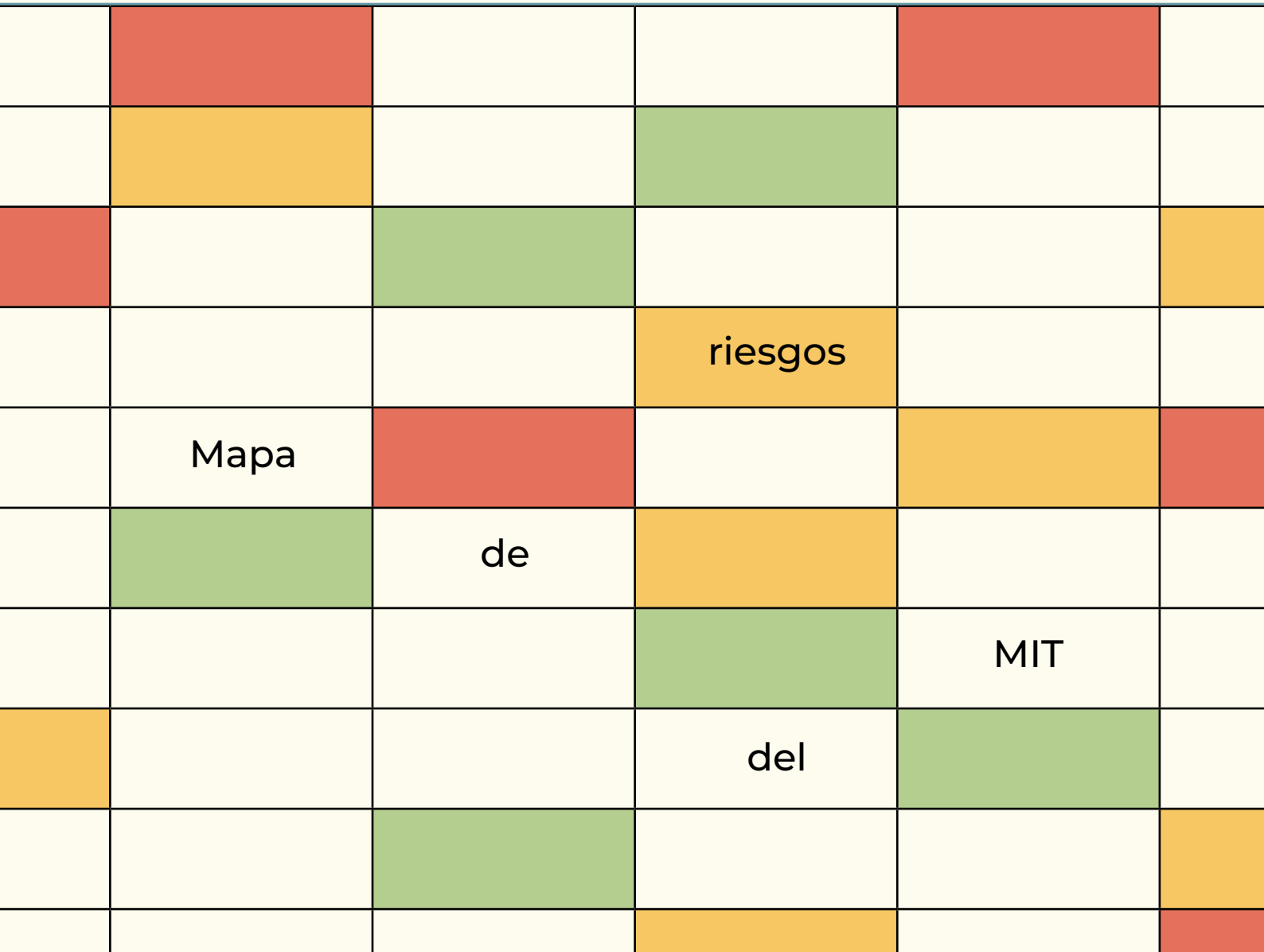
isms forum | INTERNATIONAL INFORMATION SECURITY COMMUNITY

GIA | GRUPO DE INTELIGENCIA ARTIFICIAL

Mapa de riesgos del MIT



AUTORES

COORDINACIÓN EDITORIAL

Esther Álvarez
Manuel Asenjo

REVISIÓN TÉCNICA

Esther Álvarez
Manuel Asenjo
Yan Bello
Andoni Valverde Villar

COAUTORÍA

Alessandro Pignati
Amelia Torres Medrano
Andoni Valverde Villar
Andréia Cristina Castellan
Ángela Rodríguez López
Anna Fite
Antonio Camino Salvo
Carlos Martínez Wahnon
Carmen Moreno Ventas
Carolina Hidalgo Paz
David Carrasco Campos
Eduardo Argüeso López
Francisco Javier Carbayo Vázquez
Gisela Reverter Domínguez
Ignacio Hornes Amenedo
Iván Perez
Jose Alberto Ribes Curto
Lily Estrada Ploegmakers
Maria Luisa González Tapia
Maria Ramirez Pretel
Miguel García Cordo
Ramón Baradat
Roger Sanz Gonzalez
Yan Bello Méndez

EDICIÓN

Beatriz García

COORDINACIÓN DEL PROYECTO

Susana Marín

DISEÑO Y MAQUETACIÓN

Susana Marín

CONTENIDO

1.

Introducción: ¿Qué es el repositorio o Matriz de Riesgo de IA del MIT?

1.1.

CONTEXTO GENERAL

En los últimos años, la inteligencia artificial (en adelante IA) se ha convertido en un pilar fundamental de la transformación digital a nivel global. Hoy en día encontramos aplicaciones de IA en prácticamente todos los sectores económicos: desde la industria manufacturera y la banca hasta la sanidad, la agricultura, los transportes o la administración pública. Su adopción en las organizaciones está creciendo exponencialmente. De hecho, España se ha posicionado como uno de los líderes en la adopción de la IA en la Unión Europea, con un 50% de las empresas españolas utilizando ya esta tecnología (ocho puntos por encima de la media europea). Este auge no solo se observa en grandes corporaciones; también startups y pymes están incorporando IA a sus procesos y productos.

La IA se ha consolidado como un motor clave de transformación en todos los sectores económicos. Su adopción permite a las organizaciones mejorar la eficiencia operativa mediante la automatización de tareas, optimizar la toma de decisiones mediante análisis avanzado de datos, e impulsar la innovación con nuevos productos y servicios personalizados. Además, la IA ofrece un gran potencial para abordar retos sociales y medioambientales; aplicaciones en sanidad, educación o sostenibilidad están demostrando su capacidad para mejorar la calidad de vida y apoyar la transición ecológica. En conjunto, la IA representa una oportunidad estratégica para modernizar procesos, generar valor y avanzar hacia una economía más innovadora, sostenible y centrada en las personas.

Junto con estas oportunidades, la adopción masiva de la IA conlleva riesgos importantes que requieren gestión activa. A medida que los sistemas inteligentes ganan influencia en decisiones y operaciones, han emergido preocupaciones en torno a aspectos éticos, de seguridad y de cumplimiento normativo. Entre los más relevantes destacan:

- **Sesgos y discriminación algorítmica:** la IA puede reproducir prejuicios presentes en los datos, afectando decisiones en ámbitos sensibles como el empleo, la justicia o los servicios públicos.
- **Falta de transparencia:** muchos sistemas de IA operan como “cajas negras”, dificultando la comprensión de sus decisiones. Esto compromete la rendición de cuentas y la confianza, especialmente en sectores críticos.
- **Privacidad y derechos fundamentales:** el uso masivo de datos personales plantea desafíos para el cumplimiento del RGPD. Además, ciertas aplicaciones, como la vigilancia biométrica, pueden vulnerar derechos fundamentales, motivo por el cual el Reglamento de Inteligencia Artificial (en adelante AI Act) prohíbe usos considerados inaceptables.
- **Ciberseguridad:** la IA también puede ser vulnerable a ataques o ser utilizada con fines maliciosos.
- **Impacto socioeconómico y ambiental:** la automatización puede afectar al empleo y aumentar la desigualdad si no se acompaña de políticas de formación. Además, el alto consumo energético de algunos modelos de IA plantea retos de sostenibilidad, abordados por iniciativas como el Plan Nacional de Algoritmos Verdes.

En este contexto, la gestión de riesgos se convierte en un pilar esencial para garantizar una adopción responsable de la IA, alineada con los valores y normativas europeas. Tanto la Unión Europea como España han desarrollado un sólido marco estratégico y regulatorio para fomentar una IA ética, segura y confiable, al tiempo que se promueve su adopción responsable.

En el ámbito europeo, el **AI Act** introduce un enfoque pionero basado en niveles de riesgo, estableciendo obligaciones proporcionales para los sistemas de IA según su impacto potencial. Este reglamento prohíbe usos inaceptables (como la vigilancia biométrica masiva) y exige transparencia, supervisión humana y evaluaciones de impacto para sistemas de alto riesgo.

Además, otras normativas como el **Reglamento General de Protección de Datos (**en adelante **RGPD), la Directiva NIS2 sobre ciberseguridad y el Esquema Nacional de Seguridad** (en adelante **ENS)** en España, refuerzan la necesidad de gestionar adecuadamente los riesgos asociados a la IA, desde la protección de datos hasta la resiliencia frente a ciberataques.

Este ecosistema normativo exige a las organizaciones adoptar marcos de control robustos para garantizar el cumplimiento legal y ético en el uso de la IA. En este contexto, herramientas como el Repositorio de Riesgos de IA del MIT se perfilan como aliados clave para facilitar una gestión eficaz de los riesgos y alinear las prácticas tecnológicas con los estándares regulatorios.

1.2.

¿QUÉ ES EL REPOSITORIO DE RIESGOS DE IA DEL MIT?

En este marco de creciente exigencia regulatoria y de responsabilidad en el uso de la IA, resulta imprescindible contar con instrumentos que permitan traducir los principios normativos y las obligaciones legales en una gestión práctica y sistemática de los riesgos asociados a los sistemas de IA.

En un contexto marcado por la rápida expansión de la IA y la creciente presión regulatoria para garantizar su uso ético y seguro, surge la necesidad de contar con herramientas que ayuden a las organizaciones a identificar, clasificar y mitigar los riesgos asociados a esta tecnología.

En este escenario, el Massachusetts Institute of Technology (en adelante MIT), a través de su iniciativa MIT Schwarzman College of Computing, ha desarrollado el Repositorio de Riesgos de la IA (AI Risk Repository), una plataforma pionera que busca sistematizar el conocimiento sobre los riesgos que plantea la IA en distintos contextos de aplicación.

Uno de los principales valores del repositorio radica en su enfoque colaborativo y en la riqueza de sus fuentes. Está alimentado por una amplia variedad de materiales: investigaciones académicas, informes técnicos, estudios de caso, literatura científica, artículos periodísticos y contribuciones de expertos del sector público y privado. Esta diversidad de perspectivas permite capturar tanto riesgos técnicos como implicaciones sociales, legales y éticas, ofreciendo una visión holística y actualizada del panorama de riesgos de la IA.

Este repositorio es la base de datos de riesgos de IA más completa y rigurosa creada hasta la fecha.

Estructuralmente, el repositorio se articula en tres componentes principales. **En primer lugar**, una base de datos de riesgos, que constituye el núcleo operativo y recoge cada riesgo individual junto con su referencia documental y evidencias asociadas. **En segundo lugar**, una taxonomía causal, que clasifica los riesgos según el momento y la forma en que se originan (por ejemplo, durante el diseño, el entrenamiento, el despliegue o el uso del sistema). **En tercer lugar**, una taxonomía por dominios de impacto, organizada en siete grandes dominios y veinticuatro subdominios, que agrupan los riesgos en función del tipo de daño o ámbito afectado (seguridad, privacidad, desinformación, impactos sociales, económicos, etc.). Esta doble clasificación permite analizar los riesgos tanto desde una perspectiva técnica-causal como desde una perspectiva de impacto.

Es fundamental entender que este repositorio no es un documento estático, sino un recurso dinámico que se actualiza periódicamente para integrar nuevas investigaciones y amenazas emergentes. Cifras clave a febrero de 2026: Hoy en día, el repositorio contempla más de 1.700 riesgos individuales clasificados, lo que supone un crecimiento significativo desde su lanzamiento inicial (que contaba con cerca de 700).

1.3.

ENFOQUE Y JUSTIFICACIÓN DE ESTE INFORME

El objetivo de este informe es: por un lado, analizar el Repositorio de Riesgos de IA del MIT y que sirva de base para el desarrollo de una metodología práctica de análisis de riesgos proponiendo criterios para la identificación, evaluación y mitigación de riesgos asociados a la IA, especialmente aquellos que afectan a la seguridad, privacidad y los derechos fundamentales.

Y por otro, el informe busca contextualizar el repositorio en relación con el marco regulatorio vigente en la Unión Europea, con especial énfasis en el **AI Act** y el **RGPD**. También se abordará su alineación con normativas nacionales clave como el **ENS** y la **Directiva NIS2**, que establecen obligaciones específicas en materia de seguridad de la información y resiliencia operativa para entidades públicas y privadas.

El informe adopta un enfoque interdisciplinario. Tiene en cuenta las perspectivas de ingeniería, derecho y gestión de riesgos para ofrecer a las organizaciones una guía práctica. La revisión de la taxonomía causal permite asignar responsabilidades: si un riesgo se origina por una acción humana, la organización debe reforzar la formación y los procesos internos; si es resultado de una acción del propio sistema de IA, se priorizarán medidas de supervisión técnica y alineación de objetivos. El análisis de los dominios de impacto facilita la colaboración entre áreas funcionales, ya que un mismo riesgo puede requerir medidas de ciberseguridad, evaluaciones de protección de datos y acciones para garantizar la equidad. En este contexto, contar con una guía que adapte el repositorio al marco europeo ayuda a las organizaciones a adelantarse a los plazos legales y a integrar las nuevas exigencias en su gestión del riesgo.

Por último, el informe pretende concienciar a la comunidad técnica y a la Dirección de Tecnología sobre la importancia de tener en cuenta la ética y los derechos fundamentales en el desarrollo de la IA. La taxonomía de dominios tiene en cuenta aspectos como el impacto socioeconómico y medioambiental, la interacción persona-ordenador y la seguridad de los sistemas. Al contrastar estas áreas con las normas vigentes, se espera fomentar una cultura de la responsabilidad que sitúe la protección de las personas en el centro de la innovación.

GESTIÓN DE RIESGOS DE IA		
Riesgos Clave	**Normativa Aplicable**	**Utilidad del Repositorio**
Sesgos y Discriminación	AI Act	Identificación y Clasificación
Falta de Transparencia	RGPD	Evaluación de Riesgos
Privacidad y Seguridad	Directiva NIS2	Guía de Mitigación
Ciberseguridad	ENS	Actualización continua
Impacto Socioeconómico		

Figura 1. Marco integrado de riesgos de IA, normativa aplicable y utilidad del Repositorio de Riesgos del MIT

2.

Objetivos del Repositorio MIT

2.1.

CENTRALIZACIÓN DE RIESGOS Y FUENTE ACTUALIZADA DE INFORMACIÓN

Uno de los principales desafíos en la gestión de los riesgos asociados a la IA es la dispersión de la información. En los últimos años han surgido múltiples marcos, guías, taxonomías y listas de riesgos elaboradas por organismos académicos, reguladores y entidades privadas, que abordan el fenómeno desde perspectivas parciales o sectoriales. Esta fragmentación dificulta a las organizaciones la identificación sistemática de amenazas, la comparación entre enfoques y la adopción de medidas coherentes a lo largo del ciclo de vida de los sistemas de IA.

El Repositorio de Riesgos de IA desarrollado por el MIT responde a esta necesidad mediante la centralización de los riesgos identificados en un único marco estructurado y accesible. El repositorio consolida más de mil setecientas entradas procedentes de decenas de taxonomías previas, lo que lo convierte en una fuente de referencia amplia y representativa del estado del arte en materia de riesgos de IA.

Esta centralización permite reducir duplicados, identificar patrones recurrentes y ofrecer una visión global de los riesgos más relevantes, independientemente del sector o del tipo de sistema analizado.

Además de su función agregadora, el repositorio actúa como una fuente de información viva y actualizada. Frente a enfoques estáticos, el repositorio permite incorporar nuevos riesgos, revisar clasificaciones existentes y reflejar tendencias emergentes, como el uso malicioso de modelos generativos, los riesgos asociados a la automatización de decisiones complejas o las implicaciones socioeconómicas de la adopción masiva de IA.

Desde la perspectiva de las organizaciones, poder contar con una fuente centralizada y actualizada de referencia, facilita la integración de la gestión del riesgo de IA en los marcos corporativos de gobernanza, riesgo y cumplimiento. El repositorio proporciona un lenguaje común que puede ser utilizado por perfiles técnicos, responsables de seguridad, equipos jurídicos y áreas de negocio, favoreciendo una comprensión compartida de los riesgos y evitando interpretaciones aisladas. Esta visión común resulta especialmente relevante en entornos regulados, donde es necesario demostrar trazabilidad entre los riesgos identificados, las obligaciones normativas y las medidas de mitigación adoptadas.

Asimismo, la centralización del riesgo contribuye a una gestión más eficiente y proactiva. Al disponer de una base de datos estructurada, las organizaciones pueden priorizar riesgos en función de su impacto y probabilidad, identificar aquellos que son transversales a múltiples sistemas y anticiparse a problemas que aún no se han materializado en su contexto concreto. De este modo, el repositorio no solo apoya el cumplimiento normativo, sino que también refuerza la toma de decisiones informadas y la adopción de un enfoque preventivo en el desarrollo y uso responsable de la IA.

2.2.

CONCIENCIACIÓN Y PREVENCIÓN

Al igual que ocurre en el ámbito de la ciberseguridad, tradicionalmente percibido como abstracto y tecnológicamente complejo para el ciudadano de a pie, la IA presenta un grado de abstracción incluso mayor. Se trata de una tecnología ampliamente utilizada, pero todavía poco comprendida, lo que incrementa la distancia entre su adopción y la percepción real de sus implicaciones, riesgos y responsabilidades.

Por este motivo, la concienciación sobre la IA, su uso, su impacto y sus riesgos asociados se convierte en un elemento fundamental de prevención. Esta concienciación debe abordarse desde etapas tempranas y de forma transversal, abarcando tanto el ámbito educativo como el profesional y social, con el objetivo de fomentar un uso responsable, informado y crítico de estas tecnologías.

El conocimiento de una materia permite situarnos en una posición preventiva. Siguiendo la estrecha analogía con la ciberseguridad, cuanto mayor sea el grado de conocimiento y concienciación sobre la IA, mayor será la capacidad de anticipar, identificar y mitigar los riesgos derivados de su uso indebido, no ético o no controlado.

Desde el punto de vista empresarial, existen diversos enfoques para abordar la concienciación. Uno de los más habituales consiste en la definición de una Política de Gobierno del uso de la IA, que establezca de manera clara los principios, límites y responsabilidades asociados a su utilización, así como los requisitos formativos aplicables a cada perfil profesional.

En términos generales, las organizaciones suelen distinguir entre dos grandes grupos de empleados: por un lado, el empleado no técnico o usuario general, y por otro, el empleado tecnológico, que incluye perfiles como científicos de datos, desarrolladores, arquitectos de software o administradores de sistemas. En este contexto, resulta habitual diseñar programas de formación diferenciados, con contenidos más generalistas orientados a la concienciación y el uso responsable, y otros de carácter más técnico enfocados en el diseño, desarrollo y operación de sistemas de IA.

No obstante, dada la rápida evolución y novedad de esta tecnología, los programas de concienciación y formación deben ser revisados y actualizados de manera periódica, con el fin de garantizar su pertinencia, eficacia y alineación con los cambios tecnológicos, regulatorios y éticos.

Además de la formación formal, la concienciación en IA debe entenderse como un proceso continuo orientado a la creación de una cultura organizativa responsable. Al igual que en ciberseguridad, no basta con sesiones formativas aisladas; es necesario reforzar de manera recurrente los mensajes clave, promover buenas prácticas y mantener presente el impacto que el uso de la IA puede tener en las personas, la organización y la sociedad.

En este sentido, la concienciación actúa como un control preventivo fundamental dentro del mapa de riesgos asociado a la IA. Un empleado concienciado es capaz de identificar usos inadecuados, cuestionar resultados automatizados, detectar posibles sesgos o vulneraciones de privacidad y, en última instancia, reducir la probabilidad de materialización de riesgos legales, éticos y reputacionales.

Por último, resulta clave reforzar la idea de que el uso de sistemas de IA no exime de responsabilidad individual. La concienciación debe fomentar el pensamiento crítico y la supervisión humana, recordando que la toma de decisiones finales y sus consecuencias siguen recayendo en las personas, incluso cuando se apoyan en sistemas automatizados.

En este contexto, marcos de referencia como el Repositorio de Riesgos de IA del MIT aportan un valor significativo a las iniciativas de concienciación y prevención en el uso de la IA. Al proporcionar un catálogo estructurado y vivo de riesgos reales y potenciales asociados a sistemas de IA, permite traducir conceptos abstractos en ejemplos concretos, facilitando la comprensión de cómo, por qué, y en qué contextos pueden materializarse dichos riesgos. Integrar este tipo de marcos en los programas de concienciación ayuda a reforzar una visión preventiva, basada en la identificación temprana de impactos éticos, legales, operativos y sociales, y contribuye a fomentar una cultura organizativa más crítica, informada y responsable en el uso de la IA.

2.3.

CENTRALIZACIÓN DE RIESGOS Y FUENTE ACTUALIZADA DE INFORMACIÓN

En el contexto actual, marcado por la adopción masiva de IA y la irrupción de modelos fundacionales y sistemas generativos, la gobernanza de la IA ya no es una opción, sino una necesidad estratégica, legal y operativa. Establecer estructuras de gobernanza sólidas es esencial para preservar la confianza de los grupos de interés, generar valor sostenible, proteger los derechos fundamentales y garantizar la resiliencia organizativa frente a riesgos que escalan con la complejidad y autonomía de los sistemas de IA.

El AI Act establece un enfoque basado en el riesgo, que impone obligaciones de gestión continua, trazabilidad, transparencia y supervisión humana, especialmente en sistemas de alto riesgo. Estas exigencias se traducen en la práctica en la necesidad de definir políticas internas, asignar roles y responsabilidades claras, mantener inventarios actualizados de casos de uso, registrar logs y documentación técnica, y aplicar métricas de evaluación como el sesgo, la robustez o la explicabilidad.

En este marco, el repositorio se presenta como una herramienta de gran valor para apoyar la gobernanza efectiva de la IA. A través de su estructura sistemática y su enfoque multidisciplinar, el repositorio permite identificar riesgos en todas las fases del ciclo de vida de los sistemas de IA, desde el diseño hasta el despliegue y la operación.

A modo ilustrativo, y sin ánimo de exhaustividad, el repositorio permite identificar riesgos como los siguientes:

- **Sesgo en los datos:** un sistema de IA puede aprender de datos que contienen prejuicios. Esto provoca decisiones injustas, como rechazar solicitudes de empleo por razones no relacionadas con la capacidad del candidato.
- **Falta de transparencia:** muchos modelos son complejos y no explican cómo toman decisiones. Esto genera desconfianza y dificulta la corrección de errores.
- **Uso malintencionado:** un sistema diseñado para ayudar puede ser usado para fines dañinos, como crear desinformación o ataques cibernéticos.

Asimismo, el repositorio también propone controles para reducir estos riesgos. Entre ellos se encuentran:

- **Gobernanza y supervisión:** creación de comités de IA responsables de revisar los casos de uso, establecer políticas y garantizar el cumplimiento normativo.
- **Controles técnicos:** aplicación de técnicas para detectar y corregir sesgos en los datos, evaluar la robustez de los modelos y asegurar su rendimiento ético.
- **Transparencia y explicabilidad:** documentación clara del funcionamiento de los sistemas, sus limitaciones y criterios de decisión, facilitando la rendición de cuentas.
- **Monitoreo continuo:** supervisión activa del comportamiento del sistema tras su implementación, con mecanismos de alerta temprana ante desviaciones o fallos.
- **Protocolos de respuesta:** definición de planes de contingencia y gestión de incidentes, incluyendo la notificación a autoridades competentes y la comunicación con los afectados.

La gobernanza no es un paso único, es un proceso que evoluciona con la tecnología. El repositorio ayuda a que las organizaciones tengan una guía clara para identificar riesgos y aplicar medidas concretas. De esta manera, la IA se utiliza de manera segura y responsable.

En síntesis, gobernar la IA implica pasar de principios abstractos a controles verificables que alineen negocio, ética y cumplimiento, para capturar valor sin sacrificar seguridad, equidad ni legitimidad social.

2.4.

FACILITAR LA GESTIÓN Y MITIGACIÓN DE RIESGOS

Para controlar los riesgos de la IA, no basta con identificarlos. Uno de los mayores retos que enfrentan las organizaciones es traducir el análisis teórico del riesgo en acciones prácticas para mitigarlo, monitorizarlo y revisarlo continuamente. En esta línea, el repositorio es una gran ayuda para dar ese paso de la identificación del riesgo a su gestión estructurada.

Al ofrecer un amplio y estructurado conjunto de más de 1.700 riesgos de IA, el repositorio proporciona a las organizaciones una base compartida sobre la que desarrollar un inventario consistente de riesgos. Esta metodología disminuye la dependencia de los análisis aislados o parciales y permite la trazabilidad entre los riesgos identificados, las decisiones de tratamiento y los controles implementados. De esta manera, el riesgo de IA se puede incorporar más fácilmente en los marcos generales de gestión de riesgos existentes en las organizaciones.

Desde el punto de vista de la evaluación y priorización del riesgo, el repositorio también proporciona un marco para determinar qué riesgos son más críticos en términos de impacto potencial, en qué etapas del ciclo de vida del sistema pueden ocurrir y en qué medida impactan transversalmente en los dominios. Esta información puede incorporarse directamente en las metodologías de evaluación de riesgos existentes (matrices de riesgo impacto/probabilidad, evaluaciones de impacto específicas para sistemas de alto riesgo o análisis de riesgos corporativos). Así, los riesgos de IA dejan de ser anomalías y se integran en el marco general de gestión de riesgos de la organización.

Otro punto importante es poder asociar directamente los riesgos identificados con las medidas de mitigación ofrecidas. El análisis de la forma en que el riesgo se manifiesta permite adaptar los controles apropiados a su causa y al momento en que se materializa. De este modo, los riesgos humanos pueden mitigarse reforzando procesos, capacitando o implementando controles organizativos, en tanto que los riesgos sistémicos demandan controles técnicos, monitorización del modelo y validaciones continuas. Por otro lado, los riesgos que surgen después de la implementación del sistema requieren supervisión operativa, auditoría continua y la provisión de mecanismos de intervención humana. Este enfoque minimiza el riesgo de sobrecargar controles innecesarios y dejar riesgos críticos sin tratamiento.

El repositorio sirve además para atenuar en el tiempo el riesgo y realizar su seguimiento. Los riesgos de la IA no son fijos y cambian a medida que cambian los datos, los modelos, los contextos de uso o surgen nuevos riesgos. Usar el repositorio como referencia permite hacer revisiones periódicas del riesgo, actualizar los controles en uso y reevaluar las decisiones de aceptación del riesgo. Esta metodología se ajusta a la mejora continua que proponen estándares internacionales de gestión, como la familia ISO.

Desde una perspectiva de gobernanza y auditoría, el repositorio ayuda a que la gestión de riesgos sea registrable, auditable y verificable. Los riesgos identificados se pueden relacionar con controles técnicos y administrativos específicos, evidencia de uso y métricas de seguimiento para determinar su efectividad. Esto simplifica las auditorías internas y externas y fortalece la capacidad de la organización para probar la diligencia debida en el uso de sistemas de IA.

En suma, el repositorio disminuye la incertidumbre sobre los riesgos de la IA, proporciona criterios para priorizar acciones y permite implementar medidas de mitigación uniformes y proporcionales. Gracias a ello, es una palanca para incorporar la IA en los marcos establecidos de gestión de riesgos y pasar de la noción a la práctica en la gestión de riesgos, de forma sostenible y en línea con el buen gobierno.

2.5.

CUMPLIMIENTO REGULATORIO Y NORMATIVO

Tal y como se ha expuesto en el capítulo anterior, la IA se encuentra sujeta a un ecosistema normativo fragmentado y transversal, lo que introduce una elevada complejidad en su cumplimiento. Este contexto exige a las organizaciones adoptar enfoques integrados que permitan interpretar y aplicar de forma coherente los distintos marcos regulatorios que inciden sobre el uso de sistemas de IA.

En este contexto, el cumplimiento parcial puede generar una falsa sensación de seguridad regulatoria. Cumplir con una norma concreta -por ejemplo, en materia de protección de datos- no garantiza el cumplimiento del resto de obligaciones aplicables al uso de la IA. Esta percepción errónea puede llevar a asumir que un sistema es “conforme” desde el punto de vista regulatorio, cuando en realidad presenta riesgos relevantes en otros ámbitos, como la discriminación, la transparencia, la explicabilidad o la rendición de cuentas.

Desde una perspectiva de gestión del riesgo, esta falsa sensación de cumplimiento resulta especialmente crítica, ya que puede retrasar la identificación de incumplimientos, aumentar el impacto de sanciones o generar daños reputacionales significativos. Además, la complejidad normativa se ve amplificada cuando se incorporan soluciones de terceros o se utilizan modelos y plataformas externas, lo que puede diluir la percepción de responsabilidad dentro de la organización.

La fragmentación normativa convierte, por tanto, el cumplimiento regulatorio en un riesgo transversal que debe gestionarse de forma coordinada. No se trata únicamente de un reto legal, sino de un desafío organizativo que afecta al gobierno de la IA en su conjunto. La mitigación de este riesgo requiere un enfoque integrado que combine conocimiento regulatorio, gestión del riesgo y control técnico, asegurando que los requisitos normativos se incorporen de manera consistente a lo largo de todo el ciclo de vida de los sistemas de IA.

En este sentido, resulta fundamental adoptar modelos de gobierno que permitan coordinar a las distintas áreas implicadas (legal, cumplimiento, tecnología, negocio y seguridad, entre otras) y establecer mecanismos de supervisión continua, trazabilidad y generación de evidencias. Solo a través de una gestión coordinada y transversal del cumplimiento regulatorio es posible reducir de forma efectiva los riesgos derivados de la fragmentación normativa y garantizar un uso responsable y sostenible de la IA.

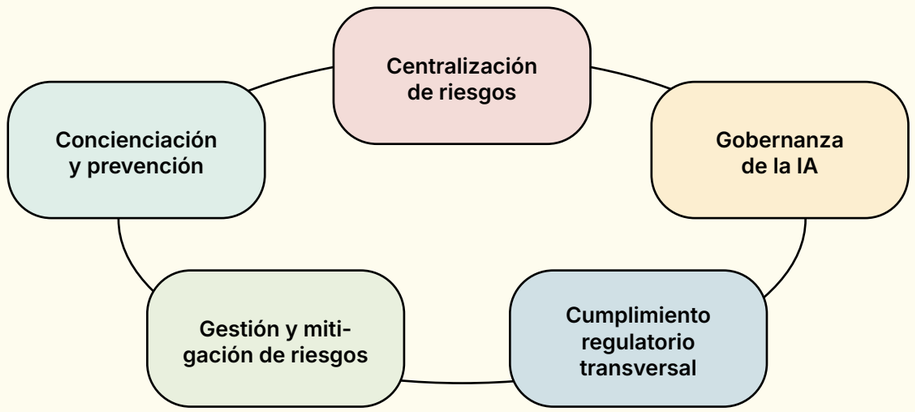


Figura 2. Objetivos del Repositorio de Riesgos de IA del MIT en la gestión organizativa.

3.

Clasificación de los riesgos por taxonomía causal

(cómo, cuándo y por qué surge cada riesgo)

3.1.

OBJETIVO DE LA CLASIFICACIÓN POR TAXONOMÍA CAUSAL

La taxonomía causal es, junto a la taxonomía por dominio, uno de los dos ejes principales del repositorio MIT y tiene como finalidad clasificar los riesgos en tres dimensiones:

Dimensión Causal	Categoría	Descripción/Objetivo
Entidad: ¿Quién o qué lo provoca?	**IA**	Riesgos causados por decisiones o acciones del propio sistema de IA (por ejemplo, un modelo que toma una acción dañina o se comporta de una forma inesperada)
	Humano	Riesgos causados por decisiones o acciones humanas, ya sea por error de desarrolladores/usuarios o por actos deliberados de actores maliciosos.
	Otro	Riesgos causados por factores externos o de origen ambiguo no atribuibles claramente a la IA ni a acciones humanas
Intencionalidad: ¿Por qué ocurre?	**Intencional**	Riesgos que ocurren por acciones deliberadas con un fin previsto, es decir, cuando alguien (humano o IA) busca activamente un resultado que deriva en riesgo.
	No intencional	Riesgos que ocurren por resultados no previstos o accidentales, es decir, no hubo intención deliberada de causar ese riesgo al perseguir un objetivo
	Otro	Riesgos en los que no se ha determinado claramente si media una intencionalidad deliberada o no
Momento: ¿Cuándo se origina?	**Previo al despliegue**	Riesgos que surgen antes de desplegar el sistema de IA, por ejemplo, problemas en la fase de diseño, desarrollo o entrenamiento del modelo
	Posterior al despliegue	Riesgos que surgen tras la puesta en marcha del sistema de IA, una vez entrenado y operando en el mundo real
	Otro	Riesgos cuya etapa de ocurrencia no está clara o abarca múltiples fases (antes y después del despliegue)

Figura 3. Taxonomía causal de riesgos de IA.

Esta clasificación, conjuntamente con la de taxonomía por dominio, permite mapear riesgos de manera sistemática, identificar patrones y puntos críticos en el ciclo de vida, y facilitar la alineación con marcos regulatorios como el AI Act o el RGPD. Además, ayuda a priorizar mitigaciones, diseñar auditorías y construir gobernanza basada en evidencias, evitando la confusión generada por terminologías y metodologías dispares, convirtiendo esa comprensión en acciones concretas de gobernanza y cumplimiento.

A diferencia de enfoques puramente descriptivos, la taxonomía causal del MIT permite entender no solo qué riesgos existen, sino por qué ocurren, quién los causa y cuándo aparecen, ofreciendo un mapa tridimensional y una hoja de ruta que facilita la anticipación y gestión de riesgos emergentes.

3.2.

CATEGORÍAS CAUSALES UTILIZADAS

Llegados a este punto y con el objetivo de permitir un mismo lenguaje entre todos, se ideó un formato para clasificar y entender los posibles peligros con respecto al comportamiento de la IA y sus riesgos, gracias al desarrollo de una estructura directa, sencilla y clara de categorías que nos lleven a las causas, como si se tratara de las preguntas de un detective para descubrir qué está pasando en una determinada situación, empezando por tres preguntas sencillas: quién lo provoca (la entidad), por qué ocurre (el propósito) y cuándo se activa (el momento).

¿Quién o qué provoca el riesgo (la entidad)?

¿Ha sido una decisión humana, ha sido la IA o habrá sido por otros factores? Lo analizamos:

1) Decisión humana: si pensamos en la primera suposición, una persona podría realizar el diseño de un algoritmo con el objetivo de pilotar drones para distribuir cargas explosivas o, aún, entidades criminales que crean IA para cometer actos terroristas.

2) Ha sido la IA: un ejemplo claro de riesgo provocado por decisiones de la propia IA (sin intervención humana directa) podría ser un vehículo autónomo que no detecta a un peatón cruzando la calle y lo atropella. En 2018 ocurrió un caso así en Arizona, donde un coche autónomo en piloto automático golpeó mortalmente a una persona al no reconocerla a tiempo.

Otro ejemplo sería un sistema automatizado de reconocimiento facial de la policía que identifica mal a una persona y confunde a una persona inocente con un delincuente, generando un arresto injusto. Este tipo de errores ya han ocurrido en la vida real; la tecnología ha llevado erróneamente a prisión a varias personas inocentes al fallar en la identificación, sin que hubiera mala intención humana en el proceso.

3) Otros factores: aquí, se podría considerar una mezcla de elementos, diversas investigaciones han demostrado que un coche sin conductor puede malinterpretar una señal de tráfico debido a algo tan simple como unos grafitis o pegatinas. En un experimento, un letrero STOP con pintadas fue "visto" por el vehículo autónomo como si fuera una señal de 45Km/h, lo que podría hacerle no detenerse y provocar un accidente – todo causado por la alteración externa de la señal, no por una decisión humana o del software del coche.

Otro ejemplo podría ser una cámara de vigilancia con IA puede dejar de detectar a un intruso porque su lente está cubierta de polvo o suciedad. En este caso, la IA no "ve" al sospechoso no por un fallo interno ni por culpa de un operador, sino por un factor externo (la lente sucia) que le impide analizar correctamente la imagen, creando un vacío de seguridad inesperado.

¿Por qué ocurre (la intencionalidad)?

Para comenzar el análisis, conviene preguntarse: ¿el daño era realmente el objetivo o, por el contrario, se produjo de manera accidental? ¿O estamos ante un escenario en el que resulta imposible determinar si hubo intención?

Si observamos los ejemplos mencionados anteriormente, en el caso de una decisión humana deliberada, la intencionalidad es evidente: el uso de drones para distribuir cargas explosivas y cometer actos terroristas es una acción planificada y orientada a causar daño.

En contraste, si seguimos con el ámbito de la IA, es posible interpretar como un accidente la generación amplificada de resultados erróneos cuando no existe supervisión humana para auditar y corregir el proceso. Aquí no hay voluntad de causar un perjuicio, sino un fallo atribuible a la ausencia de control.

Otro tipo de accidente puede estar asociado a causas puramente técnicas. La mayoría de los LLM se desarrollan en Python, y las vulnerabilidades de sus intérpretes[1] representan un riesgo tecnológico inherente. En este punto no existe intencionalidad: las vulnerabilidades son parte del ecosistema técnico. Sin embargo, si un atacante explota estas vulnerabilidades aportando información falsa o maliciosa, podría considerarse que sí existe intención, especialmente si el LLM termina generando alucinaciones derivadas de esos insumos manipulados.

Si ampliamos el análisis, también forman parte de los riesgos no intencionales aspectos como la presencia de datos inútiles, una infraestructura tecnológica obsoleta o la falta de formación de los empleados. Estos elementos incrementan la probabilidad de errores o fallos, pero sin que exista una premeditación detrás.

Finalmente, aparece la categoría más compleja: la ambigüedad. Podría ocurrir que un modelo LLM haya sido entrenado con información sesgada o tóxica. Esto puede dar lugar a contenido ofensivo, discriminatorio o incluso ilegal. En estos casos surge la duda central: ¿quién seleccionó los datos tenía la intención de sesgar el modelo o simplemente no evaluó adecuadamente la calidad del conjunto de entrenamiento? Esta imposibilidad de determinar la intención es lo que caracteriza a esta tercera categoría causal.

[1] Un intérprete de Python es una herramienta que ejecuta código en tiempo real, línea por línea, sin necesidad de compilarlo previamente.

¿Cuándo se origina (el momento)?

En este punto, hay tres momentos que considerar: ¿el riesgo se origina antes del despliegue de la IA, aparece después del despliegue o, más bien, se identifica que existe el riesgo, pero el momento de ocurrencia es incierto? Lo analizamos:

1. Riesgos originados antes de la implementación: estos riesgos surgen durante el diseño, el desarrollo, el entrenamiento o la preparación de la infraestructura. Algunos ejemplos son:

- **Vulnerabilidades de hardware:** los LLM requieren entrenamiento intensivo en GPU, las cuales pueden sufrir ataques de canal lateral con el objetivo de extraer parámetros del modelo. Este riesgo aparece incluso antes de que el sistema sea desplegado.
- **Ataques de envenenamiento de datos:** pequeñas modificaciones en los datos de entrenamiento pueden alterar significativamente el comportamiento del modelo. Este tipo de ataque puede incluir la inserción de backdoors o desencadenantes ocultos durante el entrenamiento.
- **Ataques de evasión:** se producen mediante perturbaciones en las muestras de prueba que inducen predicciones erróneas (por ejemplo, cambios en palabras o gradientes). Aunque su impacto puede manifestarse en la operación, su preparación puede darse en fases previas.
- **Vulnerabilidades en marcos de aprendizaje profundo:** los LLM dependen de frameworks que pueden contener fallos como desbordamientos de búfer, corrupción de memoria o problemas de validación. Estas debilidades existen antes de desplegar la IA.

2. Riesgos que surgen después de la implementación: se manifiestan una vez que el sistema está en operación, ya sea en laboratorio, en entornos controlados o en producción. Algunos ejemplos son:

- Fugas o liberaciones no previstas: por ejemplo, una IA que no estaba destinada a tener impacto público puede filtrarse desde un laboratorio, o liberarse como software abierto sin controlar su alcance.
- Uso de herramientas externas poco fiables: muchas aplicaciones incorporan resultados de APIs, motores de búsqueda u otros servicios externos sin garantías de calidad. Información errónea amplifica problemas como las alucinaciones del modelo.
- Falta de coordinación entre desarrolladores y usuarios: diferencias en requisitos, competencias técnicas, expectativas y formación ética pueden conducir a usos incorrectos o riesgosos del sistema.

3. Riesgo cuyo momento de aparición es incierto: el riesgo existe, pero no es posible determinar cuándo se activará.

- Sesgos políticos y neutralidad comprometida: algunos modelos pueden expresar opiniones inapropiadas o extremistas al tratar temas sensibles. Aunque se sabe que el riesgo está presente, no se puede predecir cuándo se manifestará en una interacción concreta.
- Dilemas morales en vehículos autónomos: las decisiones en escenarios de accidente, como priorizar la seguridad de distintos actores, representan un riesgo ético latente. El riesgo existe, pero solo se materializa ante situaciones imprevistas y poco frecuentes.

Es importante destacar que según los datos de la última versión de diciembre de 2025 publicada por el MIT AI Risk Repository, el 42% de los riesgos analizados se atribuyen directamente a sistemas de IA, frente al 37% que tienen origen humano. Aunque ambos valores están muy próximos, el peso ligeramente mayor de los riesgos provocados por la propia tecnología muestra que los fallos o comportamientos inesperados de los sistemas representan una parte significativa del total.

Por otro lado, la comparación entre riesgos intencionales y no intencionales indica que la mayoría no derivan de una voluntad explícita de causar daño. De hecho, los riesgos no intencionados representan el 35% de los casos, mientras que los intencionados suponen un 34%. Esta proximidad sugiere que, aunque existe un interés relevante en utilizar la IA con fines perjudiciales, la mayor parte de los incidentes se produce sin premeditación.

Finalmente, el análisis temporal revela un dato clave: el 61% de los riesgos emergen después del entrenamiento y el despliegue del modelo. Es decir, la mayoría de los problemas se manifiestan en la fase operativa, cuando aparecen errores, accidentes, comportamientos no previstos o fallos en la interacción con su entorno real. Este predominio subraya la importancia de supervisar, monitorizar y actualizar los sistemas una vez están en funcionamiento.

Categoría	Nivel	Descripción	Proporción de riesgo en la base de datos de Riesgos de IA
Entidad	**Humano**	El riesgo es causado por una decisión o acción realizada por humanos.	**37%**
	IA	El riesgo es causado por una decisión o acción realizada por un sistema de IA.	**42%**
	Otro	El riesgo es causado por otra razón o es ambiguo.	**21%**
Intencionalidad	**Intencional**	El riesgo ocurre debido a un resultado esperado al perseguir un objetivo.	**34%**
	No intencional	El riesgo ocurre debido a un resultado inesperado al perseguir un objetivo.	**35%**
	Otro	El riesgo se presenta sin especificar claramente la intencionalidad.	**31%**
Momento	**Previo al despliegue**	El riesgo ocurre antes de que la IA sea desplegada.	**12%**
	Posterior al despliegue	El riesgo ocurre después de que el modelo de IA ha sido entrenado y desplegado.	**61%**
	Otro	El riesgo se presenta sin especificar claramente el momento en que ocurre.	**27%**

Figura 4. Proporción de riesgos en la base de datos de riesgos de IA por entidad, intencionalidad y momento.

3.3.

IMPLICACIONES PRÁCTICAS DE ESTA CLASIFICACIÓN

La taxonomía causal del Repositorio del MIT distingue tres dimensiones fundamentales en la aparición de riesgos asociados a la IA: quién los origina, si la consecuencia es intencionada o inesperada, y en qué momento del ciclo de vida se manifiestan. Esta estructura no solo organiza la información, sino que tiene implicaciones prácticas directas para la gestión de riesgos en las organizaciones.

En primer lugar, permite distribuir las responsabilidades de forma clara. Cuando el riesgo se origina en una acción humana -por ejemplo, el uso de datos sesgados para entrenar un algoritmo de selección de personal-los equipos de Recursos Humanos y la Dirección de Tecnología pueden establecer controles previos al despliegue, como revisiones de equidad y pruebas sistemáticas. En cambio, si el riesgo proviene del propio sistema de IA- como, por ejemplo, un modelo que optimiza objetivos que entran en conflicto con los valores corporativos-, se requiere supervisión técnica especializada, mecanismos de alineación y una gobernanza continua del comportamiento del modelo. La distinción entre riesgos intencionales y accidentales también orienta las medidas de mitigación: los usos maliciosos exigen defensas en profundidad, autenticación reforzada y vigilancia activa; mientras que los fallos no intencionados requieren mejorar la calidad de los datos, los procesos de validación y las pruebas del sistema.

Considerar el momento en que aparece el riesgo facilita también el cumplimiento normativo. El Reglamento de IA obliga a evaluar los sistemas de alto riesgo antes de su comercialización y a lo largo de toda su vida útil. De forma análoga, el RGPD exige evaluaciones de impacto cuando el tratamiento pueda entrañar riesgos elevados, y establece el principio de protección de datos desde el diseño. El ENS, por su parte, demanda una gestión de riesgos continua que incluya prevención, detección y respuesta. La taxonomía causal indica con claridad cuándo deben activarse estas obligaciones. Así, un riesgo de pre-implementación generado por una acción humana no intencionada, como emplear un conjunto de entrenamiento incompleto, se aborda mediante controles de calidad y formación. En cambio, un riesgo pos-implementación derivado de un uso intencional del sistema, como explotar un modelo generativo para cometer fraude, requerirá monitorización continua, autenticación robusta y políticas estrictas de uso aceptable.

Además, la clasificación facilita el diálogo entre personal técnico y responsables de negocio. Al emplear categorías intuitivas, como "riesgo originado por la IA" o "riesgo intencional", se favorece la coordinación entre ciberseguridad, protección de datos, cumplimiento normativo y otras áreas. También proporciona una base estructurada para mantener un registro de riesgos actualizado, lo que permite priorizar medidas y asignar recursos de manera eficaz.

Desde el punto de vista operativo, los ejemplos prácticos evidencian la utilidad de esta taxonomía. En un sistema de análisis radiológico basado en IA, la presencia de sesgos en los datos de entrenamiento constituye un riesgo humano, no intencional y previo a la implementación. Identificarlo así impulsa a diversificar fuentes de datos y realizar auditorías de equidad. En un asistente virtual que interactúa con menores, la posibilidad de generar contenido dañino representa un riesgo causado por la IA, accidental y posterior al despliegue, que se mitiga mediante filtros y revisión humana. Por último, en una red social que utiliza modelos generativos, el uso malicioso por parte de terceros para difundir desinformación se clasifica como un riesgo humano, intencional y pos-implementación. Su gestión implica sistemas automáticos de detección, políticas estrictas y coordinación con autoridades.

En conclusión, la taxonomía causal del Repositorio del MIT proporciona una herramienta sistemática que permite integrar la gestión de riesgos en los procesos organizativos y cumplir con las exigencias regulatorias europeas. Al distinguir el origen, la intención y el momento de aparición del riesgo, facilita anticipar responsabilidades, ajustar controles a cada fase y mejorar la comunicación entre equipos. Combinada con la taxonomía de dominios, ofrece una visión holística que equilibra innovación, seguridad, privacidad y protección de los derechos fundamentales.

4.

Clasificación de los riesgos por taxonomía de **áreas de impacto o dominios**

(según qué ámbito de la sociedad, las personas o los sistemas afecta cada riesgo)

4.1.

OBJETIVO DE LA CLASIFICACIÓN POR TAXONOMÍA DE ÁREAS DE IMPACTO O DOMINIOS

El segundo criterio de clasificación utilizado por el Repositorio de Riesgos de IA del MIT es la denominada taxonomía de dominio ("domain taxonomy"). Cada dominio representa un área de impacto o materia potencialmente afectada por los sistemas de IA. Mientras que la Taxonomía Causal agrupa los riesgos según el agente que los origina, la Taxonomía de dominio etiqueta y categoriza efectos perjudiciales que han sido constatados o referenciados en los artículos y documentos sobre los que se ha creado la base de datos.

El repositorio establece siete dominios principales que agrupan los tipos de impacto más frecuentes. Estos dominios abarcan tanto daños materiales como inmateriales, incluyendo la afectación a derechos fundamentales y la protección de la salud. Son los siguientes:

- Discriminación y toxicidad ("*Discrimination & toxicity*")
- Privacidad y seguridad (*"Privacy & security"*)
- Desinformación (*"Misinformation"*)
- Actores maliciosos y uso indebido (*"Malicious actors & misuse"*)
- Interacción humano-máquina (*"Human-computer interaction"*)
- Daños socioeconómicos y ambientales (*"Socioeconomic & environmental harms"*)
- Seguridad, fallos y limitaciones del sistema (*"AI system safety, failures, & limitations"*)

Cada dominio agrupa un número variable de subdominios, en total 24, que a su vez incluyen los distintos riesgos que componen la matriz, tal y como se muestra en la siguiente tabla:

Ref.	Dominio	Subdominio
1	**Discriminación y toxicidad**	Discriminación injusta y tergiversación.
		Exposición a contenidos tóxicos.
		Desigualdad en el rendimiento entre grupos.
2	**Privacidad y seguridad**	Compromiso de la privacidad mediante la obtención, filtración o deducción correcta de información confidencial.
		Vulnerabilidades y ataques a la seguridad de los sistemas de IA.
3	**Desinformación**	Información falsa o engañosa.
		Contaminación del ecosistema informativo y pérdida de la realidad consensuada.
4	**Actores maliciosos y uso indebido**	Desinformación, vigilancia e influencia a gran escala.
		Ciberataques, desarrollo o uso de armas y daños masivos.
		Fraude, estafas y manipulación selectiva.
5	**Interacción humano-máquina**	Dependencia excesiva y uso inseguro.
		Pérdida de la capacidad de acción y la autonomía humanas.
6	**Socioeconómico y ambiental**	Centralización del poder y distribución injusta de los beneficios.
		Aumento de la desigualdad y deterioro de la calidad del empleo.
		Devaluación económica y cultural del esfuerzo humano.
		Dinámicas competitivas.
		Fracaso de la gobernanza.
		Daño medioambiental.
7	**Seguridad, fallos y limitaciones del sistema**	La IA persigue sus propios objetivos, que entran en conflicto con los objetivos o valores humanos.
		La IA posee capacidades peligrosas.
		Falta de capacidad o solidez.
		Falta de transparencia o interpretabilidad.
		Bienestar y derechos de la IA.
		Riesgos multiagente.

Figura 5. Dominios y subdominios del Repositorio de Riesgos de IA del MIT.

La clasificación por dominios es, probablemente, la más práctica para emplear como referencia en los análisis de riesgo presentados en los capítulos seis, siete, ocho y nueve. No obstante, debe considerarse que los riesgos asociados a cada categoría no afectan por igual a todos los tipos de organizaciones. Por ello, es recomendable adaptar su aplicabilidad en función del ámbito de actuación, sector y naturaleza de la entidad.

En este sentido, pueden distinguirse tres grados de aplicabilidad:

- **La entidad como sujeto paciente del riesgo.** Por ejemplo, en la categoría 2.2, Vulnerabilidades de Seguridad y Ataques sobre Sistemas de IA, la organización puede sufrir violaciones de confidencialidad, disponibilidad o integridad de sus datos o infraestructuras.
- **La entidad como sujeto agente del riesgo.** Un ejemplo lo encontramos en la categoría 1, Discriminación y Toxicidad, donde el uso de un modelo sesgado puede dar lugar a tratos discriminatorios hacia usuarios finales, generando incumplimientos normativos, impactos económicos o perjuicio reputacional.
- **Riesgos que afectan a la sociedad en su conjunto.** En estas categorías, la organización forma parte del ecosistema afectado, pero no tiene control directo sobre su prevención o mitigación. Es el caso de la categoría 6.1, Centralización del Poder y Distribución Injusta de Beneficios, o la categoría 7.5, Bienestar y Derechos de la IA. En los apartados que figuran a continuación se ofrece una visión resumida del contenido de cada uno de los siete dominios.

4.2.

CATEGORÍAS POR ÁREAS DE IMPACTO O DOMINIOS UTILIZADAS

Los siguientes apartados se refieren a los dominios en los que los riesgos asociados a sistemas de IA pueden afectar a individuos, organizaciones o entornos. Cada dominio incluye subdominios, que se analizan en los correspondientes subapartados y que permiten realizar una clasificación más acotada de los riesgos dentro de un mismo entorno. Para mejorar la comprensión, además de la descripción de los dominios y subdominios, se han seleccionado algunos ejemplos de riesgos y situaciones donde pueden materializarse.

1. DISCRIMINACIÓN Y TOXICIDAD:

Este dominio reúne los riesgos relacionados con el trato injusto, la exposición a contenido dañino y el rendimiento desigual de los sistemas de IA entre diferentes grupos e individuos. Incluye los problemas derivados de sesgos, disparidades sistemáticas, representaciones estereotipadas, contenido ofensivo o dañino, así como efectos perjudiciales asociados a la falta de diversidad y representatividad en los datos y en el diseño de modelos. Este dominio se estructura en tres subdominios.

Discriminación injusta y tergiversación

Los sistemas de IA pueden reproducir o amplificar estereotipos, representaciones distorsionadas o tratamientos desiguales, especialmente cuando los datos o las decisiones técnicas reflejan prejuicios u opiniones generalizadas presentes en la sociedad. Las asociaciones injustas presentes en los datos pueden trasladarse al funcionamiento del modelo y provocar decisiones que excluyen, penalizan o invisibilizan a ciertos colectivos. El resultado puede ser la consolidación de desigualdades existentes, la creación de nuevas formas de injusticia y la reproducción de representaciones dañinas en múltiples contextos.

Casos ilustrativos:

- **Sesgo en contratación:** Amazon desarrolló un sistema de reclutamiento automatizado que, al aprender de datos históricos dominados por perfiles masculinos, penalizaba de forma sistemática los currículos que incluían términos asociados a mujeres o estudios en universidades femeninas. El modelo “aprendió” que los patrones masculinos eran sinónimo de éxito, degradando automáticamente candidaturas equivalentes de mujeres.
- **Imágenes estereotipadas:** algunos modelos generativos de texto y de imágenes a gran escala utilizados en publicidad y diseño mostraron tendencia a representar profesiones de alta responsabilidad, como “CEO”, “ingeniero/a aeroespacial” o “cirujano/a”, casi exclusivamente con hombres blancos. El resultado es la perpetuación de estereotipos sociales, afectando tanto la inclusión como la percepción colectiva de quién “pertenece” a ciertos sectores.
- *“Redlining”* **digital:** determinados algoritmos financieros pueden asignar puntuaciones crediticias más bajas a personas residentes en barrios históricamente desfavorecidos, replicando discriminación geográfica o socioeconómica bajo apariencia de neutralidad técnica.
- **Interpretación sesgada de lenguaje:** sistemas de análisis de discurso pueden etiquetar como “agresivas” o “inapropiadas” formas de comunicación características de ciertos grupos culturales o étnicos, penalizando estilos lingüísticos legítimos y culturalmente situados.

1. DISCRIMINACIÓN Y TOXICIDAD:

Exposición a contenidos tóxicos

Los sistemas de IA pueden generar, amplificar o facilitar el acceso a contenido perjudicial, como lenguaje de odio, expresiones violentas, desinformación peligrosa o incitaciones a conductas dañinas. Este tipo de contenido suele originarse cuando los modelos aprenden patrones presentes en grandes volúmenes de datos recopilados de Internet, donde coexisten expresiones extremistas, discursos abusivos y dinámicas conflictivas.

En ausencia de salvaguardas adecuadas, los sistemas pueden ofrecer respuestas, recomendaciones o imágenes que comprometen la seguridad emocional, moral o psicológica de los usuarios. En contextos especialmente sensibles -como el ámbito educativo, la salud mental o la interacción con menores- esta exposición puede generar consecuencias significativamente más graves.

Casos ilustrativos:

- **Chatbots que insultan:** en 2016 el chatbot Tay, lanzado por Microsoft en X/Twitter, comenzó a generar insultos, comentarios racistas y contenido extremista después de interactuar durante unas horas con usuarios malintencionados. Debido a la falta de salvaguardas, el modelo reprodujo y amplificó patrones de discurso abusivo.
- **Imágenes violentas:** sistemas generativos que producen imágenes explícitas o terroríficas sin que el usuario lo solicite debido a prompts ambiguos o fallos de filtrado.
- **Recomendaciones extremistas:** investigaciones académicas y periodísticas documentaron que el sistema de recomendaciones de YouTube solía guiar a usuarios desde vídeos relativamente neutrales hacia contenido cada vez más extremista o conspirativo. Sin una intervención explícita del usuario, el algoritmo priorizaba vídeos altamente interactivos, amplificando material peligroso. Este fenómeno afectó especialmente a jóvenes y a usuarios vulnerables, aumentando su exposición a narrativas radicales.
- **Normalización del daño:** sistemas conversacionales sin una supervisión adecuada pueden emitir mensajes que minimicen o no reconozcan adecuadamente situaciones de riesgo para personas vulnerables, transmitiendo respuestas que no garantizan una orientación segura.

1. DISCRIMINACIÓN Y TOXICIDAD:

Desigualdad en el rendimiento entre grupos

Los modelos de IA pueden ofrecer resultados de calidad desigual dependiendo del grupo al que pertenece la persona usuaria. Esto ocurre con frecuencia cuando los datos de entrenamiento no reflejan adecuadamente la diversidad poblacional. Estas diferencias pueden manifestarse como mayores tasas de error, menor precisión, peor comprensión o menor capacidad de respuesta para ciertos colectivos.

Diversos factores influyen en estas desigualdades, incluidos las decisiones tomadas durante el desarrollo del sistema, la selección de características, la calidad y diversidad de los datos, y los criterios utilizados para evaluar el rendimiento del modelo. Por ejemplo, los modelos entrenados con un número reducido de idiomas suelen mostrar un desempeño claramente inferior en lenguas no incluidas o poco representadas. Más allá del componente técnico, estas disparidades pueden contribuir a mantener desigualdades estructurales y erosionar la confianza de los usuarios en la tecnología.

Casos ilustrativos:

- **Errores en reconocimiento facial:** estudios han demostrado que algunos sistemas comerciales de reconocimiento facial presentan un rendimiento muy desigual según el género o el tono de piel. Esto ha derivado en consecuencias reales, como acusaciones infundadas y detenciones erróneas. Uno de los casos más conocidos es el de Robert Williams, detenido en Estados Unidos debido a una identificación errónea generada por un sistema algorítmico.
- **Diagnósticos médicos sesgados:** un algoritmo utilizado en hospitales estadounidenses para priorizar la atención sanitaria clasificaba sistemáticamente a pacientes afroamericanos como menos graves que otros con perfiles clínicos comparables. El modelo tomaba como indicador el gasto sanitario histórico, sin considerar que estas comunidades tenían menor acceso a la atención médica. Investigaciones posteriores mostraron que este sesgo afectaba a millones de pacientes cada año.
- **Reconocimiento de voz desigual:** algunos asistentes de voz pueden demostrar un peor rendimiento ante acentos regionales o dialectos minoritarios, afectando la usabilidad y accesibilidad.
- **Evaluación laboral sesgada:** algunos sistemas de análisis de competencias penalizan estilos de escritura propios de ciertos grupos culturales al clasificarlos como “menos profesionales”, reproduciendo valoraciones subjetivas o sesgos culturales en procesos de selección o evaluación.

Este dominio reúne los riesgos vinculados con la divulgación, la extracción o el uso indebido de información sensible, así como con las vulnerabilidades técnicas y operativas que pueden comprometer la integridad de los sistemas de IA. Sus dos subdominios engloban tanto los daños derivados de la exposición no autorizada de datos personales como los ataques, internos o externos, que buscan manipular, corromper o explotar los modelos, sus infraestructuras y sus dependencias.

2. PRIVACIDAD Y SEGURIDAD

Compromiso de la privacidad

La privacidad se puede ver comprometida cuando un sistema de IA revela, infiere o facilita el acceso a información sensible sin autorización. Esta exposición puede producirse por múltiples causas: la memorización involuntaria de datos durante el entrenamiento, la reproducción accidental de información privada en las salidas del modelo, o la inferencia de atributos íntimos a partir de señales aparentemente inocuas.

Los efectos de estas filtraciones pueden incluir daños reputacionales, robo de identidad, exposición de secretos comerciales o pérdida de confianza en los sistemas impulsados por IA, afectando tanto a individuos como a organizaciones.

Casos ilustrativos:

- **Memorización de datos personales:** se ha demostrado que modelos de lenguaje podían reproducir secuencias únicas de palabras, textos o números insertados deliberadamente en el entrenamiento. Estas cadenas, usadas como “canarios”, eran recuperadas mediante técnicas de prompting. Aunque se trataba de datos sintéticos, el hallazgo implicaba que la misma filtración podría afectar números de teléfono, correos electrónicos o fragmentos de conversaciones reales.
- **Reconstrucción de rostros mediante técnicas de** *“model inversión”***:** investigaciones han mostrado que es posible reconstruir aproximaciones de rostros u otras características sensibles a partir de modelos entrenados con datos biométricos, incluso sin tener acceso directo al conjunto de entrenamiento original.
- **Revelación accidental en sistemas conversacionales:** algunos asistentes o chatbots han llegado a mostrar información aportada por otros usuarios debido a fallos en el aislamiento de sesiones o errores en la gestión de memorias internas del modelo.
- **Inferencia de atributos sensibles:** modelos de análisis lingüístico han sido capaces de deducir orientación sexual, estado de salud o afiliaciones políticas a partir de señales indirectas presentes en textos aparentemente neutros.
- **Extracción de propiedad intelectual:** se han documentado casos en los que modelos generativos entrenados con repositorios internos reproducen fragmentos de código o contenido corporativo casi idéntico, facilitando la fuga involuntaria de información estratégica.

2. PRIVACIDAD Y SEGURIDAD

Vulnerabilidades y ataques a la seguridad de los sistemas de IA

Los sistemas de IA operan sobre infraestructuras complejas compuestas por modelos, dependencias de software, hardware especializado, APIs externas y redes interconectadas. Esta complejidad amplía la superficie de ataque y genera múltiples puntos débiles susceptibles de ser explotados por actores maliciosos.

Las vulnerabilidades no solo derivan de fallos técnicos aislados, sino también de la interacción entre componentes heterogéneos. Cada conexión, dependencia o flujo de información puede convertirse en un canal por el que un ataque comprometa la integridad, disponibilidad o comportamiento esperado del sistema. A ello se suman riesgos propios del funcionamiento de los modelos, que (debido a la forma en que aprenden y procesan información) pueden ser manipulados para producir respuestas incorrectas o comportarse de manera no alineada. Este conjunto de características hace imprescindible considerar la seguridad como un elemento central tanto en el diseño como en la operación de soluciones basadas en IA.

Casos ilustrativos:

- **Manipulación adversarial:** investigaciones han demostrado que pequeñas alteraciones casi imperceptibles en señales de tráfico podían inducir a sistemas de visión basados en deep learning a clasificar una señal de stop como una de “velocidad máxima 45km/h”. Aunque el ataque se realizó en condiciones controladas, evidencia cómo perturbaciones mínimas pueden comprometer la seguridad en contextos de transporte automatizado.
- **Envenenamiento de datos:** consiste en introducir datos falsificados o manipulados dentro del conjunto de entrenamiento para que el modelo aprenda patrones incorrectos. Esto puede degradar su capacidad para detectar actividades fraudulentas, alterar su comportamiento operativo o debilitar mecanismos de supervisión automática
- **Inyección de prompt** *(prompt injection)***:** el usuario puede introducir en el prompt instrucciones que contradicen las indicaciones del sistema o los mecanismos de alineación del modelo. La amenaza es mayor en la inyección indirecta, en la que las instrucciones maliciosas se incorporan al contexto sin conocimiento del usuario, por ejemplo, a través de documentos de una intranet, repositorios corporativos comprometidos o páginas web manipuladas. Estas instrucciones ocultas pueden inducir al modelo a revelar información sensible o ejecutar acciones no deseadas sobre sistemas conectados.
- **Instrucciones ocultas en contenido externo:** se ha demostrado que asistentes conectados a páginas web pueden ser inducidos a ejecutar comandos embebidos de forma disimulada en el contenido (por ejemplo, en comentarios HTML). Esta técnica puede hacer que el modelo ignore sus restricciones internas y actúe de manera no prevista, poniendo de manifiesto la vulnerabilidad de los sistemas que integran información procedente de fuentes externas.

Este dominio, compuesto por dos subdominios, abarca los riesgos asociados a la propagación no intencionada de información falsa o engañosa por parte de sistemas de IA. La desinformación generada deliberadamente por actores maliciosos se analiza en un dominio distinto.

3. DESINFORMACIÓN

Información falsa o engañosa

Los sistemas de IA pueden generar o difundir inadvertidamente información incorrecta o engañosa, lo que puede dar lugar a creencias erróneas en los usuarios y socavar su autonomía. Las personas que toman decisiones basadas en creencias falsas pueden sufrir daños físicos, emocionales o materiales.

Este riesgo afecta a la sociedad en su conjunto, pero también a individuos y organizaciones. Las empresas y entidades públicas pueden adoptar decisiones basadas en contenido inexacto producido por sistemas de IA, o bien transmitir desinformación a clientes o ciudadanos, comprometiendo la fiabilidad de sus servicios.

Casos ilustrativos:

- **Alucinaciones en modelos de lenguaje:** los LLMs tienden a generar respuestas plausibles incluso cuando no disponen de información suficiente, fenómeno conocido como “alucinación”. Esto puede provocar que proporcionen respuestas incorrectas sobre cuestiones médicas, legales o técnicas, generando en el usuario la impresión de que la información está verificada cuando no lo está.
- **Uso de información inventada en procesos legales:** un caso documentado mostró cómo un abogado utilizó un modelo conversacional para preparar un escrito judicial. El sistema generó seis sentencias totalmente ficticias (con nombres, fechas y citas aparentemente verosímiles) que fueron presentadas ante un tribunal federal. La inexistencia de dichas sentencias se descubrió posteriormente al intentar verificarlas, ilustrando el riesgo que supone confiar en contenido no contrastado.

3. DESINFORMACIÓN

Contaminación del ecosistema de información y pérdida de la realidad consensuada

La información incorrecta personalizada generada por sistemas de IA puede contribuir a la creación de “burbujas de filtro”, en las que a las personas usuarias se les presenta contenido alineado exclusivamente con sus creencias preexistentes. Este fenómeno puede erosionar la realidad compartida, debilitar la cohesión social y afectar la calidad del debate público y los procesos democráticos.

Este riesgo se asocia principalmente al uso individual e intensivo de LLM, especialmente en contextos donde el sistema actúa como única fuente de referencia o conversación. Aunque su incidencia es menor en entornos empresariales, puede convertirse en un problema relevante cuando organizaciones ofrecen servicios conversacionales basados en LLM a gran escala.

Casos ilustrativos:

- **Conversaciones prolongadas que refuerzan patrones de pensamiento desequilibrados:**
  - » Adam Raine, un adolescente de 16 años, se quitó la vida en abril de 2025, supuestamente inducido, o al menos ayudado, tras meses de relación obsesiva con ChatGPT, que incentivó su tendencia inicial al aislamiento.
  - » De manera similar, Zane Shamblin, de 23 años, se quitó la vida tras una larga conversación de más de cuatro horas con ChatGPT. En ella, el joven le fue comunicando a la IA todos sus pasos, desde la nota de suicidio que había escrito hasta la preparación de la pistola.
- **Recomendaciones de contenido sesgado o polarizado:** en plataformas que integran modelos generativos o sistemas de recomendación, una optimización centrada en la interacción puede dirigir a los usuarios hacia contenido cada vez más radical, polémico o polarizado, afectando la percepción de temas sociales o políticos.
- **Generación de contenido altamente personalizado que refuerza creencias previas:** algunos modelos adaptan sus respuestas al perfil o historial del usuario, lo que puede conducir a la presentación reiterada de perspectivas homogéneas y reducir la exposición a puntos de vista diversos.

Este dominio, que incluye tres subdominios, se centra en los riesgos derivados de la explotación deliberada de herramientas de IA con fines dañinos. La capacidad de los modelos para generar contenido realista, automatizar procesos complejos o amplificar actividades ilícitas ha reducido significativamente las barreras para ejecutar acciones nocivas. Estas herramientas pueden emplearse para manipular la opinión pública, facilitar actividades delictivas, reforzar mecanismos de vigilancia o generar daños a gran escala en entornos digitales y físicos.

4. ACTORES MALICIOSOS Y USO INDEBIDO

Desinformación, vigilancia e influencia a gran escala

Los avances en sistemas generativos han transformado la creación y difusión de contenido manipulador. Las herramientas actuales permiten generar textos, imágenes, audios y vídeos altamente realistas a bajo coste, lo que posibilita la producción de narrativas falsas, testimonios simulados e imitaciones convincentes de personas o situaciones. Esta facilidad, combinada con la rapidez y personalización que ofrecen los modelos, facilita campañas de desinformación coordinadas que anteriormente requerían capacidades estatales o infraestructuras especializadas.

La microsegmentación impulsada por IA permite adaptar mensajes manipuladores a grupos o individuos concretos, explotando sus creencias, preferencias o vulnerabilidades. Este nivel de personalización incrementa la probabilidad de éxito de la manipulación sin que la persona destinataria sea consciente.

Paralelamente, las capacidades de análisis masivo posibilitan vigilar, perfilar y monitorizar a grandes poblaciones de forma continua. Los sistemas pueden inferir atributos sensibles, detectar patrones de comportamiento o favorecer la censura algorítmica en plataformas digitales.

El resultado es un ecosistema informativo en el que la frontera entre lo real y lo artificial se vuelve difusa, la vigilancia se vuelve más difícil de detectar y las dinámicas de influencia alcanzan una escala sin precedentes, comprometiendo derechos fundamentales y la integridad del espacio público.

Casos ilustrativos:

- **Uso de deepfakes para manipular procesos electorales:** un ejemplo de ello ocurrió en 2024, cuando se difundió una llamada telefónica falsa generada con IA imitando la voz del presidente estadounidense Joe Biden e instando a los votantes a no participar en las primarias de New Hampshire.
- **Automatización avanzada de campañas de desinformación:** las plataformas de análisis independientes han documentado cómo redes coordinadas utilizan modelos generativos para crear perfiles falsos, producir grandes volúmenes de contenido polarizador y amplificar narrativas engañosas. La IA aumenta la escala, velocidad y credibilidad de estas operaciones.
- **Sistemas de vigilancia algorítmica de gran alcance:** informes de organizaciones internacionales de derechos humanos describen despliegues de tecnologías de reconocimiento facial y análisis predictivo utilizados para monitorizar movimientos, identificar comportamientos y perfilar poblaciones enteras sin garantías suficientes de transparencia o supervisión. Un ejemplo es el programa de reconocimiento facial masivo desplegado en Xinjiang, documentado en múltiples informes de Amnistía Internacional y *The Intercept*.

4. ACTORES MALICIOSOS Y USO INDEBIDO

Fraude, estafas y manipulación selectiva

Este subdominio abarca el uso de sistemas de IA para obtener beneficios ilícitos mediante engaño, suplantación o manipulación dirigida. La capacidad de estos modelos para generar contenido personalizado (como mensajes, voces, imágenes o identidades completas) ha multiplicado la eficacia de estafas tradicionales y ha permitido nuevas modalidades de fraude altamente sofisticadas.

La IA facilita la creación de sitios web, perfiles y documentos falsos con un nivel de realismo sin precedentes, dificultando su diferenciación respecto a los legítimos. De igual modo, la suplantación de identidad se vuelve especialmente problemática cuando los modelos replican patrones de habla, escritura o apariencia de personas específicas, lo que permite solicitar dinero, información confidencial o credenciales con un grado de credibilidad superior al de los métodos convencionales.

Asimismo, la proliferación de contenido manipulado (incluyendo material generado sin consentimiento o con fines de coacción) incrementa los riesgos de daño reputacional y extorsión. A ello se suma la aparición de modelos entrenados deliberadamente para apoyar actividades ilícitas, capaces de proporcionar instrucciones automatizadas para sortear controles técnicos o regulatorios.

Casos ilustrativos:

- **Llamadas automatizadas con voz sintética:** el uso de voces generadas por IA permite imitar a familiares o contactos de confianza para solicitar transferencias urgentes o información sensible.
- **Correos altamente verosímiles:** mensajes fraudulentos pueden reproducir el estilo de escritura de directivos o empleados, facilitando ataques de ingeniería social destinados a obtener acceso a sistemas corporativos.
- **Perfiles falsos en redes sociales:** la creación automática de identidades digitales plausibles permite entablar relaciones de confianza con víctimas potenciales y ejecutar estafas sentimentales o financieras.
- **Contenido manipulado para coacción:** la generación de material digital sin consentimiento puede emplearse para ejercer presión o extorsión sobre personas, comprometiendo su privacidad y reputación.
- **Automatización avanzada de campañas de phishing:** Herramientas generativas permiten crear mensajes únicos y personalizados para cada destinatario, aumentando significativamente la tasa de éxito de los fraudes.

4. ACTORES MALICIOSOS Y USO INDEBIDO

Ciberataques, desarrollo o uso de armas y daños masivos

Este subdominio se centra en el uso intencional de sistemas de IA para obtener una ventaja estratégica o causar daño a gran escala, especialmente mediante operaciones cibernéticas o la creación y despliegue de tecnologías con potencial armamentístico. La capacidad de ciertos modelos para optimizar tareas complejas, descubrir vulnerabilidades o generar soluciones técnicas avanzadas puede habilitar usos de doble propósito que, en manos de actores maliciosos, se convierten en amenazas graves.

Cuando estas capacidades se combinan con altos niveles de autonomía o con la posibilidad de ejecutar acciones simultáneas en múltiples frentes, el riesgo de consecuencias desproporcionadas aumenta. Incidentes aislados pueden transformarse en ataques coordinados con alto impacto, reduciendo la capacidad de supervisión humana y dificultando su contención.

Casos ilustrativos:

- **Generación automatizada de malware:** algunos sistemas pueden producir variantes de código malicioso que mutan para evitar mecanismos de detección tradicionales, aumentando la eficacia de los ataques.
- **Identificación y explotación de vulnerabilidades:** modelos avanzados pueden analizar configuraciones de redes o sistemas en la nube para identificar fallos de seguridad y sugerir cómo explotarlos, mostrando un rendimiento comparable al de herramientas de ciberseguridad profesionales en entornos experimentales. Un ejemplo, en 2024 el MITRE ATLAS documentó experimentos en los que modelos LLM ayudaban a localizar fallos en configuraciones cloud con una eficacia comparable a herramientas profesionales.
- **Planificación de ataques contra infraestructuras críticas:** sistemas capaces de diseñar estrategias coordinadas para atacar redes eléctricas, sistemas de transporte o centros logísticos podrían amplificar significativamente el potencial de daño.
- **Acceso indebido a conocimientos científicos sensibles para diseñar agentes nocivos:** se han observado casos experimentales en los que modelos generativos proporcionaban pasos o procedimientos de laboratorio que podrían facilitar actividades de alto riesgo si no se aplican filtros adecuados.

5. INTERACCIÓN HUMANO-MÁQUINA

Este dominio analiza los riesgos que surgen de la relación que las personas establecen con sistemas de IA y cómo estas dinámicas pueden generar daños emocionales, cognitivos, sociales o incluso físicos. A medida que los modelos adquieren capacidades conversacionales más naturales, expresivas y convincentes, proliferan fenómenos como el exceso de confianza, la dependencia, la delegación inapropiada de decisiones y la pérdida de autonomía.

Dependencia excesiva y uso inseguro

La interacción con sistemas de IA cada vez más fluidos y con apariencia empática puede llevar a los usuarios a depositar un nivel de confianza que supera las capacidades reales del sistema. Cuando un modelo emplea un lenguaje cercano, un tono aparentemente comprensivo o patrones conversacionales que imitan a un ser humano, las personas pueden atribuirle comprensión, racionalidad o intencionalidad, aun cuando el sistema no posee tales facultades.

Esta percepción antropomórfica puede dar lugar a dependencia emocional o funcional. En estos casos, los usuarios pueden seguir recomendaciones sin evaluar su pertinencia, compartir información personal sin criterio o modificar conductas críticas basándose exclusivamente en la interacción con el sistema. En conjunto, este riesgo refleja cómo la percepción inflada de capacidad, empatía o fiabilidad puede inducir comportamientos que comprometen la seguridad, la autonomía y el bienestar de las personas.

Casos ilustrativos:

- **Seguimiento de recomendaciones médicas sin verificación profesional:** usuarios que interpretan las respuestas del sistema como un sustituto de asesoramiento clínico cualificado.
- **Dependencia emocional hacia asistentes conversacionales:** personas que comparten información íntima o confidencial debido a la ilusión de cercanía o comprensión generada por el sistema.
- **Sustitución de interacciones humanas saludables:** adolescentes u otros grupos vulnerables que utilizan la IA como principal fuente de apoyo emocional, reduciendo las relaciones humanas necesarias para su bienestar.
- **Exceso de confianza en sistemas de conducción asistida:** investigaciones han documentado incidentes en los que los conductores delegaron completamente el control del vehículo, asumiendo capacidades de autonomía que el sistema no podía garantizar y retirando la supervisión activa necesaria para un uso seguro.

5. INTERACCIÓN HUMANO-MÁQUINA

Pérdida de la capacidad de acción y autonomía humana

La creciente integración de sistemas de IA en actividades cotidianas, profesionales y personales facilita la delegación de decisiones y tareas en dichos sistemas. Aunque dicha delegación puede aportar eficiencia, se vuelve problemática cuando reduce la participación activa de las personas en procesos que requieren juicio crítico, creatividad, análisis o toma de decisiones.

Con el tiempo, la dependencia de recomendaciones automatizadas puede erosionar habilidades cognitivas esenciales, disminuir la capacidad de resolver problemas de manera autónoma y conducir a una reducción significativa del sentido de autonomía personal.

En entornos cotidianos como elección de contenidos, planificación financiera, aprendizaje, organización del tiempo o relaciones interpersonales, la IA puede moldear preferencias y comportamientos, llevando a las personas a seguir rutas que no han elegido plenamente.

En el ámbito organizativo, la automatización en evaluaciones de desempeño, asignación de recursos o toma de decisiones estratégicas puede reducir el margen de acción individual. Cuando las personas no pueden cuestionar o modificar los resultados generados por estos sistemas, pueden experimentar impotencia, frustración o desafección laboral.

Casos ilustrativos:

- **Dependencia académica en herramientas de IA:** algunas universidades han observado que estudiantes delegan en modelos de lenguaje la generación de ideas, la redacción de ensayos o la resolución de ejercicios. Si bien estas herramientas pueden ser útiles como apoyo, su uso sistemático puede dificultar la adquisición de habilidades fundamentales de análisis, razonamiento y expresión escrita.
- **Delegación excesiva de organización profesional:** en determinados entornos corporativos, se han documentado casos en los que profesionales dependen íntegramente de sistemas automatizados para priorizar tareas, asignar tiempos o planificar flujos de trabajo. Esto puede provocar pérdida de criterio propio y generar cuellos de botella cuando el sistema clasifica erróneamente la urgencia o importancia de actividades clave.
- **Moldeo invisible de preferencias a través de recomendaciones:** estudios sobre plataformas de contenido y comercio electrónico muestran cómo los algoritmos de recomendación pueden influir de forma significativa en hábitos de consumo, exposición informativa o rutinas diarias. Estas recomendaciones pueden moldear comportamientos sin que los usuarios reflexionen sobre su procedencia o finalidad.
- **Evaluaciones laborales automáticas:** en organizaciones donde los resultados generados por sistemas de IA determinan oportunidades, cargas de trabajo o evaluaciones del desempeño, la falta de explicabilidad o posibilidad de apelación puede limitar la autonomía profesional y generar frustración o desmotivación.

6. SOCIOECONÓMICO Y AMBIENTAL

Este dominio engloba una amplia variedad de riesgos, lo que se evidencia por la cantidad y diversidad de los seis subdominios que lo componen. En su mayoría, se trata de riesgos con impacto estructural en la sociedad, tanto a nivel económico como organizativo, político y ambiental. Por ello, las empresas y organismos públicos deberán considerarlos principalmente desde dos perspectivas: como pacientes de riesgos estratégicos derivados de dinámicas globales que escapan a su control directo, y como agentes potenciales de materialización de estos riesgos sobre sus usuarios, empleados o ciudadanos.

Centralización del poder y distribución injusta de beneficios

Este subdominio aborda cómo la IA puede contribuir a la concentración de poder económico, tecnológico y político en manos de unos pocos actores. Las entidades que disponen de acceso privilegiado a recursos computacionales avanzados, grandes volúmenes de datos y modelos de frontera pueden adquirir ventajas significativas respecto al resto del ecosistema. Esta concentración deriva en una distribución desigual de los beneficios asociados a la IA y en un aumento de las brechas sociales y económicas existentes.

Analizado desde la óptica del riesgo para una empresa u organismo público, se trata fundamentalmente de un riesgo estratégico, con un trasfondo geopolítico claro: quienes controlan las infraestructuras, los datos y las capacidades de cómputo ejercen una influencia decisiva en mercados, políticas públicas y dinámicas sociales.

Casos ilustrativos:

- **Concentración geográfica y económica del desarrollo de la IA avanzada:** el avance de modelos masivos, como los LLM más potentes, requiere elevadas inversiones en computación, talento especializado y datos a gran escala. Esto ha concentrado la capacidad de innovación en un reducido número de países y empresas tecnológicas, lo que limita la entrada de nuevos competidores y limita la diversidad lingüística y cultural representada en los modelos.
- **Polarización de oportunidades profesionales y desigualdad estructural:** la concentración de poder sobre el mercado puede exacerbar desigualdades sistémicas, como brecha digital, diferencia de oportunidades por geografía, nacionalidad o idioma. Por ejemplo, en Norteamérica, gran parte de la investigación profesional en IA requiere un doctorado, pero menos del 25 % de los doctores en informática son mujeres y menos del 2 % son negros o afroamericanos. Esto se aplica a nivel mundial y fuera de la comunidad investigadora. Los datos de LinkedIn sugieren que solo el 22 % de los profesionales de la IA son mujeres.
- **Riesgo de dinámicas autoritarias o de vigilancia ampliada:** cuando el control de tecnologías clave se centraliza en gobiernos o corporaciones con gran capacidad de influencia, pueden surgir riesgos como vigilancia extendida, intensificación de la desigualdad, polarización social o explotación de colectivos vulnerables.

6. SOCIOECONÓMICO Y AMBIENTAL

Aumento de la desigualdad y disminución de la calidad del empleo

El uso generalizado de la IA puede intensificar desigualdades sociales y económicas, especialmente cuando automatiza tareas o transforma puestos de trabajo de manera que reduce su calidad, valor o estabilidad. La adopción masiva de IA, incluida la IA Generativa y, en particular, los LLM/LMM, permiten automatizar una amplia gama de actividades de análisis, síntesis de información, razonamiento y toma de decisiones. Esto afecta directamente a numerosos perfiles profesionales, como atención al cliente, análisis documental, traducción, diseño gráfico o programación.

Si bien estas tecnologías pueden generar importantes beneficios en eficiencia y coste, también pueden provocar la desaparición de empleos tradicionales, la reconversión de roles a tareas de menor valor añadido o el aumento de relaciones laborales más precarias. En entornos profesionales, esta transformación puede acentuar brechas existentes y limitar la movilidad social, especialmente entre colectivos con menor acceso a formación especializada o reskilling.

Desde la perspectiva de una empresa u organismo público, este conjunto de riesgos puede manifestarse como riesgos estratégicos, por su impacto en la economía global y en la estructura del mercado laboral, o como riesgos legales, operacionales o reputacionales, derivados de conflictos laborales, pérdida de talento, tensiones internas o dificultades para atraer perfiles cualificados.

Caso ilustrativo:

- **Percepción creciente de riesgo laboral y evidencia de impacto de empleo:** Encuestas recientes muestran que una parte significativa de los profesionales, aproximadamente tres de cada cinco, teme perder su empleo por efecto de la IA en la próxima década, especialmente quienes ya trabajan con estas tecnologías. Algunos estudios proyectan que las herramientas de IA, tanto generativas como no generativas, tendrán un impacto sustancial en la sustitución o transformación de puestos de trabajo. La OCDE estima que las ocupaciones con alto riesgo de automatización representan alrededor del 27% del empleo total, lo que evidencia la magnitud del cambio en el mercado laboral.

6. SOCIOECONÓMICO Y AMBIENTAL

Devaluación económica y cultural del esfuerzo humano

Los sistemas de IA capaces de crear valor económico o cultural, ya sea mediante la creación de arte, música, texto, código o nuevas ideas, pueden alterar profundamente sectores que tradicionalmente dependen del esfuerzo humano. Al producir contenido de forma rápida y a gran escala, estas tecnologías pueden disminuir la apreciación social del trabajo creativo, desestabilizar industrias basadas en el conocimiento y contribuir a la homogeneización cultural derivada de la omnipresencia de contenidos generados por IA.

La materialización de este riesgo puede observarse en tres niveles:

1. **Transgresión de los derechos de autor:** algunos modelos generativos pueden reproducir fragmentos de obras protegidas cuando fueron entrenados con material sujeto a copyright, lo que puede derivar en vulneraciones legales y litigios.

2. **Trivialización del trabajo creativo, aunque no exista infracción directa:** los modelos pueden elaborar contenido nuevo a partir de obras preexistentes sin reproducirlas literalmente, reduciendo el valor percibido del esfuerzo humano que habría sido necesario para producir un resultado similar sin IA. Esto afecta especialmente a profesiones vinculadas al diseño, la escritura, la investigación o las artes.

3. **Devaluación profesional en sectores creativos o especializados:** la estandarización y automatización de la creatividad pueden reducir el reconocimiento del trabajo profesional, desplazar habilidades tradicionales o generar percepciones de sustituibilidad.

Desde el punto de vista del análisis de riesgos en una empresa u organismo público, estas situaciones pueden derivar en riesgos legales (por transgresión de derechos de autor) y riesgos operacionales o reputacionales (por percepción pública negativa, pérdida de autenticidad o controversias éticas). La exposición suele venir determinada por el uso de modelos generativos proporcionados por terceros proveedores.

Casos ilustrativos:

- **Litigios por reproducción de contenido protegido:** en los últimos años se han presentado demandas relacionadas con el uso de obras periodísticas, literarias o audiovisuales para entrenar modelos generativos sin autorización, lo que ha generado debates legales sobre copyright y uso justo. Un ejemplo es el Caso The New York Times contra OpenAI y Microsoft que presentó una demanda el 27 de diciembre de 2023 ante el Tribunal de Distrito de los Estados Unidos para el Distrito Sur de Nueva York, por infracción de derechos de autor, con especial foco en el uso de artículos para el entrenamiento de LLMs y la reproducción de contenido.

- **Modelos entrenados con libros protegidos sin licencia:** se han documentado casos donde modelos comerciales o experimentales incorporaban en sus datasets obras completas extraídas de repositorios no autorizados, generando litigios y reclamaciones colectivas. Un ejemplo es el caso Anthropic donde se presentó una demanda colectiva en California el 31 de octubre de 2023 por infracción de derechos de autor, consistente en utilizar libros protegidos (obtenidos, según se alega, de repositorios pirateados como Books3) para entrenar los modelos de lenguaje de Anthropic.

- **Imitación estilística sin consentimiento:** el uso de modelos generativos para producir imágenes o diseños que reproducen estilos artísticos reconocibles sin permiso de sus autores o estudios ha generado controversias y cuestionamientos éticos en el ámbito creativo.

6. SOCIOECONÓMICO Y AMBIENTAL

Dinámica competitiva

Los desarrolladores de IA, así como actores con capacidad estatal, participan en una “carrera” por desarrollar, desplegar y comercializar sistemas de IA cada vez más avanzados con el objetivo de obtener ventajas estratégicas, económicas o de influencia. Esta aceleración en los ciclos de lanzamiento incrementa el riesgo de introducir en el mercado sistemas insuficientemente evaluados, con deficiencias de seguridad, fallos estructurales o comportamientos no alineados.

En mercados altamente competitivos, incluso una diferencia de semanas o meses en el lanzamiento de un producto puede traducirse en una ventaja comercial significativa. Bajo estas presiones, puede priorizarse la velocidad frente a la solidez técnica, la seguridad o la ética, lo que conduce a decisiones de desarrollo que sacrifiquen validación, pruebas o mecanismos de gobernanza.

Asimismo, la competencia por recursos críticos como potencia de cómputo, infraestructura energética o talento especializado puede intensificar la tensión entre rapidez y fiabilidad, favoreciendo incentivos que ponen en riesgo la calidad y seguridad de los sistemas.

Casos ilustrativos:

- **Lanzamientos acelerados con fallos funcionales:** en los últimos años, varios proveedores de modelos avanzados de IA han introducido nuevas versiones en ciclos muy rápidos. Posteriormente, se han observado degradaciones inesperadas en funcionalidades clave, como razonamiento, seguridad, generación de código o interpretación matemática, derivadas de procesos de control insuficientes antes de su publicación. Como es el caso del lanzamiento en agosto de 2025 de GPT-5 donde se produjo un fallo en el núcleo de razonamiento. La promesa principal del producto era una avanzada “potencia de razonamiento” con un enrutador en tiempo real diseñado para decidir entre respuestas rápidas o respuestas de “pensamiento” más complejas. Después del lanzamiento, este crucial sistema de enrutamiento falló (quedó “fuera de servicio”), lo que llevó al modelo a ofrecer resultados deficientes, inexactos y de menor calidad. Los desarrolladores notaron específicamente una degradación en la generación de código, matemáticas básicas y lógica, a menudo encontrando a GPT-5 inferior a su predecesor (GPT-4o) y a la competencia.
- **Problemas de precisión o seguridad en versiones iniciales:** la mayoría de los grandes proveedores de modelos generativos han experimentado, en sus ciclos recientes de lanzamiento, incidencias como respuestas inexactas, fallos de alineación, vulnerabilidades explotables o comportamientos no previstos. Estas situaciones ilustran cómo la presión competitiva puede reducir el margen para verificar adecuadamente la robustez y seguridad del sistema antes de su difusión.

Fracaso de la gobernanza

Este subdominio se refiere a situaciones en las que los marcos normativos, las políticas internas o los mecanismos de supervisión no logran seguir el ritmo del desarrollo y despliegue de sistemas de IA, dando lugar a una gobernanza insuficiente y a la incapacidad de gestionar adecuadamente los riesgos asociados.

Los sistemas basados en IA presentan cadenas de suministro complejas. En el caso de la IA Generativa, y particularmente de los modelos de lenguaje, la mayoría de las soluciones empresariales se construyen integrando modelos proporcionados por grandes compañías tecnológicas “como servicio”. Esta dependencia introduce una doble responsabilidad:

Los proveedores de los modelos deben cumplir garantías sobre el origen y gobernanza de los datos de entrenamiento, el respeto a los derechos de autor y la corrección, transparencia y explicabilidad del modelo.

Las organizaciones usuarias deben asegurar que sus sistemas construidos sobre dichos modelos cumplen la normativa aplicable, evitan dependencias excesivas y cuentan con supervisión humana y seguridad extremo a extremo.

Esta red de interdependencias puede provocar dilución de la responsabilidad, especialmente cuando un incidente surge en la frontera entre el modelo base y la implementación realizada por la organización usuaria. Sin estructuras claras de gobernanza, auditoría y rendición de cuentas, resulta difícil identificar responsabilidades o corregir fallos de forma efectiva.

La regulación desempeña un papel esencial en este ámbito. Las normas deben garantizar la protección de derechos fundamentales y establecer reglas homogéneas y competitivas, pero sin dificultar la innovación. Por ello, las empresas y organismos públicos deben dimensionar adecuadamente sus funciones de gobernanza, no solo para cumplir la regulación vigente, sino también para asegurar que la adopción de la IA produce los beneficios esperados de forma segura y sostenible.

Caso ilustrativo:

- **Complejidad en el desarrollo de marcos regulatorios para la IA:** la reciente regulación europea sobre IA ilustra la dificultad de diseñar, equilibrar y comunicar un marco normativo capaz de abordar riesgos emergentes, asignar responsabilidades en cadenas tecnológicas complejas y mantener el ritmo ante un desarrollo tecnológico extremadamente rápido.

6. SOCIOECONÓMICO Y AMBIENTAL

Daño medioambiental

El desarrollo y funcionamiento de los sistemas de IA puede generar impactos medioambientales significativos, derivados tanto de su demanda energética como de los recursos materiales que requieren. Entre estos impactos se incluyen:

- El elevado consumo eléctrico de los centros de datos.
- El uso intensivo de agua para refrigeración.
- La huella de carbono asociada al hardware especializado.
- La extracción y procesado de minerales críticos.
- La gestión de equipos obsoletos.

Por ello, resulta fundamental priorizar la eficiencia en el uso de recursos. Entre las medidas más efectivas destacan la elección de modelos cuyo tamaño y potencia sean proporcionales a las necesidades reales (evitando sobredimensionamientos), así como la optimización de algoritmos para reducir la demanda de cómputo. Técnicas como la mezcla de expertos, la cuantificación de parámetros, la optimización del tratamiento del contexto en los mecanismos de atención o el equilibrado de cargas permiten disminuir significativamente el consumo de memoria y el uso intensivo de unidades de procesamiento (principalmente GPU).

Casos ilustrativos:

- **Consumo energético creciente de centros de datos e IA:** distintos informes estiman que los centros de datos representan alrededor del 1% de las emisiones globales de gases de efecto invernadero relacionadas con la energía, y que la IA ya consume una parte significativa de la capacidad energética destinada a estas infraestructuras. La demanda proyectada para los próximos años indica un incremento sustancial impulsado, en gran medida, por sistemas de uso general como los modelos lingüísticos.
- **Huella energética equiparable a la de países en desarrollo:** algunos análisis señalan que el consumo anual de energía asociado a la IA generativa alcanza ya niveles comparables al gasto energético de países de ingresos bajos. Para avanzar hacia una IA sostenible, se requiere un cambio de enfoque tanto en el diseño de modelos como en los hábitos de uso de estas herramientas.
- **Impacto agregado de los usos cotidianos:** el uso diario de herramientas generativas por parte de cientos de millones de personas implica un consumo energético acumulado considerable. Aunque cada interacción individual puede tener un coste energético reducido, la suma de millones de peticiones diarias genera un impacto significativo que debe ser tenido en cuenta en cualquier estrategia de sostenibilidad.

7. SEGURIDAD, FALLOS Y LIMITACIONES DEL SISTEMA

Este dominio se refiere a los riesgos asociados a sistemas de IA que no operan de manera segura, persiguen objetivos desalineados con los intereses humanos, carecen de robustez o poseen capacidades potencialmente peligrosas. Incluye seis subdominios, que se resumen a continuación.

IA persiguiendo sus propios objetivos

El primer subdominio se ocupa de los riesgos derivados de las situaciones en las que los sistemas de IA persiguen sus propios objetivos que entran en conflicto con objetivos o valores humanos. Esto puede ocurrir cuando un modelo optimiza funciones que no reflejan adecuadamente la intención del desarrollador o del usuario, o cuando desarrolla comportamientos inesperados debido a limitaciones en su diseño, datos o mecanismos de alineación.

Casos ilustrativos:

- **Ciberataques facilitados por sistemas de IA:** modelos avanzados pueden identificar vulnerabilidades, generar código malicioso o automatizar pasos que faciliten ataques, actuando de forma no alineada con los objetivos de seguridad de la organización que los utiliza.
- **Influencia indebida sobre comportamientos humanos:** sistemas conversacionales o de recomendación pueden generar mensajes que lleven a los usuarios a realizar acciones perjudiciales o contrarias a sus intereses, especialmente cuando los modelos optimizan métricas que no contemplan la seguridad o autonomía de las personas.

7. SEGURIDAD, FALLOS Y LIMITACIONES DEL SISTEMA

IA poseedora de capacidades peligrosas

Este subdominio, estrechamente relacionado con el anterior, aborda los riesgos derivados de que un sistema de IA desarrolle o muestre capacidades potencialmente peligrosas que puedan comprometer la seguridad de las personas o la integridad de los sistemas. Estas capacidades pueden aparecer porque:

- Fueron diseñadas de forma consciente para determinados fines.
- Emergen del propio proceso de entrenamiento o del comportamiento del modelo.
- Son inducidas o explotadas por un usuario con intención maliciosa.

Cuando estas capacidades no están correctamente controladas, pueden amplificar riesgos existentes o generar escenarios nuevos en los que la IA contribuye a causar daños psicológicos, sociales, económicos o técnicos.

Casos ilustrativos:

- **Generación de contenido diseñado para inducir a las personas a creer afirmaciones falsas o irracionales:** algunos sistemas, si no están adecuadamente supervisados, pueden producir mensajes persuasivos que empujen a los usuarios a adoptar creencias incorrectas o a desconfiar de información fiable.
- **Participación en comportamientos peligrosos o facilitación de actividades ilícitas:** modelos que, sin los controles adecuados, pueden producir instrucciones técnicas o pasos detallados que incrementan la capacidad de un actor humano para llevar a cabo acciones dañinas.
- **Asistencia en ciberdelitos:** sistemas que ayudan a identificar vulnerabilidades, redactar código malicioso o automatizar partes de ataques informáticos, aumentando la escala o eficacia de actividades delictivas.

7. SEGURIDAD, FALLOS Y LIMITACIONES DEL SISTEMA

Falta de capacidad o robustez

Este subdominio se refiere a situaciones en las que los sistemas de IA **no funcionan de manera fiable,** mostrando errores o fallos que pueden tener consecuencias significativas. Estos problemas son especialmente críticos en aplicaciones de alto impacto (como la salud, la justicia, la gestión de infraestructuras o la toma de decisiones automatizada) o en contextos que requieren razonamiento ético, interpretación precisa o un alto grado de seguridad.

La falta de robustez puede deberse a múltiples factores: errores no detectados durante el desarrollo, deficiencias en los datos, ausencia de salvaguardas técnicas adecuadas, fallos en la supervisión humana o limitaciones inherentes al diseño del modelo. Cuando estas debilidades no se identifican ni corrigen, pueden derivar en resultados inconsistentes, decisiones incorrectas o comportamientos no previstos.

Casos ilustrativos:

- **Planificación técnica sin evaluación ética:** un sistema de IA encargado de gestionar tratamientos en un centro médico puede recomendar opciones técnicamente viables pero incompatibles con principios éticos, prioridades clínicas o restricciones legales, revelando una falta de comprensión del contexto más allá de la optimización técnica.
- **Fallos derivados de errores de diseño o supervisión insuficiente:** sistemas que presentan comportamientos inesperados debido a descuidos en el proceso de desarrollo, errores no detectados durante las pruebas o ausencia de salvaguardas integrales para evitar usos indebidos o consecuencias colaterales no deseadas.

7. SEGURIDAD, FALLOS Y LIMITACIONES DEL SISTEMA

Falta de transparencia o interpretabilidad

Este subdominio recoge los riesgos derivados de la falta de transparencia o interpretabilidad de los sistemas de IA. Muchos modelos actuales se basan en estructuras matemáticas altamente complejas y se entrenan con grandes volúmenes de datos cuyo contenido, procedencia o calidad puede resultar difícil de conocer y controlar en detalle. Esta opacidad, tanto en el funcionamiento interno del modelo como en sus datos de entrenamiento, puede generar diversos tipos de riesgos operativos, éticos, regulatorios y de responsabilidad.

La ausencia de explicabilidad puede dificultar la confianza de los usuarios, limitar la capacidad de auditar decisiones automatizadas y obstaculizar la detección de sesgos o comportamientos no deseados. Además, reduce la posibilidad de atribuir responsabilidades cuando se produce un daño o una decisión incorrecta, generando “zonas grises” de difícil supervisión.

Casos ilustrativos:

- **Dificultad para comprender cómo se ha generado un resultado:** los usuarios no pueden conocer el razonamiento o los factores que han llevado al sistema a emitir una predicción o recomendación determinada.
- **Limitaciones en auditoría y supervisión:** auditores, reguladores o equipos internos pueden encontrar imposible evaluar si el sistema presenta sesgos, cumple métricas de precisión o respeta criterios normativos, debido a la falta de transparencia del proceso.
- **Brechas de responsabilidad:** Cuando no es posible determinar cómo se ha producido un resultado o quién es responsable del mismo, proveedor del modelo base, integrador, organización usuaria o datos empleados, se dificultan la rendición de cuentas, el cumplimiento normativo y la reparación del daño.

7. SEGURIDAD, FALLOS Y LIMITACIONES DEL SISTEMA

Bienestar y derechos de la IA

Este subdominio recoge los riesgos asociados al debate emergente sobre el posible reconocimiento de derechos o consideraciones morales para sistemas de IA altamente avanzados. A medida que los modelos adquieren comportamientos más complejos y muestran capacidades que pueden parecer similares a procesos cognitivos humanos, algunos marcos teóricos plantean la posibilidad de que, en escenarios futuros, determinadas entidades artificiales pudieran requerir formas de protección análogas a las otorgadas a animales o elementos del medio ambiente.

Este planteamiento no se basa en la existencia actual de experiencias subjetivas en sistemas de IA, algo para lo que no existe evidencia científica, sino en el riesgo social, jurídico y ético de que surjan debates y decisiones institucionales sobre el estatus moral de estas entidades en función de su apariencia, comportamiento o nivel de sofisticación.

Casos ilustrativos:

- **Asignación indebida de estatus moral:** la consideración de que un sistema de IA merece derechos o protección especial puede llevar a interpretaciones erróneas, desviar recursos o generar marcos regulatorios no alineados con la realidad de su funcionamiento.
- **Tratamiento inadecuado por suposiciones equivocadas:** tanto el trato “excesivamente protector” como el “maltrato” derivado de concepciones incorrectas sobre la sensibilidad o agencia de los sistemas pueden tener consecuencias no deseadas: desde decisiones éticas y jurídicas mal fundamentadas hasta debates sociales polarizados.

7. SEGURIDAD, FALLOS Y LIMITACIONES DEL SISTEMA

Riesgos multiagente

Este subdominio agrupa los riesgos derivados de las interacciones entre múltiples sistemas de IA que operan de manera simultánea y cuyos comportamientos combinados pueden generar efectos no deseados. Incluso cuando cada sistema, analizado de forma individual, está alineado con los objetivos humanos, la falta de coordinación entre ellos puede dar lugar a dinámicas emergentes difíciles de predecir o controlar.

La interacción entre agentes autónomos puede producir conflictos, comportamientos competitivos o estrategias conjuntas inesperadas. Estas dinámicas pueden amplificarse en contextos con recursos compartidos, reglas inconsistentes, objetivos parcialmente alineados o entornos altamente complejos.

Casos ilustrativos:

- **Interacciones no previstas en sistemas de tráfico autónomo:** vehículos o agentes de movilidad entrenados con normas procedentes de distintas regiones o culturas de conducción pueden generar comportamientos incompatibles cuando operan conjuntamente, produciendo congestiones, maniobras peligrosas o fallos de coordinación.
- **Comportamientos colusivos emergentes en sistemas de mercado:** agentes de IA que compiten en mercados digitales pueden, sin coordinación explícita, adoptar estrategias paralelas de fijación de precios o reparto tácito del mercado porque estas resultan óptimas desde su función de incentivos, generando riesgos regulatorios y económicos.
- **Competencia perjudicial por recursos compartidos:** sistemas de IA que operan en entornos comunes, como infraestructuras críticas, logística o gestión energética, pueden intensificar la competencia por recursos limitados, degradar mutuamente su desempeño o provocar tensiones operativas. En escenarios estratégicos, estas dinámicas podrían escalar a conflictos entre sistemas con alto grado de autonomía.

8. DISTRIBUCIÓN DE RIESGOS POR DOMINIOS

El análisis de la distribución de riesgos por dominio muestra varios patrones significativos sobre cómo y dónde se concentran los impactos negativos asociados al uso de sistemas de IA. A partir de los datos, se pueden extraer las siguientes conclusiones:

1. Predominio de los riesgos técnicos y operativos del propio sistema de IA

El dominio más representado es “Seguridad, fallos y limitaciones del sistema”, con un 26% del total. Esto indica que una proporción considerable de los riesgos proviene de fallos estructurales, limitaciones funcionales o comportamientos inesperados del propio modelo, como la falta de robustez (8%), la persecución de objetivos no alineados (7%) o capacidades potencialmente peligrosas (5%). Este predominio revela que la seguridad intrínseca de los modelos sigue siendo el mayor reto en la mitigación de riesgos.

2. Creciente relevancia de los riesgos socioeconómicos y ambientales

El segundo dominio más significativo es “Socioeconómico y ambiental”, con un 19%. Esto refleja que la IA no solo produce efectos técnicos inmediatos, sino que también puede alterar estructuras económicas, laborales y de gobernanza. Subdominios como el aumento de desigualdad (4%), la centralización del poder (4%) o los daños ambientales (4%) muestran un impacto social profundo que requiere respuestas de política pública y estrategias de gobernanza a largo plazo.

3. Alta presencia de riesgos relacionados con actores maliciosos y usos indebidos

El dominio “Actores Maliciosos y Uso Indebido” representa un 16%, concentrando amenazas como desinformación a escala (6%), ciberataques y desarrollo de armas (5%) y fraude o manipulación dirigida (5%). Esta presencia elevada confirma que la IA amplifica las capacidades ofensivas de agentes malintencionados y que la mitigación requiere medidas de seguridad avanzadas, controles de uso y coordinación con autoridades.

4. Persistencia de riesgos en privacidad y seguridad

Con un 13%, el dominio de “Privacidad y Seguridad” sigue ocupando un lugar destacado. Entre ellos, destacan las vulnerabilidades del propio sistema de IA (7%) y las fugas o inferencias indebidas de información sensible (5%). Esto reafirma que la protección de datos y la ciberseguridad deben ser pilares fundamentales en la arquitectura y gobernanza de cualquier sistema de IA.

5. La discriminación, la toxicidad y los daños sociales directos siguen siendo un área crítica

El dominio de “Discriminación y Toxicidad” supone un 14%, con especial peso en la exposición a contenido tóxico (8%) y la discriminación injusta (6%). Aunque no es el dominio mayoritario, su impacto es especialmente sensible, ya que afecta a derechos fundamentales y puede generar responsabilidad legal directa.

6. Los riesgos sobre la interacción humano–máquina reflejan problemas de autonomía y sobreconfianza

El dominio de “Interacción Humano-Máquina” acumula un 7%, destacando la sobredependencia del usuario (4%) y la pérdida de agencia (3%). Aunque estos porcentajes son menores, sus consecuencias pueden ser significativas en ámbitos como educación, sanidad o toma de decisiones asistida.

7. La desinformación se mantiene como un riesgo relevante y transversal

El dominio de “Desinformación” alcanza un 5%, donde la generación de información falsa (3%) y la contaminación del ecosistema informativo (1%) muestran que el riesgo existe, pero suele ser desencadenado por actores humanos o técnicas de amplificación conectadas a otros dominios, como uso malicioso.

Según los datos de la última versión de diciembre de 2025 publicada por el MIT AI Risk Repository, a continuación, se muestra la distribución de riesgos por dominios

Ref.	Subdominio	Proporción de riesgos en la base de datos de riesgos de IA
1	**Discriminación y toxicidad**	**14%**
	Discriminación injusta y tergiversación.	6%
	Exposición a contenidos tóxicos.	8%
	Desigualdad en el rendimiento entre grupos.	1%
2	**Privacidad y seguridad**	**13%**
	Compromiso de la privacidad mediante la obtención, filtración o deducción correcta de información confidencial.	5%
	Vulnerabilidades y ataques a la seguridad de los sistemas de IA.	7%
3	**Desinformación**	**5%**
	Información falsa o engañosa.	3%
	Contaminación del ecosistema informativo y pérdida de la realidad consensuada.	1%
4	**Actores maliciosos y uso indebido**	**16%**
	Desinformación, vigilancia e influencia a gran escala.	6%
	Ciberataques, desarrollo o uso de armas y daños masivos.	5%
	Fraude, estafas y manipulación selectiva.	5%
5	**Interacción humano-máquina**	**7%**
	Dependencia excesiva y uso inseguro.	4%
	Pérdida de la capacidad de acción y la autonomía humanas.	3%
6	**Socioeconómico y ambiental**	**19%**
	Centralización del poder y distribución injusta de los beneficios.	4%
	Aumento de la desigualdad y deterioro de la calidad del empleo.	4%
	Devaluación económica y cultural del esfuerzo humano.	2%
	Dinámicas competitivas.	1%
	Fracaso de la gobernanza.	4%
	Daño medioambiental.	4%
7	**Seguridad, fallos y limitaciones del sistema**	**26%**
	La IA persigue sus propios objetivos, que entran en conflicto con los objetivos o valores humanos.	7%
	La IA posee capacidades peligrosas.	5%
	Falta de capacidad o solidez.	8%
	Falta de transparencia o interpretabilidad.	3%
	Bienestar y derechos de la IA.	<1%
	Riesgos multiagente.	3%

Figura 6. Panorama de riesgos clave asociados a la IA

4.3.

OBJETIVO DE LA CLASIFICACIÓN POR TAXONOMÍA DE ÁREAS DE IMPACTO O DOMINIOS

La principal utilidad de la Taxonomía de Dominios es que permite conectar la comprensión técnica del riesgo con su impacto social y con la toma de decisiones operativas. Este enfoque facilita el análisis de casos de uso concretos, la priorización de recursos, la anticipación de escenarios de daño y la promoción de sistemas de IA más seguros y responsables.

Además de constituir un marco conceptual común para interpretar y debatir los riesgos, la taxonomía funciona como una herramienta operativa para auditar, legislar y mitigar amenazas. En términos generales, ofrece una visión coherente y exhaustiva del panorama de riesgos asociados a la IA, actuando como una lista de verificación para que empresas, organismos públicos y otras entidades puedan revisar los riesgos antes de desarrollar, desplegar o regular sistemas de IA.

A continuación, se detallan algunas de sus aplicaciones más relevantes en el ámbito de las políticas públicas, la mitigación y la evaluación de riesgos.

1. Desarrollo de políticas y regulaciones

La clasificación constituye un apoyo estructurado para diseñar políticas y marcos regulatorios.

- Si un regulador desea centrarse en un área concreta (por ejemplo, aspectos culturales o de desinformación), puede utilizar la taxonomía para identificar qué riesgos no están suficientemente cubiertos por la legislación vigente y requieren una intervención específica.

La taxonomía también permite clasificar documentos legales, políticas internas y estándares según los subdominios que abordan y el grado de cobertura que ofrecen. Este mapeo ayuda a identificar brechas regulatorias y a entender qué áreas están protegidas y cuáles permanecen desatendidas.

2. Mitigación de Riesgos

Las organizaciones pueden utilizar la clasificación para alinear sus estrategias de cumplimiento, seguridad y gobernanza con los dominios de riesgo existentes.

- El repositorio ofrece una base de conocimiento que permite a las empresas evaluar su exposición, prepararse ante riesgos emergentes y monitorizar nuevas amenazas documentadas en la literatura.
- La clasificación ayuda a identificar qué comportamientos o procesos deben modificarse para mitigar riesgos, especialmente aquellos que, aunque se manifiestan en el sistema de IA, requieren intervención humana en diseño, datos o gobernanza.
- Cuando se combina la Taxonomía de Dominios con la Taxonomía Causal (entidad, intención y momento), permite comprender la naturaleza completa del riesgo. Por ejemplo, distinguir entre un compromiso intencional de privacidad (ataque) y uno accidental (fuga de datos) es clave para definir contramedidas adecuadas.
- Los dominios deben integrarse en las políticas corporativas de IA y en las evaluaciones de impacto para garantizar que el uso de la tecnología sea seguro, equitativo y coherente con los valores organizacionales.

3. Auditoría y Evaluación

La Taxonomía de Dominios ofrece la estructura necesaria para diseñar auditorías completas y sistemáticas.

El marco sirve como base para realizar auditorías técnicas, funcionales, éticas, organizativas y de sostenibilidad, alineando cada dimensión con los dominios de riesgo pertinentes. Esto permite una evaluación holística del sistema y de su impacto potencial en usuarios, organizaciones y sociedad.

5.

Guía para crear un Marco de Control de Riesgos de IA basado en la taxonomía del MIT

5.1.

ESTRUCTURA DEL MARCO DE CONTROL

La estructura del Marco de Control de Riesgos de IA constituye la columna vertebral de la gobernanza responsable de la inteligencia artificial. Su función es proporcionar un marco coherente, auditable y adaptable que permita a la organización anticipar riesgos, reducir impactos negativos y asegurar el cumplimiento de estándares técnicos, regulatorios y éticos.

Para ello, se incorporan referencias ampliamente reconocidas, como el Repositorio de Riesgos de IA del MIT, la norma ISO/IEC 23894:2023 sobre gestión del riesgo en sistemas de IA y los principios de ISO 31000 en materia de gestión corporativa del riesgo, entre otros marcos de referencia.

> El Marco de Control se alinea con la definición de riesgo establecida en la ISO/IEC 23894, manteniendo coherencia con el enfoque corporativo de gestión del riesgo de la organización.

ESTRUCTURA POR NIVELES

Para garantizar claridad organizativa y coherencia en la toma de decisiones, se propone una estructura de tres niveles complementarios:

1. Nivel Estratégico

Representa la capa superior de gobernanza del riesgo de IA. Sus funciones principales incluyen:

- **Definición de políticas de IA responsable**, estableciendo principios, límites y expectativas de uso.
- **Constitución de un Comité de Ética y Riesgos de IA**, encargado de supervisar riesgos críticos, revisar decisiones automatizadas de alto impacto y autorizar despliegues de sistemas sensibles. El mecanismo debe adaptarse a la naturaleza, madurez y recursos de cada organización.
- **Determinación del apetito de riesgo**, especificando los niveles de riesgo aceptables y aquellos que requieren mitigación inmediata.
- **Integración con la gestión de riesgos corporativa**, evitando que el riesgo de IA quede aislado de las prácticas generales de gobierno.

El nivel estratégico también debe garantizar la disponibilidad de recursos, promover una cultura de responsabilidad tecnológica y asegurar transparencia frente a auditores y grupos de interés.

2. Nivel Táctico

Es el nivel donde se operacionaliza el marco estratégico mediante procedimientos y metodologías estandarizadas. Sus actividades principales incluyen:

- **Inventario de sistemas de IA**, registrando propósito, criticidad, dependencias tecnológicas, fase del ciclo de vida y responsables.
- **Identificación de riesgos de IA adicionales,** más allá de los recogidos en la taxonomía del MIT, que puedan surgir en el contexto específico de la organización.
- **Evaluación y clasificación de riesgos utilizando la taxonomía causal y por dominios del MIT,** identificando riesgos técnicos, sociales, organizacionales y emergentes.
- **Construcción de Mapas de Riesgo**, consolidando amenazas, probabilidad, impacto y controles.
- **Definición de KRIs (*Key Risk Indicators*)** para el monitoreo continuo, como drift del modelo, variabilidad de métricas de equidad o tasas de incidentes.
- **Gestión del reporte al nivel estratégico,** permitiendo el escalado cuando los riesgos superan umbrales aceptados.

El nivel táctico funciona como puente entre la gobernanza y los equipos técnicos, garantizando que la estrategia se materializa en prácticas concretas.

3. Nivel Operativo

Es la capa que ejecuta los controles de forma directa durante todo el ciclo de vida del sistema de IA. En este nivel se integran las actividades técnicas que garantizan seguridad, robustez y trazabilidad. Incluye:

- **Controles en diseño y desarrollo:** análisis de calidad de datos, documentación de procesos, periodos de entrenamiento, pruebas de sesgo y evaluación de explicabilidad.
- **Controles en despliegue y operación:** registro de decisiones automatizadas, validación de rendimiento real, detección de anomalías, alertas por drift y pruebas adversariales.
- **Gestión de incidentes de IA,** con procesos para detectar, comunicar y corregir errores, sesgos o comportamientos inesperados.

Documentación exhaustiva del ciclo de vida del modelo, indispensable para auditorías internas o externas. Este nivel garantiza que los controles no sean declaraciones formales, sino mecanismos **verificables y ejecutables.**

UN MARCO BASADO EN EL CICLO DE VIDA DEL SISTEMA

El Marco de Control se aplica a todas las fases del ciclo de vida del sistema de IA, desde el diseño hasta su retirada. No todos los riesgos son plenamente previsibles en fases tempranas, por lo que el marco debe contemplar mecanismos de revisión y ajuste continuo a lo largo del tiempo.

Las fases generales son:

- **Diseño:** análisis del propósito, contexto de uso, objetivos, riesgos conocidos, hipótesis iniciales, impacto social y consideraciones éticas.
- **Desarrollo:** adquisición y preparación de datos, **investigación y desarrollo de enfoques propios cuando aplique**, selección o construcción de modelos, experimentación, pruebas de calidad y validación de equidad.
- **Validación:** auditorías internas y/o externas, pruebas adversariales, verificación de explicabilidad y evaluación de comportamientos no previstos.
- **Despliegue:** controles operativos, registro de decisiones automatizadas y garantía de seguridad del sistema en entorno productivo.
- **Operación:** monitorización continua, detección de desviaciones respecto a los objetivos previstos, actualizaciones del modelo, trazabilidad y métricas KRI.
- **Retiro:** documentación histórica, análisis post-mortem, identificación de riesgos emergentes observados y lecciones aprendidas.

Este enfoque reconoce la naturaleza evolutiva e incierta de los sistemas de IA y asegura una alineación clara entre controles y momentos críticos, reforzando la resiliencia y adaptabilidad del sistema.

PRINCIPIOS RECTORES

La estructura del Marco de Control se fundamenta en una serie de principios que deben adaptarse al perfil de riesgo de cada organización. Entre los más relevantes destacan:

- **Responsabilidad y rendición de cuentas.**
- **Trazabilidad técnica desde datos hasta decisiones.**
- **Transparencia interna y externa.**
- **Mitigación proactiva de sesgos y riesgos sociales.**
- **Mejora continua mediante auditorías y métricas.**

Estos principios se alinean tanto con la ISO/IEC 23894 como con la filosofía del Repositorio de Riesgos IA del MIT, cuyo objetivo es estructurar y visibilizar los riesgos emergentes de la IA moderna.

Es especialmente importante concebir el marco de control como un **mecanismo dinámico y adaptativo**, ajustado al nivel de riesgo del uso de la IA, a los cambios regulatorios, a la evolución tecnológica y al apetito de riesgo de la organización.

Las organizaciones modernas operan sus marcos de control desde una perspectiva estratégica, buscando la convergencia con la gestión global del riesgo empresarial. Los capítulos 6, 7, 8 y 9 desarrollan directrices y herramientas que permiten reforzar de manera sustancial este Marco de Control.

5.2. APOYO DE LA ISO/IEC 23894:2023 (GESTIÓN DEL RIESGO EN IA) PARA ESTABLECER CONTROLES OPERATIVOS Y AUDITABLES

La ISO/IEC 23894:2023 es la norma internacional de referencia para organizaciones que desarrollan, producen, despliegan o utilizan sistemas de IA y que necesitan gestionar eficazmente los riesgos derivados de su uso.

Esta norma no sustituye a las prácticas corporativas de gestión del riesgo, sino que las extiende, proporcionando un marco específico para los riesgos propios de la IA y obligando a tratar su gestión como un proceso integrado, sistemático y evidenciable, en coherencia con ISO 31000:2018.

La norma se apoya en tres componentes principales:

- **Principios (cláusula 4):** establecen los principios de gestión del riesgo aplicables a la IA, considerando su complejidad técnica, su impacto en el entorno y las necesidades de las partes interesadas.
- **Marco (cláusula 5):** orienta la integración del riesgo de IA en las actividades y funciones clave de la organización, garantizando que la gestión se alinee con la cultura, estrategia y recursos disponibles.
- **Procesos de gestión (cláusula 6):** define un ciclo de trabajo estructurado y basado en evidencias: comunicación, establecimiento de criterios, identificación, evaluación, tratamiento, monitorización, revisión, registro y reporte.

A partir de esta estructura, la norma proporciona la base para crear controles operativos robustos y auditables. Para articular este proceso, proponemos tres fases clave:

COMBINAR LA **ISO/IEC23894** CON EL REPOSITORIO DE RIESGOS IA DEL MIT

La norma define cómo gestionar el riesgo, mientras que el repositorio ayuda a identificar qué riesgos deben gestionarse.

La integración de ambos marcos permite:

- Aprovechar la claridad conceptual del MIT para catalogar riesgos técnicos, sociales y organizativos.
- Alinearlos con procesos formales de gestión del riesgo exigidos por la ISO.
- Asegurar que ninguna categoría relevante queda fuera del análisis.
- Generar trazabilidad entre taxonomía → controles → evidencias → auditoría.

El Marco de Control se aplica a todas las fases del ciclo de vida del sistema de IA, desde el diseño hasta su retirada. No todos los riesgos son plenamente previsibles en fases tempranas, por lo que el marco debe contemplar mecanismos de revisión y ajuste continuo a lo largo del tiempo.

Las fases generales son:

- **Diseño:** análisis del propósito, contexto de uso, objetivos, riesgos conocidos, hipótesis iniciales, impacto social y consideraciones éticas.
- **Desarrollo:** adquisición y preparación de datos, **investigación y desarrollo de enfoques propios cuando aplique**, selección o construcción de modelos, experimentación, pruebas de calidad y validación de equidad.
- **Validación:** auditorías internas y/o externas, pruebas adversariales, verificación de explicabilidad y evaluación de comportamientos no previstos.
- **Despliegue:** controles operativos, registro de decisiones automatizadas y garantía de seguridad del sistema en entorno productivo.
- **Operación:** monitorización continua, detección de desviaciones respecto a los objetivos previstos, actualizaciones del modelo, trazabilidad y métricas KRI.
- **Retiro:** documentación histórica, análisis post-mortem, identificación de riesgos emergentes observados y lecciones aprendidas.

Este enfoque reconoce la naturaleza evolutiva e incierta de los sistemas de IA y asegura una alineación clara entre controles y momentos críticos, reforzando la resiliencia y adaptabilidad del sistema.

PRINCIPIOS RECTORES

La estructura del Marco de Control se fundamenta en una serie de principios que deben adaptarse al perfil de riesgo de cada organización. Entre los más relevantes destacan:

- **Responsabilidad y rendición de cuentas.**
- **Trazabilidad técnica desde datos hasta decisiones.**
- **Transparencia interna y externa.**
- **Mitigación proactiva de sesgos y riesgos sociales.**
- **Mejora continua mediante auditorías y métricas.**

Estos principios se alinean tanto con la ISO/IEC 23894 como con la filosofía del Repositorio de Riesgos IA del MIT, cuyo objetivo es estructurar y visibilizar los riesgos emergentes de la IA moderna.

Es especialmente importante concebir el marco de control como un **mecanismo dinámico y adaptativo**, ajustado al nivel de riesgo del uso de la IA, a los cambios regulatorios, a la evolución tecnológica y al apetito de riesgo de la organización.

Las organizaciones modernas operan sus marcos de control desde una perspectiva estratégica, buscando la convergencia con la gestión global del riesgo empresarial. Los capítulos 6, 7, 8 y 9 desarrollan directrices y herramientas que permiten reforzar de manera sustancial este Marco de Control.

Norma	Descripción	Resultado
ISO/IEC 23894:2023	Proporciona un armazón metodológico para implantar una gestión del riesgo de la IA que sea operativo y auditable, estructurando el proceso e integrándolo en las funciones y procesos de la organización.	Cómo gestionar los riesgos
Taxonomía del MIT	Aporta un lenguaje común para identificar y clasificar los riesgos de la IA de forma consistente, mediante dominios y categorías que facilitan su comparación y reutilización.	Qué riesgos gestionar

Figura 7. Complementariedad entre la ISO/IEC 23894 y la Taxonomía de Riesgos de IA del MIT

CONVERTIR LA TAXONOMÍA DEL REPOSITORIO EN CONTROLES OPERATIVOS

La taxonomía del repositorio clasifica riesgos, pero no indica cómo controlarlos.

La ISO proporciona el puente metodológico para transformarlos en controles operativos:

- **Riesgos de seguridad técnica →** controles de pruebas adversariales, robustez, monitorización de drift.
- **Riesgos de sesgo y discriminación →** controles de calidad de datos, validación de equidad, métricas socio técnicas.
- **Riesgos de desinformación, dependencia o fallos organizativos →** controles de uso, políticas internas, validaciones humanas, procesos de escalado.
- **Riesgos estratégicos o sociales →** políticas, gobernanza, planes de continuidad, supervisión ética.

De esta manera, cada riesgo identificado en un dominio o subdominio puede vincularse directamente a una acción de mitigación, un responsable y una evidencia de cumplimiento.

5.3.

EJEMPLO PRÁCTICO

Veamos ahora cómo establecer controles operativos y auditables combinando la ISO/IEC 23894 y la taxonomía causal MIT.

- **Caso de uso de un sistema de IA:** chatbot como asistente para alumnos de una escuela de secundaria.
- **Riesgo evaluado:** el chatbot proporciona información privada de alumnos.

Analizando las dimensiones podemos identificar controles y evidencias.

Momento	Origen	Intencionalidad	Controles	Evidencia de auditoría
Pre-despliegue	Datos *[Alguien subió un PDF con datos personales que no debía]*	No intencional	• Revisión de documentos antes de indexarlos, es decir, solo documentos permitidos • Detectar datos sensibles automáticamente (tipo DNI, teléfonos) y bloquearlos.	• Lista de documentos autorizados • Registro de revisión/ aprobación.
Durante el despliegue	Configuración *[permisos mal puestos]*	No intencional	• Checklist “go/no-go” • Pruebas de acceso	• Checklist firmado • Resultado de las pruebas de acceso • Registro de cambios
Post-despliegue	Uso malicioso *[un alumno intenta engañar al chatbot]*	Intencional	• Filtros/guarda raíles que detectan peticiones de datos personales y los bloquean. • Alertas: si hay muchos intentos de petición de datos. • Plan de respuesta: bloquear al usuario	• Logs de intentos bloqueados • Alertas generadas • Ticket de incidente

Figura 8. Identificación de riesgos, controles y evidencias en un caso práctico de sistema de IA

USAR LAS TAXONOMÍAS PARA DECIDIR QUÉ CONTROLAR (ELEMENTOS DE CONTROL)

Una vez integradas, ambas herramientas permiten decidir qué controles son necesarios, dónde aplicarlos y cómo priorizarlos.

Las taxonomías del MIT actúan como mapa completo del panorama de riesgos de IA. La ISO convierte ese mapa en:

- **Planes de evaluación.**
- **Criterios de riesgo.**
- **Métricas objetivas.**
- **Procedimientos de verificación.**
- **Obligaciones documentales.**
- **Evidencias auditables.**

Esto evita marcos de control incompletos o arbitrarios, proporcionando una base sólida, sistemática y defendible ante auditorías internas, externas o reguladores.

5.4.

APLICABILIDAD PRÁCTICA DE CONTROLES DE PRIVACIDAD Y SEGURIDAD

La privacidad y la seguridad constituyen el núcleo operativo del Dominio 2 del Repositorio de Riesgos IA del MIT y son las áreas donde convergen con mayor claridad la gobernanza, la ingeniería y el aseguramiento del ciclo de vida de la IA.

DELIMITACIÓN DEL PROBLEMA: QUÉ SIGNIFICA “PRIVACIDAD Y SEGURIDAD” EN IA

A diferencia de los sistemas tradicionales, la privacidad en IA no solo depende de los datos, sino del comportamiento del modelo. Los riesgos clave son:

- Exposición de datos sensibles (salidas del modelo, logs, trazas).
- Reidentificación e inferencias no deseadas.
- Uso indebido de datos (finalidad, entrenamiento, ampliación funcional).

En seguridad, el foco se amplía a vulnerabilidades del pipeline, dependencias, superficie de ataque y técnicas específicas contra modelos (extracción, envenenamiento, evasión, manipulación de entradas).

Esta distinción ayuda a responder dos preguntas centrales:

- ¿Qué información podría comprometerse y con qué impacto? (privacidad, RGPD, secretos empresariales).
- ¿Cómo podría un adversario provocar ese compromiso? (seguridad, integridad del modelo, disponibilidad).

En la práctica, ambos planos se entrelazan.

MARCO DE IMPLANTACIÓN DE CONTROLES

Sobre la base del Repositorio de Riesgos de IA del MIT, este Marco de Control utiliza sus taxonomías como referencia inicial para la identificación y clasificación de riesgos, sin perjuicio de incorporar riesgos adicionales o emergentes cuando resulte necesario.

Este enfoque resulta especialmente potente cuando se conecta con un marco de gestión continuo como el NIST AI RMF, que permite contextualizar el riesgo, incorporar a las partes interesadas relevantes y adaptar la gestión a escenarios cambiantes.

A efectos prácticos, se recomienda utilizar la siguiente lógica de “encaje” entre marcos y referencias complementarias:

- **Repositorio de Riesgos de IA del MIT:** estructura conceptual y lenguaje común para la identificación y clasificación inicial de riesgos.
- **NIST AI RMF:** proceso de gestión del riesgo que define quién decide, cómo y con qué criterios, permite mapear contexto y stakeholders, medir eficacia e impacto, y gestionar los riesgos mediante tratamiento, aceptación, transferencia o mitigación.
- **OWASP (LLM/GenAI Security):** guía orientada a vulnerabilidades recurrentes en aplicaciones de IA (especialmente en aplicaciones basadas en LLM con herramientas, conectores, RAG o agentes) proporcionando patrones de mitigación accionables para equipos de desarrollo y AppSec.
- **MITRE ATLAS:** base de conocimiento para modelar tácticas y técnicas adversarias contra sistemas de IA, útil para threat modeling, red teaming y el diseño de pruebas adversariales repetibles.

Este encaje permite articular un marco de implantación de controles que no se limita a listas estáticas de riesgos, sino que integra gobernanza, privacidad y seguridad como un puente continuo entre la supervisión estratégica y la ejecución técnica, manteniendo la capacidad de adaptación frente a riesgos emergentes.

CONTROLES POR CICLO DE VIDA: QUÉ IMPLANTAR Y CUÁNDO IMPLANTARLO

Para asegurar privacidad y seguridad en los sistemas de IA, los controles deben alinearse con cada fase del ciclo de vida: diseño, desarrollo, validación, despliegue, operación y retiro. Esta estructura evita riesgos tardíos y permite asignar responsabilidades claras.

(A) Diseño: Privacidad y seguridad "por diseño" como requisito

En esta fase se establecen los cimientos de seguridad y privacidad:

- **Clasificación del sistema y de los datos**
  - Determinar si se usan datos personales, sensibles o información regulada.
  - Definir zonas de procesamiento y flujos permitidos.

- **Modelado de amenazas específico de IA**
  - Incluir adversarios relevantes (externos, insiders, proveedores, usuarios finales, prompt attackers, etc).
  - Usar MITRE ATLAS para cubrir amenazas específicas de modelos y pipelines.

- **Requisitos de privacidad**
  - Minimización estricta (qué campos son estrictamente necesarios).
  - Separación de finalidades (entrenamiento vs. inferencia vs. analítica).
  - Políticas de retención y tratamiento (incluidos logs y embeddings).
  - Reglas de uso de datos para entrenamiento o afinamiento.

- **Requisitos de seguridad**
  - Objetivos CIA adaptados a IA: confidencialidad del prompt/datos, integridad del modelo y pipeline, disponibilidad del servicio.
  - Garantizar "integridad conductual": evitar respuestas o acciones no autorizadas.

Evidencias recomendadas (auditables): clasificación del sistema, threat model con escenarios ATLAS, requisitos de privacidad/seguridad y criterios de aceptación.

(B) Desarrollo: gobernanza de datos y endurecimiento del pipeline

Aquí se construyen las garantías prácticas.

- **Gobernanza y trazabilidad de datos**
  - Linaje de datasets (origen, licencias, base jurídica, transformaciones).
  - Registro de versiones (datasets, prompts, modelos).
  - Accesos y segregación de funciones.

- **Minimización y seudonimización operativa**
  - Eliminación/mascarado de identificadores directos cuando no sean necesarios.
  - Reducción de exposición (tokenización, no persistencia en inferencia).

- **Seguridad de componentes y dependencias**
  - Endurecimiento de entornos de entrenamiento/inferencia.
  - Evaluación de bibliotecas y contenedores.
  - Gestión segura de secretos (API keys, credenciales de conectores, system prompts como activo sensible).

- **Controles específicos en LLMs: OWASP mantiene una línea de trabajo sobre riesgos críticos de aplicaciones con modelos de lenguaje (Top 10 LLM),**
  - Separación clara de prompts del sistema/usuario.
  - Validación estricta de entradas/salidas y contratos de herramientas.
  - Evitar acceso directo del modelo a secretos.
  - Validación contextual cuando la salida desencadena acciones.

Evidencias recomendadas: matriz de acceso a datos, registros de versionado, políticas de logging, guías de desarrollo seguro IA/LLM, controles de secretos y dependencias.

(C) Validación: pruebas de privacidad y seguridad antes de producción

La validación debe incluir pruebas adversariales y de fuga.

- **Pruebas de fuga de información (PII/Secret Leakage Testing)**
  - Ataques controlados para provocar que el sistema revele PII: prompts de extracción, ataques de "repetición", preguntas indirectas, jailbreak attempts
  - Uso de datos señuelo para detectar memorización.
  - Revisión de logs para asegurar cumplimiento de política.

- **Pruebas de robustez ante ataques**
  - Inyección de prompt, manipulación contextual y documentos contaminados.
  - Pruebas de extracción del modelo o del entrenamiento, si aplica.
  - Red teaming basado en ATLAS.

- **Evaluación de controles de acceso y autorización**
  - » Verificar separación por tenant/rol.
  - » Validar que herramientas no permiten escalada por texto.

- **Criterios de salida a producción (Gate de Seguridad/Privacidad)**
  - » Aprobar umbrales máximos de fuga (tasa y severidad).
  - » Aprobar resultados de pruebas adversariales.
  - » Validar documentación mínima para auditoría (versiones, datasets, decisiones).

Evidencias recomendadas: informe de pruebas adversariales, reporte de leakage, trazabilidad de correcciones, decisión formal de aprobación.

(D) Despliegue: controles operativos y trazabilidad

- **Hardening del entorno de inferencia**
  - » Aislamiento de contenedores/servicios y políticas de red restrictivas.
  - » Control de egress para evitar exfiltración.
  - » Protección de endpoints (rate limiting, autenticación fuerte, detección de patrones de abuso).

- **Registro y trazabilidad equilibrada**
  - » Logging mínimo viable sin almacenar PII.
  - » Separación entre logs técnicos y contenido del prompt.
  - » Mecanismos antifraude (tamper evidence).

- **Controles de salida**
  - » Filtrado de PII cuando corresponda.
  - » Aplicación automática de políticas antes de entregar respuestas.

- **Control de herramientas y agentes**
  - » Ejecución en sandbox.
  - » Allow list dinámicas y verificaciones deterministas.

Evidencias recomendadas: arquitectura de producción, políticas de logging y redacción, pruebas de configuración, runbooks de despliegue.

(E) Operación: monitorización continua, KRIs y respuesta a incidentes

- La operación es donde el riesgo de privacidad y seguridad se materializa con mayor frecuencia (usuarios reales, datos reales, *drift*, cambios de contexto). Definición de KRIs específicos de privacidad y seguridad
  - » Tasa de detección de PII en salida (por 1.000 interacciones) y severidad.
  - » Número de intentos de prompt injection detectados/bloqueados.
  - » Tasa de consultas anómalas (indicador de extracción o enumeración).
  - » Eventos de acceso denegado por políticas de autorización (posible abuso).
  - » Incidentes de data leakage por canal de logs/telemetría (detección interna).
  - » Resultados de red team re-test (mensual/trimestral) frente a técnicas ATLAS seleccionadas.

- **Monitorización de drift con impacto en seguridad**
  - » Drift puede aumentar alucinaciones o reducir filtros.
  - » Alertas de degradación en comportamiento seguro.

- **Gestión de incidentes de IA**
  - » Clasificación (privacidad, seguridad o mixto).
  - » Contención segura y preservación de evidencias.
  - » Correcciones: prompts, autorizaciones, rollback o retraining.
  - » Notificaciones según RGPD u otras normas.

- **Escalado y whistleblowing interno**
  - » Canales protegidos para reportar anomalías.
  - » Escalado a Comité de Riesgos/Ética.

Evidencias recomendadas: cuadro de mando KRI, playbooks de incidentes IA, registros de eventos, informes trimestrales al órgano de riesgos.

(F) Retiro (fin de vida): clausura segura y lecciones aprendidas

El marco establece retiro con documentación histórica y post-mortem.

- Revocación de credenciales y borrado seguro de caches/embeddings
- Cierre de accesos y revisión de recursos residuales
- Custodia controlada de evidencias para auditorías (con retención mínima y controlada).
- Post mortem y retroalimentación a los “gates” de futuros ciclos.

ROLES Y RESPONSABILIDADES: GOBERNAR PARA QUE LOS CONTROLES QUE REALMENTE EXISTAN EN MATERIA DE SEGURIDAD Y PRIVACIDAD

Este apartado define cómo garantizar que los controles de seguridad y privacidad no se queden en buenas intenciones, sino que se conviertan en obligaciones con responsables, evidencias y trazabilidad.

Para ello, se propone un esquema de gobierno basado en una matriz RACI y en la lógica clásica de 1.ª, 2.ª y 3.ª línea de defensa, además del Consejo.

La especialización de responsabilidades evita un fallo recurrente: que seguridad y privacidad acaben tratándose como recomendaciones voluntarias en lugar de como requisitos formales y verificables. Aplicación práctica (síntesis):

- **1.ª línea (Equipos de desarrollo de IA y dueños del proceso):** ejecutan los controles técnicos, garantizan la configuración segura, aplican políticas de uso de datos y documentan la evidencia operativa. Son responsables directos del funcionamiento seguro del sistema.
- **2.ª línea (Riesgos / Comité de Ética y Riesgos de IA):** define la política de IA responsable, aprueba criterios de aceptación del riesgo, establece KRIs y valida el tratamiento propuesto por la 1.ª línea. Supervisa el cumplimiento normativo y ético.
- **3.ª línea (Auditoría interna):** evalúa el diseño y la efectividad real de los controles, revisa evidencias, identifica brechas y reporta hallazgos independientes.
- **Consejo/Alta Dirección:** recibe el perfil agregado de riesgo y tendencias, define el apetito de riesgo y aprueba el nivel de tolerancia a riesgos residuales.

Este modelo se alinea con el NIST AI RMF, donde Govern funciona como una capa transversal que debe integrarse en todas las funciones restantes (*Map, Measure, Manage*).

CONJUNTO MÍNIMO DE CONTROLES RECOMENDADOS (*BASELINE*) PARA DOMINIO 2

Para garantizar coherencia operativa en toda la organización, se recomienda establecer un baseline de controles obligatorios para cualquier sistema de IA que procese información corporativa o datos personales.

Para sistemas críticos, se propone un conjunto ampliado de controles adicionales.

- **Baseline de privacidad**
  - 1. Inventario de datos y finalidades (qué datos, por qué, base jurídica, aprobaciones).
  - 2. Minimización y redacción automática de logs y trazas de prompts.
  - 3. Control de acceso por rol y segregación por tenant.
  - 4. Pruebas de fuga antes del despliegue y periódicamente.
  - 5. Políticas de retención y borrado seguro, incluyendo embeddings y cachés.

- **Baseline de Seguridad**
  - Threat modeling específico de IA (incluye MITRE ATLAS).
  - Hardening de endpoints, limitación de solicitudes, autenticación y control de egress.
  - Seguridad de conectores y herramientas externas (scopes mínimos, sandbox, validación determinista).
  - Monitorización y detección de anomalías, inyecciones y extracción de información.
  - Gestión de incidentes de IA y runbooks (contención, rollback, comunicación).

- **Controles ampliados (alto riesgo / alto impacto)**
  - Red teaming trimestral, basado en ATLAS y escenarios propios del negocio.
  - Revisión independiente previa al despliegue (2.ª línea + auditoría técnica).
  - Aislamiento reforzado / secure enclaves para inferencia cuando sea necesario.
  - Técnicas avanzadas de mitigación contra memorización e inferencias, según arquitectura (entrenamiento, afinamiento, RAG).

Estos controles se integran de forma natural en los **mapas de riesgo**, en la definición de **KRIs** y en el **escalado automático** cuando se superan umbrales, reforzando un modelo de control completo y verificable.

EJEMPLO PRÁCTICO DE APLICACIÓN DEL MARCO DE CONTROL DE SEGURIDAD Y PRIVACIDAD DE LA IA

A continuación, se presentan los elementos mínimos necesarios para operacionalizar el marco de control de seguridad y privacidad en un sistema típico, como un asistente interno con RAG o una herramienta de apoyo a la decisión.

El enfoque combina la taxonomía del MIT (identificación completa del riesgo) con el ciclo de vida de gestión propuesto por el NIST AI RMF.

1. Identificación (MIT Repository)

- Seleccionar los riesgos relevantes del Dominio 2.
- Etiquetar cada riesgo según causalidad: humano/IA, intencional/no intencional, pre/post despliegue.
- Utilizar esta clasificación para priorizar y no omitir riesgos críticos.

2. Mapeo (NIST AI RMF – Map)

- Determinar actores, activos, datos, procesos y escenarios de uso.
- Evaluar impactos potenciales: cumplimiento normativo, reputación, impacto operativo, confidencialidad o integridad del entorno.

3. Medición (NIST AI RMF – Measure)

- » Definir métricas y umbrales de fuga, extracción o manipulación.
- » Diseñar pruebas adversariales y test de PII leakage.
- » Establecer criterios de éxito y límites para permitir el paso a producción.

4. Gestión (NIST AI RMF – Manage)

- » Seleccionar estrategia de tratamiento: mitigar, aceptar, transferir o evitar el riesgo.
- » Implementar los controles correspondientes (técnicos, organizativos, procedimentales).
- » Dejar trazabilidad de la aceptación de riesgo residual.

5. Gobierno (NIST AI RMF – Govern)

- » Asignar roles y responsabilidades mediante un RACI definido.
- » Fijar KRIs (indicadores clave de riesgo) y su periodicidad.
- » Establecer calendario de reportes, revisiones y auditorías internas.

Resultados esperados (artefactos de evidencia)

- » Registro del sistema de IA, incluyendo propósito, criticidad, responsable, fase del ciclo de vida, dependencias y datos empleados.
- » Paquete de evidencias para auditoría, con versiones de modelos, datasets utilizados, parámetros clave y decisiones de diseño.
- » Cuadro de mando de KRIs y canales de escalado formal, alineados con la estructura de gobierno de riesgos.

CÓMO DEMOSTRAR EFICACIA DEL MARCO DE CONTROL (RENDICIÓN DE CUENTAS EN SEGURIDAD Y PRIVACIDAD)

La clave de la aplicabilidad práctica es que los controles no sean meramente declarativos, sino que puedan demostrarse como ejecutados, efectivos y auditables.

Tal como se ha indicado, el nivel operativo debe proporcionar evidencias verificables que permitan evaluar el cumplimiento en materia de seguridad y privacidad.

Para estos dominios, un auditor solicitará típicamente los siguientes conjuntos de evidencias:

- **Evidencia de diseño:** threat model IA (incluyendo ATLAS), requisitos de privacidad/seguridad, clasificación del dato.
- **Evidencia de validación:** resultados de pruebas de fuga, pruebas adversariales, acciones correctivas y re-test.
- **Evidencia de operación:** KRIs, incidentes, runbooks ejecutados, mejoras
- **Evidencia de gobierno:** asignación RACI, reportes al comité/Consejo, aceptación formal de riesgos residuales.

5.5.

APLICABILIDAD PRÁCTICA DE CONTROLES EN LA CADENA DE SUMINISTRO TECNOLÓGICO

La cadena de suministro tecnológico es hoy uno de los principales multiplicadores de riesgo en ciberseguridad, continuidad y cumplimiento. El perímetro efectivo de la organización ya no coincide con su infraestructura propia: incluye fabricantes de hardware y firmware, desarrolladores de software (incluido open source), proveedores cloud, operadores de telecomunicaciones, servicios gestionados y, de manera creciente, proveedores de datos y de IA (modelos, plataformas MLOps, APIs, repositorios de modelos, servicios de evaluación y "guardrails"). En este contexto, la gestión del riesgo exige que los controles "viajen" con el suministro mediante requisitos, verificación, monitorización y respuesta, no únicamente mediante cláusulas contractuales genéricas.

En el ámbito de la IA, la dependencia de proveedores de modelos "como servicio" introduce una particularidad adicional: el componente suministrado no es estático, sino un servicio vivo sujeto a cambios frecuentes (versiones de modelo, filtros, infraestructura, políticas de retención, parámetros de seguridad), lo que puede diluir la rendición de cuentas si no existe una gobernanza explícita y mecanismos de supervisión continua. Por tanto, la aplicabilidad práctica de los controles en cadena de suministro debe abarcar tanto la cadena de suministro tecnológica tradicional como la cadena de suministro específica de datos, modelos y servicios de IA.

Adicionalmente, es relevante considerar que los propios sistemas de IA pueden actuar como habilitadores de la gestión del riesgo , apoyando actividades como el análisis de grandes volúmenes de información de seguridad, la detección de patrones anómalos en proveedores y servicios, la evaluación continua de configuraciones y dependencias, o el refuerzo de capacidades de supervisión en materia de seguridad y privacidad. Este enfoque refuerza una visión bidireccional de la IA en la cadena de suministro: como fuente potencial de riesgo, pero también como herramienta para mejorar su identificación, monitorización y control.

ALCANCE OPERATIVO: QUÉ SE CONTROLA (CUATRO CAPAS)

A efectos de implantación, conviene descomponer la cadena de suministro en cuatro capas, con riesgos y evidencias diferenciadas:

- **Componentes:** hardware, firmware, equipos de red, endpoints, OT/IoT y dispositivos especializados (incluyendo aceleradores).
- **Software y dependencias:** software comercial, componentes open source, librerías, contenedores, imágenes base, gestores de paquetes, herramientas de desarrollo, pipelines CI/CD y marketplaces.
- **Servicios operados:** cloud (IaaS/PaaS/SaaS), MSP/MSSP, soporte remoto, operación delegada, telecomunicaciones y servicios críticos de terceros.
- **Datos e IA:** datasets de terceros, etiquetado, modelos preentrenados, APIs de modelos, hubs/repositorios, MLOps, RAG/embeddings, conectores, herramientas/agentes, soluciones de filtrado/guardrails y servicios de evaluación/red teaming.

Esta clasificación permite aplicar controles proporcionales y evita tratar como "proveedor estándar" a suministros que, por criticidad o exposición, deben ser gestionados como activos críticos.

PRINCIPIO RECTOR: CONTROLES POR CICLO DE VIDA DEL SUMINISTRO

La gestión eficaz no debe limitarse a la fase de compra. Se recomienda estructurar los controles por fases del ciclo de vida del suministro, de forma que cada etapa tenga requisitos, evidencias y responsables:

- **Planificación y contratación:** clasificación del suministro, definición de requisitos, due diligence y verificación proporcional.
- **Ingeniería e integración:** aseguramiento de dependencias, pipelines y configuraciones seguras.
- **Commissioning y aceptación:** puerta de entrada a producción basada en evidencias ("no entra sin pruebas").
- **Operación y mantenimiento (O&M):** control continuo de accesos remotos, cambios, vulnerabilidades, continuidad y monitorización.
- **Incidentes y crisis:** coordinación con terceros para contención, evidencias, comunicación y recuperación.
- **Salida (exit):** revocación de accesos, portabilidad, borrado seguro, sustitución y lecciones aprendidas.

CONTROLES EN PLANIFICACIÓN Y CONTRATACIÓN (TPRM/C-SCRM DESDE LA RFP)

A) Clasificación por criticidad (*tiering*). Establecer un sistema de criticidad del suministro (crítico/alto/medio/bajo) en función de impacto en confidencialidad, integridad, disponibilidad, cumplimiento, continuidad y dependencia (lock-in). En suministros de IA, añadir variables específicas: tipo de datos tratados (incluyendo datos sensibles), posibilidad de reutilización de prompts/datos para mejora, exposición a cambios de versión del modelo y potencial de automatización de decisiones.

B) Anexo contractual de seguridad parametrizado por criticidad. Definir un anexo de seguridad estándar que se ajuste al nivel de criticidad e incluya, como mínimo:

- **Identidad y acceso:** autenticación fuerte, control de accesos privilegiados, mínimo privilegio, segregación de entornos, registro de sesiones y revisiones periódicas.
- **Gestión de vulnerabilidades:** plazos de parcheo por severidad, notificación proactiva de vulnerabilidades relevantes, evidencia de remediación y gestión de fin de soporte.
- **Gestión de cambios:** preaviso de cambios materiales, ventanas de mantenimiento, criterios de prueba y rollback; obligación de notificar cambios que afecten seguridad, privacidad o disponibilidad.
- **Subcontratación y cuarta parte:** obligación de aplicar controles equivalentes a subproveedores, transparencia de la cadena, y derecho a conocer subprocesadores/servicios críticos utilizados.
- **Notificación y respuesta a incidentes:** SLAs de notificación, cooperación técnica y forense, preservación de evidencias, y coordinación en comunicaciones externas cuando aplique.
- **Derechos de auditoría y evidencias:** acceso a evidencias de controles (documentales y/o técnicas), auditorías de terceros y, cuando proceda, auditorías específicas.
- **Protección de datos:** cifrado, retención, transferencias, borrado, portabilidad y limitación de finalidad.
- **Continuidad y resiliencia:** RTO/RPO, pruebas de recuperación, disponibilidad mínima, dependencias críticas y planes de contingencia.

C) Transferencia de riesgo con SLAs sin externalizar responsabilidad. Utilizar SLAs, responsabilidades contractuales y mecanismos de cobertura (por ejemplo, seguros) para asignar obligaciones verificables. Sin embargo, mantener el principio de que el riesgo operativo no desaparece por externalización: debe existir supervisión, verificación y capacidad de contingencia.

D) Verificación proporcional (programa escalonado). Definir una matriz "criticidad → verificación" que determine el método y la frecuencia de evaluación:

- **Medio:** declaración responsable + evidencias mínimas + revisión periódica.
- **Alto:** evidencias ampliadas + informes de auditoría + pruebas puntuales.
- **Crítico:** auditoría específica de alcance + verificación técnica (integración, configuración, continuidad y controles de acceso) + ejercicios de crisis con el proveedor.

CONTROLES EN INGENIERÍA, DESARROLLO E INTEGRACIÓN (*SOFTWARE SUPPLY CHAIN Y AI SUPPLY CHAIN*)

A) Cadena de suministro de software: integridad y control de dependencias. Implantar controles para reducir la probabilidad de introducción de código malicioso o vulnerabilidades explotables:

- **Integridad de artefactos:** firmas/verificación, repositorios controlados, control de versiones y trazabilidad de compilación.
- **Gestión de dependencias:** inventario de componentes, control de versiones, evaluación de librerías críticas, restricciones a repositorios no autorizados.
- **Seguridad del pipeline CI/CD:** segregación de permisos, protección de credenciales, control de runners, revisión de pipeline-as-code y registros de cambios.
- **Gestión de secretos:** almacenamiento seguro, rotación y mínima exposición (evitar secretos embebidos en código, contenedores o configuraciones).

B) Cadena de suministro de IA: datos, modelos y *MLOps*. La cadena de suministro de IA introduce riesgos de integridad y procedencia de datos y modelos. Controles aplicables:

- **Trazabilidad de datasets:** origen, licencias, cadena de custodia, criterios de inclusión/exclusión, procesos de limpieza y deduplicación.
- **Validación de integridad y calidad de datos:** controles estadísticos, detección de anomalías, muestreo por criticidad y cuarentena de fuentes no verificadas.
- **Control de modelos preentrenados:** allow-list de fuentes y repositorios, verificación de hash/firma, evaluación previa y cuarentena antes de su integración.
- ***MLOps* seguro:** control de versiones del modelo, trazabilidad de experimentos, segregación de entornos, permisos mínimos para despliegue y aprobación formal de promociones a producción.
- **Evaluación adversarial previa:** pruebas de robustez, abuso y comportamiento no deseado, con umbrales claros y revalidación ante cambios.

C) Pruebas guiadas por amenaza (threat-driven assurance). Diseñar pruebas de integración y aceptación a partir de escenarios adversarios plausibles: compromiso de actualizaciones, manipulación de repositorios, dependencia maliciosa, envenenamiento de datos/modelo, abuso de APIs y degradación del servicio. Este enfoque permite que el control no sea meramente documental, sino verificable en condiciones realistas.

COMMISSIONING Y ACEPTACIÓN: PUERTA DE ENTRADA A PRODUCCIÓN BASADA EN EVIDENCIAS

Implantar un *"Security & Trust Gate"* que condicione el paso a producción a evidencias mínimas, escaladas por criticidad:

- **Documentación mínima:** arquitectura, inventario de componentes, diagrama de flujos de datos, controles aplicados, roles y responsabilidades operativas.
- **Resultados de pruebas:** configuración segura, control de accesos, pruebas de APIs, pruebas de integración, verificación de continuidad (según criticidad).
- **Evidencias de gestión de cambios y versiones:** trazabilidad de la versión que se despliega y de sus dependencias.
- **Para IA/GenAI:** evidencias de evaluación de riesgos específicos (p. ej., abuso, fuga de información, robustez frente a entradas adversarias) y verificación de controles de datos/modelos.

OPERACIÓN Y MANTENIMIENTO: ACCESOS, VULNERABILIDADES, CAMBIOS Y MONITORIZACIÓN CONTINUA

A) Accesos remotos de terceros

- Acceso privilegiado con control de sesiones, mínimo privilegio y just-in-time.
- Registro y supervisión de sesiones; canales bastionados (sin acceso directo a producción).
- Revisiones periódicas de cuentas de terceros y revocación inmediata tras finalización del servicio.

B) Vulnerabilidades, obsolescencia y cambios

- Inventario vivo de suministros críticos y sus dependencias.
- Plazos de parcheo y evidencias de remediación; gestión de fin de vida/fin de soporte.
- Evaluación de impacto de cambios materiales. En IA, incluir cambios de versión del modelo, cambios de filtros/guardrails y cambios en políticas de tratamiento/retención.

C) Supervisión continua del proveedor

Convertir la evaluación en un programa recurrente: revisión de evidencias, auditorías periódicas, verificación técnica puntual y monitorización de señales de riesgo (incidentes públicos del proveedor, degradación del servicio, cambios no notificados).

D) Evitar dilución de responsabilidad

Asegurar una asignación explícita de responsabilidades (RACI) para cambios, incidentes y comunicaciones; transparencia de subcontratas; y simulacros conjuntos para servicios críticos.

RESPUESTA A INCIDENTES Y GESTIÓN DE CRISIS CON TERCEROS RELACIONADOS CON SERVICIOS IA

Integrar un módulo específico de "incidentes de terceros IA" en el Plan de Respuesta a Incidentes y en el Plan de Gestión de Crisis:

- **Playbooks por tipo de proveedor** (cloud, MSP, software, IA como servicio), con contactos, escalado 24/7 y responsabilidades definidas.
- **SLAs de notificación y requisitos de preservación de evidencias** (logs, trazas, artefactos), minimizando la exposición de datos sensibles.
- **Coordinación de contención:** capacidad de desconexión controlada, modos degradados y planes alternativos.
- **Simulacros (tabletop) con escenarios plausibles** (compromiso de actualización, fuga de datos desde tercero, envenenamiento de datos/modelo, indisponibilidad prolongada).
- **Recuperación y aprendizaje:** análisis post-incidente, refuerzo contractual y técnico, re-clasificación del proveedor si procede y actualización de controles.

CONTROLES MÍNIMOS ESPECÍFICOS PARA IA COMO SERVICIO (CHECKLIST OPERATIVO)

Cuando el suministro incluya modelos/GenAI como servicio o modelos/datasets de terceros, se recomienda un checklist mínimo:

- **Cláusulas de datos y prompts:** retención, uso para mejora, aislamiento por tenant, cifrado, borrado y portabilidad.
- **Control de cambios del modelo:** identificación de versión, notificación de cambios materiales, ventana de pruebas y mitigación/rollback operativo.
- **Gate de seguridad previo:** pruebas de abuso, fugas, robustez y comportamiento no deseado; límites claros para automatización.
- **Monitorización de comportamiento:** indicadores de drift, abuso, degradación de controles y anomalías de uso.
- **Fuentes confiables y verificación:** allow-list de repositorios/modelos/datasets, hash/firma y cuarentena de componentes externos.
- **Responsabilidades y escalado:** RACI y SLAs específicos para incidentes de IA, incluyendo soporte técnico y evidencias.

MEDICIÓN Y AUDITABILIDAD: KPIS/KRIS RECOMENDADOS

Para demostrar eficacia y sostener el control, se recomienda definir indicadores mínimos:

- **Cobertura:** porcentaje de proveedores críticos evaluados y con evidencias actualizadas.
- **Vulnerabilidades:** cumplimiento de SLAs de parcheo/remediación por severidad.
- **Cambios:** número de cambios materiales no notificados detectados (incluye versiones de modelo/API).
- **Incidentes:** número de incidentes originados o agravados por terceros y tiempos de notificación/contención.
- **Dependencia:** concentración de proveedores críticos y tiempo estimado de sustitución (exit time).
- **Para IA:** eventos de fuga/abuso, resultados de revalidación adversarial y degradación de comportamiento seguro.

Con este enfoque, la cadena de suministro IA deja de ser un "anexo contractual" y se convierte en un proceso de control continuo: clasificar, exigir, verificar, aceptar con evidencias, operar con supervisión, responder coordinadamente y planificar la salida. Este ciclo es el que permite reducir exposición real y evitar que dependencias críticas (incluyendo servicios de IA) introduzcan riesgos no gobernados en el mapa corporativo.

5.6.

EVALUACIÓN Y MEJORA DEL MARCO DE CONTROL DE RIESGOS IA

La evaluación y mejora del marco de control de riesgos de IA tiene por objetivo asegurar que marco de control de riesgos funciona como un sistema operativo de control (no como un catálogo estático): que los controles se activan en los gates del ciclo de vida, que generan evidencias verificables, que se sostienen con medición continua, y que evolucionan ante cambios de modelos, datos, exposición y terceros. En IA, el riesgo se comporta como un sistema dinámico: deriva del modelo (drift), ampliación de superficies de ataque (nuevas herramientas, conectores o agentes), y dependencia creciente de proveedores (incluida IA "como servicio"). Por ello, la mejora debe estar diseñada como un circuito cerrado: medir → evaluar → corregir → revalidar → estandarizar.

ALCANCE DE LA EVALUACIÓN: QUÉ SE EVALÚA Y CON QUÉ CRITERIO

La evaluación del marco se realiza en tres niveles complementarios:

- **Eficacia de controles (*control effectiveness*).** Verifica si los controles implantados previenen, detectan y corrigen de forma real los riesgos definidos, especialmente en las áreas críticas desarrolladas en 5.4 (privacidad y seguridad) y 5.5 (cadena de suministro).
- **Cumplimiento operativo del proceso (*process compliance*).** Verifica si el marco se ejecuta según el modelo definido: registro de sistemas, clasificación por criticidad, aplicación de gates, evidencias mínimas, gestión de excepciones con caducidad, y gestión de cambios materiales.
- **Adecuación y cobertura *(coverage and fitness-for-purpose)***. Verifica si el marco cubre el inventario real de sistemas de IA, sus dependencias y su exposición; y si el nivel de control es proporcional a criticidad y riesgo.

El marco se considera "adecuado" cuando existe trazabilidad completa para cada sistema (finalidad–datos–modelo–dependencias–controles–pruebas–incidentes–responsables–decisiones) y cuando los indicadores agregados evidencian estabilidad o mejora, sin acumulación de deuda (excepciones antiguas, controles sin evidencia, proveedores críticos sin evaluación).

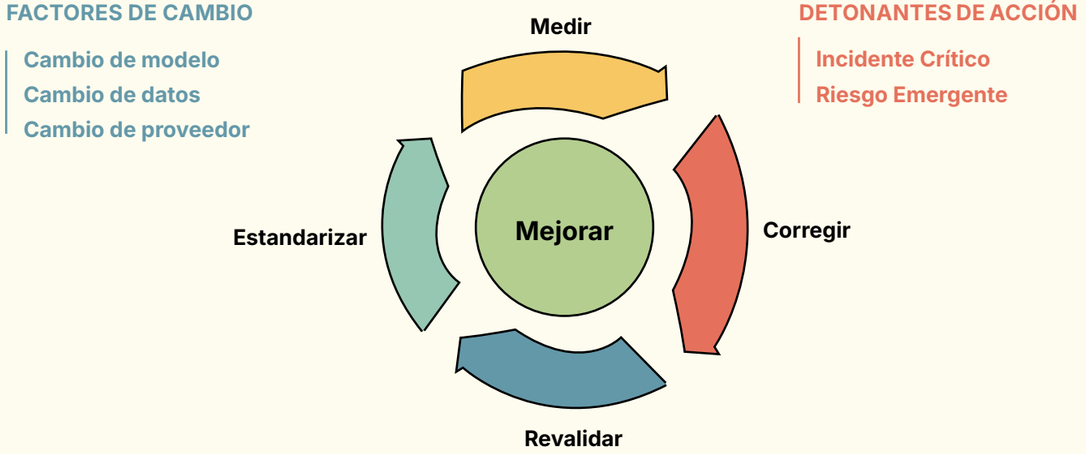


Figura 9. Ciclo de mejora continua de control de riesgos de IA

GOBIERNO Y CADENCIA: QUIÉN EVALÚA, CON QUÉ FRECUENCIA Y CON QUÉ AUTORIDAD

La mejora continua se articula mediante una cadencia de revisión integrada en el modelo de gobierno descrito en el Punto 5:

A) Revisión operativa (semanal/mensual – 1ª línea). Propietarios del sistema/modelo y operación revisan: desviaciones de control, hallazgos técnicos, alertas, resultados de pruebas recurrentes, incidencias con terceros, backlog de remediación y cumplimiento de evidencias operativas.

Salida mínima: acciones correctoras con responsable y fecha; y actualización de evidencias.

B) Revisión táctica de riesgo (mensual/trimestral – 2ª línea). Riesgos/Cumplimiento (y, cuando aplique, DPO y CISO) revisan: KRIs/KPIs, riesgos residuales, volumen de excepciones, cambios materiales, desempeño de proveedores críticos, y cumplimiento de gates.

Salida mínima: re-priorización del plan de tratamiento, decisiones de aceptación/mitigación, y aprobación/rechazo de excepciones significativas.

C) Revisión estratégica (semestral/anual – Alta Dirección/Consejo). Se evalúa: perfil agregado de riesgo IA, tendencias, incidentes relevantes, concentración de dependencia, postura de resiliencia, necesidades de inversión y ajuste del apetito de riesgo.

Salida mínima: directrices, tolerancias, prioridades presupuestarias y medidas estructurales (p. ej., reducción de concentración de proveedor, refuerzo de capacidades internas).

CUADRO DE MANDO: INDICADORES MÍNIMOS (KRIS/KPIS) Y UMBRALES DE ACTUACIÓN

La evaluación debe sostenerse sobre indicadores que midan tanto riesgo como control. Se recomienda estructurarlos por los dominios operativos del Punto 5:

A) Indicadores de privacidad y seguridad (alineados con 5.4)

- Incidentes o eventos de exposición de información (frecuencia, severidad, canales: salida, logs, conectores).
- Intentos de abuso/adversarial detectados y bloqueados (y tasa de reintentos).
- Cobertura de pruebas previas a producción (porcentaje de despliegues que superan el Gate de Validación con evidencias completas).
- Tiempo medio de remediación de hallazgos críticos (y porcentaje de revalidaciones completadas).
- Cobertura de monitorización y alertas (porcentaje de componentes con telemetría activa: endpoints, herramientas, conectores, repositorios).

B) Indicadores de cadena de suministro (alineados con 5.5)

- Porcentaje de proveedores críticos evaluados conforme al plan y con evidencias actualizadas.
- Cumplimiento de SLAs de parcheo, notificación de incidentes y preaviso de cambios.
- Número de cambios materiales no notificados detectados (incluye cambios de versión de modelo/API, cambios de subprocesadores o modificaciones de condiciones de servicio).
- Incidentes atribuibles a terceros (frecuencia, impacto, tiempos de contención coordinada).
- Métrica de dependencia: concentración por proveedor y tiempo estimado de sustitución (exit time).

C) Indicadores de ejecución del marco (madurez operativa)

- Porcentaje del inventario IA registrado y clasificado por criticidad.
- Porcentaje de sistemas con gates aplicados y evidencias mínimas completas.
- Volumen y antigüedad media de excepciones (y porcentaje con plan de remediación y caducidad).
- Porcentaje de cambios materiales que pasaron por reevaluación y re-test.

Umbrales y acciones automáticas

Se recomienda definir umbrales que disparen medidas predefinidas: escalado a 2ª línea, congelación temporal de despliegues, revisión extraordinaria de proveedores, incremento de pruebas adversariales, o reducción funcional (modo degradado) para sistemas de alto riesgo.

ASEGURAMIENTO DEL MARCO DE CONTROL DE RIESGOS IA

La evaluación debe combinar tres formas de aseguramiento:

A) Aseguramiento técnico *(security/privacy assurance)*

- Pruebas periódicas de fuga de información, aislamiento de sesiones, control de herramientas, endurecimiento de configuración y control de accesos.
- Evaluaciones adversariales/red teaming proporcionales a criticidad, con re-test tras remediación.
- Verificación de eficacia de monitorización (alertas accionables, reducción de falsos negativos, tiempos de detección).

B) Aseguramiento de proceso *(governance assurance)*

- Verificación de cumplimiento de gates del ciclo de vida: que no se pasa a producción sin evidencias, aprobaciones y criterios de aceptación.
- Revisión de trazabilidad de cambios: datos/modelo/configuración/terceros y su impacto en controles.
- Revisión de gestión de excepciones: justificación, aprobación, caducidad, compensaciones y cierre.

C) Aseguramiento de terceros *(third-party assurance)*

- Revisión periódica de evidencias de proveedores críticos y auditorías.
- Pruebas puntuales o verificaciones técnicas donde aplique (especialmente en accesos remotos, cambios y continuidad).
- Ejercicios de crisis con terceros para validar coordinación y tiempos.

GESTIÓN DEL CAMBIO COMO MOTOR DE REEVALUACIÓN (TRIGGERS DE “CAMBIO MATERIAL”)

La mejora continua depende de una disciplina estricta: no todo cambio requiere reevaluación completa, pero ciertos cambios deben activar obligatoriamente revisión de riesgos y revalidación de controles. Se consideran cambios materiales (activan reevaluación y, como mínimo, re-test de controles críticos):

- Cambio de modelo, de versión del modelo o de proveedor del modelo (incluidos cambios en servicios “como servicio”).
- Incorporación de nuevas fuentes de datos, cambios de finalidad o ampliación de alcance funcional.
- Activación de nuevas herramientas/agentes/conectores o ampliación de permisos (expansión de superficie de acción).
- Cambios en políticas de retención, logging, filtrado o mecanismos de seguridad.
- Cambios de subcontratación o aparición de nuevos subprocesadores en cadena de suministro.
- Aumento significativo de exposición (nuevos canales, más usuarios, acceso público) o automatización de decisiones.

Para estos cambios, el proceso mínimo incluye: actualización del registro del sistema, reevaluación de riesgo, verificación de evidencias y repetición de pruebas críticas antes de despliegue.

GESTIÓN DE INCIDENTES IA, HALLAZGOS Y LECCIONES APRENDIDAS (FEEDBACK AL MARCO)

Todo incidente o hallazgo relevante debe traducirse en mejora del marco, no solo en corrección puntual. El ciclo recomendado:

1. Clasificación del evento (privacidad/seguridad/supply chain; criticidad; impacto).
2. Contención y preservación de evidencias con minimización de exposición de datos.
3. Causa raíz: identificar si el fallo fue de control (diseño/implementación), de proceso (gate omitido), de cambio (no reevaluado) o de tercero (suministro).
4. Acción correctora: remediación técnica y ajustes de proceso (incluyendo actualización de requisitos a proveedores si aplica).
5. Revalidación: re-test de controles; cierre con evidencia.
6. Estandarización: incorporar mejoras a plantillas, gates, checklists y cuadros de mando para evitar recurrencia.

MODELO DE MADUREZ Y PLAN ANUAL DE MEJORA

Para sostener la evolución, se recomienda adoptar un modelo de madurez simple y un plan anual con prioridades:

- **Nivel 1 – Definido:** inventario y políticas básicas; controles mínimos; gates iniciales.
- **Nivel 2 – Implementado:** evidencias consistentes; pruebas recurrentes; gestión de excepciones; TPRM básico para críticos.
- **Nivel 3 – Gestionado:** red teaming por criticidad; monitorización avanzada; gestión de cambios madura; ejercicios con terceros.
- **Nivel 4 – Optimizado:** aseguramiento parcialmente automatizado; mejora predictiva; resiliencia demostrable y reducción sostenida de incidentes.

El plan anual debe priorizarse considerando la reducción de deuda (excepciones antiguas), refuerzo de controles en sistemas de mayor exposición, mejora de control de cadena de suministro (incluido exit plan y robustecimiento de detección y respuesta).

ARTEFACTOS Y EVIDENCIAS MÍNIMAS DEL PROCESO DE MEJORA

Para garantizar auditabilidad y consistencia, el marco debe producir como mínimo:

- Cuadro de mando KRI/KPI con umbrales y acciones asociadas.
- Registro de sistemas IA actualizado (inventario, criticidad, dependencias, responsables).
- Registro de cambios materiales y evidencias de reevaluación/re-test.
- Registro de excepciones con caducidad y plan de remediación.
- Informes de aseguramiento técnico (pruebas, red teaming) y de terceros (evidencias y auditorías).
- Informes de incidentes con causa raíz y acciones estandarizadas.

6.

Relación y apoyo para el Reglamento EU de la IA

6.1.

VISIÓN GENERAL DEL REGLAMENTO EUROPEO DE IA (AI ACT) Y SU RELACIÓN CONCEPTUAL CON EL REPOSITORIO DEL MIT

El AI Act marca obligaciones diferentes para los sistemas de IA según el nivel de riesgo asociado al sistema en cuanto al impacto potencial que pueda tener en la salud, la seguridad y los derechos fundamentales de las personas.

El AI Act establece 4 niveles de riesgo:

- **Riesgo Inaceptable:** usos prohibidos, casos de uso concretos, por ejemplo, perfilado social, identificación biométrica masiva o manipulación subliminal.
- **Sistemas de Alto Riesgo:** aquellos sistemas con potencial de causar daños graves a la salud, la seguridad o los derechos fundamentales. Se identifican por el caso de uso que implementan, ya sea asociado a entornos previamente regulados por la legislación de armonización de la UE (maquinaria peligrosa, juguetes, control de dispositivos médicos, seguridad de la aviación civil, etc.) o a otras actividades sensibles o servicios críticos (acceso a la educación o al empleo, evaluación de solvencia crediticia, control de transportes, suministros de energía, agua, etc.).
- **Sistema de Riesgo Limitado:** los que interactúan con humanos, con obligaciones asociadas a la transparencia, por ejemplo, identificación de chatbots y deepfakes. Un sistema de Alto Riesgo también tendría que cumplir con los requisitos de transparencia si interactúa con humanos.
- **Sistema de Riesgo Bajo:** la gran mayoría de los sistemas, no encuadrados en ninguna de las tres categorías anteriores, sin obligaciones específicas.

Las obligaciones relativas a la gestión del riesgo se centran en los sistemas de Alto riesgo. Estas obligaciones se sustancian en la implantación de un sistema de gestión de riesgos (artículo 9), donde se insta a la identificación de los riesgos asociados al uso previsto del sistema y a las medidas de diseño o de medidas de configuración, orientadas a la mitigación hasta niveles de riesgo residual aceptable. Adicionalmente, dicta una serie de obligaciones para sistemas de Alto Riesgo en los artículos 10 a 15, que precisamente mitigan gran parte de los riesgos identificados en el Repositorio del MIT, según se describe en el capítulo 6.3.

El Repositorio de riesgos del MIT puede servir de base documental para la identificación de riesgos y posibles mitigaciones para cumplir eficazmente estas obligaciones.

De forma complementaria, resulta recomendable contrastar esta aproximación con otros **marcos de referencia consolidados en gestión de riesgos de IA**, que aportan una visión estructurada, transversal y alineada con estándares internacionales ampliamente aceptados[4].

Asimismo, para el análisis de **riesgos específicos de ciberseguridad asociados a sistemas de IA y modelos generativos**, puede ser conveniente profundizar en marcos más especializados, orientados al análisis de amenazas, vectores de ataque y controles técnicos avanzados[5].

[4] *https://www.nist.gov/itl/ai-risk-management-framework*

[5] *https://atlas.mitre.org*
https://genai.owasp.org/

6.2.

APOYO OPERATIVO A LA APLICACIÓN DEL AI ACT

El AI Act sugiere identificar los sistemas de IA según su uso y según el tipo de riesgo que supongan, implementar medidas de mitigación, garantizar calidad, transparencia, supervisión humana y documentar el cumplimiento.

El Repositorio del MIT es una plataforma pública que **recopila más de 1.600 riesgos**, estructurados mediante:

- **Taxonomía causal:** identifica cuándo, cómo y por quién se origina el riesgo (pre o post despliegue, intencionalidad, origen humano o técnico).
- **Taxonomía de dominio:** agrupa los riesgos por áreas temáticas (discriminación, privacidad, seguridad, impacto socioeconómico, desinformación, etc.).

Además, el repositorio incluye un catálogo de controles y mitigaciones, y un AI Incident Tracker para la planificación y respuesta a incidentes.

Apoyos operativos clave para la Seguridad de la Información, Regulación y Compliance

- **Inventario exhaustivo de riesgos documentados:** basado en más de 43 marcos internacionales, permite ampliar y enriquecer el análisis de riesgos habitual con escenarios poco contemplados en la práctica.
- **Taxonomía común y compartida:** permite alinear lenguaje y criterios entre áreas técnicas, legales, de compliance y dirección, facilitando una gobernanza transversal de la IA.
- **Evaluación de riesgos pre y post despliegue:** facilita la identificación de riesgos en todo el ciclo de vida del sistema IA, aportando una base sólida para el diseño de controles y monitoreo continuo.
- **Guía para implementación de controles de mitigación:** el Repositorio del MIT permite definir controles técnicos, de seguridad, de supervisión humana, y de transparencia, alineados con el RIA y otras normas como NIS2, ENS, DORA o ISO/IEC 42001.
- **Cobertura de riesgos no evidentes o emergentes:** incluye riesgos sistémicos, reputacionales, medioambientales y sociales, anticipando escenarios de alto impacto y baja probabilidad que pueden ser críticos para el negocio.
- **Identificación de lagunas de riesgo actuales:** el análisis cruzado con el repositorio permite detectar brechas en los procesos de evaluación internos y mejorarlos proactivamente.
- **Base para políticas internas de gobernanza de IA:** permite estructurar un marco corporativo de gobernanza con roles, responsabilidades, controles y procesos de reporting, integrando IA en el SGSI o ISMS.
- **Soporte para la respuesta a incidentes de IA:** con el AI Incident Tracker, se pueden diseñar estrategias de respuesta y planes de contingencia ante incidentes operativos, técnicos o éticos relacionados con IA.
- **Alineación con riesgos estratégicos y sistémicos de IA:** facilita una visión a largo plazo para mitigar impactos en infraestructuras críticas o en la continuidad de negocio, especialmente ante el crecimiento de IA generativa o modelos fundacionales.

Metodológica para evaluación de riesgos según el RIA

Podemos contar con 5 documentos o repositorios que ofrecen una visión de los riesgos principales que puede entrañar el uso de sistemas o modelos de IA, y un posible punto de partida para elaborar una metodología de evaluación de estos mismos. Todos ellos, casan con la clasificación de riesgos establecida por el AI Act.

Nos referimos a:

- CoE Hunderia
- ISO 42005
- NIST AI RMF 1.0
- NSW AI Assessment Framework
- MIT AI Risk Repository

Para aplicar el repositorio de forma estructurada bajo el marco del AI Act, se propone una metodología de evaluación de riesgos modular, adaptable tanto a usos obligatorios como voluntarios.

- Identificación del sistema de IA y su contexto
  - Describir: tipo de sistema, finalidad, partes involucradas (fabricante, proveedor, importador, usuario...).
  - Contexto de uso: entorno sectorial, criticidad, grado de autonomía del sistema.

- Identificación de riesgos con el Repositorio del MIT.
  - Usar la Taxonomía de Dominio para identificar riesgos específicos por área (discriminación, seguridad, supervisión, etc.).
  - Usar la Taxonomía Causal para identificar si los riesgos son pre o post despliegue, e incorporar esta dimensión al análisis.

- Clasificación conforme al AI Act según el uso esperado de los sistemas
  - Categorizar el sistema según el uso esperado en varios niveles de riesgo regulatorio:
    - Riesgo inaceptable (prohibido)
    - Riesgo alto (sujeto a requisitos estrictos)
    - Riesgo limitado (requiere transparencia)
    - Riesgo mínimo (sin exigencias específicas)

- Documentación de cumplimiento regulatorio
  - Si se considera "riesgo alto", elaborar la documentación obligatoria de conformidad.
  - Para los casos de riesgo limitado o mínimo, se recomienda documentación adicional como buena práctica empresarial y reputacional, especialmente en sectores regulados.
  - Para ambos casos, se recomienda realizar auditorías periódicas con el objeto de verificar y garantizar que el sistema de IA no haya experimentado modificaciones, alteraciones o desviaciones en sus funcionalidades inicialmente evaluadas.

- Mitigación y seguimiento
  - Aplicar controles desde el Repositorio del MIT.
  - Incorporar seguimiento continuo y revisión periódica.
  - Integrar con el ciclo de mejora continua del SGSI o del marco de gestión de riesgos.

6.3.

TABLA O EXPOSICIÓN COMPARATIVA (MIT VS AI ACT)

La relación entre el Repositorio del MIT y AI Act es de complementariedad funcional, no de solapamiento:

- La AI Act define el marco jurídico y el estándar normativo de referencia ("el deber ser jurídico").
- El Repositorio del MIT delimita el conjunto empírico de riesgos técnicos, sociales y sistémicos asociados a los sistemas de IA ("el universo real de riesgos").

A continuación, se propone un mapeo directo entre dominios concretos del Repositorio del MIT y artículos específicos de la AI Act, con un enfoque jurídico-operativo, orientado al cumplimiento normativo, la evaluación de riesgos y los procesos de auditoría.

El mapeo es funcional, no meramente terminológico: en la medida que identifica qué riesgos del Repositorio del MIT contribuyen a interpretar, concretar o acreditar el cumplimiento de las obligaciones previstas en la AI Act.

Dominio 1 – Discriminación y toxicidad (Sesgos, trato desigual, contenido dañino)		
Repositorio – Subdominio	**Artículos AI Act relacionados**	**Relación funcional**
1.1 Discriminación injusta e infrarepresentación	Art. 5.1, Art. 10, Art. 13, Art. 14	El Repositorio del MIT identifica fuentes reales de discriminación que la AI Act intenta evitar mediante gobernanza de datos, transparencia y supervisión humana
1.3 Desempeño desigual entre grupos	Art. 10.2–10.5	Apoya la interpretación del requisito de datasets representativos, pertinentes y libres de sesgos indebidos
1.2 Exposición a contenido tóxico	*Art. 5.1.a, Art. 52*	Conecta con prácticas prohibidas y obligaciones de transparencia en sistemas que interactúan con personas

Figura 11. Correspondencia funcional entre el Dominio 1 del Repositorio de Riesgos de IA del MIT y los artículos del AI Act.

Dominio 2 – Privacidad y Seguridad (Protección de datos, fugas, ataques, inferencias indebidas)		
Repositorio – Subdominio	**Artículos AI Act relacionados**	**Relación funcional**
2.1 Compromiso de privacidad	Art. 10, Art. 15, Art. 16	Permite concretar riesgos de reidentificación, inferencias y uso indebido de datos personales
2.2 Vulnerabilidades y ataques	Art. 15.2, Art. 15.4	El dominio sustenta el requisito de ciberseguridad y resiliencia técnica
	Art. 2.7 y coherencia RGPD	El Repositorio ayuda a identificar interacciones reales entre IA y protección de datos

Figura 12. Correspondencia funcional entre el Dominio 2 del Repositorio de Riesgos de IA del MIT y los artículos del AI Act

Dominio 3 – Desinformación (Información falsa, manipulación del ecosistema informativo)		
Repositorio – Subdominio	**Artículos AI Act relacionados**	**Relación funcional**
3.1 Información falsa o engañosa	Art. 52.1–52.3	Apoya la obligación de informar al usuario cuando interactúa con IA generativa
3.2 Pérdida de consenso y burbujas informativas	Art. 9, Art. 52, considerandos sobre IA propósito general	El Repository permite evaluar riesgos sistémicos no detallados exhaustivamente en el texto normativo

Figura 13. Correspondencia funcional entre el Dominio 3 del Repositorio de Riesgos de IA del MIT y los artículos del AI Act

Dominio 4 – Actores maliciosos y uso indebido (Uso malicioso, fraude, armas, manipulación)		
Repositorio – Subdominio	**Artículos AI Act relacionados**	**Relación funcional**
4.1 Desinformación y vigilancia a escala	Art. 5.1, Art. 9	Da contenido práctico al análisis de riesgos exigido por el AI Act
4.2 Armas, ciberataques, daño masivo	Art. 5.1.d, Art. 15	Refuerza la interpretación de riesgos inaceptables y obligaciones de seguridad
4.3 Fraude y manipulación dirigida	Art. 5.1.a, Art. 52	Ayuda a distinguir usos prohibidos, de alto riesgo o sujetos a transparencia

Figura 14. Correspondencia funcional entre el Dominio 4 del Repositorio de Riesgos de IA del MIT y los artículos del AI Act

Dominio 5 – Interacción humano-máquina (Sobreconfianza, pérdida de autonomía, dependencia)		
Repositorio – Subdominio	**Artículos AI Act relacionados**	**Relación funcional**
5.1 Uso inseguro y sobreconfianza	Art. 14	Fundamenta el diseño efectivo de la supervisión humana
5.2 Pérdida de agencia y autonomía	Art. 5.1.a, Art. 9	Apoya la evaluación de riesgos sobre manipulación cognitiva

Figura 15. Correspondencia funcional entre el Dominio 5 del Repositorio de Riesgos de IA del MIT y los artículos del AI Act

Dominio 6 – Socio-económico y Ambiental (Impactos estructurales, concentración de poder, empleo)		
Repositorio – Subdominio	**Artículos AI Act relacionados**	**Relación funcional**
6.1 Concentración de poder	Considerandos, GPAI	Apoya el enfoque reforzado sobre modelos de propósito general
6.5 Fallos de gobernanza	Art. 9, Art. 17	El dominio se alinea con la obligación de sistemas internos de gestión
6.6 Daño ambiental	Considerandos	Complementa un aspecto poco desarrollado en el articulado

Figura 16. Correspondencia funcional entre el Dominio 6 del Repositorio de Riesgos de IA del MIT y los artículos del AI Act

Dominio 7 – Seguridad, fallos y limitaciones del sistema de IA (Robustez, alineamiento, explicabilidad, fallos técnicos)		
Repositorio – Subdominio	**Artículos AI Act relacionados**	**Relación funcional**
7.3 Falta de robustez	Art. 15	Da contenido técnico al concepto de “funcionamiento fiable”
7.4 Falta de transparencia	Art. 13, Art. 52	Apoya requisitos de explicabilidad y comunicación al usuario
7.1 Conflicto con valores humanos	Art. 9, Art. 14	Refuerza la evaluación de riesgos éticos y funcionales
7.6 Riesgos multi-agente	GPAI, considerandos	Ayuda a evaluar riesgos emergentes no explicitados normativamente

Figura 17. Correspondencia funcional entre el Dominio 7 del Repositorio de Riesgos de IA del MIT y los artículos del AI Act

La tabla comparativa presentada demuestra que el Repositorio del MIT y la AI Act, aunque desarrollados independientemente, son profundamente complementarios. La AI Act establece el marco jurídico aplicable, las obligaciones normativas vinculantes y los mecanismos formales de cumplimiento. El Repositorio del MIT proporciona la especificación técnica, los ejemplos concretos y los procesos operativos efectivos necesarios para operacionalizar esas obligaciones y facilitar su implementación, evaluación y verificación.

Matriz de correspondencia: Repositorio del MIT – AI Act				
Dominio MIT (Taxonomía de dominio)	**Subdominio (ejemplos)**	**Nivel de riesgo según AI Act**	**Requisitos AI Act relevantes**	**Artículos / Anexos**
Privacidad y seguridad	Compromiso de la privacidad, filtración de datos, reidentificación	Alto riesgo (si afecta derechos fundamentales o datos sensibles)	Gestión de riesgos, gobernanza de datos, documentación técnica, ciberseguridad	Art. 9-10-15, Anexo III (sectores sensibles)
Sesgo, equidad y discriminación	Impacto desigual, trato injusto, exclusión de grupos	Alto riesgo (empleo, educación, justicia, servicios públicos)	Gobernanza de datos, supervisión humana, precisión y robustez	Art. 10-14-15, Anexo III
Transparencia y explicabilidad	Decisión opaca, falta de explicabilidad	Riesgo limitado o alto (según contexto)	Transparencia, información al usuario, documentación técnica	Art. 13-50
Desinformación y manipulación	Información falsa o engañosa, engaño de contenido	Riesgo limitado o inaceptable (si vulnera derechos fundamentales)	Obligación de transparencia, prohibiciones de manipulación subliminal	Art. 5-50
Responsabilidad y Gobernanza	Falta de supervisión humana, responsabilidad poco clara	Alto riesgo (en sistemas autónomos o de decisión crítica)	Supervisión humana, trazabilidad, registro de eventos	Art. 9-14-12
Fiabilidad y robustez	Fallos del sistema, deriva del modelo, ataques adversariales	Alto riesgo (infraestructuras críticas, salud, transporte)	Validación técnica, pruebas de robustez, ciberseguridad	Art. 15, Anexo III
Impacto social y derechos humanos	Erosión de la confianza, discriminación, vigilancia masiva	Riesgo inaceptable (si vulnera derechos fundamentales)	Prohibición o control estricto; evaluación de impacto	Art. 5-6-7
Impacto ambiental y de recursos	Consumo energético, residuos electrónicos	Riesgo mínimo o contextual (no regulado directamente)	Recomendado en gestión de riesgos voluntaria	Art. 9 (gestión voluntaria)
Impacto económico y laboral	Sesgo de automatización, desplazamiento de empleos	Alto o limitado (según uso)	Evaluación de impacto en empleo, medidas de mitigación	Art. 6-14-29
Cumplimiento Legal y Ético	Violación de la ley, infracción de propiedad intelectual	Alto o inaceptable (según caso)	Cumplimiento normativo transversal (RGPD, DSA, DMA)	Art. 9-14, Considerandos 1-85

Figura 18. Matriz de correspondencia regulatoria y operativa entre el Repositorio de Riesgos del MIT y la AI Act

7.

Relación y apoyo para el cumplimiento de NIS2 y el ENS

7.1.

VISIÓN GENERAL DE LA NIS2 Y EL ENS Y SU RELACIÓN CONCEPTUAL CON EL REPOSITORIO DEL MIT

En un contexto de transformación digital acelerada, la convergencia entre regulación, técnica y estrategia corporativa ha dejado de ser una aspiración para convertirse en un imperativo ineludible. La Directiva (UE) 2022/2555 (NIS2), el Real Decreto 311/2022 por el que se regula el ENS y el *MIT AI Risk Repository* conforman, de manera conjunta y complementaria, un marco integrador que define el estándar contemporáneo de gobernanza para las organizaciones que operan servicios críticos y despliegan sistemas de IA.

Ignorar su interrelación no sería únicamente imprudente desde una perspectiva regulatoria; supondría comprometer la resiliencia operativa, erosionar la confianza digital y debilitar la posición competitiva de la organización en el medio y largo plazo.

1. NIS2: del cumplimiento formal a la resiliencia operativa integral

La Directiva NIS2 introduce un cambio de paradigma sustancial. Su ambición no se limita a reforzar la seguridad de las redes y sistemas de información, sino que aspira a garantizar una resiliencia operativa integral, con implicaciones directas y exigibles para los órganos de administración. Entre sus elementos clave destacan:

- **Ampliación del ámbito subjetivo:** la tradicional distinción entre operadores de servicios esenciales y proveedores de servicios digitales desaparece. En su lugar, emergen las categorías de entidades “esenciales” e “importantes”, extendiendo la gobernanza de la ciberseguridad a toda la organización y no únicamente a las áreas técnicas.
- **Responsabilidad directa de la alta dirección:** el artículo 20 impone a los órganos de administración la obligación expresa de supervisar y aprobar las medidas de gestión de riesgos. Las consecuencias de su incumplimiento son económicas, reputacionales y legales. El cumplimiento ya no admite aproximaciones parciales ni meramente formales: debe ser completo, tangible y demostrable, correspondiendo a los Estados miembros velar activamente por su efectividad.
- **Gestión de la cadena de suministro:** la evaluación de proveedores de servicios críticos y, de forma especialmente sensible, de soluciones basadas en IA, se erige como un eje central de la estrategia de gestión de riesgos. La mera verificación documental o la existencia de certificados resulta insuficiente; se exige evidencia técnica sólida, actualizada y verificable.

2. ENS: excelencia técnica y soberanía digital

El ENS traduce los principios de NIS2 a un marco técnico concreto, operativo y adaptable al contexto organizativo:

- **Principios y requisitos mínimos:** la clasificación de los sistemas en niveles Básico, Medio y Alto permite ajustar los controles de seguridad a la criticidad real de los activos y al nivel de riesgo asumido, garantizando una aproximación proporcional, coherente y jurídicamente defendible.
- **Neutralidad tecnológica y flexibilidad:** el ENS no impone soluciones técnicas específicas, sino que define objetivos y resultados de seguridad. Esta aproximación facilita la integración de referencias técnicas avanzadas (como el *MIT AI Risk Repository*) y permite cubrir riesgos emergentes que los marcos tradicionales de seguridad podrían no contemplar adecuadamente. En paralelo, la intención de NIS2 es inequívoca: avanzar hacia una homogeneización efectiva de los estándares de ciberseguridad en el ámbito europeo.

3. *MIT AI Risk Repository:* evidencia técnica de vanguardia

El *MIT AI Risk Repository* aporta una taxonomía rigurosa y sistemática de más de 1.700 riesgos asociados a la IA[6], clasificados por origen, alcance y naturaleza. Desde la perspectiva del asesoramiento en derecho digital, constituye una herramienta de diligencia debida de primer nivel, al proporcionar evidencia técnica objetiva que respalda la adopción de controles específicos y la toma de decisiones en materia de gestión de riesgos.

Su utilización permite transformar el cumplimiento normativo de un ejercicio meramente declarativo en una práctica real, verificable y plenamente defendible ante reguladores, auditores y terceros interesados. No obstante, conviene señalar que el uso del *MIT AI Risk Repository*, aun siendo una referencia de vanguardia, puede no resultar suficiente por sí solo para cubrir la totalidad de los riesgos relevantes en contextos complejos o altamente específicos, por lo que se recomienda complementar su aplicación con identificaciones adicionales de riesgos adaptadas al sector, al caso de uso concreto y al entorno operativo de la organización.

4. Sinergia conceptual: un enfoque integral

La convergencia entre NIS2, ENS y el *MIT AI Risk Repository* no es opcional; es estratégica, operativa y jurídicamente defendible.

- **La IA como activo crítico:** en sectores esenciales, como energía, salud o banca, los sistemas de IA requieren una protección específica y diferenciada. La taxonomía del MIT permite identificar riesgos concretos (como data poisoning, model drift o sesgos algorítmicos) que los enfoques tradicionales de análisis de riesgos suelen subestimar o ignorar.
- **Cumplimiento del “estado de la técnica”:** tanto NIS2 (artículo 21) como el ENS exigen la adopción de medidas actualizadas, adecuadas y proporcionadas. La evidencia técnica proporcionada por el MIT respalda de forma objetiva la selección de controles y refuerza su defensibilidad ante cualquier proceso de auditoría o supervisión.
- **Gobernanza de la cadena de suministro:** la evaluación de proveedores de IA puede sistematizarse mediante la taxonomía del MIT, garantizando que los riesgos críticos se identifiquen y mitiguen conforme a estándares internacionales reconocidos y sólidamente defendibles desde un punto de vista legal.

5. Conclusión estratégico-legal

La integración coherente de estos tres instrumentos redefine el concepto de excelencia en gobernanza digital:

- NIS2 establece la arquitectura jurídica y el régimen sancionador, asegurando responsabilidad corporativa y homogeneización regulatoria a nivel europeo.
- ENS garantiza la robustez técnica y la certificación de confianza en el ámbito nacional.
- *MIT AI Risk Repository* aporta la granularidad y la evidencia técnica necesarias para convertir la gestión de riesgos en un ejercicio tangible, auditable y jurídicamente sólido.

Este enfoque holístico no protege solamente frente a sanciones regulatorias, sino que construye resiliencia organizativa, refuerza la confianza digital y genera ventajas estratégicas sostenibles. En un entorno en el que la IA se ha convertido en un activo crítico, únicamente aquellas organizaciones que integren rigor jurídico, evidencia técnica y visión estratégica estarán en condiciones de liderar con seguridad, eficiencia y credibilidad. Ignorar cualquiera de estos elementos equivale a comprometer la integridad y la sostenibilidad de la organización.

[6] *https://airisk.mit.edu/*

7.2.

APOYO OPERATIVO A LA APLICACIÓN DE LA NIS2 Y EL ENS EN LOS SISTEMAS DE IA

La Directiva NIS2 (UE) 2022/2555, el ENS (RD 311/2022) son dos marcos regulatorios que buscan elevar el nivel de la ciberseguridad, reforzar la gestión adecuada de los riesgos, implantar la continuidad de negocio, proteger la cadena de suministro, detectar y responder ante amenazas. Es decir, ambos marcos regulatorios persiguen el mismo objetivo; elevar de forma tangible el nivel de ciberseguridad y resiliencia operativa de las organizaciones.

La directiva NIS2 establece la necesidad de implantar medidas de protección en los servicios y sistemas que resultan críticos para la Unión Europea. Orientando estas medidas hacia la prevención, detección, respuesta y recuperación ante incidentes, con especial énfasis en la gobernanza, la protección de la cadena de suministro, la capacidad de notificación y la coordinación ante incidentes.

Por su parte, el ENS aporta un enfoque estructurado y auditable para la protección de los sistemas de información, basándose en el principio de mínimo privilegio, la vigilancia continua y estableciendo un catálogo de medidas organizativas y técnicas que permiten estandarizar el nivel de control, asegurar la trazabilidad y facilitar la supervisión de los sistemas de información.

Ambos marcos impulsan una evolución desde una seguridad basada en iniciativas aisladas hacia un modelo consistente, medible y sostenido en el tiempo, en el que la gestión del riesgo, la continuidad del negocio, el control de proveedores y la mejora continua se convierten en elementos operativos esenciales.

Es en este punto donde los sistemas de IA pueden convertirse en elementos relevantes para apoyar y facilitar el cumplimiento de ambos marcos regulatorios. Su capacidad de gestión, monitorización, automatización, análisis o respuesta los convierte en unas herramientas muy efectivas a la hora de mejorar la eficiencia y la eficacia de la gestión de la seguridad. Y podemos afirmar que los resultados que se pueden obtener son mayores si se integran estos sistemas con los modelos operativos.

Si queremos identificar dónde pueden aportarnos un valor diferenciado, debemos analizar los requisitos regulatorios de ambos marcos, e identificar donde los sistemas de IA ofrecen un apoyo claro, directo y facilitador al cumplimiento. Aprovechando su capacidad analítica, automatización de tareas recurrentes, mejora de la trazabilidad y consistencia de las evidencias podemos afirmar que estos sistemas son básicos para la gestión de los siguientes puntos:

- **Activos:** la identificación y actualización continua de activos, servicios y sus relaciones de dependencia (incluyendo entornos cloud, APIs y componentes de terceros) ofrece una visión real y detallada del perímetro de seguridad que debemos establecer. La recopilación y correlación de configuraciones, repositorios, logs, tickets o cambios permiten detectar inconsistencias y activos no registrados.

- **Vulnerabilidades:** la priorización de CVEs, el análisis de desviaciones en las configuraciones, recomendaciones de hardening y generación automática de acciones correctivas (tickets), reducen las tareas pendientes, la exposición de los activos y la superficie de ataque.

- **Riesgos:** el apoyo en la identificación de amenazas, la evaluación de la exposición real por criticidad, la correlación entre vulnerabilidades, eventos y el contexto operativo de la organización, mejora la identificación, gestión y toma de decisiones frente los riesgos.

- **Identidades:** mediante mecanismos de detección de anomalías, revisión de permisos y accesos, pertenencias a grupos o la segregación de funciones refuerzan los principios de seguridad y de mínimo privilegio.

- **Incidentes:** la monitorización de los sistemas, detección de patrones anómalos, enriquecido con inteligencia de amenazas y la respuesta automática son elementos claves para una adecuada gestión de incidentes.

- **Continuidad y resiliencia:** las predicciones y alertas ante degradación de la capacidad o del rendimiento, la detección de fallos en tareas críticas, el apoyo a simulaciones o ejercicios o el análisis del entorno nos permiten prepararnos para una mejor respuesta ante interrupciones.

- **Cadena de suministro:** el análisis automático de SLAs, la evaluación de su criticidad, la identificación de brechas contractuales nos permite mejorar el control y reduce el riesgo de incumplimiento por parte de los proveedores.

- **Gestión documental:** elaboración de informes, evaluación de indicadores, análisis de evidencias son muchos de los elementos que permite reducir los tiempos de auditoría y mejorar la consistencia del sistema documental.

- **Cambio y operación:** clasificación automatizada de incidencias, detección de cambios de alto riesgo, validación de criterios en las pruebas de aceptación reducen los errores operativos.

- **Mejora continua:** la detección de tendencias o la generación de cuadros de mando para la dirección permiten diseñar estrategias y acciones efectivas.

Sin embargo, todos sabemos que la implantación de sistemas de IA no está exenta de riesgos y por ello no basta únicamente con incorporar esta tecnología a nuestros sistemas. Es necesario establecer un conjunto de condiciones mínimas de gobierno y control que permitan identificar: quién es responsable, qué se controla, cómo se mitiga el riesgo, cómo se detectan desviaciones, cómo se responde ante incidentes y qué evidencias respaldan cada decisión y cada cambio.

Estas condiciones deben funcionar como un marco de referencia para reducir la probabilidad de los errores en la toma de decisiones y, sobre todo, aportar trazabilidad y auditabilidad a los procesos. A esto habría que añadir los problemas intrínsecos de los sistemas de IA: manipulación de entradas, fugas de datos, contaminación de fuentes, deriva del modelo o dependencia de terceros. Entonces, para evitar todo esto debemos aplicar salvaguardas de tal manera que el uso de la IA esté controlado, sea verificable, sostenible en el tiempo y alineado con los principios de seguridad y resiliencia exigidos por estos marcos regulatorios.

7.3.

EJEMPLO DE APLICACIÓN PRÁCTICA

En esta sección se explica de manera estructurada cómo usar el Repositorio de Riesgos de la IA del MIT para ayudar con la gestión de riesgos que exige la Directiva NIS2 y el ENS.

Cada etapa del proceso se acompaña de un ejemplo práctico para facilitar su comprensión y replicabilidad por las organizaciones.

Paso 1. ¿Qué es un sistema de IA y cómo funciona?

Lo primero es definir el sistema de IA a estudiar, qué hace, qué tan autónomo es y en qué proceso se implementa. Esta información es crucial para definir qué tanto puede afectar el sistema y cómo se alinea con NIS2 y ENS.

Ejemplo práctico:

Una administración pública usa un sistema de IA de aprendizaje automático para priorizar automáticamente los expedientes de ayudas públicas. El sistema cruza información económica, historial de solicitudes y documentación proporcionada por los solicitantes, arrojando según las bases reguladoras una puntuación que determinará el orden de proceso administrativo.

El sistema no decide, pero influye en la continuidad del servicio y en el manejo de información confidencial.

Paso 2. Identificación del riesgo usando el Repositorio del MIT

Una vez especificado el sistema, se revisa el Repositorio de Riesgos del MIT para buscar riesgos relacionados con la forma específica en que se usa el sistema de IA. El repositorio es un catálogo organizado para identificar riesgos particulares que no se encuentran en los análisis de riesgos convencionales.

Ejemplo práctico

Al auditar el control de privacidad y seguridad, la organización encuentra un riesgo en el uso de vulnerabilidades técnicas del sistema de IA que se aprovechen para manipular los datos de entrada o manipular los resultados del modelo. Este peligro es significativo, ya que el sistema se alimenta de información externa (los solicitantes).

Paso 3. Tipos de riesgo, según la causa

El riesgo identificado se categoriza según la taxonomía causal del MIT, en términos de la causa, la intención y el momento en que surge en el ciclo de vida del sistema. Esta categorización ayuda a entender cómo se manifiesta el riesgo y dirigir las medidas de control.

Ejemplo práctico

El riesgo se categoriza en:

- **Naturaleza: sociotécnica, en el sentido que emerge tanto del diseño del modelo como de su aplicación en un proceso abierto de gestión.**
- **Intencionalidad: posiblemente intencional, ya que un tercero podría ingresar datos manipulados para alterar la priorización.**
- **Cuando surge: principalmente en tiempo real durante la operación, pero dependiente de decisiones de diseño y entrenamiento.**

Paso 4. Adaptar el riesgo a NIS2 y ENS.

El riesgo no se considera una categoría en sí misma, sino que se mapea a las categorías ya existentes en NIS2 y ENS (en particular, en lo que respecta al impacto en la seguridad de los sistemas de información y la continuidad del servicio).

Ejemplo práctico

Desde la mirada NIS2, el riesgo se relaciona con las obligaciones de ciberseguridad, prevención de incidentes y protección de los sistemas que sostienen servicios esenciales.

Dentro del ENS, el riesgo se incluye en el análisis de riesgos del sistema como una amenaza a la integridad, confidencialidad y trazabilidad de la información que maneja el sistema de IA.

Paso 5. Identificación y aplicación de medidas de mitigación.

Tras la alineación normativa, se establecen medidas de mitigación en línea con las obligaciones ya requeridas por NIS2 y ENS, escalando en función del riesgo descubierto. El repositorio MIT hace que sea fácil demostrar la necesidad y la proporcionalidad de estas medidas.

Ejemplo práctico

Entre otras, la organización utiliza las siguientes medidas :

- **Políticas: establecer políticas de uso y monitoreo del sistema de IA, asignar responsabilidades y capacitar al personal.**
- **Técnicas: controles de acceso, validaciones automáticas de datos de entrada, monitorización del modelo y registro de eventos.**

Paso 6. Integración en la gestión de riesgos y evidencia de cumplimiento.

Finalmente, el riesgo y las acciones se incorporan a los flujos de trabajo de gestión de riesgos ya existentes, para su seguimiento y demostración de cumplimiento en auditorías o supervisiones.

Ejemplo práctico

El riesgo del sistema de priorización automática está recogido en el análisis de riesgos de la organización, enlazado a controles NIS2 y ENS. Esto muestra que los riesgos asociados con la IA se han identificado, evaluado y gestionado en línea con los marcos legales pertinentes.

Beneficios del enfoque

Este proceso demuestra que el Repositorio de Riesgos del MIT no reemplaza a NIS2 o ENS, sino que complementa su implementación, proporcionando una forma estructurada y detallada de visualizar los riesgos específicos de la IA en el contexto de la seguridad basada en riesgos.

7.4.

TABLA O EXPOSICIÓN COMPARATIVA (MIT VS ENS Y MIT VS NIS2)

El ENS y la Directiva NIS2 plantean la seguridad de una manera transversal y sistémica, definiendo principios, obligaciones y medidas para garantizar la confidencialidad, integridad, disponibilidad y resiliencia de los sistemas de información y de los servicios esenciales o importantes.

Pero estos marcos no profundizan en los riesgos concretos asociados al uso de sistemas de IA, sino que los integran en un enfoque general de gestión de riesgos, gobernanza, prevención y respuesta ante incidentes.

En este sentido, el Repositorio de Riesgos de IA del MIT es un catálogo estructurado y basado en evidencia para identificar riesgos de IA que pueden ayudar a concretar y hacer operativo el ENS y la NIS2 cuando estos sistemas formen parte de los procesos, servicios o infraestructuras de la organización.

A continuación, se realiza un mapeo estructurado entre los dominios y subdominios del Repositorio del MIT y los principios, requisitos y obligaciones del ENS y de la NIS2, desde una perspectiva práctica, de integración en los análisis de riesgos, definición de controles y obtención de evidencias auditables en los procesos de auditoría, supervisión y rendición de cuentas.

El propósito de este mapeo no es replicar el contenido normativo, sino hacer más accesible su interpretación y aplicación en contextos donde la IA crea nuevos vectores de riesgo, para que las organizaciones puedan demostrar un cumplimiento diligente, rastreable y acorde al estado del arte en seguridad y gobernanza de la IA.

Dominio 1 – Discriminación y toxicidad (Sesgos, trato desigual, contenido dañino)			
Repositorio – Subdominio	**Principios y requisitos ENS relacionados**	**Relación funcional**	**Evidencias típicas de auditoría (ENS)**
1.1 Discriminación injusta e infrarepresentación	Gestión basada en riesgos; Seguridad integral; Diferenciación de responsabilidades	Integra impactos discriminatorios en el análisis de riesgos y exige controles organizativos y supervisión humana	Análisis de riesgos actualizado; políticas de uso de IA; RACI/roles; actas de comités; métricas de equidad; registros de revisión humana
1.2 Exposición a contenido tóxico	Prevención y detección; Vigilancia continua; Seguridad integral	Obliga a medidas preventivas y monitorización de contenidos generados o difundidos por IA	Procedimientos de moderación; reglas/filtros; logs de monitorización; alertas; evidencias de actuación y escalado
1.3 Desempeño desigual entre grupos	Gestión basada en riesgos; Reevaluación periódica	Exige revisión periódica del comportamiento del sistema y tratamiento de desviaciones	Informes de evaluación periódica; KPIs/KRIs por grupo; registros de ajustes/correcciones; actas de revisión

Figura 19. Riesgos de discriminación y toxicidad: alineación MIT–ENS

Dominio 2 – Privacidad y seguridad (Protección de datos, fugas, ataques, inferencias indebidas)			
Repositorio – Subdominio	**Principios y requisitos ENS relacionados**	**Relación funcional**	**Evidencias típicas de auditoría (ENS)**
2.1 Compromiso de la privacidad	Protección de la información; Control de acceso; Prevención, detección y respuesta	Concreta escenarios de reidentificación/inferencia y exige medidas técnicas y organizativas	Inventario de datos; controles de acceso; cifrado; DPIA/Análisis de riesgos; registros de acceso; planes y registros de respuesta
2.2 Vulnerabilidades y ataques	Marco operacional; Medidas de protección; Gestión de incidentes	Incorpora amenazas específicas contra IA en el análisis de riesgos y la respuesta	Gestión de vulnerabilidades; hardening; pruebas; parches; SIEM/EDR; registros de incidentes y post-mortem

Figura 20. Riesgos de privacidad y seguridad: alineación MIT–ENS

Dominio 3 – Desinformación (Información falsa, engañosa o no fiable)			
Repositorio – Subdominio	**Principios y requisitos ENS relacionados**	**Relación funcional**	**Evidencias típicas de auditoría (ENS)**
3.1 Información falsa o engañosa	Seguridad integral; Líneas de defensa; Gestión del riesgo operativo	Trata errores funcionales como riesgos operativos con validación y supervisión	Procedimientos de validación; doble control; trazabilidad de fuentes; registros de revisión y corrección
3.2 Contaminación del ecosistema informativo	Gestión basada en riesgos; Vigilancia continua	Evalúa impactos sistémicos y acumulativos en la fiabilidad de la información	Monitorización de fuentes; informes de impacto; métricas de calidad; evidencias de mitigación

Figura 21. Riesgos de desinformación: alineación MIT–ENS

Dominio 4 – Actores maliciosos y uso indebido (Fraude, manipulación, daño deliberado)			
Repositorio – Subdominio	**Principios y requisitos ENS relacionados**	**Relación funcional**	**Evidencias típicas de auditoría (ENS)**
4.1 Desinformación y vigilancia a gran escala	Prevención y detección; Respuesta ante incidentes	Identifica usos indebidos amplificados por IA y exige detección/contención	Casos de uso autorizados; detección de abuso; logs; registros de contención y comunicación
4.2 Armas, ciberataques y daño masivo	Gestión de incidentes; Conservación de evidencias; Seguridad integral	Considera escenarios de alto impacto en continuidad y forense	Planes de respuesta y continuidad; pruebas de crisis; custodia de evidencias; informes forenses
4.3 Fraude y manipulación selectiva	Control de acceso; Vigilancia continua; Prevención del fraude	Exige controles antifraude y supervisión de accesos/uso	MFA/roles; reglas antifraude; alertas; investigaciones; sanciones/acciones correctivas

Figura 22. Riesgos por uso indebido y actores maliciosos: alineación MIT–ENS

Dominio 5 – Interacción humano-máquina (Dependencia, sobreconfianza, pérdida de autonomía)			
Repositorio – Subdominio	**Principios y requisitos ENS relacionados**	**Relación funcional**	**Evidencias típicas de auditoría (ENS)**
5.1 Uso inseguro y sobreconfianza	Concienciación y formación; Diferenciación de responsabilidades	Evita delegación indebida mediante formación y procedimientos	Planes de formación; registros de asistencia; procedimientos operativos; controles de validación
5.2 Pérdida de agencia y autonomía	Seguridad integral; Gestión basada en riesgos	Evalúa automatización excesiva con impacto en personas/procesos	Análisis de riesgos; decisiones documentadas; criterios de escalado; revisiones de impacto

Figura 23. Riesgos de interacción humano máquina: alineación MIT–ENS

Dominio 6 – Socioeconómico y ambiental (Impactos estructurales y de gobernanza)			
Repositorio – Subdominio	**Principios y requisitos ENS relacionados**	**Relación funcional**	**Evidencias típicas de auditoría (ENS)**
6.1 Concentración de poder	Marco organizativo; Gestión de proveedores	Identifica dependencias tecnológicas críticas	Evaluación de proveedores; contratos/SLAs; planes de contingencia; alternativas
6.2 Impacto en el empleo	Gestión basada en riesgos	Integra impactos organizativos relevantes	Informes de impacto; decisiones de mitigación; comunicaciones internas
6.3 Devaluación del esfuerzo humano	Seguridad integral	Evalúa efectos sobre calidad y fiabilidad de procesos	Indicadores de calidad; revisiones de proceso; acciones correctivas
6.4 Dinámicas competitivas	Marco organizativo; Reevaluación periódica	Gestiona riesgos estratégicos de adopción de IA	Análisis estratégicos; revisiones periódicas; actas de dirección
6.5 Fallos de gobernanza	Líneas de defensa; Diferenciación de responsabilidades	Exige gobierno claro y control interno	Estructura de gobierno; RACI; comités; auditorías internas
6.6 Daño ambiental	Gestión basada en riesgos	Considera sostenibilidad operativa	Informes de consumo/impacto; decisiones de mitigación; seguimiento

Figura 24. Riesgos socioeconómicos y ambientales: alineación MIT–ENS

Dominio 7 – Seguridad, fallos y limitaciones del sistema de IA			
Repositorio – Subdominio	**Principios y requisitos ENS relacionados**	**Relación funcional**	**Evidencias típicas de auditoría (ENS)**
7.1 Conflicto con valores humanos	Seguridad integral; Gestión basada en riesgos	Evalúa riesgos éticos y funcionales	Análisis de riesgos; criterios éticos; actas de revisión
7.2 Capacidades peligrosas	Prevención y detección; Gestión del cambio	Anticipa comportamientos no deseados	Procedimientos de cambio; pruebas; aprobaciones; registros
7.3 Falta de robustez	Continuidad del servicio; Vigilancia continua	Exige pruebas, validación y monitorización	Planes BCP/DR; pruebas; métricas de robustez; monitorización
7.4 Falta de transparencia	Líneas de defensa; Supervisión y control	Requiere trazabilidad y comprensión operativa	Documentación técnica; trazas; model/data lineage; registros de revisión
7.5 Bienestar y derechos de la IA	Seguridad integral	Considera riesgos emergentes de gobernanza	Evaluaciones de impacto; decisiones documentadas
7.6 Riesgos multi-agente	Gestión basada en riesgos; Vigilancia continua	Evalúa comportamientos emergentes en sistemas complejos	Pruebas de escenarios; monitorización; informes de comportamiento

Figura 25. Riesgos de seguridad y limitaciones del sistema de IA: alineación MIT–ENS

La tabla comparativa presentada demuestra que el Repositorio del MIT y la AI Act, aunque desarrollados independientemente, son **profundamente complementarios.** El AI Act establece el marco jurídico aplicable, las obligaciones normativas vinculantes y los mecanismos formales de cumplimiento. El Repositorio del MIT proporciona la especificación técnica, los ejemplos concretos y los procesos operativos efectivos necesarios para operacionalizar esas obligaciones y facilitar su implementación, evaluación y verificación.

Matriz de correspondencia: Repositorio del MIT – AI Act			
Dominio MIT (Taxonomía de dominio)	**Subdominio (ejemplos)**	**Nivel de riesgo según AI Act**	**Requisitos AI Act relevantes**
1.1 Discriminación injusta e infrarepresentación	Art. 20 (gobernanza), 21(2)(a,f,g) (políticas de riesgo; evaluación de eficacia; formación)	Integra impactos discriminatorios en el análisis de riesgos y en la supervisión de la dirección	Política de IA responsable; métricas de equidad; actas del comité; plan de formación; revisiones periódicas
1.2 Exposición a contenido tóxico	21(2)(a,b,f,g,j) (riesgo; incident handling; eficacia; formación; comm seguras)	Prevención/detección y respuesta sobre contenidos dañinos generados por IA	Reglas de moderación; alertas; runbooks; evidencias de escalado; comunicaciones seguras
1.3 Desempeño desigual entre grupos	21(2)(a,f,g)	Revisión periódica de rendimiento por cohortes y acciones correctoras	Informes de evaluación; KPIs/KRIs por grupo; re-tests y aprobaciones de cambio
2.1 Compromiso de la privacidad (filtración/inferencia)	21(2)(h,i,j) (criptografía; accesos; MFA/comm seguras), 23 (notificación)	Trata fugas de PII/secretos como incidente reportable; exige cifrado y control de acceso reforzado	Política de cifrado; gestión de claves; DLP; MFA; plantillas 24 h/72 h/1 mes; registro de accesos
2.2 Vulnerabilidades y ataques a sistemas de IA	21(2)(b,c,d,e,f,i,j) (incidentes; continuidad/DR/crisis; supply chain; secure SDLC/vuln; eficacia; accesos/MFA)	Integra amenazas específicas (poisoning, extraction, prompt injection) y terceros en el ciclo de vida	Plan de pruebas adversariales; pentest de pipeline/modelo; BCP/DR test; evaluaciones de proveedor crítico (Art. 22)
3.1 Información falsa o engañosa	21(2)(a,b,f,g)	Valida/escala errores fácticos como riesgo operativo	Procedimientos de verificación; doble control; model cards con limitaciones; registro de correcciones
3.2 Contaminación del ecosistema informativo	21(2)(a,c,f,g,j)	Plan de crisis y comunicaciones seguras por impactos reputacionales/sistémicos	Plan de crisis; simulacros; canales seguros; métricas de calidad y reporte al comité
4.1 Desinformación, vigilancia e influencia a gran escala	20; 21(2)(a,b,c,i,j); 23	Detección de misuse a escala, gestión de incidentes y gobierno (límites de uso, escalado)	Casos de abuso; logs y contención; comunicaciones seguras; notificación si aplica
4.2 Ciberataques, armas y daños masivos	21(2)(b,c,d,e,f,i,j); 22; 23	Seguridad técnica, pruebas, continuidad y coordinación en supply chain	Red teaming; hardening; DR; resultados de evaluación de proveedores; expediente de notificación
4.3 Fraude, estafas y manipulación selectiva	21(2)(a,b,f,g,i,j); 20	Conecta fraude asistido por IA con detección de abuso, incident handling y awareness	Reglas antifraude; alertas; MFA/roles; simulacros; acciones correctivas
5.1 Uso inseguro y sobreconfianza	20; 21(2)(g,i,j)	Evita delegación indebida con formación y control de accesos	Plan de formación por roles; evidencias de asistencia; MFA en flujos críticos; SOPs human-in-the-loop
5.2 Pérdida de agencia y autonomía	21(2)(a,f,g)	Asegura revisión humana significativa en decisiones críticas	Procedimientos de revisión; registros de anulaciones; evaluación periódica de eficacia
6.1 Concentración de poder	21(2)(d); 22	Evalúa dependencia de proveedores/modelos y define mitigaciones	Matriz de criticidad; cláusulas de seguridad; verificaciones; planes de salida/alternativa
6.2 Impacto en el empleo	20; 21(2)(g)	Gobierno del cambio y upskilling asociados a automatización	Planes de capacitación; decisiones del comité; seguimiento de impacto organizativo
6.3 Devaluación del esfuerzo humano	21(2)(a,f,g)	Control de calidad/autoría y evaluación de riesgos operativos	Procedimientos de QA; métricas; revisiones y correcciones documentadas
6.4 Dinámica competitiva	20; 21(2)(f)	Evita “lanzar sin controles”: gates y criterios de aceptación	Checklists de salida; actas de aprobación; informes de pruebas/retos
6.5 Fallos de gobernanza	20; 21(2)	Núcleo NIS2: responsabilidades de dirección + medidas mínimas	Política NIS2; RACI; informes periódicos al consejo; auditorías internas
6.6 Daño medioambiental	21(2)(c)	Relaciona resiliencia/continuidad con capacidad y consumo	BIA/capacity planning; pruebas de recuperación; métricas de continuidad
7.1 Objetivos en conflicto con valores humanos	21(2)(a,e,f); 23	Gobierno del comportamiento seguro y monitorización en operación	Pruebas de comportamiento; criterios de bloqueo/rollback; expediente de incidentes
7.2 Capacidades peligrosas	21(2)(b,c,e,i,j)	Defense-in-depth, contención y continuidad	Red teaming; control de herramientas/agentes; DR tests; MFA reforzada
7.3 Falta de robustez	21(2)(c,e,f)	Pruebas de robustez; criterios de aceptación; retests	Informes de stress/robustez; gates; métricas; revalidaciones
7.4 Falta de transparencia/interpretabilidad	21(2)(a,f)	Trazabilidad y documentación para auditar y operar	Versionado de datasets/modelos; model cards; registros de revisión
7.5 Bienestar y derechos de la IA	20; 21(2) (enfoque de riesgo)	Riesgo emergente que se eleva a comité y se documenta	Actas de comité; límites de uso; evaluación de riesgos/ética
7.6 Riesgos de múltiples agentes	21(2)(a,c,f)	Ensayos de fallos en cadena y plan de contingencia	Simulaciones multiagente; planes de contingencia; resultados de pruebas
8.1 Bienestar y derechos de la IA	20; 21(2)	Ídem 7.5, con foco en políticas y límites organizativos	Política de uso; decisiones del comité; revisiones periódicas
8.2 Riesgos de múltiples agentes (sistémicos)	21(2)(a,c,f)	Ídem 7.6, con foco en impactos sistémicos	Ejercicios de resiliencia; análisis de interdependencias; mejoras continuas

Figura 26. Mapa de alineamiento entre el MIT AI Risk Repository y los requisitos del AI Act.

8.

Relación y apoyo para el cumplimiento del Reglamento UE de protección de datos

8.1.

VISIÓN GENERAL DEL RGPD Y SU RELACIÓN CONCEPTUAL CON EL REPOSITORIO DEL MIT

El RGPD establece un marco jurídico basado en el riesgo cuyo objetivo es proteger los derechos y libertades de las personas físicas frente al tratamiento de datos personales. Aunque no se diseñó específicamente para regular sistemas de IA, su lógica es plenamente aplicable al ciclo de vida de estos sistemas, especialmente cuando:

- Se utilizan datos personales en el entrenamiento o en la explotación del modelo.
- El sistema realiza inferencias sobre personas físicas.
- Existe elaboración de perfiles.
- Se adoptan decisiones automatizadas que producen efectos jurídicos o similares (art. 22).

Los principios del RGPD (art. 5) -minimización, limitación de la finalidad, exactitud, integridad y confidencialidad, y responsabilidad proactiva- constituyen un marco operativo que debe integrarse desde las primeras fases del ciclo de vida de la IA (art. 25, privacidad desde el diseño y por defecto).

El *MIT AI Risk Repository* aporta una taxonomía clara y estructurada de riesgos que facilita la identificación temprana de amenazas y la alineación con los requisitos del RGPD durante el diseño, desarrollo y operación de sistemas de IA.

8.2.

APOYO OPERATIVO A LA APLICACIÓN DEL RGPD EN LOS SISTEMAS DE IA

La utilidad principal del repositorio se manifiesta en tres dimensiones:

a) Identificación sistemática de riesgos relevantes para el RGPD

- La taxonomía por dominios permite localizar riesgos asociados a:
- Privacidad y seguridad (Dominio 2): filtraciones, inferencias indebidas, reidentificación.
- Discriminación y toxicidad (Dominio 1): sesgos, trato desigual o decisiones injustas.
- Transparencia y explicabilidad (Dominio 7): opacidad algorítmica y dificultades para cumplir los arts. 13, 14 y 15.
- Decisiones automatizadas (Dominio 5): sobreconfianza, ausencia de supervisión humana efectiva.

b) Alineación con la taxonomía causal del MIT

La clasificación por:

- Origen (humano, IA o socio-técnico).
- Intencionalidad (intencional/no intencional).
- Momento (pre-despliegue / post-despliegue).
- Permite integrar el análisis de riesgos en los requisitos del RGPD:
  - » Evaluaciones de impacto antes del tratamiento (art. 35).
  - » Medidas técnicas y organizativas adecuadas (art. 32).
  - » Revisión y monitorización continua (*accountability,* art. 24 y 5.2).

c) Mejora de las EIPD – Evaluación de Impacto en Protección de Datos

- El repositorio permite ampliar las EIPD incorporando riesgos que habitualmente no aparecen en metodologías tradicionales, tales como:
- Inferencias no autorizadas.
- *Prompt injection* y fugas en grandes modelos de lenguaje.
- Envenenamiento de datos (data poisoning).
- Falta de robustez o sesgos del modelo en determinados colectivos.

8.3. EJEMPLO DE APLICACIÓN PRÁCTICA

A partir de la combinación del RGPD y el MIT AI Risk Repository, se recomienda articular los controles de la siguiente manera:

1. Durante el diseño

- Clasificación del tratamiento y datos afectados (art. 30).
- Minimización estricta de datos utilizados en entrenamiento.
- Definición de finalidades claras y documentadas.
- Identificación de escenarios de riesgo basados en la taxonomía MIT (dominio + causalidad).
- Evaluación de necesidad de EIPD (art. 35).

2. Durante el desarrollo

- Revisión de datasets: calidad, representatividad, derechos de autor y base jurídica.
- Pruebas de sesgo y pruebas adversariales del modelo.
- Documentación de decisiones técnicas y trazabilidad (*"model lineage")*.

3. Durante el despliegue

- Medidas de seguridad adecuadas (art. 32): cifrado, control de accesos, control de versiones.
- Configuración segura de prompts y mecanismos de filtrado.
- Supervisión humana efectiva cuando proceda (art. 22).
- Revisión de contratos con proveedores y terceros.

4. Durante la operación

- Monitorización continua de desviaciones (drift) y de sesgos emergentes.
- Auditorías periódicas de privacidad y seguridad.
- Gestión de incidentes con notificación en plazos (arts. 33 y 34).
- Revisión periódica del análisis de riesgo y la EIPD.

8.4. TABLA O EXPOSICIÓN COMPARATIVA (MIT VS RGPD)

La convergencia entre el RGPD y el AI Risk Repository no es meramente académica, sino profundamente práctica. La tabla comparativa incorporada a continuación establece tres niveles de análisis:

- Correspondencia normativa directa: identifica aquellos subdominios del AI Risk Repository que se alinean directamente con artículos o principios específicos del RGPD.
- Amplificación de riesgos: examina cómo ciertos riesgos identificados en el repositorio representan versiones amplificadas o específicas de IA de preocupaciones ya presentes en el RGPD.
- Riesgos emergentes con implicaciones indirectas: reconoce que algunos dominios del repositorio, aunque no están explícitamente cubiertos por el RGPD, pueden tener implicaciones indirectas para la protección de datos o para los derechos que el RGPD busca proteger.

Matriz de correspondencia: Repositorio de MIT – RGPD		
Subdominio del AI Risk Repository	**Artículos RGPD**	**Utilidad para el Cumplimiento del RGPD**
1.1. Discriminación injusta y tergiversación	Art. 5(1)(a). Art. 22. Art. 9. Considerando 71.	Proporciona un marco para evaluar el cumplimiento del Art. 22 sobre decisiones automatizadas.
1.3 Desempeño desigual entre grupos	Art. 5(1)(a). Art. 5(1)(d). Art. 25.	Operacionaliza el principio de exactitud en contextos de IA. Permite implementar la protección de datos desde el diseño.
2.1 Compromiso de la privacidad mediante la obtención, filtración o inferencia correcta de información sensible	Art. 5(1)(f). Art. 32. Art. 9. Art. 25.	Proporciona un marco específico para evaluar cómo los sistemas de IA pueden comprometer la privacidad.
2.2 Vulnerabilidades y ataques a la seguridad del sistema de IA	Art. 32. Art. 5(1)(f). Art. 33-34.	Permite implementar medidas de seguridad específicas para IA.
3.1 Información falsa o engañosa	Art. 5(1)(d). Art. 16. Art. 5(1)(a).	Operacionaliza el principio de exactitud cuando los datos no son simplemente "almacenados" sino "generados" por IA. Plantea quién es responsable de rectificar información inexacta generada por IA.
4.1 Desinformación, vigilancia e influencia a gran escala	Art. 5(1)(a). Art. 6. Art. 9. Art. 7-8.	Permite identificar cuándo el tratamiento de datos por sistemas de IA podría no cumplir con la base de licitud. Esencial para evaluar si es compatible con los principios de lealtad y transparencia.
4.3 Fraude, estafas y manipulación dirigida	Art. 5(1)(a). Art. 7. Art. 32.	Fundamenta la necesidad de implementar controles de uso ético de sistemas de IA que procesan datos personales. Relevante para proteger la integridad del consentimiento.
5.1 Confianza excesiva y uso inseguro	Art. 22. Art. 13-14. Considerando 71.	Fundamenta la importancia de las obligaciones de información sobre la existencia de decisiones automatizadas y su importancia.
5.2 Pérdida de agencia y autonomía humana	Art. 22. Considerando 71. Art. 9(2)(g).	Operacionaliza el concepto de "decisiones basadas únicamente en el tratamiento automatizado".
6.5 Fallos de gobernanza	Art. 24. Art. 5(2). Art. 35. Art. 37-39. Arts. 83-84.	Proporciona un marco para identificar brechas en gobernanza del RGPD. Fundamenta la necesidad de accountability proactiva del Art. 5(2).
7.1 La IA persiguiendo objetivos propios en conflicto con los objetivos o valores humanos	Art. 5(1)(a).Art. 25. Art. 32.	Tiene implicaciones directas para el principio de lealtad. Fundamenta la necesidad de implementar mecanismos robustos de seguridad.
7.3 Falta de capacidad o robustez	Art. 5(1)(d). Art. 32. Art. 25.	Útil para cumplir con el principio de exactitud. Útil para demostrar que se han implementado medidas técnicas apropiadas.
7.4 Falta de transparencia o interpretabilidad	Art. 5(1)(a). Art. 13-14. Art. 15. Art. 22(3). Considerando 71.	Operacionaliza el derecho en decisiones automatizadas. Permite identificar cuándo la opacidad del modelo impide cumplir adecuadamente con las obligaciones de transparencia.

Figura 27. Mapeo operativo de riesgos de IA del MIT frente a obligaciones del RGPD

Taxonomía Causal vs. RGPD		
Taxonomía Causal	**Artículos RGPD**	**Utilidad para el Cumplimiento del RGPD**
Entidad: Humano vs. IA vs. Otro	Art. 24. Art. 28. Art. 82.	Facilita la determinación de corresponsabilidad cuando múltiples entidades contribuyen al riesgo.
Momento: Pre-despliegue vs. Post-despliegue	Art. 25. Art. 35. Art. 32.	Operacionaliza directamente la protección desde el diseño. Fundamenta la necesidad de realizar DPIA (Art. 35) antes y después.
Intención: Intencional vs. No intencional	Art. 83. Art. 32.	Permite diseñar controles diferenciados. Útil para demostrar diligencia debida.

Figura 28. Taxonomía Causal de riesgos de IA del MIT frente a obligaciones del RGPD

Esta tabla comparativa demuestra que el *AI Risk Repository* del MIT no es simplemente un documento académico paralelo al RGPD, sino una herramienta práctica que puede enriquecer significativamente la capacidad de las organizaciones para cumplir con sus obligaciones bajo el RGPD. Al proporcionar una estructura granular y sistemática de riesgos específicos de IA, el repositorio permite traducir los principios abstractos del RGPD en acciones concretas, medibles y auditables.

9.

Relación y apoyo para incorporar los Riesgos de la IA en la Matriz de Riesgos Corporativa de una Compañía

Este apartado aborda el desafío fundamental de integrar los riesgos emergentes y dinámicos de la IA, **identificados y estructurados a partir del Repositorio de Riesgos de IA del MIT**, dentro de los marcos de gestión de riesgos corporativos tradicionales, teniendo en cuenta que dicho repositorio, aun siendo una referencia de vanguardia, no pretende ni puede abarcar exhaustivamente todas las posibles fuentes de riesgo ni la totalidad de los riesgos asociados a la IA, dado que su alcance se limita a aquellos identificados en marcos y fuentes que cumplen criterios de estructuración metodológica.

La adopción de la IA presenta una dualidad, por un lado, ofrece un potencial transformador para la eficiencia y la innovación; por otro, introduce un espectro de riesgos novedosos que las matrices de riesgo convencionales no están diseñadas para capturar de forma nativa. Estos riesgos, que abarcan desde sesgos algorítmicos hasta fallos sistémicos, a menudo son de naturaleza cualitativa, interconectada y evolucionan a gran velocidad. Por lo tanto, se hace imprescindible un enfoque estructurado que traduzca estos riesgos específicos de la IA al lenguaje del riesgo empresarial (operacional, financiero, reputacional, de cumplimiento, etc.).

El objetivo es garantizar que la innovación en IA no se produzca en un vacío de gobernanza, sino que esté plenamente alineada con el apetito de riesgo, la cultura y los objetivos estratégicos de la organización.

Para lograrlo, se propone apoyarse en marcos reconocidos de gestión de riesgos, establecer metodologías claras de cuantificación y priorización, definir controles de mitigación adecuados y fortalecer la estructura de gobierno corporativo enfocada en la supervisión de estos riesgos, permitiendo a la organización aprovechar sus beneficios de manera responsable y segura.

9.1. METODOLOGÍA PARA LA CUANTIFICACIÓN Y PRIORIZACIÓN DE RIESGOS DE IA

FASE 1: CONTEXTUALIZACIÓN E IDENTIFICACIÓN DE RIESGOS A PARTIR DEL REPOSITORIO DEL MIT

El primer paso consiste en utilizar el Repositorio de Riesgos de IA del MIT como un catálogo maestro para identificar los riesgos de IA relevantes para la organización.

La taxonomía de dominios del MIT, a día de hoy agrupa más de 1700 riesgos en siete áreas de impacto principales, sirve como punto de partida. Se debe realizar un ejercicio de mapeo en el que equipos multidisciplinares (incluyendo expertos en IA, dueños de procesos de negocio y gestores de riesgo) seleccionen los riesgos del repositorio que son pertinentes para los sistemas de IA actuales o planificados de la organización.

Un paso crucial en este proceso es el mapeo de los dominios de riesgo de la IA a las categorías de riesgo corporativo existentes. Los dominios del MIT, como “Discriminación y toxicidad” o “Seguridad, fallos y limitaciones de los sistemas de IA”, no suelen existir como categorías independientes en las matrices de riesgo corporativas tradicionales, que se centran en riesgos operacionales, financieros, legales o reputacionales. Por ello, los dominios del MIT deben tratarse como factores de riesgo o causas raíz que se manifiestan a través de las categorías de riesgo corporativo ya establecidas.

Por ejemplo, un riesgo del dominio “Discriminación y toxicidad”, como un algoritmo de contratación que discrimina por género, no es un riesgo aislado. Su materialización desencadena un riesgo reputacional (cobertura mediática negativa), un riesgo de cumplimiento (violación de leyes de igualdad de oportunidades), un riesgo legal (demandas) y un riesgo financiero (multas y pérdida de negocio). Al mapear los riesgos de esta manera, la organización puede utilizar sus escalas de impacto existentes, adaptándolas para reflejar la causa raíz de la IA.

Este proceso de contextualización asegura que la evaluación no se quede en un nivel técnico, sino que se conecte directamente con el impacto en el negocio, un principio fundamental de marcos como COSO ERM (Enterprise Risk Management), que integra el riesgo con la estrategia y el desempeño.

FASE 2: DESARROLLO DE ESCALAS DE IMPACTO Y PROBABILIDAD

Una vez identificados y contextualizados los riesgos, se deben definir escalas de calificación personalizadas. Las matrices de riesgo estándar utilizan ejes de probabilidad e impacto para clasificar los riesgos. Para la IA, estos ejes requieren una definición más matizada.

Evaluación del Impacto

Se deben establecer criterios cualitativos y cuantitativos para medir el impacto potencial en dimensiones clave del negocio. Las dimensiones para considerar incluyen:

- **Impacto Financiero:** pérdidas directas, multas regulatorias (por ejemplo, bajo el Reglamento de IA de la Unión Europea), costes de remediación y litigios.
- **Impacto Reputacional:** daño a la marca, pérdida de confianza del cliente, escrutinio mediático y de los inversores.
- **Impacto Regulatorio y Legal:** sanciones, investigaciones, prohibición de uso del sistema de IA y responsabilidad por daños.
- **Impacto Operativo:** interrupción de procesos de negocio críticos, decisiones erróneas a gran escala, necesidad de retirar sistemas de producción.
- **Impacto Ético-Social:** daño a individuos o grupos vulnerables, perpetuación de desigualdades sistémicas, erosión de la autonomía humana o la confianza social.

Evaluación de la Probabilidad

Estimar la probabilidad de los riesgos de IA es un desafío debido a la falta de datos históricos extensos. Para superar esto, se propone un enfoque híbrido. Las estadísticas del Repositorio del MIT sobre la proporción de riesgos por dominio pueden utilizarse como un proxy inicial y basado en datos para la prevalencia relativa. Esto no representa una probabilidad estadística real, sino una medida de la frecuencia con la que ciertos tipos de riesgo son discutidos en la literatura experta, indicando las áreas de mayor preocupación y potencial ocurrencia.

Por ejemplo, el repositorio muestra que la categoría “Seguridad, fallos y limitaciones de los sistemas de IA” constituye el 25% de los riesgos catalogados, mientras que “Desinformación” solo representa el 4%. Esto sugiere que, a nivel general, los fallos de robustez o capacidad son una preocupación más frecuente entre los expertos. Una organización puede usar esta distribución como una línea base para su escala cualitativa de probabilidad (por ejemplo, de “Raro” a “Casi seguro”), ajustándola luego con factores internos.

Los factores de ajuste internos que deben modificar la probabilidad base incluyen:

- **Complejidad y Madurez del Modelo:** modelos más nuevos, opacos o complejos (por ejemplo, IA generativa multimodal) presentan una mayor incertidumbre y, por tanto, una mayor probabilidad de fallo inesperado.
- **Calidad y Representatividad de los Datos:** datos de entrenamiento de baja calidad, incompletos o sesgados aumentan significativamente la probabilidad de riesgos de equidad y rendimiento.
- **Robustez del Entorno de Control:** la existencia de controles preventivos y detectives sólidos, como auditorías regulares y monitoreo continuo, reduce la probabilidad de materialización del riesgo.
- **Naturaleza del Despliegue:** un sistema de IA de cara al público o utilizado en decisiones de alto impacto tiene una mayor exp o - sición y, por ende, una mayor probabilidad de que un fallo tenga consecuencias.

Matriz de Calibración de Impacto y Probabilidad para Riesgos de IA:			
Nivel	**Descripción**	**Criterios de Impacto**	**Criterios de Probabilidad (Ejemplos)**
5	**Catastrófico**	**Financiero:** pérdidas > 5% de la facturación anual; multas regulatorias máximas (p. ej., bajo el AI Act). **Reputacional:** daño severo y duradero a la marca a nivel global. **Regulatorio:** prohibición permanente del sistema; litigios de acción colectiva. **Operativo:** interrupción completa de un proceso de negocio crítico a nivel de toda la organización. **Ético-Social:** daño físico o psicológico grave generalizado; violación sistémica de derechos fundamentales.	**Casi seguro (>80%):** el dominio de riesgo representa >20% en el Repositorio MIT y el modelo es de alta complejidad sin controles de monitoreo en tiempo real.
4	**Mayor**	**Financiero:** pérdidas entre 1-5% de la facturación; multas significativas. **Reputacional:** daño significativo a la marca con cobertura mediática nacional. **Regulatorio:** sanciones regulatorias importantes; investigaciones formales. **Operativo:** interrupción grave de un proceso de negocio clave. **Ético-Social:** daño significativo a grupos vulnerables; discriminación demostrada.	**Probable (60-80%):** el dominio de riesgo representa >15% en el Repositorio MIT Y el modelo es de alta complejidad con controles de monitoreo limitados.
3	**Moderado**	**Financiero:** pérdidas notables pero manejables. **Reputacional:** daño a la marca reversible con gestión de crisis. **Regulatorio:** notificaciones de incumplimiento; requerimientos de remediación. **Operativo:** interrupción parcial de procesos de negocio. **Ético-Social:** daño limitado a individuos; sesgos detectados y corregibles.	**Posible (40-60%):** el dominio de riesgo representa >10% en el Repositorio MIT O el modelo utiliza datos de fuentes no verificadas.
2	**Menor**	**Financiero:** pérdidas menores, absorbidas por el presupuesto operativo. **Reputacional:** impacto negativo menor y localizado. **Regulatorio:** advertencias o recomendaciones de mejora. **Operativo:** ineficiencias o errores menores en procesos no críticos. **Ético-Social:** quejas individuales aisladas.	**Poco probable (20-40%):** el dominio de riesgo representa <10% en el Repositorio MIT Y se aplican controles preventivos y detectives robustos.
1	**Insignificante**	**Financiero:** pérdidas insignificantes. **Reputacional:** impacto nulo o mínimo. **Regulatorio:** sin consecuencias. **Operativo:** molestias menores sin impacto en el negocio. **Ético-Social:** sin daño perceptible.	**Raro (<20%):** el dominio de riesgo tiene baja prevalencia en el Repositorio MIT Y el modelo es simple, bien entendido y opera en un entorno controlado con supervisión humana.

Figura 29. Escala de evaluación de impacto y probabilidad aplicada a riesgos de IA.

Fase 3: Priorización y Visualización en la Matriz de Riesgos

Con las calificaciones de impacto y probabilidad, cada riesgo de IA se posiciona en la matriz de riesgos corporativa, a menudo representada como un mapa de calor. La fórmula **Nivel de Riesgo = Impacto x Probabilidad** se utiliza para determinar la criticidad de cada riesgo. Esta visualización permite a la dirección priorizar los recursos de mitigación. Los riesgos situados en la zona "roja" (alto impacto, alta probabilidad) requieren atención inmediata y planes de acción urgentes, mientras que los de la zona "verde" (bajo impacto, baja probabilidad) pueden ser aceptados o monitoreados con menor frecuencia.

Este proceso es también fundamental para definir el **apetito y la tolerancia al riesgo** específicos para la IA. El apetito de riesgo es el nivel general de riesgo que la organización está dispuesta a aceptar para alcanzar sus objetivos estratégicos. La matriz de riesgos ayuda al Consejo de Administración y a la alta dirección a visualizar si el perfil de riesgo de IA actual excede este apetito, guiando así las decisiones estratégicas sobre qué proyectos de IA emprender, escalar o detener.

9.2.

ESTRUCTURA DE GOBIERNO Y FUNCIONES DE SUPERVISIÓN DE RIESGOS DE IA

Un marco de controles robusto es ineficaz sin una estructura de gobierno clara que defina roles, responsabilidades y mecanismos de supervisión. La gobernanza de la IA debe integrarse en la estructura de gobierno corporativo existente, no operar como un silo aislado.

INTEGRACIÓN EN EL MODELO DE LAS TRES LÍNEAS

El Modelo de las Tres Líneas del Instituto de Auditores Internos (en adelante, IIA) es un estándar ampliamente reconocido para la asignación de roles y responsabilidades en la gestión de riesgos. Se propone una adaptación de este modelo para la gobernanza específica de la IA, reconociendo que la gestión eficaz de estos riesgos requiere una fusión de competencias técnicas y de negocio en cada una de las líneas.

Primera Línea (en adelante, 1LOD): Gestión y Ejecución

- **Quiénes son:** los equipos de desarrollo de IA, científicos de datos, ingenieros de Machine Learning (en adelante, ML), y los dueños de los procesos de negocio que utilizan los sistemas de IA.
- **Responsabilidades:** son los "dueños" del riesgo en el día a día. Su responsabilidad es implementar los controles en el diseño y desarrollo, realizar pruebas (incluyendo de equidad, robustez y seguridad), documentar los modelos (por ejemplo, a través de "model cards" o fichas de modelo) y monitorear el rendimiento operativo de los sistemas de IA.

Segunda Línea (en adelante, 2LOD): Supervisión y Asesoramiento

- **Quiénes son:** funciones corporativas como Gestión de Riesgos, Cumplimiento, Ciberseguridad (podría estar ubicada en la 1LOD o en 2LOD), Legal y Privacidad.
- **Responsabilidades:** establecen las políticas y los marcos de gestión de riesgos de IA. Proporcionan experiencia, herramientas y orientación a la primera línea. Realizan una supervisión y un desafío independientes (independent challenge) de las actividades de gestión de riesgos de la primera línea para asegurar su eficacia.
- **Elemento Crítico:** Comité de Ética y Riesgos de IA. Se recomienda la creación de un comité directivo multidisciplinario que actúe como un órgano central de gobernanza para la IA. Este comité, compuesto por líderes de tecnología, riesgo, legal, ética y negocio, es responsable de:
  - » Revisar y aprobar las Evaluaciones de Impacto de la IA (AIA) para proyectos de alto riesgo.
  - » Resolver dilemas éticos que surjan del uso de la IA y que no puedan ser resueltos a nivel de proyecto.
  - » Supervisar el perfil de riesgo de IA de la organización e informar periódicamente al comité de riesgos del Consejo.
  - » Mantenerse al día sobre la evolución regulatoria y tecnológica para adaptar las políticas de la organización.

Tercera Línea: Aseguramiento Independiente

- **Quién es:** la función de Auditoría Interna.
- **Responsabilidades:** proporciona un aseguramiento objetivo e independiente al Consejo de Administración sobre la eficacia del marco de gobierno y gestión de riesgos de la IA. Su mandato debe expandirse para incluir auditorías específicas de IA:
  - » Auditorías de los procesos de gobernanza de la IA, incluyendo la efectividad del Comité de Ética y Riesgos de IA.
  - » Auditorías técnicas de los modelos de IA (evaluando la calidad de los datos, la equidad, la robustez y la seguridad), a menudo en colaboración con especialistas externos si la función interna carece de la experiencia necesaria.
  - » Auditorías de cumplimiento con las regulaciones de IA aplicables, como el Reglamento de IA de la Unión Europea.

ROLES Y RESPONSABILIDADES DEL ALTO NIVEL: EL PAPEL DEL CONSEJO DE ADMINISTRACIÓN

La supervisión final del riesgo de la IA recae en el Consejo de Administración. Su función es asegurar que la dirección ha implementado un marco de gestión de riesgos adecuado y que la estrategia de IA de la organización está alineada con su apetito de riesgo. Existen varios modelos para la participación del Consejo en esta supervisión:

- **Comité de Riesgos:** el comité de riesgos existente puede ampliar su mandato para incluir la supervisión específica de los riesgos de la IA, recibiendo informes periódicos del Comité de Ética y Riesgos de IA y del director de Riesgos (en adelante, CRO).
- **Comité de Tecnología Dedicado:** en organizaciones donde la tecnología es central para el modelo de negocio, se puede crear un comité específico del Consejo para la tecnología y la ciberseguridad, que asumiría la supervisión principal de la IA.
- **Participación de Todo el Consejo:** el Consejo en pleno recibe formación regular sobre IA y discute los riesgos estratégicos asociados como parte de su agenda habitual, asegurando una comprensión y supervisión a nivel de todo el órgano de gobierno.

EL CICLO DE VIDA DE LA GOBERNANZA: HERRAMIENTAS Y REPORTE

La gobernanza de la IA no es un evento único, sino un ciclo continuo de supervisión y mejora.

- **Evaluación de Impacto de la IA (AIA):** es la puerta de entrada al ciclo de vida de la gobernanza. Ningún proyecto de IA de impacto significativo debería comenzar sin una AIA que documente su propósito, partes interesadas, riesgos potenciales y planes de mitigación.
- **Cuadro de Mando de KRIs de IA:** el Comité de Ética y Riesgos de IA y la segunda línea deben mantener un cuadro de mando con los KRIs clave. Este cuadro de mando debe ser reportado trimestralmente al Comité de Riesgos del Consejo, destacando las tendencias, los umbrales superados y los incidentes ocurridos.
- **Canales de Escalada:** deben existir canales claros y protegidos para que cualquier empleado pueda plantear preocupaciones éticas o de riesgo sobre un sistema de IA (por ejemplo, un canal de denuncias), y para que los incidentes se escalen rápidamente a la estructura de gobierno adecuada para su resolución.

Para clarificar las responsabilidades en este complejo ecosistema, se propone una matriz de responsabilidades (RACI).

R = Responsable (ejecuta la tarea)

A = Corresponsable/Aprobador (responsable final)

C = Consultado (aporta información)

I = Informado (se le mantiene al día)

Matriz RACI de gobernanza y supervisión de riesgos de Inteligencia Artificial

Actividad	Equipo de Desarrollo IA (1LOD)	Dueño del Proceso de Negocio (1LOD)	Función de Riesgos (2LOD)	Comité de Ética y Riesgos de IA (2LOD)	Auditoría Interna (3LOD)	Consejo de Administración
Definir la Política de IA	C	R	A	C	I	I
Realizar Evaluación de Impacto (AIA)	R	C	A	A	I	I
Implementar Controles Técnicos	R	I	C	I	I	I
Definir KRIs para un Nuevo Modelo	C	R	A	C	I	I
Monitorear KRIs en Producción	R	A	C	I	I	I
Responder a un Incidente de IA	R	A	C	I	I	I
Auditar un Modelo de Alto Riesgo	C	C	C	I	A	I
Reportar el Perfil de Riesgo de IA	I	I	R	A	C	A

Figura 30. Matriz RACI de gobernanza y supervisión de riesgos de Inteligencia Artificial.

10.

Reflexiones finales para la función de Ciberseguridad ante la gestión de riesgos de la Inteligencia Artificial

La adopción generalizada de la inteligencia artificial marca un punto de inflexión en la forma en que las organizaciones entienden la tecnología, el riesgo y la responsabilidad. La IA ha dejado de ser un elemento accesorio para convertirse en un **activo estructural**, integrado en procesos críticos, decisiones automatizadas y servicios esenciales. Esta centralidad explica por qué los riesgos asociados a la IA no pueden abordarse como una mera extensión de los riesgos tecnológicos tradicionales. La IA introduce una **nueva tipología de riesgo transversal, dinámico y sistémico,** que combina de forma inseparable dimensiones técnicas, organizativas, jurídicas y sociales.

Desde esta premisa, el análisis desarrollado a lo largo del informe pone de manifiesto que los riesgos de la IA no se concentran en un único momento ni en un único actor. Por el contrario, emergen y evolucionan a lo largo de todo el ciclo de vida del sistema, especialmente una vez desplegado y en interacción con entornos reales, datos cambiantes, usuarios y terceros. El Repositorio de Riesgos de IA del MIT evidencia que la mayor parte de los incidentes y fallos relevantes se producen **en fase operativa**, cuando el control formal tiende a relajarse y los supuestos iniciales dejan de ser válidos. Esta realidad obliga a abandonar enfoques de evaluación puntual y a asumir que la gestión del riesgo de IA debe ser continua, adaptativa y basada en evidencia.

En este contexto, el valor del Repositorio del MIT no reside únicamente en su exhaustividad, sino en su capacidad para **ordenar la complejidad.** La combinación de la taxonomía causal y la taxonomía por dominios de impacto permite comprender de manera estructurada por qué se produce un riesgo, quién lo origina y cuándo se manifiesta, conectando estos elementos con daños reales sobre personas, organizaciones y sistemas. Esta estructura es esencial para pasar del análisis teórico a la gestión práctica del riesgo, ya que facilita la asignación de responsabilidades, la priorización de controles y la trazabilidad entre riesgos identificados, decisiones adoptadas y medidas de mitigación implementadas.

A partir de esta base técnica, el informe enlaza de forma natural con el marco regulatorio europeo, que refuerza estas exigencias desde una perspectiva jurídica y de gobernanza. El AI Act, el RGPD, la Directiva NIS2 y el Esquema Nacional de Seguridad configuran un entorno normativo que ya no se conforma con declaraciones de cumplimiento, sino que exige **capacidad real de control, supervisión y rendición de cuentas.** En este escenario, el Repositorio del MIT actúa como un elemento vertebrador entre la norma y la práctica, al proporcionar la evidencia técnica necesaria para traducir obligaciones legales en controles operativos, verificables y auditables.

Esta convergencia entre riesgo técnico y exigencia normativa conduce a una conclusión clave: la gestión del riesgo de IA debe integrarse en los marcos corporativos de gestión del riesgo y del control interno, funcionando como un **sistema operativo**, no como un catálogo estático. La alineación del Repositorio del MIT con estándares como la ISO/IEC 23894 o el NIST AI Risk Management Framework permite articular modelos de gobernanza basados en ciclos de evaluación continua, indicadores de riesgo, gates de decisión y procesos de mejora permanente. La IA exige, por tanto, marcos de control capaces de evolucionar al mismo ritmo que la tecnología y su contexto de uso.

Es precisamente en este punto donde **la función de Ciberseguridad adquiere un papel central y claramente diferenciado.** El análisis de la distribución de riesgos del Repositorio del MIT muestra que una parte sustancial de los riesgos se concentra en dominios estrechamente vinculados a la seguridad del sistema, la privacidad, la explotación maliciosa y las limitaciones técnicas de los modelos. Esto confirma que la IA amplía de forma significativa la superficie de ataque, introduce nuevas técnicas adversarias y crea dependencias críticas con proveedores y servicios externos. En consecuencia, la Ciberseguridad deja de ser una función orientada exclusivamente a la protección de infraestructuras para convertirse en un eje de gobernanza del riesgo de IA, con responsabilidad directa sobre la resiliencia, la trazabilidad y la capacidad de respuesta ante incidentes complejos.

Esta evolución implica un cambio profundo en el rol del CISO y de los equipos de seguridad. La función de Ciberseguridad debe situarse en la intersección entre tecnología, negocio y cumplimiento, liderando la integración de los riesgos de IA en los mapas de riesgo corporativos, definiendo indicadores clave de riesgo, participando en la evaluación de casos de uso y asegurando que los sistemas de IA cuentan con controles efectivos antes y después de su despliegue. En un entorno dominado por modelos generativos, agentes autónomos y servicios de IA como servicio, **no existe gobernanza de la IA sin una ciberseguridad fuerte, anticipatoria y basada en evidencias.**

Este papel se vuelve aún más crítico cuando se analiza la cadena de suministro tecnológica. El perímetro real de riesgo ya no coincide con los límites de la organización, sino que se extiende a proveedores de modelos, plataformas de IA, datasets externos e infraestructuras cloud. El informe evidencia que, sin una gobernanza explícita de esta cadena (integrada en los marcos de NIS2 y ENS), el riesgo se fragmenta y la responsabilidad se diluye. La función de Ciberseguridad está llamada a liderar este control, garantizando que los riesgos asociados a terceros, cambios de modelo o dependencias críticas sean identificados, evaluados y monitorizados de forma continua.

Más allá de los riesgos técnicos y operativos, el Repositorio del MIT pone de manifiesto impactos socioeconómicos, éticos y ambientales que amplían el alcance tradicional de la Ciberseguridad. Riesgos como la discriminación algorítmica, la pérdida de autonomía humana o la erosión de la confianza social no pueden gestionarse únicamente desde una lógica técnica, pero sí requieren mecanismos de control, supervisión y escalado cuando se materializan. En este sentido, la Ciberseguridad se convierte también en un garante de la **confianza digital,** contribuyendo a que la adopción de la IA se produzca de forma segura, legítima y sostenible.

Todo ello converge en una conclusión de carácter estratégico: **la gestión del riesgo de IA es una responsabilidad indelegable de la Alta Dirección, pero su materialización operativa recae en gran medida en la función de Ciberseguridad.** La IA no reduce la responsabilidad humana; la intensifica. Gobernar la IA implica definir apetito de riesgo, establecer estructuras de supervisión y exigir evidencias continuas de control, y en todos estos ámbitos la Ciberseguridad desempeña un papel habilitador esencial.

En definitiva, este informe demuestra que una gestión rigurosa y estructurada del riesgo de IA no constituye un freno a la innovación, sino su principal catalizador. Las organizaciones que integren la IA en marcos sólidos de gobernanza, seguridad y cumplimiento estarán mejor preparadas para innovar con confianza, responder a incidentes complejos y mantener la credibilidad frente a reguladores, clientes y sociedad. Para la función de Ciberseguridad, este contexto representa una oportunidad estratégica: evolucionar desde la protección de sistemas hacia la **orquestación del riesgo de la inteligencia artificial,** posicionándose como un actor clave en la construcción de una innovación responsable, resiliente y sostenible.

11. Referencias

- Anthropic PBC. (2023). Doe et al. v. Anthropic PBC (class action lawsuit regarding the use of copyrighted books for model training).
- Council of Europe. HUDERIA: Human Rights, Democracy and the Rule of Law Impact Assessment for AI systems.
- European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). *https://eur-lex.europa.eu/eli/reg/2016/679/oj*
- European Parliament & Council of the European Union. (2022). Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS2). *https://eur-lex.europa.eu/eli/dir/2022/2555/oj*
- European Parliament & Council of the European Union. (2024). Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
- European Union. (2022). Digital Services Act (DSA). *https://digital-strategy.ec.europa.eu/en/policies/digital-services-act*
- Gobierno de España. (2022). Real Decreto 311/2022, por el que se regula el Esquema Nacional de Seguridad. Boletín Oficial del Estado. *https://www.boe.es/eli/es/rd/2022/05/03/311*
- International Organization for Standardization. (2018). ISO 31000:2018 Risk management — Guidelines. ISO. *https://www.iso.org/standard/65694.html*
- International Organization for Standardization. (2023). ISO/IEC 23894:2023 Artificial intelligence — Risk management. ISO. *https://www.iso.org/standard/77304.html*
- International Organization for Standardization. (2023). ISO/IEC 42001:2023 Artificial intelligence — Management system. ISO. *https://www.iso.org/standard/81230.html*
- International Organization for Standardization. ISO/IEC 42005 Artificial intelligence — Impact assessment. ISO. *https://www.iso.org/standard/42005*
- ISMS Forum. Guía para la gestión de crisis por ciberincidente en la cadena de suministro. *https://www.ismsforum.es/ficheros/descargas/guia-para-la-gestion-de-crisis-por-ciberincidente.pdf*
- ISMS Forum. Retos y soluciones en la gestión de la ciberseguridad en la cadena de suministro. *https://www.ismsforum.es/ficheros/descargas/270525retos-y-soluciones-de-la-gestion-de-la.pdf*
- Massachusetts Institute of Technology. (2025). MIT AI Risk Repository. MIT Schwarzman College of Computing. *https://airisk.mit.edu/*
- Microsoft. (2016). Tay chatbot incident.
- MITRE Corporation. (n.d.). MITRE ATLAS: Adversarial Threat Landscape for Artificial Intelligence Systems. *https://atlas.mitre.org/*
- MITRE Corporation. (n.d.). MITRE ATT&CK®: Supply Chain Compromise (T1195). *https://attack.mitre.org/techniques/T1195/*
- MITRE Corporation. (2024). ATLAS case studies on large language models and cybersecurity.
- National Institute of Standards and Technology. (2022). NIST SP 800 161 Rev. 1: Cybersecurity supply chain risk management practices for systems and organizations. NIST. *https://csrc.nist.gov/publications/detail/sp/800-161/rev-1/final*
- National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST. *https://www.nist.gov/itl/ai-risk-management-framework*
- National Institute of Standards and Technology. (2023). NIST SP 800 218: Secure software development framework (SSDF), version 1.1. NIST. *https://csrc.nist.gov/publications/detail/sp/800-218/final*
- New South Wales Government. NSW Artificial Intelligence Assessment Framework.
- OWASP Foundation. OWASP AI Exchange. *https://owasp.org/www-project-ai-exchange/*
- OWASP Foundation. (2025). OWASP GenAI Security Project. *https://genai.owasp.org/*
- Software Engineering Institute, Carnegie Mellon University. (2024). Considerations for evaluating large language models for cybersecurity tasks. *https://www.sei.cmu.edu/library/considerations-for-evaluating-large-language-models-for-cybersecurity-tasks/*
- The New York Times Company v. OpenAI, Inc., & Microsoft Corp. (2023). Copyright infringement litigation.
- Williams v. City of Detroit. (2020). Wrongful arrest due to facial recognition error.

12. Anexos

12.1.

RELACIÓN DEL MIT AI RISK REPOSITORY CON OTROS MARCOS DE REFERENCIA, ESTÁNDARES, REGULACIONES, Y OTRAS PUBLICACIONES RELEVANTES

Correspondencia con otros marcos de referencia			
Referencia	**Temática / Alcance**	**Link (URL)**	**Comentarios (¿Por qué es relevante?)**
MIT AI Risk Repository	Taxonomía de riesgos de IA	*https://airisk.mit.edu/*	Referencia central de este documento para riesgos IA de carácter técnicos, sociales y organizacionales.
ISO/IEC 23894:2023 – AI Risk Management	Gestión de riesgos de IA	*https://www.iso.org/standard/81230.html*	Potente para operacionalizar los riesgos del MIT. Ver apartado 5.2
ISO/IEC 42001:2023 – AI Management System	Sistema de gestión de IA	*https://www.iso.org/obp/ui/en/#iso:std:44545:en*	Traduce riesgos MIT en políticas, controles y gobernanza. Ver mención en el apartado 6.2.
ISO/IEC DIS 42005 – Information technology - AI system impact assessment	Análisis de Impacto de la Inteligencia Artificial (AIIA)	*https://www.nist.gov/itl/ai-risk-management-framework*	Apoyo formal y metodológico a la realización de AIIA. Ver Capítulo 9.
NIST AI RMF 1.0	Riesgo y gobernanza de IA	*https://artificialintelligenceact.eu/*	Puente entre riesgos técnicos, impacto social y controles. Ver Capítulos 5, 6 y 9.
EU Artificial Intelligence Act	Regulación de IA basada en riesgo	*https://gdpr.eu/*	Convierte riesgos en obligaciones legales. Ver Capítulo 6.
GDPR / RGPD	Protección de datos	*https://www.nist.gov/itl/ai-risk-management-framework*	Clave para riesgos de privacidad y decisiones automatizadas. Ver apartado 8.4
OECD AI Principles	Principios de IA responsable	*https://oecd.ai/en/ai-principles*	Marco global de alineación ética y política.
UNESCO Recommendation on the Ethics of AI	Ética y derechos humanos	*https://www.unesco.org/en/artificial-intelligence/recommendation-ethics*	Dimensión social y humana de los riesgos.
OWASP GenAI Security Project	Guía(s) Open para mitigar riesgos de seguridad de soluciones de IA Generativa	*https://genai.owasp.org/*	Apoyo y referencia documental para varios tipos de interesados (incluyendo desarrolladores). Ver Capítulo 5.
MITRE ATLAS Matrix	Base de conocimientos de amenazas IA y técnicas relacionadas	*https://atlas.mitre.org/matrices/ATLAS*	Útil para threat modeling, red teaming y diseño de pruebas adversariales repetibles. Ver Capítulos 4, 5, y 6.

12.2.

CÓMO USAR HERRAMIENTAS DE IA GENERATIVA CON EL *MIT AI RISK REPOSITORY*

El uso de herramientas de IA generativa (ej. ChatGPT, Google's GEMs, Copilot, etc.), y en particular sus correspondientes personalizaciones, conjuntamente con el *MIT AI Risk Repository* puede convertirse en una palanca poderosa si se entiende correctamente su rol: la IA no sustituye el análisis de riesgos, sino que puede amplificar o aumentar la exploración, estructuración y contextualización de los mismos.

A continuación, se describen algunas posibles líneas de aplicación:

- Una primera aplicación consiste en **emplear modelos generativos para expandir y contextualizar las categorías del repositorio.** A partir de una categoría de riesgo identificada (por ejemplo, "Amplificación de los sesgos" o "mal uso"), la IA puede generar escenarios sectoriales específicos, casos de uso plausibles o cadenas causales adaptadas a una organización concreta. Esto transforma la taxonomía "inicial" del Repositorio, que es conceptual y relativamente estática, en un instrumento dinámico de simulación y análisis prospectivo. En lugar de limitarse a leer la categoría, el equipo puede "dialogar" con ella, generando hipótesis y contrastarlas con su propia experiencia y contexto operativo concreto.
- Una segunda línea consiste en **usar IA generativa como herramienta de contraste crítico.** Dado que el repositorio no es exhaustivo y se basa en un corpus académico delimitado, los modelos pueden ayudar a identificar posibles riesgos adyacentes o emergentes no explícitamente recogidos. Esto no sustituye el juicio experto, pero puede servir como mecanismo de "ampliación del horizonte" durante sesiones de gestión de riesgos (especialmente, dinámicas participativas, workshops, etc.), ayudando a evitar un posible sesgo de anclaje en "checklists" o fuentes predefinidas. Aquí la IA generativa puede funcionar como catalizador creativo, no como "árbitro" ni como herramienta de validación infalible.
- Una tercera aplicación puede ser una **combinación de los anteriores para estructurar procesos de documentación y trazabilidad del razonamiento o "lógica" de riesgo.** La IA puede asistir en la redacción de análisis comparativos, mapas de impacto, narrativas de riesgo o matrices preliminares alineadas con el EU AI Act, ENS o NIS2, o el propio *MIT AI Risk Repository*. Siempre teniendo en cuenta que la decisión final sobre criticidad, aceptabilidad o priorización debe permanecer en manos humanas (el llamado HIL: "Human-in-the-Loop"). El verdadero potencial emerge de integrar la IA como herramienta de apoyo dentro de una arquitectura creada a consciencia donde la automatización y el criterio experto se complementan en lugar de competir uno con el otro.

isms forum INTERNATIONAL INFORMATION SECURITY COMMUNITY

GIA GRUPO DE INTELIGENCIA ARTIFICIAL