosos y limita el impacto de posibles incidentes de seguridad. Este enfoque se basa en la reducción sistemática de componentes, accesos y funcionalidades innecesarias a lo largo de todo el ciclo de vida del sistema.

1. Eliminación de servicios y componentes innecesarios
    - Identificar y deshabilitar servicios, aplicaciones, APIs y procesos que no sean esenciales para el funcionamiento del sistema de IA.
    - Reducir los puntos de entrada expuestos al entorno externo.
2. Gestión de actualizaciones y parches
    - Mantener actualizados los sistemas operativos, plataformas, bibliotecas y dependencias utilizadas por los modelos de IA.

- Corregir oportunamente vulnerabilidades conocidas.

3. Configuraciones seguras por defecto
   - Establecer parámetros de seguridad restrictivos desde la fase de diseño y despliegue.
   - Evitar configuraciones que expongan datos, servicios o interfaces innecesarias.
4. Segmentación y aislamiento de red
   - Implementar segmentación y microsegmentación de red para limitar el movimiento lateral.
   - Aislar entornos de desarrollo, pruebas y producción.
5. Control de acceso estricto
   - Aplicar políticas de acceso basadas en el principio de mínimo privilegio.
   - Implementar autenticación multifactor para accesos críticos.
6. Protección perimetral y detección de intrusiones
   - Utilizar firewalls, sistemas de prevención y detección de intrusiones (IDS/IPS) y controles de tráfico.
   - Monitorear de forma continua las comunicaciones entrantes y salientes.
7. Protección de datos sensibles
   - Aplicar cifrado de datos en tránsito y en reposo.
   - Restringir el acceso a información sensible mediante controles técnicos y administrativos.
8. Auditorías y pruebas de seguridad
   - Realizar auditorías de seguridad, análisis de vulnerabilidades y pruebas de penetración de manera periódica.
   - Corregir oportunamente las debilidades identificadas.
9. Gestión controlada de cambios
   - Establecer procesos formales de gestión de cambios para modificaciones en modelos, código o infraestructura.
   - Evaluar el impacto de cada cambio en la seguridad del sistema.
10. Eliminación de código y funcionalidades obsoletas
    - Revisar y retirar componentes, librerías o funcionalidades que ya no estén en uso.
    - Reducir la complejidad técnica y el riesgo de vulnerabilidades ocultas.
11. Capacitación y concienciación del personal
    - Formar al personal sobre prácticas seguras y detección de amenazas, como la ingeniería social y el phishing.

La aplicación consistente de estas medidas permite reducir significativamente la superficie de exposición de los sistemas de inteligencia artificial, fortaleciendo su seguridad, resiliencia y confiabilidad frente a amenazas emergentes

12.3. Integridad y calidad de datos para la mitigación del envenenamiento en modelos de IA.

Garantizar la calidad de los datos para el entrenamiento de modelos de Inteligencia Artificial es fundamental para evitar ataques de envenenamiento de datos, que pueden comprometer la integridad y el rendimiento del modelo. Aquí se describen algunas recomendaciones clave:

- Definir políticas claras sobre qué datos se recopilan, cómo se almacenan y quién tiene acceso a ellos. Asegurarse de que los datos provengan de fuentes confiables.
- Implementar procesos de validación para verificar la calidad y consistencia de los datos antes de usarlos para el entrenamiento. Filtra entradas que sean sospechosas o no representativas.
- Realizar auditorías regulares de los conjuntos de datos para identificar posibles contaminaciones y evaluar la calidad general de los datos utilizados para entrenar los modelos.
- Utilizar conjuntos de datos anotados por expertos humanos para asegurar que los datos sean precisos y relevantes.
- Implementar técnicas de detección de anomalías para identificar datos que podrían ser el resultado de un ataque de envenenamiento. Esto puede incluir cambios drásticos en las estadísticas de datos o patrones inesperados.
- Asegurar que el conjunto de datos sea diverso y representativo del dominio. Esto ayuda a reducir el riesgo de que un pequeño conjunto de datos pueda influir en el modelo de manera desproporcionada.
- Capacitar a los empleados sobre los riesgos de envenenamiento de datos y las mejores prácticas para la gestión de datos, para que puedan identificar posibles problemas.
- Implementar métodos de aprendizaje robusto que sean menos sensibles a las perturbaciones en los datos de entrenamiento. Esto puede incluir el uso de técnicas de regularización y algoritmos de defensa específicos.
- Utilizar la validación cruzada para evaluar la estabilidad y el rendimiento del modelo en diferentes subconjuntos de datos, ayudando a identificar vulnerabilidades.
- Implementar un sistema de control de versiones para rastrear cambios en los conjuntos de datos y revertir a versiones anteriores si se identifica un problema de calidad.
- Mantener el registro histórico de la recopilación y el uso de datos, para poder rastrear cambios que puedan haber influido en la calidad del modelo.

Al seguir estas recomendaciones, se tiende en mejorar significativamente la calidad de los datos utilizados para el entrenamiento de modelos de IA, lo que contribuirá a evitar ataques de envenenamiento y mejorará la robustez y confiabilidad de las soluciones

12.4. Recomendaciones de capacidades técnicas, operativas, humanas y administrativas mínimas para el sector público y privado en seguridad digital.

12.4.1. Lineamientos técnicos y éticos para el ciclo de vida de IA

12.4.1.1. Diseño y planificación

- Definir bajo un propósito legítimo, específico, que sea coherente con la garantía de los derechos humanos, la protección de datos de carácter personal y el bienestar social.
- Ejecutar un proyecto de Inteligencia Artificial, se debe llevar a cabo la Evaluación de Impacto en Privacidad y Seguridad, que es la que identifica riesgos de datos de carácter personal, digitales y de derechos de los titulares.
- Diseñar la arquitectura del sistema bajo el principio de “seguridad desde el diseño o el diseño seguro” (security by design), incorporando controles de acceso, cifrado y segmentación de datos desde las etapas iniciales.
- Establecer criterios de calidad, y la pertinencia para la selección y adquisición de datos de entrenamiento, estableciendo en todo caso el uso responsable de datos o anonimizados o datos sintéticos, cuando proceda,
- Implementar procedimientos en los equipos responsables de revisión para identificar la existencia de sesgos que puedan llevar a la generación de discriminación o resultados inequitativos, documentando las medidas correctivas ejecutadas.

12.4.1.2. Desarrollo

- El código fuente, los conjuntos de datos y las versiones de los modelos deberán ser gestionados bajo sistemas de control de versiones que permitan trazabilidad, auditoría y reversión de los cambios.
- Los datos que se encuentren en reposo o en tránsito deberán ser protegidos mediante cifrado robusto conforme a estándares internacionales (por ejemplo, ISO/IEC 27001 e ISO/IEC 42001) y a las pautas del MSPI.
- El sistema deberá someterse obligatoriamente a las pruebas de robustez, resiliencia y detección de sesgos antes de pasar a producción, siendo necesario registrar los resultados de dichas pruebas y las acciones correctivas realizadas.
- Los desarrolladores deberán implementar mecanismos de transparencia funcional, utilizando un lenguaje accesible para cualquier parte interesada.
- En el caso de que el sistema interactúe con grupos vulnerables (ej.: menores de edad), deberán aplicarse controles adicionales en materia de protección de datos y de contenido, tal como se establece en la Circular SIC.

12.4.1.3. Implementación y despliegue

- Antes del paso a producción, se deberá realizar una validación exhaustiva del sistema en un entorno controlado para comprobar que sea preciso, seguro y conforme a la normativa.
- Toda implementación deberá ir acompañada de información clara y comprensible para los usuarios sobre el uso de la IA, el tratamiento de sus datos y los derechos que les asisten, conforme al principio de consentimiento informado.
- El sistema deberá asegurar que se incorporen mecanismos de explicabilidad que permitan a los usuarios y a las autoridades comprender la forma en la que se obtiene el resultado o se llega a la decisión.
- El entorno de despliegue deberá cumplir con los controles de seguridad necesarios.
- Los registros de la configuración y los parámetros críticos deberán triplicarse de forma segura e incorporar mecanismos de integridad para evitar su manipulación.

12.4.1.4. Operación y uso

- Los sistemas deben operar bajo un régimen de supervisión humana significativa, manteniendo la capacidad de intervenir o revertir decisiones automatizadas cuando se considere necesario.
- Se debe garantizar que las salidas del sistema no induzcan a la manipulación o toma de decisiones perjudiciales para el usuario.
- El sistema debe contar con planes de mantenimiento y actualización continua, que contemplen la aplicación periódica de parches de seguridad y mejoras funcionales.
- Se debe implementar el monitoreo constante de desempeño, riesgos y anomalías, y la capacidad de respuesta inmediata ante incidentes de seguridad o privacidad de la información.
- Se debe guardar y conservar el registro de actividades (logs) con fines de trazabilidad, de investigación y de auditoría.

12.4.1.5. Monitoreo y auditoría

- Se deberán designar los responsables en cada una de las fases del ciclo de vida del sistema.
- Se deberá elaborar reportes periódicos sobre riesgos, rendimiento y acciones correctivas para asegurar la transparencia hacia usuarios y las autoridades competentes.
- Se deberán realizar auditorías mínimo una vez al año, para revisar el código del sistema de Inteligencia Artificial incluido conjuntos de datos, modelos y decisiones automatizadas.
- Se deberán implementar indicadores clave (KPI) de desempeño y riesgo, revisados de manera periódica para ajustar la operación del sistema.
- Se deben registrar y conservar la documentación de todos los incidentes de seguridad asociados para posterior análisis y mejora del sistema.

12.4.1.6. Retiro o desactivación

- Antes de retirar un sistema de IA deberá comunicarse a los usuarios, indicando motivos, impactos y alternativas.

- Antes del retiro definitivo del sistema de IA, se deberán desactivar de forma controlada todas las interfaces, APIs y accesos asociados, evitando riesgos de seguridad.
- Se debe asegurar la eliminación de los datos personales almacenados de manera que sea segura o que permita la anonimización irreversible, en cumplimiento de lo definido en la Ley 1581 de 2012 y normas complementarias.
- Se debe documentar el proceso de retiro del sistema de IA, incluyendo la migración de procesos críticos a sistemas alternativos para evitar afectaciones a usuarios o servicios internos.

12.4.2. Capacidades técnicas

Hardware:

- Utilizar CPUs, NPUs y GPUs potentes para acelerar el procesamiento de datos y el entrenamiento de modelos.
- Implementar soluciones de almacenamiento rápido y escalable como SSDs y sistemas de almacenamiento en la nube para gestionar grandes volúmenes de datos.
- Asegurar una infraestructura de red robusta para facilitar la transferencia rápida de datos.

Software:

- Emplear plataformas tradicionales (tales como TensorFlow, PyTorch), e integrar marcos modernos para LLM y RAG (tales como LangChain, Hugging Face y bases de datos vectoriales), garantizando un desarrollo seguro.
- Implementar software de seguridad para proteger los datos y sistemas de IA contra accesos no autorizados y ataques cibernéticos.
- Emplear plataformas de análisis avanzadas para obtener perspectivas valiosas y evaluar el rendimiento de los modelos.
- Definir procedimientos claros para la recopilación, procesamiento, y almacenamiento de datos de manera segura.
- Asegurar que las soluciones de IA se integren, de manera fluida, con los sistemas existentes, siguiendo los lineamientos del Marco de Interoperabilidad de MinTIC.
- Desarrollar planes de contingencia para garantizar la operación continua de las soluciones de IA.

12.4.3. Capacidades humanas

- Contratar y formar a empleados con habilidades en ciencia de datos, ingeniería de software, y ciberseguridad.
- Ofrecer programas de formación para actualizar constantemente las habilidades del personal en nuevas tecnologías y metodologías de IA.
- Educar al personal sobre principios éticos en el diseño y uso de IA.

12.4.4. Capacidades administrativas

- Designar líderes responsables de la dirección estratégica y el éxito de las iniciativas de IA.

- Establecer políticas claras para la gobernanza, cumplimiento normativo y gestión de riesgos asociados con la IA.
- Realizar evaluaciones de impacto regulares para medir el retorno de inversión y la contribución al cumplimiento de objetivos organizacionales.

13. Bibliografía.

Access Now. (2024). Radiografía normativa: ¿Dónde, qué y cómo se está regulando la Inteligencia Artificial en América Latina? https://www.accessnow.org/wpcontent/uploads/2024/02/LAC-Reporte-regional-de-politicas-de-regulacion-a-laIA.pdf

European Commission. (2025). AI Act. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Fiil-Flynn, S., Butler, B., Carroll, M., Cohen-Sasson, O., Craig, C., Guibault, L., et al. (2022). Legal reform to enhance global text and data mining research. Science, 378(6623), 951–953.

Future of Life Institute. (2025). High-level summary of the AI Act. https://artificialintelligenceact.eu/high-level-summary/

International Organization for Standardization & International Electrotechnical Commission. (2022). ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology.

International Organization for Standardization & International Electrotechnical Commission. (2023). ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9475.

Ministerio de Ciencia, Tecnología e Innovación. (2023). Hoja de ruta para la adopción ética y sostenible de la Inteligencia Artificial en Colombia. https://minciencias.gov.co/sites/default/files/upload/noticias/hoja_de_ruta_adopcion_etica_y_sostenible_de_inteligencia_artificial_colombia_0.pdf

Ministerio de Tecnologías de la Información y las Comunicaciones. (2025a). Marco de Interoperabilidad. Portal de Gobierno Digital. https://gobiernodigital.mintic.gov.co/portal/Iniciativas/Marco-de-Interoperabilidad/

Ministerio de Tecnologías de la Información y las Comunicaciones. (2025b). Modelo de Seguridad y Privacidad de la Información.

Ministerio de Tecnologías de la Información y las Comunicaciones. (2025c). Uso responsable de la inteligencia artificial. Guía Ética para la Implementación, Desarrollo y Uso de Sistemas de Inteligencia Artificial en Entidades Públicas de Colombia. Gobierno de Colombia. https://www.mintic.gov.co/portal/715/articles-425888_recurso_1.pdf

MITRE. (2024). MITRE ATLAS™ (Adversarial Threat Landscape for AI Systems). https://atlas.mitre.org/

National Institute of Standards and Technology (NIST). (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce.

National Institute of Standards and Technology (NIST). (2025). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (NIST AI 100-2e2025). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-2e2025

National Institute of Standards and Technology (NIST). (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1

Organisation for Economic Co-operation and Development (OECD). (2019). Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL0449

Organisation for Economic Co-operation and Development (OECD). (2024a). Facts not fakes: Tackling disinformation, strengthening information integrity. OECD Publishing. https://www.oecd.org/es/publications/hechos-frente-a-falsedades-fortaleciendo-la-democracia-a-traves-de-la-integridad-de-la-informacion_06f8ca41-es.html

Organisation for Economic Co-operation and Development (OECD). (2024b). Global Partnership on Artificial Intelligence. https://www.oecd.org/en/about/programmes/global-partnership-on-artificial-intelligence.html

Organisation for Economic Co-operation and Development (OECD) & Banco de Desarrollo de América Latina (CAF). (2022). Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina y el Caribe. OECD Publishing. https://www.oecd.org/es/publications/2022/03/the-strategic-and-responsible-use-of-artificial-intelligence-in-the-public-sector-of-latin-america-and-the-caribbean_17c90e5e.html

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). (2019). Estudio preliminar sobre un posible instrumento normativo relativo a la ética de la Inteligencia Artificial. https://unesdoc.unesco.org/ark:/48223/pf0000369455_spa

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). (2021). Recomendación sobre la ética de la Inteligencia Artificial. https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). (2023). UNESCO's recommendation on the ethics of Artificial Intelligence: Key facts.

https://www.unesco.org/en/articles/unescos-recommendation-ethics-artificialintelligence-key-facts

OWASP. (2025). The OWASP Top 10. https://owasp.org/Top10/2025/

OWASP. (2025). OWASP Gen AI Security Project. Recuperado el 16 de abril de 2026, de https://genai.owasp.org/resource/top-10-2025-de-riesgos-y-mitigaciones-para-llms-y-aplicaciones-de-ia-generativa/

Parlamento Europeo. (2021). Resolución del Parlamento Europeo, de 6 de octubre de 2021, sobre la Inteligencia Artificial en el Derecho penal y su utilización por las autoridades policiales y judiciales en asuntos penales (2020/2016(INI)). https://www.europarl.europa.eu/doceo/document/TA-9-2021-0405_ES.html

Parlamento Europeo. (2022). Resolución del Parlamento Europeo, de 3 de mayo de 2022, sobre la Inteligencia Artificial en la era digital (2020/2266(INI)) (P9_TA(2022)0140). https://www.europarl.europa.eu/doceo/document/TA-9-2022-0140_ES.pdf

UNICEF. (2021). Policy guidance on AI for children (Version 2.0). https://www.unicef.org/innocenti/media/1341/file/UNICEF-Global-Insight-policy-guidance-AI-children-2.0-2021.pdf