cisco isms forum GIA | GRUPO DE INTELIGENCIA ARTIFICIAL

Gobierno de la IA

Cisco, a través de su programa de inversiones CDA "Country Digital Acceleration", llamado DIGITALIZA en España, está comprometido en acelerar la digitalización de nuestro país en los distintos ámbitos de nuestra economía y sociedad. En particular, el desarrollo del conocimiento y concienciación de la práctica de ciberseguridad en las organizaciones españolas es un punto clave. En este caso, el análisis se centra en establecer un marco integral de gobierno de la Inteligencia Artificial, abordando desde los principios éticos y regulatorios hasta la gestión de riesgos, la protección de datos y la definición de roles y responsabilidades. El documento proporciona directrices prácticas para garantizar una IA confiable, segura y alineada con la normativa vigente, facilitando la toma de decisiones estratégicas y la adaptación organizativa ante los retos tecnológicos y regulatorios emergentes.

Índice

Miembros del proyecto

Coordinadores:

Angel Ortiz
Sergio Padilla

Gestión del proyecto:

Beatriz García

Participantes:

Alberto López
Alberto Pinedo
Alberto Torralba
Carlos A. Sáiz
Elena Mora
Enrique Cervantes
Jose Ramón Monleón
Ignacio Cagiga
Jaime Requejo
Javier Pinillos
Jesús Alonso
Jesús Muñoz
Manuel Barrios
Manuel Ruiz del Corral
Mario Encinas
Marta Martínez
Rafael Fernández
Rafael Tenorio

Diseño y maquetación

Susana Marín

Revisión:

Angel Pérez
Carlos A. Sáiz
Francisco Lázaro

Introducción

01.

1.1. Objetivo del Documento de Gobierno de IA

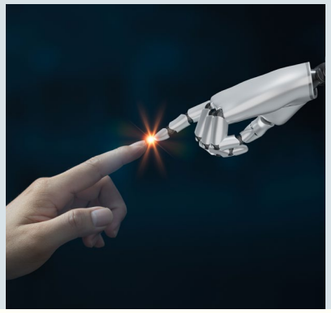
El objetivo de este documento es establecer el marco fundamental y las directrices esenciales para el **Gobierno de la Inteligencia Artificial (en adelante, IA)**, sirviendo como una hoja de ruta estructurada y completa para que las organizaciones puedan desarrollar, desplegar y utilizar sistemas de IA de manera **segura, resiliente, responsable, ética y conforme a la normativa vigente.**

Aspira a ser **mandato estratégico y operativo** para asegurar la **continuidad del negocio** y la **resiliencia cibernética** en la era de la IA. La rápida adopción de la IA -gracias, entre otros, al auge de los Grandes Modelos de Lenguaje (LLM) y la Inteligencia Artificial Generativa (GenAI)-, está transformando procesos clave para el negocio, las operaciones, la seguridad de la información o el servicio al cliente, y se espera un incremento significativo en su adopción en los próximos tres años. Sin embargo, este crecimiento exponencial y su potencial disruptivo conllevan importantes desafíos, incluyendo legales, éticos y sociales. No en vano, la gran mayoría de las organizaciones percibe un **riesgo potencial considerable** (el 90.8% lo califica de moderado a extremo), con los riesgos más críticos concentrados en el ámbito de **operaciones y seguridad de la información.**

Así, un objetivo clave de este marco es proporcionar un **Modelo de Gobierno IA** que no solo permita la adecuada gestión de los riesgos legales y éticos, sino que garantice que el desarrollo y despliegue de la IA se realice de forma segura, **confiable y robusta**, gestionando los riesgos a lo largo de **todo el ciclo de vida del servicio o sistema.**

Fomentar una IA **confiable, ética y segura**

El documento se enfoca en la necesidad de que los sistemas de IA sean con(fiables), lo cual implica que deben ser, simultáneamente, **lícitos, éticos y robustos.**



01.

Enfoque centrado en el ser humano y la ética

La IA debe centrarse en el bienestar y la seguridad de las personas. Esto requiere la integración de valores éticos desde las etapas iniciales del diseño (ethics by design). Los principios éticos fundamentales que se deben abordar incluyen el respeto a la **autonomía humana**, la **prevención del daño**, la **equidad** (evitando sesgos y discriminación) y la **explicabilidad y transparencia**.

02.

Mitigación de riesgos

Es necesaria una gestión continua de los riesgos derivados, y ampliamente percibidos, del uso de la IA. Se debe identificar, evaluar y tratar los riesgos en todo el ciclo de vida de la IA, lo cual es crucial para sistemas con alto grado de autonomía y capacidad de toma de decisiones.

03.

Transparencia y explicabilidad

Los sistemas de IA deben funcionar con un nivel de transparencia suficiente para que los responsables de su despliegue puedan interpretar y utilizar correctamente sus resultados. La obligación incluye facilitar la comprensión de la lógica aplicada en la toma de decisiones, garantizando el cumplimiento del principio de transparencia del RGPD y evitando los sistemas de “caja negra”.

Desprendiéndose de lo anterior, el riesgo inherente a la IA exige un enfoque que integre la ciberseguridad desde las etapas más tempranas de un proyecto (security by design), gestionando proactivamente el riesgo y minimizando en lo posible el riesgo operacional derivado. Por ello, los objetivos de seguridad, para los responsables de seguridad de la información, se centran en:

Garantizar la robustez y la resiliencia

La fiabilidad de la IA (incluyendo validez, robustez y ciberseguridad) es una obligación crítica, especialmente para los sistemas de alto riesgo (Art. 15 del Reglamento de Inteligencia Artificial, en adelante nos referiremos al mismo como RIA)[1]. El sistema debe ser resistente a errores, fallos e incoherencias, y ser capaz de resistir intentos no autorizados de alteración, como los ataques de **envenenamiento de datos** o los **ejemplos adversarios** (evasión de modelos).

Mitigar amenazas específicas

Yendo más allá de los ciberataques tradicionales. Es imperativo desarrollar un **catálogo integral de escenarios de ataque específicos de IA** (como los detallados en MITRE ATLAS y OWASP Top 10 for LLM) y diseñar simulaciones realistas para evaluarlos.

Definir la rendición de cuentas en seguridad (RACI)

Establecer una **Matriz de Responsabilidad (RACI)** clara donde los equipos de seguridad de la información de una entidad sean responsables directos (en inglés, **responsible y accountable -R, A-)**, alineando la monitorización, el registro de auditoría y la gestión de riesgos, con las mejores prácticas.

Establecer control continuo y trazabilidad

Necesidad de implementar **mecanismos de monitorización continua** para identificar actividades anómalas (como el Shadow AI o uso no autorizado de sistemas públicos) y garantizar la existencia de **archivos de registro (logs)** automáticos y detallados a lo largo de todo el ciclo de vida, lo cual es obligatorio para sistemas de alto riesgo (Art. 12 RIA).

[1] Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y se modifican determinados actos legislativos de la Unión (DOUE-L-2024-81079). https://www.boe.es/buscar/doc.php?id=DOUE-L-2024-81079

Alineamiento con el **marco regulatorio y cumplimiento normativo**

Como parte de los objetivos del documento, es importante guiar a la organización hacia el cumplimiento del creciente y exigente panorama regulatorio, evitando las consecuencias económicas y reputacionales derivadas del incumplimiento, haciendo imprescindible un marco de gobernanza sólido.

01

Adaptación al Reglamento de IA (RIA)

Uno de los objetivos principales de cualquier empresa u organización debe ser garantizar el cumplimiento del **Reglamento (UE) 2024/1689 de Inteligencia Artificial (RIA)**, considerado la primera legislación exhaustiva a nivel mundial en este ámbito. El RIA establece normas armonizadas basadas en un enfoque de riesgos (riesgo inaceptable, alto, limitado y mínimo).

02

Requisitos de alto riesgo

Se busca asegurar que los sistemas clasificados como de alto riesgo, que pueden **impactar significativamente en la seguridad, salud o derechos fundamentales,** cumplan con los requisitos obligatorios del RIA, incluyendo la gestión de riesgos, la gobernanza de datos, la documentación técnica, la trazabilidad, la transparencia, la supervisión humana y la fiabilidad. También es muy relevante la alta probabilidad de que utilicemos modelos y servicios de IA externos (aprovisionamiento como consumidor o servicio SaaS...), por lo que se debe aclarar la responsabilidad, exigiendo al proveedor garantías suficientes.

03

Integración normativa

No solo es necesario asegurar el cumplimiento de la normativa específica de IA, sino también de otras regulaciones esenciales, como el **Reglamento General de Protección de Datos (en adelante, RGPD[2]),** la normativa de **Propiedad Intelectual/Industrial o cualesquiera normas o requisitos obligatorios relativos a la seguridad de la información y la ciberseguridad.** El marco debe demostrar algo ya comentado como es la obligación de accountability y responsabilizarse proactivamente de los sistemas utilizados o desarrollados

04

Marco contractual robusto

El documento busca, también, establecer un marco contractual robusto para regular las condiciones y responsabilidades de las partes implicadas en la contratación de sistemas y servicios de IA, siendo especialmente relevante para la gestión de proveedores de sistemas de IA.

[2] Parlamento Europeo y Consejo de la Unión Europea. (2016). Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (Reglamento General de Protección de Datos). Diario Oficial de la Unión Europea, L119, 1-88. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32016R0679

Proporcionar un **marco operativo y práctico**

El documento está destinado a un público diverso (CISO, CAIO, CDO, CIO y otros CXO relevantes e involucrados en cada organización, DPO, ingenieros, desarrolladores, equipos legales...) y tiene como fin ser un recurso práctico y adaptable.

01.

Estructuras de gobierno y políticas internas

Una cuestión de importancia de esta guía es orientar a las organizaciones para que definan un **Modelo de Gobierno IA** que incluya la adaptación de las estructuras internas, la definición de roles y responsabilidades claros y la elaboración de **políticas y procedimientos internos.**

02.

Guía metodológica

Se proporciona una metodología para **incrustar la gobernanza a lo largo de todo el ciclo de vida de la IA** (diseño, desarrollo, despliegue, seguimiento/control y retirada final). Esto incluye directrices sobre la calidad del dato, la prevención de sesgos y la documentación técnica.

03.

Cultura y capacitación

Se debe impulsar una cultura organizacional consciente, transversal y empoderada, con programas de formación específicos sobre seguridad, ética y riesgos de la IA para todo el personal, ya que la falta de capacidad interna y la resistencia cultural son motivos comunes de fracaso en la implementación.

En resumen, el documento busca capacitar a las organizaciones para **aprovechar las fortalezas de la IA mitigando sus riesgos,** proporcionando una metodología estructurada y clara para que el desarrollo y uso de la IA se realice dentro de un marco de confianza, legalidad y responsabilidad.

Evolución desde la guía de 2023 **1.2.**

La evolución del Gobierno de la IA desde la publicación de la "Guía de ISMS Forum de 2023, Modelo de Gobierno de Inteligencia Artificial " hasta la actualidad, se caracteriza por un cambio fundamental: pasar de la **identificación conceptual de la necesidad** de gobernar la IA a la **implementación operativa y regulatoria obligatoria.**

Si bien, volviendo a 2023, el debate se centraba en establecer los cimientos de una **Estrategia de AI Management.** Hoy, la urgencia radica en la **adaptación inmediata** de las estructuras de gobierno, riesgo y cumplimiento (GRC) para manejar amenazas que han evolucionado en complejidad y volumen, alineándose con un marco normativo de alto impacto que ya ha sido publicado y cuyo cronograma de aplicación está en marcha.

La Guía de 2023 reconocía el potencial de la IA en la estrategia empresarial y la necesidad de establecer principios rectores. La evolución queda definida por la convergencia de la materialización de esos principios en requisitos legales obligatorios (RIA), la aceleración tecnológica impulsada por los Grandes Modelos de Lenguaje (LLM) y la IA Generativa y los riesgos de seguridad emergentes inherentes a estos sistemas:

01.

La consolidación del enfoque basado en el riesgo

2023

En 2023, ya se postulaba que el modelo de gobierno debía establecerse sobre la base de principios como la licitud, la ética y la robustez, y se reconocía la necesidad de gestionar los riesgos derivados de la IA de manera continua

2026

La principal evolución es la ratificación del RIA, que transforma la gestión de riesgos en una obligación legal estructurada y jerarquizada en sistemas de alto riesgo (SAR) y en modelos de riesgo sistémico, al clasificar los modelos de IA de propósito general (GPAI o GPM) en nivel ordinario y de riesgo sistémico.

[3] ISMS Forum. (2023). Modelo de Gobierno de la Inteligencia Artificial. Asociación Española para el Fomento de la Seguridad de la Información. https://www.ismsforum.es/ficheros/descargas/modelo-de-gobiernoconverted1700037593.pdf

02. *El Salto de la IA Generativa (GenAI)*

Aunque la IA Generativa ya había irrumpido en 2023 (con el fenómeno ChatGPT, principalmente), su adopción y sus riesgos han escalado rápidamente:

Previsión de adopción

2023

La encuesta de 2023 indicaba que la gran mayoría de las organizaciones **(88.89%)** esperaban un incremento del uso de GenAI en los siguientes 3 años.

2026

El enfoque ha cambiado de ser una oportunidad tecnológica a ser también un desafío de seguridad crítico.

Riesgos emergentes personalizados

2023

Los riesgos se agrupaban en categorías generales **(discriminación, falta de ética, riesgo operativo).**

2026

La evolución requiere ahora centrarse en riesgos específicos de LLM y GenAI, tal como los define el **OWASP Top 10 para LLM**, incluyendo los ***ataques de Inyección de Prompt***, la ***extracción de Modelos (Model Extraction)*** o el ***envenenamiento de Datos (Data Poisoning)***, que son descritos más adelante, en los capítulos pertinentes.

Gobernanza por diseño (AIGD)

2023

Recomendación de aplicación de los principios de **“Privacidad por Diseño” (Privacy by Design)**.

2026

La complejidad de los sistemas autónomos (Agentic AI Systems) y la GenAI han impulsado el enfoque de **AI Governance by Design (AIGD, del inglés AI Governance by Design)**. Esta evolución directa de los principios de “Privacidad por Diseño”, que ahora son aplicados de manera integral. La AIGD busca integrar consideraciones éticas, legales y sociales, directamente, en el desarrollo de los sistemas desde su concepción, evitando los costosos ajustes de los enfoques reactivos tradicionales.

Avances **tecnológicos, regulatorios** y **organizativos** 1.3.

El contexto actual del Gobierno de la IA está marcado por una **confluencia o convergencia de avances tecnológicos exponenciales, una intensa actividad regulatoria global y la necesidad imperativa de reestructurar las capacidades organizacionales** para gestionar el riesgo resultante. La situación ha pasado de ser un debate teórico sobre principios éticos a un **imperativo estratégico de cumplimiento legal y ciberseguridad** con plazos definidos.

Aquí nace, por lo tanto, la necesidad de una **reconfiguración profunda de la madurez organizacional** para lograr la ciber-resiliencia, como se irá viendo a lo largo del documento.

Panorama regulatorio global: **de la ética a la imposición legal**

El contexto regulatorio se transforma radicalmente, desde un enfoque inicial de publicación de directrices éticas (como las de la UNESCO o la OCDE) a la imposición de normativas con fuerza de ley, siendo el RIA el catalizador del cambio global.

1. La centralidad del Reglamento (UE) 2024/1689 de Inteligencia Artificial (RIA)

El RIA es considerado la primera legislación exhaustiva a nivel mundial diseñada específicamente para regular los sistemas de IA. Enfoque basado en el riesgo, dicta la estrategia de cumplimiento de las organizaciones multinacionales, como se irá viendo.

2. El marco regulatorio global fragmentado

Mientras la UE adopta un enfoque prescriptivo, el resto del mundo avanza con diferentes velocidades y filosofías:

Estados Unidos

Históricamente, el enfoque ha sido más flexible y no prescriptivo, centrado en la autorregulación sectorial y la incentivación de la innovación. El gobierno ha emitido órdenes ejecutivas buscando liderazgo global en IA a través de la cooperación internacional.

China y Brasil

Jurisdicciones en reacción. China presentó un proyecto de ley con foco en la investigación y el desarrollo. Brasil también tiene legislación en debate (PL 2.338/2023), distinguiendo sistemas de riesgo excesivo y alto riesgo.

Desafío Multinacional

Esta fragmentación obliga a las organizaciones con presencia global a desarrollar un enfoque de cumplimiento en múltiples jurisdicciones, adaptando los contratos y la tecnología al estándar más alto (que a menudo es el RIA de la UE).

3. La convergencia normativa con leyes existentes

El contexto actual no se limita al RIA. La gobernanza de la IA es una tarea transversal que exige la integración total con normativas preexistentes:

Privacidad y protección de datos personales

Históricamente, el enfoque ha sido más flexible y no prescriptivo, centrado en la autorregulación sectorial y la incentivación de la innovación. El gobierno ha emitido órdenes ejecutivas buscando liderazgo global en IA a través de la cooperación internacional.

Gobernanza y acceso a los datos (Data Governance Framework):

Más allá de la protección de datos personales, es fundamental asegurar una gobernanza de datos robusta, ya que los datos deben ser confiables y estar debidamente gobernados. Esto incluye, por ejemplo, la necesidad de considerar la nueva normativa europea que regula el intercambio y acceso a datos, como la Ley de Gobernanza de Datos (DGA), y la Ley de Datos (Data Act), que rigen la disponibilidad y uso compartido de la información, siendo cruciales para la calidad del dato y su trazabilidad.

Ciberseguridad

Normativas como **DORA** (Reglamento sobre la resiliencia operativa digital del sector financiero), la **Directiva NIS 2** y el **Reglamento de Ciberresiliencia** impactan directamente en la robustez y seguridad exigida a los sistemas de IA (Art. 15 RIA).

Propiedad intelectual, secretos empresariales y responsabilidad civil

Se requiere un enfoque contractual explícito para proteger la **Propiedad Intelectual** (licenciamiento de software, derechos de autor sobre inputs y outputs de GenAI) y los s**ecretos empresariales**, especialmente cuando se introduce **información confidencial** en sistemas de IA de terceros.

Normativa sectorial

Es indudable que, la gobernanza debe ser adaptada a la industria. En sectores altamente regulados, la IA debe cumplir con la normativa sectorial específica.

Por ejemplo, en el sector salud, cualquier sistema de IA que conlleve el procesamiento de datos de salud, se considera, al menos, un sistema de alto riesgo, sujeto a validaciones exhaustivas bajo regulaciones como el Reglamento de Productos Sanitarios (reglamento (UE) 2017/745), además de los requisitos estrictos de acceso a la historia clínica. El marco debe demostrar la obligación de accountability y responsabilizarse proactivamente de los sistemas utilizados o desarrollados.

Reconfiguración organizacional y la **gobernanza proactiva:** madurez, roles y gestión de proveedores

Es un gran desafío pasar de la intención a la operacionalización de la gobernanza, lo cual requiere nuevas estructuras, procesos y una cultura adaptada.

El contexto tecnológico y regulatorio, caracterizado por la volatilidad tecnológica y la regulación del RIA, obliga a las organizaciones a madurar rápidamente sus procesos de Gobernanza, Riesgo y Cumplimiento (GRC). La estrategia debe pivotar hacia la gobernanza proactiva y la resiliencia cibernética para evitar consecuencias económicas, reputacionales y operacionales de la disrupción del ecosistema de la IA.

Panorama actual de la IA

02.

La IA se ha consolidado como una tecnología revolucionaria que redefine múltiples sectores y aspectos de la vida cotidiana. Un punto de inflexión significativo ha sido la introducción de la **IA generativa y los modelos fundacionales**, como GPT (de OpenAI), Claude AI (de Anthropic) y Gemini (de Google), entre otros. Estos modelos, entrenados con vastas cantidades de datos de distintos dominios, se han erigido en pilares esenciales de las tecnologías más innovadoras, ofreciendo una **versatilidad y potencia extraordinarias** en tareas que van desde la generación de texto e imágenes hasta el análisis complejo de datos y la interacción natural con los usuarios. No obstante, esta creciente complejidad también plantea ***desafíos importantes*** en cuanto a la interpretabilidad de sus decisiones, la privacidad de los datos, la seguridad, la opacidad de los LLM y la necesidad de marcos éticos y regulatorios sólidos, como el RIA, para garantizar un desarrollo y uso responsable y fiable.

El panorama actual de la IA se caracteriza por una evolución rápida y transformadora, impulsada por los avances en la capacidad de almacenamiento, procesamiento y conectividad de datos, y de ahí la urgencia por definir su adecuada gobernanza.

2.1. Impacto transformador de los modelos fundacionales

El término modelo fundacional, que hoy puede resultar habitual o cotidiano, fue acuñado por primera vez por la Universidad de Stanford en agosto de 2021, y se utiliza para describir modelos de IA que han sido entrenados con **grandes cantidades de datos sin etiquetar provenientes de diferentes dominios, generalmente mediante aprendizaje autosupervisado.** Este método de aprendizaje consiste en que el modelo aprende a partir de datos sin etiquetas, generando sus propias señales de supervisión, en lugar de depender de etiquetas manuales, que son costosas y requieren mucho tiempo, de forma que el modelo se entrena para predecir una parte de los datos a partir de otras partes de los mismos datos.

Este tipo de aprendizaje les confiere la capacidad de **adaptarse a una amplia gama de tareas** y de aprender representaciones generales a partir de los datos, lo que los dota de una gran **versatilidad y potencia.** Los modelos fundacionales pueden luego ser ajustados con una cantidad mínima de datos etiquetados para tareas específicas (un proceso conocido como aprendizaje por transferencia), **reduciendo así drásticamente el coste y el tiempo de desarrollo de soluciones de IA.**

Dejando de lado la revolución que han supuesto en la forma en que hoy en día se desarrollan nuevos sistemas de IA, el impacto transformador de estos modelos es significativo y se puede apreciar en una gran diversidad de áreas. Algunas cuestiones fundamentales que destacar son:

→ **Evolución exponencial del uso de la IA:** el nivel de detalle, completitud y riqueza en las respuestas de los modelos fundacionales ha mejorado cualitativa y cuantitativamente las aplicaciones basadas en IA, y ha impulsado exponencialmente el uso de estas por parte de la sociedad, facilitando el proceso de digitalización, y llegando a considerarse un fenómeno disruptivo como el nacimiento de Internet.

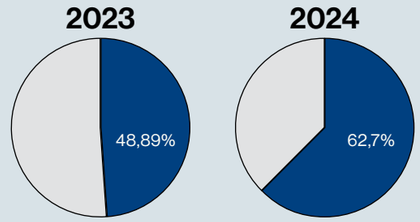
→ **Aumento de la productividad:** su alta capacidad de procesamiento y conectividad en la resolución de problemas complejos, y su semejanza con el razonamiento humano, se está haciendo visible en el impulso de la productividad, y de la eficiencia de los profesionales que lo usan como soporte en su día a día.

→ **Versatilidad en tareas:** desde su concepción, los modelos fundacionales están diseñados para poder adaptarse a un amplio abanico de tareas, lo que ha favorecido su rápida adopción en labores cotidianas como pueden ser la **generación de textos** (son excelentes para crear contenido escrito, responder preguntas, resolver problemas, y asistir en la redacción, o en la generación de código), el **análisis y resumen de textos complejos**, la **generación y procesamiento de imágenes**, el **análisis de grandes volúmenes de datos**, o las **búsquedas de información e investigaciones.**

→ **Ventaja competitiva:** la IA está brindando a las organizaciones una ventaja competitiva crucial, mejorando la experiencia del cliente y la eficiencia operativa.

→ **Reducción de costes:** la adopción de modelos fundacionales y la IA generativa está directamente vinculada a la reducción de costes operativos y organizacionales, un factor clave que impulsa su implementación en el tejido empresarial. En 2024[4], el **62,7%** de las compañías ya estaban empleando IA generativa o planeando hacerlo a corto plazo, respecto del **48,89%** de empresas que declaró lo mismo en 2023.

Uso de IA Generativa en las compañías



En resumen, los modelos fundacionales han redefinido el panorama de la IA, haciendo que la tecnología sea más accesible, versátil y capaz de integrarse en una multitud de aplicaciones y procesos, lo que lleva a una transformación significativa en la forma en que se desarrolla, utiliza y entiende la IA.

[4] ISMS Forum. (2024). Análisis II encuesta: Adopción y Gobierno de la Inteligencia Artificial. Asociación Española para el Fomento de la Seguridad de la Información. https://www.ismsforum.es/ficheros/descargas/encuesta-ia-final3-31731310534.pdf

2.2.

Riesgos emergentes asociados al uso avanzado de IA: deepfakes, automatización de ataques, exposición por LLM y evolución de amenazas tradicionales

En la propia definición del riesgo emergente se define como aquellos eventos o factores nuevos o cambiantes que, con baja probabilidad de materialización, pueden tener un impacto alto, ya que plantean nuevas formas de presentarse. Con la IA y el auge de esta en muchos escenarios, **emergen riesgos sobre la ciberseguridad y la protección de la información en las organizaciones.**

Según se mejoran los motores de IA, con nuevos algoritmos y optimización de los procesos de cálculo, es más sencillo incorporar las capacidades de IA en todos los procesos, ya sea de forma legítima y ética, o ya sea de forma menos legítima y muy poco ética. Su consecuencia es que l**os riesgos tradicionales en el ámbito de la ciberseguridad continúan evolucionando, tanto cualitativamente como cuantitativamente.** Tradicionalmente, una persona estaba detrás de una amenaza o ataque. Posteriormente, se evolucionó hacia la automatización y el escalado de las tácticas, lo que ha dado lugar a una evolución clara de estas técnicas, gracias a la aplicación de la IA.

Existen cada vez más riesgos asociados, y cuya presencia se estima que irá en aumento, como son los deepfakes, automatización y sofisticación de ataques con la IA, el uso extendido de las LLM y la evolución de las amenazas tradicionales bajo la influencia de la IA. Destacamos aquí alguno, si bien se desarrollan más adelante en la guía.

01 | *DeepFakes*

Activos desde los años 2017-2018, el término se desarrolló asociándose principalmente a la creación de contenido audiovisual generado con una intención clara de engañar al receptor. Era una acción que ya se hacía anteriormente, pero que no es más que un tipo de amenaza tradicional que se ve mejorada o potenciada a través de las herramientas de IA, y no solo con audio o mensajes, ya incluso se realizan vídeos, y en tiempo real videollamadas, suplantando la identidad de casi cualquiera.

02 | *Automatización y sofisticación de ataques*

En muchas de las organizaciones se están incluyendo o potenciando las operaciones normales con la inclusión de herramientas de IA, ya sea con automatización de procesos como mejora de los resultados o alcance de nuevos objetivos antes inalcanzables.

Con esta premisa, los grupos de ciberdelincuencia incorporan las mismas herramientas de IA para mejorar sus procesos y actividades, y de igual manera que una organización legal dispone de estos aceleradores tecnológicos, un grupo de ciberdelincuencia también. Estas mejoras que aporta la IA potencian el éxito de las campañas de ataque y, de forma preocupante, cabe destacar que ya no son sólo grupos con conocimientos avanzados y con objetivos muy específicos, sino que la socialización de las herramientas de IA y su generalización han permitido a pequeños grupos alcanzar un gran volumen de objetivos y con mecanismos más elaborados y sofisticados, consiguiendo un mayor éxito en sus fraudes y sobre todo a organizaciones con menos medios en ciberseguridad o ciudadanos más vulnerables en el ámbito digital.

03 | *Uso extendido de las LLM*

En concreto la socialización de las herramientas de IA de tipo LLM (Grandes Modelos de Lenguaje), tales como ChatGPT (OpenAI), Copilot (Microsoft), Gemini (Google), DeepSeek, LLaMA (META) y otras muchas, es un elemento básico en la vida de las nuevas generaciones y también un complemento cada vez más extendido en las organizaciones y de las personas en su día a día.

Desde el punto de vista de una organización puede llegar a ser un gran riesgo que sus empleados usen estos modelos de forma extendida y sin filtros, ya que pueden comprometer información sensible y dejar de controlar dónde reside la información o datos. Aunque tampoco se trata de poner puertas al campo, es importante situar estrategias y análisis de los usos de estas IA para acometer las mejores estrategias de integración de estas herramientas.

Ante estos riesgos del uso de la IA, en particular de los LLM, lo mejor es conocer el uso que se les va a dar dentro de la organización y plantear una estrategia de, siempre atendiendo a las necesidades de negocio e intentando asegurarlo, llevar al eventual bloqueo, adopción o integración, no siendo ajenos a este contexto.

04 | *Evolución de las amenazas tradicionales*

Como ya se ha adelantado, los grupos cibercriminales, al incluir en sus actividades las herramientas de IA, potencian sus capacidades pudiendo alcanzar objetivos que sin la IA no podrían o cuyo coste era muy elevado en recursos y/o en tiempo.

Los tradicionales intentos de fraude, como el phishing, dirigido o no, se han visto, ya, potenciados con motores de IA, automatizando estos ataques, potenciando estrategias de mayor potencia y abordaje más efectivo de sus víctimas. El grado de éxito aumenta y permite al grupo cibercriminal mejorar sus capacidades y obtener mayores beneficios.

Por todo ello, cualquier organización, independientemente de su tamaño, se enfrenta ya a amenazas cada vez más sofisticadas gracias al uso de IA, siendo difícil detectarlas y con un impacto potencial muy elevado.

05 | *Abuso en el uso de las herramientas IA*

En ocasiones, la inclusión de funciones de IA a marchas forzadas en mucha de las tecnologías presentes puede conllevar un aumento de los riesgos, con una probabilidad de materializarse también cada vez mayor. Ya es posible verlo en sistemas autónomos, en teléfonos móviles, vehículos y en un sinfín de plataformas y software comercial. Estas funcionalidades otorgan en muchos de los casos la toma de decisiones básicas agilizando las ejecuciones y los flujos de trabajo. Esto que claramente es un beneficio y una mejora notable en los productos, implica que fallos en ciertas decisiones por parte de estas herramientas puedan llegar a ocasionar errores graves, sobre todo cuando el impacto sea sobre la integridad de las personas.

A la hora de abordar estas soluciones, se deberá analizar si realmente son necesarias y si los beneficios compensan los posibles daños o impactos, ya que en muchas ocasiones no es así.

Ecosistema de roles en la gobernanza de la IA

03.

La gobernanza eficaz de la IA requiere la participación activa y coordinada de distintos perfiles profesionales. En este capítulo se describen las principales figuras implicadas, sus responsabilidades clave, las dinámicas de colaboración necesarias entre ellos y una propuesta de matriz de roles y responsabilidades.

Principales **figuras implicadas** 3.1.

La gobernanza de la IA es una disciplina transversal que requiere la interacción fluida entre múltiples roles estratégicos. Si bien cada organización tendrá un modelo de gobernanza acorde a su naturaleza, tamaño o madurez, aquí se describen, pedagógicamente, los principales perfiles que pueden conformar este ecosistema.

Chief AI Officer (CAIO)

Figura relativamente nueva, ya existente en organizaciones modernas. Su responsabilidad principal no es solo supervisar proyectos específicos de IA, sino definir la visión estratégica general que tendrá la IA dentro de la empresa. El CAIO responde a preguntas como: ¿para qué vamos a usar IA?, ¿qué valor aportará al negocio?, ¿cómo lo haremos de forma ética y sostenible?

El CAIO debe asegurar que la IA se utilice de manera segura, justa y conforme a los valores de la organización. Su rol incluye conectar el potencial técnico con los objetivos del negocio, identificando oportunidades, gestionando riesgos y garantizando que la IA no solo funcione de forma confiable y responsablemente.

Rol que actúa como un puente entre la alta dirección y los equipos operativos, colaborando estrechamente con perfiles clave como el CIO, el CISO, el CDO y el DPO, y otras figuras relevantes, alineando la estrategia de IA con las políticas corporativas, normativas legales y capacidades tecnológicas de la empresa.

Dentro de sus **funciones** estarían:

- **Responsable de definir e implementar la estrategia global de IA en la organización.**
- **Supervisa el uso ético, seguro y eficiente de los sistemas de IA.**
- **Colabora, al menos, con CIO, CISO, CDO y DPO para asegurar alineación con políticas corporativas y normativas.**

Chief Information Officer (CIO)

El CIO es la figura encargada de garantizar que la infraestructura tecnológica de la empresa esté preparada para soportar la IA.

Entre sus responsabilidades están asegurar la idoneidad, escalabilidad, seguridad y continuidad de los servidores, redes, plataformas y servicios en la nube y/o en los centros de datos propios para el despliegue de modelos de IA. También debe evaluar la interoperabilidad de sistemas nuevos y antiguos, con vistas a facilitar la integración fluida de soluciones basadas en IA dentro de los procesos del negocio.

El CIO trabaja con el CAIO para transformar digitalmente la organización, facilitando que la IA se implemente con eficacia, sin fricciones técnicas ni cuellos de botella.

Sus principales **funciones** son:

- **Responsable de las infraestructuras tecnológicas que soportan los sistemas de IA.**
- **Garantiza la viabilidad, escalabilidad e interoperabilidad del ecosistema tecnológico.**
- **Coordina con el CAIO para integrar la IA en los procesos de negocio.**

Chief Information Security Officer (CISO)

El CISO vela por la seguridad de la información, incluyendo todos los sistemas de información, de la organización, y en el contexto de IA su papel se vuelve aún más crucial. Un modelo de IA mal protegido puede ser manipulado, vulnerado o utilizado de forma maliciosa, comprometiendo los datos y/o las decisiones que de él se deriven.

El CISO debe implementar controles de seguridad, garantizando la protección de los modelos frente a ataques adversarios, fugas de información o manipulaciones durante el entrenamiento. También participa activamente en la evaluación de riesgos inherentes al uso de IA, estableciendo medidas preventivas y planes de respuesta ante incidentes.

En coordinación con el CAIO y el CDO, asegura que los datos usados para alimentar los modelos estén debidamente protegidos y gestionados conforme a las mejores prácticas de ciberseguridad

Sus principales **funciones** son:

- **Supervisa la seguridad de los sistemas de IA, incluyendo protección frente a amenazas como ataques adversarios o manipulación de modelos.**
- **Establece controles de seguridad y participa en la evaluación de riesgos.**
- **Coordina con el CAIO y el CDO para proteger la confidencialidad, integridad y disponibilidad del dato.**

Chief Data Officer (CDO)

El CDO es responsable de todo lo relacionado con los datos dentro de la organización: desde su adquisición y almacenamiento, hasta su calidad, acceso, uso y gobernanza. En el contexto de IA, el dato es el combustible esencial para entrenar modelos, por lo que la misión del CDO adquiere una dimensión crítica.

Debe definir políticas claras sobre cómo se recogen, clasifican, almacenan y utilizan los datos, especialmente aquellos sensibles o personales. Además, trabaja para asegurar que los datos sean de alta calidad, relevantes, no sesgados y representativos, aspectos fundamentales para que un modelo de IA funcione correctamente.

El CDO colabora estrechamente con el CAIO para que los modelos se construyan sobre datos adecuados, y con el CISO y DPO para garantizar, más allá del cumplimiento legal, en todo lo relativo a la seguridad del dato, privacidad y protección de datos personales.

Sus principales **funciones** son:

- **Responsable del gobierno del dato y su calidad, aspecto crítico en el entrenamiento de modelos de IA.**
- **Define políticas para la gestión, acceso y uso ético del dato.**
- **Colabora estrechamente con el CAIO, el CDO y el DPO para garantizar la conformidad normativa.**

Data Protection Officer (DPO)

El DPO es el garante del respeto a la privacidad dentro de la organización. Su misión principal es asegurar que todo tratamiento de datos personales –incluido el que se produce en sistemas de IA– cumpla con la normativa aplicable, como el RGPD.

En proyectos de IA, el DPO debe evaluar si se están tratando datos personales, sus riesgos para los derechos de las personas, y si es necesario realizar una Evaluación de Impacto en Protección de Datos (DPIA o RIA).

Trabaja en estrecha colaboración con el CAIO, el CISO, el CDO y el equipo legal para evitar que el uso de IA derive en violaciones de privacidad o sanciones regulatorias.

Sus principales **funciones** son:

- **Supervisa la conformidad con las normativas de protección de datos (ejemplo: RGPD).**
- **Evalúa los impactos de privacidad de los sistemas de IA (ejemplo: RIA - Risk Impact Assessments) y puede realizar las Evaluaciones de Impacto en los Derechos Fundamentales (FRIA) cuando es requerido por el RIA.**
- **Colabora con CAIO, CDO y el equipo legal para asegurar el cumplimiento normativo.**

Además, en función de la organización y si los sistemas de IA pueden tratar datos personales, será el rol que pueda asumir las funciones asociadas al cumplimiento del RIA, así como garantizar un uso responsable del sistema de IA, aprovechando las estructuras y procesos ya creados para dar cumplimiento al RGPD.

Para un análisis detallado de estas cuestiones, resulta imprescindible consultar los Libros Blancos del DPO[5] publicado por ISMS Forum.

Compliance Officer y equipo legal

Son la brújula normativa dentro del ecosistema de IA. Interpretan y aplican la legislación vigente, evaluando si los usos que se pretende dar a la IA están dentro del marco legal aplicable. Revisan los contratos con proveedores de tecnología, las condiciones de uso de plataformas de IA generativa, la protección de la propiedad intelectual de modelos y datos, así como las implicaciones jurídicas de automatizar decisiones que puedan afectar a personas.

Estos profesionales son clave para asegurar que la organización no incurra en riesgos legales ni éticos, especialmente a medida que entran en vigor nuevas normativas.

Su rol es complementario al del DPO, pero con un foco más amplio en la legalidad, ética empresarial y cumplimiento regulatorio en general.

Sus principales **funciones** son:

- **Aseguran que los sistemas de IA se desarrollen y utilicen conforme a las leyes y regulaciones vigentes.**
- **Evalúan cláusulas contractuales, derechos de propiedad intelectual, y marcos regulatorios emergentes y establecidos.**
- **Asesoran sobre la legalidad de usos específicos de IA, incluyendo decisiones automatizadas.**

[5] ISMS Forum. (2019). El Libro Blanco del DPO: Funciones, retos y buenas prácticas del Delegado de Protección de Datos. https://www.ismsforum.es/ficheros/descargas/el-libro-blanco-del-dpo---isms-forum-y-data.pdf
ISMS Forum. (2024). Libro Blanco del DPO: Funciones, retos y buenas prácticas del Delegado de Protección de Datos. https://www.ismsforum.es/ficheros/descargas/libroblancodpofinal1739382530.pdf

Dinámicas de Colaboración y **Coordinación Interdepartamental**

3.2.

La naturaleza transversal de la IA exige dinámicas fluidas entre estos roles.

Para ilustrar cómo estas interacciones se materializan en la práctica, se ha incluido el **Anexo 1: Casos de Uso y Desafíos Comunes de Coordinación**, donde se presentan ejemplos concretos que muestran la necesidad de una colaboración efectiva entre algunas de las áreas nombradas.

Principales mecanismos de colaboración:

→ **Comités o Grupos conjuntos:** reuniones periódicas para revisar casos de uso, riesgos, cumplimiento y avances.

→ **Protocolos de escalado:** para la gestión de incidentes o decisiones críticas.

→ **Espacios de trabajo colaborativos:** uso de herramientas compartidas como Teams para el seguimiento de proyectos de IA.

→ **Cultura de transparencia:** documentación accesible sobre el diseño, entrenamiento y uso de modelos de IA.

3.3. Matriz de **Roles y Responsabilidades RACI** (Responsible, Accountable, Consulted, Informed)

Actividad / Tarea	CAIO	CIO	CISO	CDO	DPO	Equipo Legal & Compliance	Equipo InfoSec	Equipos Técnicos
Definición de estrategia de IA	R	C	C	C	C	I	I	I
Diseño de arquitectura técnica de soluciones IA	R	A	C	I	I	I	I	R
Evaluación de riesgos de seguridad en IA	C	C	A	C	I	I	R	C
Supervisión del cumplimiento normativo (AI Act, RGPD)	C	I	I	C	A	R	I	I
Gobierno del dato usado para entrenamiento	C	I	C	A	C	I	I	R
Desarrollo e implementación de modelos IA	A	C	C	C	I	I	I	R
Evaluación ética y de impacto social	A	I	I	C	C	R	I	I

→ **R (Responsible):** quien ejecuta la tarea

→ **A (Accountable):** quien toma la decisión final y responde ante ella

→ **C (Consulted):** quien debe ser consultado

→ **I (Informed):** quien debe ser informado

Esta matriz RACI permite clarificar expectativas y responsabilidades, reducir ambigüedades y asegurar la coordinación eficiente entre los distintos actores implicados en la gobernanza de la IA.

Indicadores y **Métricas de Gobernanza**

La gobernanza de la IA requiere mecanismos de medición que permitan evaluar su grado de madurez, cumplimiento y efectividad.

Los indicadores y métricas proporcionan una base objetiva para valorar cómo se están gestionando los riesgos éticos, legales, técnicos y de seguridad asociados a la IA, y sirven además para evidenciar la trazabilidad y la mejora continua del marco de gobernanza, así como demostrar el cumplimiento de los principios de transparencia, trazabilidad, supervisión humana y gestión del riesgo a lo largo de todo el ciclo de vida de los sistemas de IA.

Sin datos medibles, las políticas y principios quedan en el plano declarativo, sin capacidad real de seguimiento o de toma de decisiones informadas.

En este sentido, disponer de un marco de métricas permite convertir esos requisitos regulatorios en mecanismos de seguimiento continuo, ofreciendo una visión objetiva del grado de control, eficacia y mejora del sistema de gobernanza. Además, facilita la alineación con marcos internacionales de referencia y respalda la rendición de cuentas frente a auditores, autoridades y partes interesadas.

Recomendaciones **Prácticas**

A partir de estas métricas, es posible plantear un conjunto de recomendaciones que fortalezcan la gobernanza y alineen el uso de la IA con estándares éticos y regulatorios:

→ **Establecer una política de IA alineada con el RIA:** contar con una política corporativa que incorpore sus requisitos desde el inicio permitirá anticiparse a las obligaciones legales y evitar sanciones.

→ **Crear un comité de gobernanza de IA multidisciplinar:** la IA no es solo un asunto técnico. Involucrar a expertos en ética, derecho, seguridad, negocio y tecnología garantiza una visión integral, reduce sesgos y promueve decisiones más equilibradas.

→ **Aplicar principios de transparencia, equidad y rendición de cuentas:** la confianza en la IA se construye con prácticas claras. Esto implica no solo documentar los modelos y sus usos, sino también garantizar resultados auditables y responsables identificados para cada fase del ciclo de vida.

→ **Adoptar marcos como NIST AI RMF o ISO/IEC 42001:** apoyarse en estándares internacionales facilita la homogeneidad de prácticas, mejora la interoperabilidad y aporta legitimidad frente a auditores y reguladores.

→ **Formar a los equipos en riesgos y buenas prácticas de IA:** la gobernanza no puede depender únicamente de normativas o procesos formales. La comprensión el riesgo por toda la plantilla y su rol activo en la gestión responsable de la IA es clave.

Estas recomendaciones constituyen una hoja de ruta práctica para que las organizaciones avancen hacia una gobernanza robusta, no solo cumpliendo con la normativa, sino también generando confianza y valor sostenible en el uso de la IA.

Gobernanza de la IA: principios, modelos y marcos operativos

04.

La gobernanza de la IA comprende principios, reglas, procesos y controles que orientan el diseño, desarrollo, despliegue y supervisión de esta tecnología en las organizaciones. Su propósito es doble, por un lado, **impulsar los beneficios económicos y sociales** que esta tecnología aporta; y por otro, **contener los riesgos éticos, técnicos y legales** que pueden emerger cuando algoritmos y modelos influyen en decisiones que afectan a personas, servicios y recursos públicos o privados. Este capítulo ofrece una visión estructurada que permite a una organización pasar de las ideas generales a la práctica diaria. Se incorporan referencias a estándares internacionales, como la ISO/IEC 42001 para sistemas de gestión de IA, la ISO/IEC 23894:2023 que proporciona una orientación específica para gestionar riesgos de IA en organizaciones de cualquier sector, la ISO/IEC TR 24368 sobre ética y preocupaciones sociales en IA, la serie IEEE 7000, el Marco de Gestión de Riesgos de IA del Instituto Nacional de Estándares y Tecnología de Estados Unidos (en adelante, "NIST AI RMF"), los Principios de la Organización para la Cooperación y el Desarrollo Económicos (en adelante, "OCDE") y el RIA.

Principios fundamentales de la gobernanza de IA

4.1.

La base de la gobernanza de la IA descansa en un conjunto de principios fundamentales concebidos para orientar el diseño, el desarrollo y el uso de sistemas inteligentes. Diversos organismos internacionales, como la OCDE, la UNESCO o la Unión Europea, han propuesto una serie de valores y directrices éticas convergentes. Estos principios buscan asegurar que la IA sea confiable, es decir, que respete los derechos humanos fundamentales, que sea técnicamente robusta, transparente, equitativa y esté sujeta a rendición de cuentas:

01.

Transparencia y explicabilidad

Claridad en la toma de decisiones algorítmicas, comprensibles para los seres humanos. La transparencia, además, facilita la trazabilidad en las decisiones automatizadas y permite identificar errores o sesgos. Una IA explicable implica proporcionar información significativa, que esté adecuada al contexto, que identifique las fuentes de datos en las que se apoya, y que muestre los criterios y la lógica utilizados por el modelo.

02.

Equidad y no discriminación

Promover la justicia social y la inclusión, evitando intensificar prejuicios o desigualdades. Los sistemas no tratarán de forma injusta por motivos de raza, género, origen socioeconómico, discapacidad u otras categorías protegidas por la ley. Además, las organizaciones deben poder detectar y mitigar los sesgos algorítmicos, garantizando que los beneficios de la IA sean accesibles para todos los sectores de la sociedad. Este principio se vincula con la idea de una IA centrada en el ser humano.

03.

Privacidad y Protección de Datos

Uso de grandes volúmenes de datos que pueden incluir datos personales, por lo que se requiere salvaguardar el derecho a la privacidad, cumpliendo con las normas de protección de datos existentes (como el ya mencionado, en capítulos anteriores, RGPD). Un manejo ético de la IA incluye la minimización de datos, es decir, usar solo los datos necesarios, la anonimización cuando proceda y la realización de evaluaciones de impacto en privacidad antes de desplegar sistemas de alto riesgo.

04.

Seguridad, robustez y resiliencia

Los sistemas deben ser sólidos y seguros, comportarse según lo previsto (fiabilidad) y ser resistentes a fallos, a ataques cibernéticos o incluso a manipulaciones maliciosas. Los riesgos de seguridad deben anticiparse y mitigarse desde la fase de diseño. Un sistema robusto de IA también debe manejar adecuadamente cualquier dato erróneo o adverso, manteniendo su integridad y evitando consecuencias no deseadas o peligrosas. En entornos críticos (como la salud o las entidades esenciales e importantes), la seguridad de la IA es esencial para prevenir daños físicos o materiales.

Supervisión humana y control

se deben habilitar mecanismos de intervención y control humano, especialmente cuando se enfrenta a decisiones que pueden ser de alto riesgo, conservando las personas la autoridad última sobre dichos sistemas de IA, pudiendo anular o modificar decisiones cuando sea necesario. Este principio garantiza la autonomía humana y el criterio ético sobre la autonomía de la máquina.

05.

Sostenibilidad y bienestar social

Es fundamental medir y gestionar el impacto ambiental y social de la IA, incluyendo la huella de carbono de los modelos y sus efectos en el empleo y la cohesión social. La sostenibilidad implica alinear el desarrollo de la IA con objetivos de desarrollo sostenible, minimizando impactos negativos en el medio ambiente. Además, la IA debe contribuir positivamente al bienestar social, mejorando servicios como la salud, la educación y la accesibilidad y no agravando problemas como la desinformación o la polarización.

07.

Responsabilidad y rendición de cuentas

Todo sistema de IA debe tener responsables claramente definidos a lo largo de todo el ciclo de vida. Las organizaciones deben asumir la responsabilidad de los resultados de sus sistemas, estableciendo mecanismos de revisión, auditoría e intervención ante comportamientos no deseados. Esto implica realizar evaluaciones de impacto algorítmico, auditorías externas, y cumplir con esquemas de certificación. Es fundamental documentar las decisiones de diseño, mantener registros (logs) de las operaciones del modelo y aplicar la diligencia debida antes de adoptar cualquier sistema. Si ocurre un daño o sesgo, debe existir transparencia sobre las causas y tomarse acciones correctivas para mitigarlo. Además, deben existir canales accesibles para reclamaciones y corrección de errores detectados por usuarios o autoridades.

06.

Modelos de gobernanza de la IA

4.2.

Al trasladar los principios a mecanismos concretos, han emergido distintos modelos de gobernanza de la IA, abordando – países y entidades – el reto de no frenar la innovación. Entre dichos modelos destacan tres que priorizan diferentes aspectos, según si se privilegia la creación de normas jurídicas vinculantes o directrices voluntarias.

Modelo de gobernanza **basado en riesgos**

Es el que predomina en Europa y está alineado con el RIA. Si bien cabe mencionar que el RIA no regula un modelo de gobernanza de forma expresa ni figuras profesionales concretas. Sin perjuicio de ello, hay leyes nacionales y locales que podrían derivar en ello, así que deberá ser considerado a la hora de definir el modelo más adecuado para cada organización.

Los modelos basados en riesgos se centran en la identificación, evaluación y mitigación proactiva de riesgos, aplicando salvaguardas proporcionales al nivel de riesgo que cada sistema de IA plantea para las personas o la sociedad. Los sistemas de IA se clasifican por **categorías o taxonomías de riesgo** (por ejemplo, “riesgo inaceptable”, “riesgo alto”, “riesgo limitado” o “mínimo”), aplicándose diferentes requisitos para cada categoría. El RIA adopta justamente este esquema piramidal prohibiendo los sistemas de riesgo inadmisible (por ejemplo, sistemas de puntuación social generalizados) y regulando de forma proporcional el resto (luego se definen sistemas de alto riesgo como sistemas de IA para contratación laboral, sistemas de evaluación para créditos o diagnósticos médicos, obligados a cumplir unos requisitos muy estrictos de transparencia, trazabilidad, gestión de datos, etc.); finalmente, los sistemas de riesgo medio o bajo que tienen obligaciones más sencillas, como seguir códigos de conducta o proporcionar transparencia hacia los usuarios. El objetivo es fomentar la confianza en la IA asegurando que los controles se adecuan al potencial de daño, sin imponer cargas innecesarias a sistemas de bajo riesgo.

Modelo de gobernanza **basado en derechos**

Enfoque basado en derechos humanos, garantizando un desarrollo y uso de la IA que no vulnere derechos fundamentales (privacidad, no discriminación, libertad de expresión, debido proceso, etc.), y a la vez promoviendo la participación de aquellas personas y colectivos potencialmente afectados por la IA en las discusiones regulatorias.

En esencia, desplaza la pregunta de **“¿qué riesgos tiene esta tecnología?”** hacia **“¿cómo está impactando la IA en los derechos y la dignidad de las personas?”.**

Este enfoque sostiene que ciertos valores no son negociables ni graduables por nivel de riesgo; por ejemplo, si una aplicación de IA afecta a la igualdad ante la ley o la privacidad, debe ser abordada desde obligaciones positivas de proteger esos derechos, más que solamente como un “riesgo a mitigar”. Organismos como el Consejo de Europa y la ONU han abogado por esta visión, poniendo de relieve que se debe evitar la llamada “ética superficial” (ethics washing) para pasar a establecer garantías jurídicas que sean vinculantes. Un ejemplo de esto es la Recomendación sobre Ética de la IA de la UNESCO (adoptada por 193 Estados en 2021), la cual se autodefine como un enfoque centrado en los derechos humanos.

En la práctica, este enfoque y el basado en riesgos no son excluyentes. La legislación de la UE combina una lógica de evaluación de riesgos con disposiciones explícitas para proteger derechos, así que podemos decir que el RIA mantiene un buen equilibrio.

Enfoques basados en **principios y autorregulación o “soft-law”**

Enfoque o modelos voluntarios o no vinculantes donde se incluyen diferentes directrices éticas elaboradas por comités de expertos, asociaciones profesionales o por las propias empresas tecnológicas.

En el año 2019, la Comisión Europea publicó las **Directrices Éticas para una IA Confiable**, que están definidas por un grupo de expertos, y que enumeraron siete requisitos con carácter voluntario para guiar a las empresas en la auto certificación de sus sistemas de IA, promoviendo prácticas como las evaluaciones de impacto ético y la comunicación transparente con usuarios.

Muchas empresas tecnológicas líderes en sus respectivos sectores (Microsoft, IBM, Telefónica, Google, etc.) han publicado sus propios modelos de gobernanza, que incluyen principios y prácticas de IA estableciendo además comités internos de uso responsable de IA para supervisar los sistemas que comercializan.

Marcos operativos de gobernanza

4.3.

Gobernanza de la IA por Diseño (AI Governance by Design, AIGD)

La tendencia más avanzada en la materia es la integración de la gobernanza desde el propio diseño de los sistemas, lo que se conoce como “AI Governance by Design” (AIGD). Este enfoque implica que las consideraciones éticas, legales y sociales no se incorporan a posteriori, sino que forman parte del ciclo de vida de la IA desde su concepción, asegurando así una gobernanza proactiva y efectiva.

La AIGD implica que la gobernanza no es un añadido posterior, sino un componente esencial desde la concepción del sistema. Esto se traduce en:

Marcos de **gobernanza**

A continuación, se hace una revisión de los principales estándares, normas y marcos, tanto internacionales como nacionales, que ayudan a tener un buen marco de gobernanza de la IA, así como las prácticas organizativas recomendadas para su adopción.

ISO/IEC 42001 (2023)

de Sistema de gestión de la IA: establece los requisitos para implementar un Sistema de Gestión de la IA (AIMS) en las organizaciones. Similar a normas de gestión de calidad (ISO 9001) o seguridad de la información (ISO 27001), la ISO 42001 provee un marco estructurado basado en el ciclo de Deming (PLAN, DO, CHECK, ACT) para incorporar la gobernanza en todas las fases del ciclo de vida de la IA.

La ISO 42001 es un estándar certificable que permite demostrar mediante una auditoría independiente que una organización gestiona su IA de forma responsable y ajustada a una serie de criterios verificables. Nombra una serie de requisitos clave, siendo los más relevantes definir el contexto y alcance del uso de IA; implementar procesos continuos de gestión de riesgos; realizar evaluaciones de impacto antes del despliegue; asegurar la calidad y trazabilidad de los datos; controlar a proveedores y terceros (como servicios en la nube o modelos preentrenados); la monitorización y mejora continua de los sistemas; y el establecimiento de criterios de diseño ético y métricas para vigilarlos. En esencia, la ISO 42001 obliga a integrar la IA dentro del gobierno corporativo con un liderazgo comprometido, políticas y responsabilidades, plan de formación, documentación exhaustiva, clara y evaluación periódica del desempeño tanto de los modelos y sistemas de IA como del personal que lo gestiona.

La propia UE prevé que el cumplimiento de la ISO 42001 podría facilitar la conformidad con el RIA, por lo que su adopción dentro del marco de gobernanza de una organización es un paso muy recomendable para anticiparse a requisitos legales y facilitar la conformidad (cumplimiento).

ISO/IEC TR 24368 (2022)

Enfocado exclusivamente en los aspectos éticos y sociales de la IA y los principios clave para abordarlos. Esta guía es de carácter orientativo, no prescribe valores específicos, sino que expone consideraciones comunes para ayudar a evaluadores, tecnólogos y reguladores a entender y mitigar los retos éticos de la IA.

La importancia de la ISO TR 24368 radica en que sintetiza el consenso global existente sobre los riesgos éticos de la IA y proporciona un enfoque estructurado e inclusivo, abordando proactivamente problemas como la parcialidad algorítmica (sesgos), la opacidad de los modelos de aprendizaje profundo, o las posibles violaciones de privacidad por parte de sistemas que analizan grandes conjuntos de datos personales. Además, promueve la participación de diferentes disciplinas y actores (negocio, ingenieros, expertos en ética, juristas, sociólogos, etc.) en la evaluación de sistemas de IA, de modo que los resultados sean justos, transparentes y responsables.

En la práctica, las organizaciones pueden usar la ISO TR 24368 como guía de referencia para elaborar sus códigos éticos de IA o cuando realicen evaluaciones de impacto ético (EIA) de proyectos de IA. De hecho, la UNESCO ha desarrollado una herramienta de Evaluación de Impacto Ético[6] inspirada en principios similares, para ayudar a los equipos de proyecto a identificar los impactos sociales y éticos de una IA antes de ponerla a disposición de los usuarios finales.

NIST AI RMF del NIST (Instituto Nacional de Estándares y Tecnología de EE. UU.)[7]

Guía estructurada para identificar, medir, mitigar y monitorear riesgos a lo largo del ciclo de vida de los sistemas de IA. Este marco está organizado en cuatro grandes funciones: Mapear (contexto, mapeo de riesgos), Medir (análisis y evaluación de riesgos), Gestionar (respuesta y mitigación) y Gobernar (función transversal de supervisión continua).

Además, define las características de IA confiable (trustworthy AI) que deben perseguirse, incluyendo la validez de los sistemas (que funcionen según su propósito), la seguridad, la resiliencia, la responsabilidad, la explicabilidad, la privacidad y la ausencia de sesgos. Es un marco alineado con otros marcos y normas internacionalmente aceptadas (ISO/IEC 5338, ISO/IEC 38507, ISO/IEC 22989, ISO/IEC 24028, ISO/IEC DIS 42001 y ISO/IEC NP 42005) y tiene un enfoque muy similar al que tiene la UE para afrontar los riesgos. Parte de la premisa de que gestionar el riesgo incremental de cada aplicación de IA es la vía para evitar daños, en lugar de prohibir la tecnología.

[6] UNESCO. (2023). Evaluación del impacto ético. https://unesdoc.unesco.org/ark:/48223/pf0000386276

[7] National Institute of Standards and Technology (NIST). (2023). AI Risk Management Framework (AI RMF 1.0). https://airc.nist.gov/airmf-resources/airmf

ISO/IEC 23894:2023 (Artificial Intelligence - Guidance on risk management)[8]

Proporciona una orientación específica para gestionar riesgos de IA en organizaciones de cualquier sector. La ISO 23894 se alinea con los principios generales de ISO 31000 (gestión de riesgos corporativos) pero los adapta al contexto de la IA, describiendo procesos para identificar, evaluar, tratar y monitorear riesgos algorítmicos a lo largo de todo el ciclo de vida de un sistema de IA.

Por ejemplo, esta norma pone mucho énfasis en considerar riesgos emergentes como la manipulación adversarial de modelos (ataques adversarios), la falta de contexto en los modelos (conocidas comúnmente como alucinaciones) en IA generativa, o los impactos en derechos como parte integral del análisis de riesgo que se debe realizar. Esta guía además es muy clara y busca que sea fácil de entender para directivos no técnicos, lo que facilita la comunicación organizacional sobre riesgos de IA.

IEEE (Institute of Electrical and Electronics Engineers)

En paralelo a las normas ISO y las iniciativas promovidas por organismos públicos como el NIST, la IEEE (Institute of Electrical and Electronics Engineers) ha impulsado un repertorio bastante extenso de estándares que ayudan a orientar el desarrollo responsable de sistemas autónomos e inteligentes, así como la gobernanza de estos. Destaca la serie IEEE 7000, que abarca guías muy concretas sobre aspectos éticos.

[8] ISO/IEC 23894:2023 – ISO (2023) "Artificial Intelligence - Guidance on risk management" https://www.iso.org/es/contents/data/standard/07/73/77304.html

Implementación práctica de gobernanza de la IA

5.5.

La gobernanza de la IA no se limita a marcos normativos o principios teóricos; su verdadero valor reside en la capacidad de convertirlos en prácticas operativas dentro de las organizaciones. Estos marcos ofrecen directrices y estructuras generales, pero el desafío consiste en incorporarlos en los procesos internos y en el día a día, lo que a veces denominamos la cultura corporativa. De esta manera aseguraremos que cada proyecto de IA se desarrolle, despliegue y supervise conforme a los requisitos técnicos, éticos y legales aplicables.

Para lograr una implementación práctica, las organizaciones deben acometer una serie de tareas. En primer lugar, es fundamental diseñar **políticas corporativas de IA**, que abarcan desde la **fase de ideación** (incluyendo ***chequeos legales y requisitos éticos***) hasta la **puesta en producción** (con ***controles de cambio*** y vigilancia continua), y suelen estar integradas con las políticas existentes para la gestión de los datos y de las tecnologías de la información (TI).

En segundo lugar, la creación de **comités de ética o de gobernanza de IA**. Estos comités formados por la dirección, expertos legales, de riesgo, especialistas en IA y, en ocasiones, consultores externos, evalúan si los proyectos **cumplen con los valores de la organización y la normativa vigente**, pudiendo recomendar ajustes o vetar iniciativas

Por último, la realización de **evaluaciones de impacto algorítmico (EIA)** resulta esencial, especialmente en sistemas de IA que pueden afectar significativamente a personas. Inspiradas en las evaluaciones de impacto en privacidad (DPIA) obligatorias bajo el RGPD y exigidas por la ISO 42001 en sistemas críticos, las EIA permiten analizar las **posibles consecuencias sobre derechos y valores**, como la **discriminación**, la **opacidad**, los **sesgos** o la **privacidad** de las personas y sus datos y permite identificar medidas de mitigación. Herramientas como el Algorithmic Impact Assessment[9] de Canadá o la AI Impact Assessment propuesta por la UE y requerida bajo el artículo 27 del RIA, ejemplifican este enfoque, complementando las evaluaciones técnicas con un análisis profundo de los impactos sociales, éticos y legales de la IA.

[9] Gobierno de Canadá. (2024). Algorithmic Impact Assessment. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Gobierno y gestión del dato en la IA

05.

En su obra American Ideals, Theodore Roosevelt decía ***"Nothing in this world is worth having or worth doing unless it means effort, pain, difficulty".***

El uso de la IA para nuestros objetivos estratégicos no es una excepción.

Cierto es que los cantos de sirena que solemos escuchar (y lo que es peor, que muchos directivos suelen escuchar), no refieren demasiadas complicaciones, puesto que suelen centrarse en la última milla: el desarrollo de un modelo analítico que resolverá nuestros problemas. Digamos que es la parte en la que el atleta llega a la meta y se lleva la gloria. Lo que se suele omitir es todo el esfuerzo que ello conlleva.

Por suerte, los responsables de datos de las organizaciones bien saben que obtener valor de los datos, de manera consistente, requiere una serie de tareas que no son tan sencillas de explicar en un PowerPoint.

Y eso es lo que significa **Gobernar**; conseguir alinear a las personas y a la tecnología, alrededor del proceso de generación de valor con datos, con el objetivo de optimizar dicho valor. Y para ello es necesaria una estrategia: una **Estrategia de Datos.**

Por otra parte, la llegada de la IA generativa está logrando que lo que era terreno reservado a una minoría de desarrolladores con alta formación, pase a ser de acceso sencillo para cualquier empleado con ciertas habilidades básicas, incluyendo la franja de managers menos duchos en la materia.

Sin embargo, esto está provocando que determinados problemas de gobierno, antes circunscritos a un grupo controlable, se hayan expandido por toda la organización. Hay organizaciones que pretenden ayudar a enfrentar estos problemas con solvencia. Un claro ejemplo de ello es el Club de CDOs[10], que ha diseñado una herramienta de diagnóstico de la madurez analítica denominada dataMat. Esta se acompaña de una segunda herramienta (dataToolkit) que, sobre la base del diagnóstico del dataMat, proporciona un sistema de priorización de iniciativas, es decir, diseña la estrategia de datos de tu organización. Ambas herramientas se componen de tres capas: diagnóstico de madurez general, de madurez en ética del dato y de madurez en IA generativa.

Pues bien, en el curso de estos análisis de madurez, y especialmente en los de Ética del Dato, se identifican una serie de buenas prácticas en materia de gobierno responsable de la IA, algunas de las cuales se resumen en los siguientes puntos, agrupados en cinco bloques: estrategia, gestión de riesgos, gobierno, operación de modelos y cultura.

Estrategia y marco ético

→ Hacer que los aspectos éticos formen parte de la estrategia de la compañía, y que éstos incluyan aquellos relacionados con la IA responsable.

→ Diseñar el gobierno de la IA de forma que esté adaptado a la criticidad del caso de uso.

→ Disponer de políticas aprobadas al máximo nivel, con roles definidos, responsabilidades claras y una estructura de comités adecuada.

→ Tener identificados los dilemas éticos asociados a cada iniciativa de negocio.

Identificación y gestión de riesgos éticos

→ Disponer de criterios de tierización del riesgo ético.

→ Disponer de una medida de apetito al riesgo ético.

→ Disponer de un inventario de dilemas éticos, asociado a cada modelo o sistema de IA.

→ Inclusión del riesgo ético dentro del catálogo de riesgos de la compañía.

→ Realizar análisis periódicos de riesgo ético, asociados a los dilemas del catálogo.

→ Disponer de KPIs de medición del impacto ético.

→ Contar con mediciones del riesgo inherente y del riesgo residual (tras la aplicación de acciones de mitigación).

[10] Club Chief Data Officer. www.clubcdo.com

Gobierno organizativo y operativo

→ Existencia de equipos especializados en Gobierno de Modelos, que actúan de forma transversal a toda la organización.

→ Realizar validaciones externas independientes sobre el diseño y cumplimiento de la estrategia de ética del dato.

→ Extender el gobierno ético, tanto a los proveedores como a los clientes de datos, insights y modelos.

→ Creación de equipos internos de validación independiente (segunda línea de defensa).

→ Incluir aspectos de IA responsable dentro de los canales de denuncias.

Ciclo de vida de modelos y datos

→ Contar con un **inventario de modelos**, con responsables nominados y un nivel de riesgo asociado.

→ Conocer la **traza** en los procesos informacionales que alimentan los modelos analíticos.

→ Garantizar la **explicabilidad de los modelos**, tanto en el ámbito local (individual) como global (modelo completo).

→ Disponer de un análisis sobre la base de **criterios éticos** en el desarrollo de modelos, incluyendo el análisis de sesgos.

→ Contar con una **monitorización continua** de la degradación de los modelos.

→ Utilizar herramientas corporativas para el **control** de modelos, datos y procesos asociados.

→ Contar con procesos **Near Real Time** para el control de sesgos en cargas masivas de datos, especialmente en colectivos sensibles.

→ Disponer de un **conformity assessment** en todos los modelos.

→ Diseñar procesos de desarrollo de modelos **Ethic by Design.**

→ Incluir **supervisión humana** en los procesos de toma de decisiones automáticas.

Cultura, formación **y concienciación**

→ Disponer de un **plan formativo y de comunicación** para la concienciación en IA responsable en todos los niveles participantes, con promoción de foros de participación.

→ Promover **foros internos de discusión** sobre los riesgos éticos de los objetivos de negocio.

En resumen, la integración de procesos éticos y responsables en el desarrollo y gestión de modelos de inteligencia artificial es fundamental para garantizar su transparencia, fiabilidad y alineación con los valores de la organización. La formación continua, la supervisión humana y la medición del impacto ético contribuyen a crear una cultura sólida de IA responsable, donde los riesgos se gestionan proactivamente y los dilemas éticos se abordan con criterio y rigor. Solo así se podrá avanzar hacia una estrategia empresarial sostenible y comprometida con la sociedad. Y desde luego, sin una estrategia clara y esponsorizada al más alto nivel, todo esto no se conseguirá.

La IA como riesgo estratégico y operativo

06.

En las últimas décadas, el ámbito de las Tecnologías de la Información ha experimentado diversas revoluciones tecnológicas: la irrupción de Internet, la telefonía móvil o los servicios en la nube, entre otras. Cada una de ellas nos ha obligado a adoptar nuevas medidas de seguridad y a replantear nuestros procesos operativos.

La actual revolución de la IA sigue esa misma línea, pero introduce dimensiones adicionales que impactan directamente en la competitividad, la resiliencia, el cumplimiento normativo y la confianza de los usuarios y clientes.

Por ejemplo, la IA es capaz de emitir respuestas de forma autónoma. Esto implica el riesgo de que dichas respuestas sean incorrectas, generando consecuencias indeseadas a nivel operativo, reputacional o económico. Como en una empresa que ofrece precios a través de un asistente automatizado: si la IA proporciona un valor erróneo, se podría incurrir en una pérdida económica o un daño a la imagen corporativa.

Es importante subrayar que la IA es un sistema que aprende y evoluciona. Incluso con un entrenamiento óptimo, no existe la garantía de exactitud absoluta, ya que siempre puede producir errores.

En los documentos **“Introducción a la Inteligencia Artificial para profesionales de seguridad de la información”**[11] y **“Modelo de Gobierno de la Inteligencia Artificial”** ya se abordan los riesgos de la IA para los derechos y libertades, incluyendo aspectos como la ética y la responsabilidad legal.

Esto supone un punto diferenciador respecto a revoluciones tecnológicas anteriores, en las que la dimensión ética rara vez se ponía en primer plano.

Para un análisis más profundo de los riesgos asociados al uso de IA y las estrategias para tratarlos, se recomienda la lectura de **“Inteligencia Artificial y Ciberseguridad”**[12] publicado por ISMS Forum.

Finalmente, es importante destacar que la IA no es meramente una cuestión tecnológica, habiendo aproximaciones que hablan de ella como un componente más del negocio y, sin duda, con una fuerte dependencia de los datos, entre otras cuestiones.

[11] ISMS Forum. (2023). Introducción a la Inteligencia Artificial para profesionales de seguridad de la información. Asociación Española para el Fomento de la Seguridad de la Información. https://www.ismsforum.es/ficheros/descargas/isms-gt-ia---01-introl-a-la-ia1701173559.pdf

[12] ISMS Forum. (2024). Ética y Compliance en el uso de la Inteligencia Artificial. Asociación Española para el Fomento de la Seguridad de la Información. https://www.ismsforum.es/ficheros/descargas/isms-gt-ia-021707141605.pdf

Definición y alcance del riesgo de IA

6.1.

El riesgo de IA es la posibilidad de pérdidas, daños o incumplimientos en cualquier fase del ciclo de vida del sistema: diseño, desarrollo, despliegue, operación o retirada. Este riesgo es acumulativo y evoluciona junto con la tecnología, los datos y el entorno.

Cuando hablamos de riesgo de IA, este puede adoptar varias grandes dimensiones:

- **Riesgo estratégico:** impacta en la propuesta de valor, el modelo de negocio, la reputación y el cumplimiento de la estrategia corporativa. ***Ejemplo: un uso inadecuado de IA generativa puede erosionar la confianza de los clientes.***
- **Riesgo operativo:** afecta a procesos, personas, sistemas y proveedores, incluyendo seguridad de la información, protección de datos, continuidad de negocio y control de costes. ***Ejemplo: un fallo en un modelo de atención al cliente puede saturar los canales y aumentar los costes.***

  Transversalidad: la IA impacta, de forma directa o indirecta, a todas las áreas de la organización, desde las funciones estratégicas hasta las operativas.

  Carácter evolutivo: se trata de una tecnología en rápida transformación, lo que implica que los riesgos identificados hoy pueden cambiar, amplificarse o incluso desaparecer en cuestión de meses.

  Amplitud de exposición: la exposición al riesgo no se limita únicamente a modelos de IA desarrollados internamente. También abarca el uso de soluciones de terceros e incluso modelos abiertos o de libre acceso que se integren en los procesos corporativos.

- **Riesgos tecnológicos:** los más asociados a las vertientes clásicas de riesgos TIC, de seguridad de la información y/o ciberseguridad.

La gestión del riesgo de IA requiere un enfoque integral, considerando su transversalidad, dinamismo y la variedad de modelos utilizados. No gestionarlo adecuadamente puede afectar tanto la tecnología como la estrategia empresarial.

6.2. Mapa de riesgos **estratégicos**

Estos riesgos en la adopción y uso de la IA no son meramente técnicos, sino que pueden afectar a objetivos corporativos, posición competitiva en el mercado y la relación con clientes, reguladores y otros grupos de interés.
Algunos de los riesgos más significativos:

01. Desalineación **estratégica**

Invertir en iniciativas de IA sin un caso de negocio claro o sin una alineación explícita con los objetivos estratégicos medidos a través de OKR (Objectives and Key Results[13]) o KPI (Key Performance Indicators[14]). Esto puede dar lugar a proyectos de escaso retorno, iniciativas que no llegan a integrarse en los procesos críticos o la proliferación de pilotos que nunca se escalan, lo que comúnmente se conoce como “pilot purgatory”. Esta situación consume recursos, genera frustración interna y erosiona la credibilidad de la IA como palanca estratégica.

02. Dependencia y **lock-in de proveedores**

El uso de modelos o APIs cerradas puede condicionar severamente la capacidad de la organización para cambiar de tecnología o proveedor en el futuro. Esto implica riesgos de restricción de portabilidad, aumentos de coste no previstos o cambios en las condiciones de uso, además de la concentración excesiva en un único proveedor o en una única región (single vendor, single region), que aumenta la probabilidad de interrupciones o decisiones unilaterales del proveedor.

03. Cumplimiento normativo y **licenciamiento**

Procesos de evaluación de conformidad, elaboración de documentación técnica y auditorías periódicas son requisitos básicos para sistemas de alto riesgo. A ello se suma el riesgo de utilizar datos o resultados con restricciones de propiedad intelectual, o entrenados con datasets bajo licencias que limiten su explotación o faltos de legitimidad en origen, lo que podría dar lugar a sanciones o disputas legales.

04. Reputación **y confianza**

La imagen de marca puede verse gravemente afectada si la IA genera contenidos falsos, inexactos o engañosos (por ejemplo, deepfakes o respuestas alucinadas) que se atribuyan públicamente a la organización. La falta de transparencia sobre el funcionamiento de los modelos, la ausencia de explicabilidad o la inexistencia de mecanismos efectivos de reclamación puede debilitar la confianza de clientes, empleados y socios estratégicos.

05. Ventaja competitiva **y tiempo de mercado**

El ritmo de adopción de la IA es un factor crítico, pudiendo perder cuota de mercado frente a competidores más ágiles en la industrialización de la IA. Aunque adoptar demasiado rápido tecnologías inmaduras también conlleva riesgos (sobrecostes, bajo rendimiento o rehacer desarrollos). La clave está en sincronizar la implementación con el grado de madurez tecnológica y la capacidad organizativa para absorber el cambio.

06. Ética y responsabilidad **social**

Los sesgos algorítmicos o la toma de decisiones discriminatorias pueden impactar de forma directa en clientes, empleados o comunidades. Además, el uso de IA en actividades que no encajan con los valores corporativos y de una IA ética y responsable puede deteriorar la imagen pública y el compromiso social de la organización.

07. Continuidad y **resiliencia**

La dependencia de sistemas de IA implica asumir el riesgo de interrupciones severas por razones como cambios en los modelos externos utilizados, caídas de servicios en la nube o retirada de funcionalidades por decisión unilateral de un proveedor. La resiliencia de los procesos críticos dependerá de la capacidad para anticipar y gestionar este tipo de situaciones.

[13] Panchadsaram, R. (s.f.). What is an OKR? OKR Meaning, Definition & Examples. What Matters. https://wwwwhatmatters.com/faqs/okr-meaning-definition-exampl
[14] KPI.org. (s.f.). KPI Basics. https://www.kpi.org/kpi-basics/

6.3. Mapa de riesgos **operativos**

Los riesgos operativos de la IA están relacionados con el funcionamiento diario de los sistemas, la calidad de los datos que los alimentan, la seguridad de su ciclo de vida, la interacción con personas y procesos, así como con la gestión de terceros implicados. A continuación, se detallan las principales tipologías de riesgo, cada una de ellas con un impacto potencial en la continuidad, calidad y seguridad de los servicios.

01. Datos

La IA depende en gran medida de la calidad, integridad y procedencia de los datos. Problemas como información incompleta, desactualizada o con origen no verificable pueden degradar el rendimiento del modelo o generar resultados erróneos.

En el ámbito de la privacidad, destacan riesgos como el uso de datos personales sin una base legítima, la identificación de individuos a partir de datos aparentemente anónimos o la fuga de información sensible a través de prompts.

La soberanía de los datos añade otra dimensión: la transferencia transfronteriza y las restricciones sobre la ubicación física de la información (residencia de datos) pueden crear conflictos legales y de cumplimiento, especialmente en entornos multinacionales.

02. Seguridad **(AI Security)**

La IA introduce vectores de ataque específicos. Entre ellos, destacan los ataques al modelo, como prompt injection (inserción de instrucciones maliciosas), jailbreaks (elusión de limitaciones), envenenamiento de datos (introducción de ejemplos corruptos en el entrenamiento) o extracción de modelos (robo de parámetros o pesos).

También existen riesgos en la cadena de suministro tecnológica, donde dependencias externas, weights de modelos, contenedores o librerías pueden ser comprometidos.

La exposición de las APIs es otro frente crítico, con riesgos de abuso, elusión de límites de consumo (rate limits) o inclusión de secretos en prompts que podrían ser exfiltrados.

03. Riesgo de modelo **(Model Risk)**

Los modelos pueden perder validez debido a fenómenos como el drift de datos o conceptos, el sobreajuste (overfitting) o simplemente por falta de robustez ante condiciones no vistas.

Una gobernanza deficiente, sin documentación clara (por ejemplo, datasheets o model cards), sin trazabilidad ni control de versiones, dificulta la gestión y auditoría.

En casos sensibles, la explicabilidad resulta crítica: modelos opacos o no interpretables pueden impedir justificar decisiones ante reguladores o clientes.

04. Operaciones **(MLOps)**

Los flujos de integración y despliegue continuo (CI/CD/ML) sin puntos de control de riesgo, o con pruebas insuficientes, pueden dar lugar a modelos defectuosos en producción. También suponen riesgos los despliegues sin estrategias de “canary release” o de rollback rápido.

En operación, la monitorización debe ir más allá de la disponibilidad técnica: es necesario vigilar métricas como la latencia, el coste/token, la calidad de las respuestas, la detección de toxicidad, sesgos y plagio, para poder reaccionar y corregir a tiempo.

05. Terceros y **compras**

El uso de soluciones de IA de terceros sin una evaluación rigurosa de sus capacidades de seguridad, cumplimiento, acuerdos de nivel de servicio (SLA) y procedimientos de auditoría, expone a la organización a riesgos que, en ocasiones, no son visibles hasta que se materializa un incidente.

07. Legal/ **contractual**

El marco jurídico de la IA aún está evolucionando, pero ya plantea retos claros: limitaciones en las indemnidades contractuales por los resultados generados, asignación de responsabilidad sobre contenidos creados por IA y riesgos derivados de términos de uso que no cubran adecuadamente incidentes como producción de material ilícito o difusión de datos no autorizados, así como la generación de daños en colectivos habiéndose utilizado una IA sin respetar derechos fundamentales.

06. Personas y **procesos**

Un riesgo común es la dependencia excesiva en las respuestas generadas por IA (automation bias), que puede llevar a la pérdida gradual de pericia humana y al deterioro de la capacidad crítica para validar resultados.

También hay riesgos asociados a la “ergonomía” del trabajo con prompts: interacciones poco eficaces o mal diseñadas pueden reducir la productividad.

A nivel organizativo, la introducción de IA exige adaptar roles, planes de formación, incentivos y marcos de responsabilidad. Si esto no se gestiona de forma clara, puede derivar en una difusa responsabilidad, donde no se sabe quién es responsable de las decisiones tomadas con apoyo de IA.

6.4. Riesgos TIC y de **ciberseguridad**

El riesgo asociado a la IA puede definirse como la posibilidad de que se produzcan pérdidas o daños que generan un impacto. Entre los riesgos más reconocidos destacan:

DeepFakes

Creación de contenidos audiovisuales falsos que permiten la suplantación de identidades.

Prompt Injection

Manipulación del prompt para eludir restricciones o acceder a información sensible.

Sensitive Informtion Disclosure

Revelación accidental o forzada de información confidencial.

Model Theft

Extracción del modelo o replicación indebida de su funcionamiento.

Bias and Algoritmic Discrimination

Reproducción o amplificación de sesgos presentes en los datos.

Training Data Poisoning

Alteración de los datos de entrenamiento que compromete el rendimiento o introduce sesgos.

Insecure Output Handling

Inclusión inadvertida de código malicioso en las salidas del modelo.

Excessive Agency / Overreliance

Delegación excesiva de funciones o dependencia operativa crítica del modelo.

Intellectual Property Risk

Generación o entrenamiento con materiales sujetos a derechos de autor.

Disinformation

Producción de información inexacta o engañosa que puede erosionar la confianza.

Resumiendo, una vez identificados y comprendidos los diferentes riesgos asociados al uso de la IA, resulta más sencillo definir un proceso estructurado para su análisis, implementación y seguimiento, así como determinar las métricas adecuadas para cada caso de uso. También se deberán establecer indicadores específicos, así como KPI y KRI específicos de diferente índole asociados a los usos de la IA en la organización.

Gestión y evaluación de riesgos en IA

07.

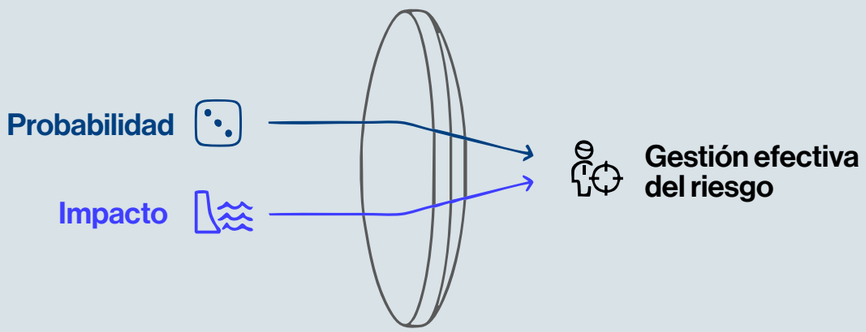
De cara a poder entender cómo funciona la IA y ser capaces de aplicarla de manera efectiva en nuestras organizaciones, así como acompañar su implantación, es preciso comprender qué riesgos están presentes y cómo gestionarlos adecuadamente.

Si bien en otros capítulos de esta guía se hace mención desde los riesgos estratégicos, hasta los operativos, pasando por aquellos más específicos y emergentes, como son los algoritmos, sus implicaciones de uso, entrenamiento y desarrollo, entre otros, en este capítulo se pretende explicar cómo deben ser gestionados.

La gestión de riesgos en la IA es el proceso de: identificar, analizar, evaluar y tratar los riesgos asociados a las tecnologías de la IA.

Antes de decidir cómo tratar un riesgo (mitigarlo, aceptarlo, etc.), necesitamos cuantificar o calificar su gravedad.

Riesgo = Probabilidad x Impacto

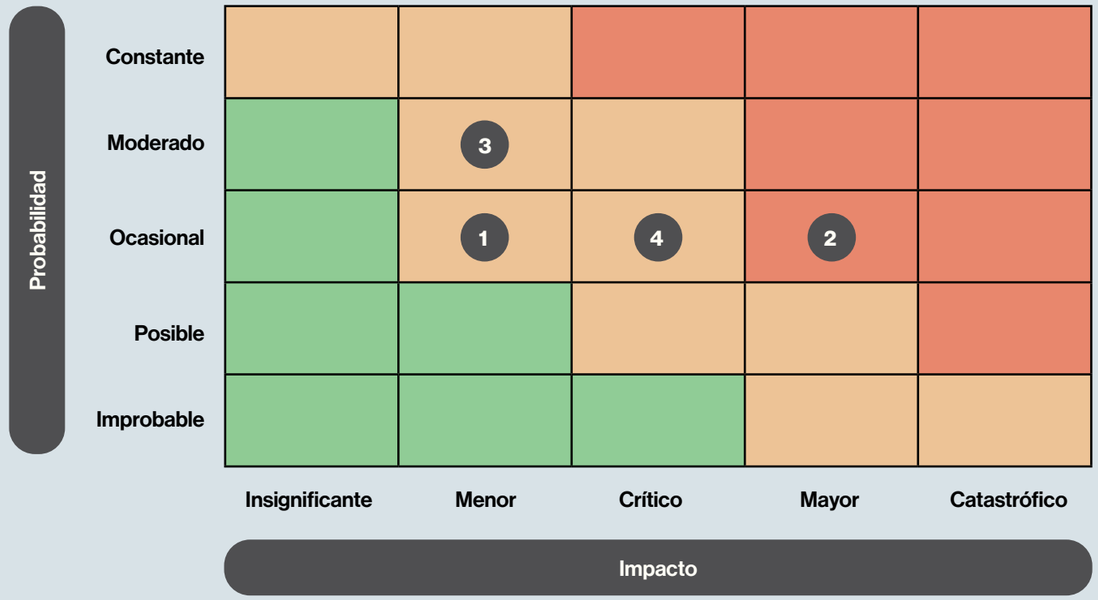


Los métodos para evaluar estos riesgos son análisis cualitativos y cuantitativos.

→ **Análisis Cualitativo:** Es un método subjetivo, basado en juicio experto, experiencia y escalas predefinidas para evaluar:
- Probabilidad
- Impacto
- Nivel de riesgo

→ **Análisis Cuantitativo:** Es un método objetivo, que asigna valores numéricos reales
- Probabilidad (porcentaje)
- Impacto (costos económicos, tiempos, etc.)
- Exposición anual al riesgo (Annualized Loss Expectancy, ALE)

Una vez evaluados los riesgos dispondremos de un sistema que nos permitirá conocer la situación de riesgo global, identificando los más críticos.



Una vez identificados comenzaremos con las tareas de mitigación.

7.1. ISO **27005**

La ISO 27005 proporciona un marco para la gestión de riesgos de seguridad de la información, detallando un proceso sistemático para identificar, analizar, evaluar y tratar los riesgos. Este proceso, que utiliza un enfoque basado en el riesgo y complementa a la norma ISO 27001, ayuda a las organizaciones a proteger sus activos de información de manera eficiente y a tomar decisiones informadas sobre la seguridad. El proceso es cíclico y se basa en el modelo **PDCA (Planificar, Hacer, Verificar, Actuar)** para garantizar la mejora continua.

- **Establecer el contexto:** definir los objetivos, criterios y el alcance de la gestión de riesgos, así como el marco de referencia.
- **Identificar los riesgos:** localizar y describir los riesgos potenciales. La norma propone dos enfoques:
    1. *Basado en eventos:* se centra en los escenarios y amenazas generales.
    2. *Basado en activos:* se enfoca en los activos de información específicos y las vulnerabilidades asociadas a ellos.
- **Analizar los riesgos:** determinar la probabilidad de que el riesgo ocurra y el impacto que tendría en la organización si se materializa.
- **Evaluar los riesgos:** comparar los riesgos analizados con los criterios de aceptación definidos en la fase de contexto para priorizarlos.
- **Tratar los riesgos:** seleccionar e implementar una o más opciones para mitigar, evitar, aceptar o transferir los riesgos que exceden el nivel aceptable.
- **Comunicar y consultar:** mantener informadas a todas las partes interesadas sobre el proceso y sus resultados.
- **Monitorear y revisar:** supervisar continuamente los riesgos, las amenazas y la eficacia de los controles implementados, así como revisar el proceso periódicamente para adaptarlo a los cambios.

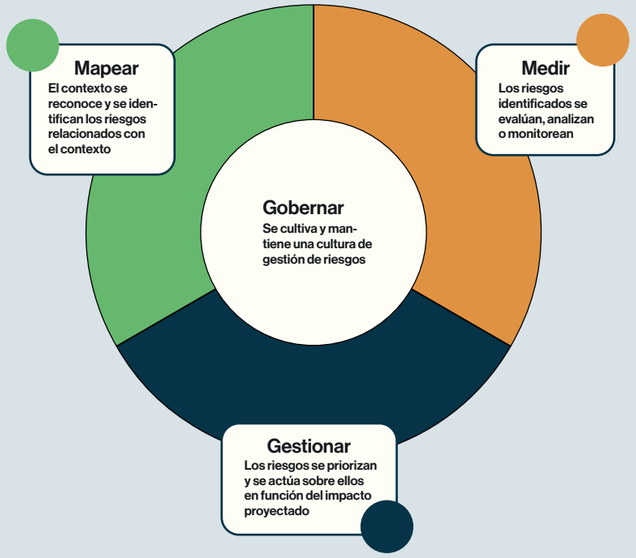
7.2. **NIST AI 100-1** (AI RMF 1.0)

El Marco de Gestión de Riesgos de IA del NIST (AI RMF) es una guía voluntaria publicada en enero de 2023 para ayudar a las organizaciones a gestionar los riesgos de la IA. Su objetivo es promover un desarrollo y despliegue responsables de sistemas de IA, centrándose en características como la confiabilidad, la equidad y la transparencia. El núcleo del marco se basa en cuatro funciones: gobernar (crear una cultura de gestión de riesgos), mapear (identificar y enmarcar riesgos en contextos empresariales), medir (analizar y evaluar los riesgos) y gestionar (abordar los riesgos identificados).

El Marco de Gestión de Riesgos de IA del NIST (AI RMF) es una guía voluntaria publicada en enero de 2023 para ayudar a las organizaciones a gestionar los riesgos de la inteligencia artificial. Su objetivo es promover un desarrollo y despliegue responsables de sistemas de IA, centrándose en características como la confiabilidad, la equidad y la transparencia. El núcleo del marco se basa en cuatro funciones: gobernar (crear una cultura de gestión de riesgos), mapear (identificar y enmarcar riesgos en contextos empresariales), medir (analizar y evaluar los riesgos) y gestionar (abordar los riesgos identificados).

Componentes clave del AI RMF del NIST

- **Objetivo:** proporcionar un enfoque estructurado para identificar, evaluar, gestionar y mitigar los riesgos de los sistemas de IA para asegurar que sean confiables y seguros.
- **Funciones principales:** las cuatro funciones principales del marco son:
    → *Gobernar:* establecer una cultura de gestión de riesgos de IA dentro de la organización.
    → *Mapear:* contextualizar los riesgos de IA a las operaciones y necesidades del negocio.
    → *Medir:* analizar y evaluar los riesgos de manera sistemática.
    → *Gestionar:* tomar acciones para abordar los riesgos que han sido mapeados y medidos.
- **Aplicabilidad:** es voluntario y está diseñado para ser utilizado por una amplia gama de organizaciones en diversas industrias y geografías.
- **Beneficios:** ayuda a las organizaciones a aprovechar los beneficios de la IA mientras mitigan daños potenciales, asegura la confianza del consumidor en sus datos personales y promueve el desarrollo continuo de la tecnología.



7.3. Reglamento de **Inteligencia Artificial**

El RIA establece un conjunto de obligaciones exigibles a quienes desarrollan, implementan o usan sistemas de IA en la UE. Entre ellas, la obligación de realizar un análisis de riesgos, especialmente si el sistema entra en la categoría de alto riesgo.
El RIA está basado en el USO que hacemos de la IA, definiendo el reglamento su propia tabla de criterios para categorizar una IA. La evidencia de que es un reglamento basado en riesgo se evidencia en que la palabra “riesgos” se menciona 181 veces.
La propuesta del RIA establece un enfoque basado en el riesgo para regular el uso de sistemas de IA. Esta clasificación determina qué obligaciones y restricciones se aplican a cada tipo de sistema.
El análisis de riesgos se realiza para las IA de Riesgo Alto.

Nivel de Riesgo	Descripción y ejemplos	Tratamiento Legal
Riesgo Inaceptable	Sistemas que contravienen valores fundamentales	**Prohibidos**
Riesgo Alto	Afectan derechos fundamentales, salud o seguridad.	**Regulados estrictamente**
Riesgo Limitado	Pueden influir en decisiones de usuarios.	**Requieren transparencia**
Riesgo Mínimo o Nulo	No suponen amenaza significativa.	**Uso libre**

El análisis de riesgos se realiza para las IA de Riesgo Alto.

Prácticas de IA prohibidas
Artículo 5 del **RIA**

- Sistemas de IA que se sirva de técnicas subliminales.
- Sistemas de IA que explote alguna de las vulnerabilidades de una persona física o un determinado colectivo de personas.
- Sistemas de IA para evaluar o clasificar a personas físicas o a colectivos de personas.
- Sistemas de IA para realizar evaluaciones de riesgos de personas físicas cometer un delito (valorar o predecir).
- Sistemas de IA que creen o amplíen bases de datos de reconocimiento facial.
- Sistemas de IA para inferir las emociones de una persona física en los lugares de trabajo y en los centros educativos.
- Sistemas de categorización biométrica que clasifiquen individualmente a las personas físicas.

Prácticas de IA de alto riesgo
Anexo III, Artículos 6 y 9 del **RIA**

Estos sistemas no están prohibidos, pero están sujetos a obligaciones exigentes de diseño, supervisión, documentación y, sobre todo, de gestión de riesgos.
Los proveedores de sistemas de IA de alto riesgo deberán:

- Garantizar que sus sistemas de IA de alto riesgo cumplen los requisitos establecidos en la **sección 2.**
- Indicar en el sistema de IA de alto riesgo o, cuando ello no sea posible, en su embalaje o en la documentación que lo acompañe, según proceda, su nombre, nombre comercial registrado o marca comercial registrada, la dirección en la que se les puede contactar.
- Disponer de un sistema de gestión de la calidad que se ajuste a lo dispuesto en el **artículo 17.**
- Conservar la documentación mencionada en el **artículo 18.**
- Cuando estén bajo su control, conservar los registros generados automáticamente por sus sistemas de IA de alto riesgo a que se refiere el **artículo 19.**
- Garantizar que el sistema de IA de alto riesgo se someta al correspondiente procedimiento de evaluación de la conformidad a que se refiere el **artículo 43**, antes de su comercialización o puesta en servicio.
- Redactar una declaración UE de conformidad con arreglo al **artículo 47.**
- Colocar el marcado CE en el sistema de IA de alto riesgo o, cuando ello no sea posible, en su embalaje o en la documentación que lo acompañe, para indicar su conformidad con el presente Reglamento, de conformidad con el **artículo 48.**
- Cumplir las obligaciones de registro a que se refiere el **apartado 1 del artículo 49.**
- Adoptar las medidas correctoras necesarias y facilitar la información exigida en el **artículo 20.**
- Previa solicitud motivada de una autoridad nacional competente, demostrar la conformidad del sistema de IA de alto riesgo con los requisitos establecidos en la **sección 2.**
- Garantizar que el sistema de IA de alto riesgo cumple los requisitos de accesibilidad de conformidad con las **Directivas (UE) 2016/2102** y **(UE) 2019/882**.
- Un marco de rendición de cuentas que establezca las responsabilidades de la dirección y del resto del personal en relación con todos los aspectos enumerados en este apartado.

Prácticas de IA de riesgo limitado
Artículo 50 y 52 del **RIA**

La obligación es la de informar al usuario de que está interactuando con una IA.

7.4. El sistema de **Gestión de Riesgos de RIA**

El RIA transforma la gestión de riesgos en una obligación legal estructurada y jerarquizada:

- **Sistemas de alto riesgo (SAR)**

Se determina un conjunto de requisitos obligatorios para los SAR. Estos incluyen la **gestión de riesgos** durante todo el ciclo de vida del sistema, la **gobernanza de datos** (calidad y representatividad), la documentación técnica, la **trazabilidad** y la **supervisión humana**.

- **Modelo de riesgo sistémico**

La legislación ha ido más allá de los sistemas de alto riesgo al clasificar los **modelos de IA de propósito general (GPM o GPAI)** en dos niveles: ordinario y de **riesgo sistémico**. Esta distinción, basada en factores como el número de parámetros, el volumen de datos de entrenamiento o la capacidad de cálculo, añade una capa regulatoria compleja que no estaba totalmente anticipada en las discusiones iniciales de 2023, que se centraban más en los sistemas de IA tradicionales.

Por tanto, el sistema de gestión de riesgos se aplica a sistemas de alto riesgo y está definido en el Artículo 9 del RIA

1. **Se establecerá, implantará, documentará y mantendrá un sistema de gestión de riesgos en relación con los sistemas de IA de alto riesgo.**
2. **El sistema de gestión de riesgos se entenderá como un proceso iterativo continuo planificado y ejecutado durante todo el ciclo de vida de un sistema de IA de alto riesgo, que requerirá revisiones y actualizaciones sistemáticas periódicas. Constará de las siguientes etapas:**

    a. La determinación y el análisis de los riesgos conocidos y previsibles que el sistema de IA de alto riesgo pueda plantear para la salud, la seguridad o los derechos fundamentales cuando el sistema de IA de alto riesgo se utilice de conformidad con su finalidad prevista;

    b. La estimación y la evaluación de los riesgos que podrían surgir cuando el sistema de IA de alto riesgo se utilice de conformidad con su finalidad prevista y cuando se le dé un uso indebido razonablemente previsible;

    c. La evaluación de otros riesgos que podrían surgir, a partir del análisis de los datos recogidos con el sistema de vigilancia poscomercialización a que se refiere el artículo 72;

    d. La adopción de medidas adecuadas y específicas de gestión de riesgos diseñadas para hacer frente a los riesgos detectados con arreglo a la letra a).

3. **Los riesgos a que se refiere el presente artículo son únicamente aquellos que pueden mitigarse o eliminarse razonablemente mediante el desarrollo o el diseño del sistema de IA de alto riesgo o el suministro de información técnica adecuada.** A la hora de determinar las medidas de gestión de riesgos más adecuadas, se procurará:

    a. Eliminar o reducir los riesgos detectados y evaluados de conformidad con el apartado 2 en la medida en que sea técnicamente viable mediante un diseño y un desarrollo adecuados del sistema de IA de alto riesgo;

    a. Implantar, cuando proceda, unas medidas de mitigación y control apropiadas que hagan frente a los riesgos que no puedan eliminarse;

    a. Proporcionar la información requerida conforme al artículo 13 y, cuando proceda, impartir formación a los responsables del despliegue.

    Con vistas a eliminar o reducir los riesgos asociados a la utilización del sistema de IA de alto riesgo, se tendrán debidamente en cuenta los conocimientos técnicos, la experiencia, la educación y la formación que se espera que posea el responsable del despliegue, así como el contexto en el que está previsto que se utilice el sistema.

4. **Los sistemas de IA de alto riesgo serán sometidos a pruebas destinadas a determinar cuáles son las medidas de gestión de riesgos más adecuadas y específicas.**

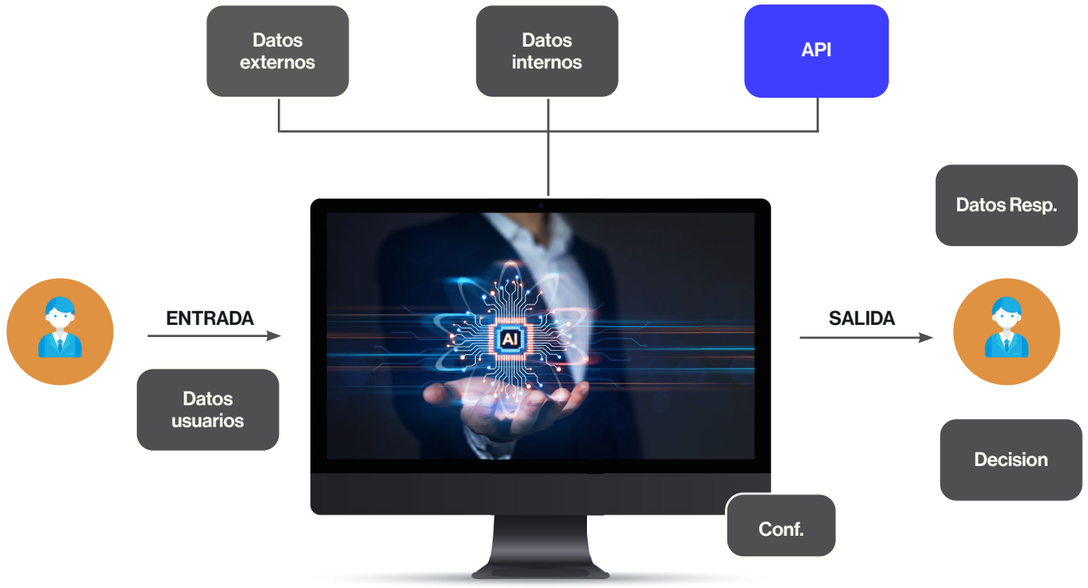
    Traslado a lenguaje de Riesgos:

    - Identifica el sistema de IA.
    - Identificación de los riesgos potenciales.
    - Evaluación de cada riesgo: probabilidad e impacto.
    - Definir e implementar medidas de mitigación.
    - Monitorizar y revisar periódicamente.
    - Documentar todo el proceso de gestión.
    - Realización de pruebas continuas.

7.5. Modelizar **el modelo de IA**

Este punto trata de estructurar el modelo de IA para poder analizar correctamente los riesgos, encajando la solución dentro de la plantilla, de forma que será mucho más sencillo identificar los riesgos.

Al ser una plantilla no se adapta a todos los modelos de IA, por lo que se puede ir actualizando en función del modelo.



Arquitectura/Infraestructura

Lo elementos que componen este modelo sencillo son:

- **Los datos/acciones de entrada.**
- **Los datos/acciones de salida.**
- **Los datos de entrenamiento.**
- **Los datos de operación.**
- **Los datos de configuración.**

De esta forma sencilla podemos caracterizar el modelo.

Una vez que disponemos del detalle de nuestra solución de IA dentro del modelo aplicamos los principales riesgos acorde a una taxonomía definida previamente. Actualmente existen varias propuestas.

En resumen, el RIA se basa en la gestión de riesgos y su análisis en los sistemas de criticidad alta. Para ello es necesario disponer de los conocimientos adecuados en gestión de riesgos para poder realizar correctamente los análisis y poder aplicar las medidas de mitigación de los mismos.

Actualmente existen múltiples metodologías de análisis de riesgos, derivadas de la seguridad y la ciberseguridad, por lo que son las más comunes a la hora de aplicarse, aunque están ya apareciendo algunas específicas de la IA.

Protección de datos y privacidad en la IA

08.

La IA ha supuesto grandes retos en el proceso de innovación y transformación digital desde diferentes órbitas a nivel social, cultural y económico. Sin duda, uno de esos retos es crear un ecosistema que permita su integración garantizando la protección de un derecho fundamental como es la protección de datos, habida cuenta que el desarrollo y uso de los sistemas de IA puede conllevar el tratamiento de datos personales.

El tratamiento de datos personales podría producirse en distintos momentos del ciclo de vida del sistema de IA y con distintas funcionalidades, por ello, desde la perspectiva de la protección de datos el desarrollo de la IA presenta grandes retos a analizar:

01. Cumplimiento del principio de **minimización de datos**

Dentro de los principios del ya mencionado, en capítulos previos, RGPD desde la órbita de IA, debemos abordar el principio de minimización de datos que se regula en su artículo 5.1c).

Según el propio RGPD, dicho principio se basa en valorar y establecer un análisis que acote la recopilación y tratamiento de datos que sean estrictamente necesarios para la finalidad para la que fueron recabados, siendo necesario efectuar un juicio de proporcionalidad del tratamiento. Tal y como señalábamos, frecuentemente los sistemas de IA requerirán el tratamiento de datos personales[15].

[15] Cotino Hueso, L., & Simón Castellano, P. (Eds.). (2024). Tratado sobre el reglamento de inteligencia artificial de la Unión Europea. Aranzadi. https://roderic.uv.es/items/cec02ec1-44a4-41b6-9df7-de3ed7d6a223

Por ello, con carácter previo al tratamiento será necesario efectuar un análisis en el que deberemos tener en cuenta las siguientes cuestiones:

- → **El tratamiento** es idóneo y adecuado para la finalidad prevista.
- → **La pertenencia**: los datos están directamente relacionados con la finalidad.
- → **Limitación del tratamiento:** no produciéndose una recopilación de datos personales no relevantes y necesarios para la finalidad.

Así, en el supuesto de sistemas de IA que traten datos de carácter personal, se debe efectuar un estudio previo que permita evaluar la proporcionalidad y necesidad del tratamiento teniendo como base las siguientes consideraciones:

- → Procurar limitar las categorías de datos que se utilizan, así como el grado de precisión de los mismos a las necesidades reales de tratamiento.
- → Acotar el volumen de datos y de interesados de los que se tratan dichos datos.
- → Establecer mecanismos que permitan limitar el acceso a las distintas categorías de datos.

A tal efecto, la AEPD nos recuerda que existen diferentes técnicas que permiten la minimización de datos, entre las que encontramos la anonimización y seudonimización[16], así como la generación de datos sintéticos (que no deberán contener información identificable) pero también otras como la supresión de conclusiones no relevantes asociadas a información personal durante el proceso de entrenamiento, por ejemplo, en el caso de entrenamiento no-supervisado.

[16] Tribunal de Justicia de la Unión Europea. (2025, 4 de septiembre). Sentencia en el asunto C-413/23 P: EDPS contra SRB (Concepto de datos personales y seudonimización). https://curia.europa.eu/jcms/upload/docs/application/pdf/2025-09/cp250107en.pdf
Agencia Española de Protección de Datos (AEPD). (2020). Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. https://www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf

02. Transparencia y **explicabilidad**

El principio de transparencia es un elemento convergencia entre el RGPD y el RIA, por el que se establecen normas armonizadas en materia de IA (en adelante, RIA), ya que abordan dicho principio desde diferentes perspectivas que resultan complementarias.

Así, en el caso del marco del RGPD el fin del principio de transparencia es dotar al individuo de información que le permita comprender el tratamiento de sus datos y dotarle de mayor autonomía en la toma de decisiones. Se establece que el tratamiento debe ser transparente (art. 5. 1ª) de manera que se garantice el derecho de los interesados a ser debidamente informados sobre la recopilación de sus datos, su uso y la conservación los mismos de manera clara, accesible y comprensible, pero, además, en el art.22 se regula el derecho a no ser objeto de decisiones basadas exclusivamente en tratamientos automatizados que puedan producir efectos jurídicos significativos.

Sin embargo, el deber de transparencia en el RIA se recoge desde la óptica de un deber técnico y organizativo de manera que se permita garantizar que los usuarios comprendan cuando interactúan con sistemas de IA y como se generan resultados, y conseguir la confiabilidad, seguridad y rendición de cuentas.

Sin embargo, la opacidad técnica de determinados modelos, así como la ponderación de los secretos comerciales que puedan proteger algoritmos, dificulta poder dar cumplimiento al ejercicio del derecho de información y transparencia. Por ello, y en aras de facilitar su gestión se pueden establecer medidas a conjugar con las obligaciones de información del RGPD, con los refuerzos establecidos en el RIA, dentro de las obligaciones establecidas en dicho cuerpo normativo que permiten establecen las obligaciones derivadas del deber de información y el principio de transparencia:

- → **Identificación de IA**. Las personas deben poder cuando están interactuando con un sistema de IA.
- → **Etiquetado de contenidos sintéticos.**
- → **Explicabilidad y trazabilidad.** Los sistemas de alto riesgo deben diseñarse de modo que su funcionamiento permita a los usuarios interpretarlos y utilizarlos de manera adecuada. Dicha información podrá incluir documentación técnica respecto instrucciones, limitaciones y riesgos.

A este respecto, y en relación con el deber de cumplimiento de transparencia, información y explicabilidad, resulta interesante la interpretación efectuada por la Sentencia de Sección Tercera de la Sala Contencioso-Administrativo del Tribunal Supremo de 11 de septiembre de 2025[17].

[17] Tomando como fundamento lo dispuesto en los artículos 14 y 16 de la ley 19/2013 de transparencia, acceso a la información y buen gobierno, se avala el acceso al código fuente de la aplicación informática BOSCO, puesta en marcha por el Ministerio de Transición Ecológica para evaluar si los solicitantes del bono social de energía eléctrica para tener la consideración de consumidor vulnerable (si bien, es necesario señalar que el sistema BOSCO no introduce propios criterios sino que mecaniza operaciones para aplicar criterios sin tomar decisiones autónomas). Dicho acceso se pondera tomando en consideración el deber de transparencia habida cuenta que se trata de una actuación administrativa articulando el derecho acceso por una fundación en defensa de colectivos vulnerables

03. Evitación de sesgos y **discriminación**

Uno de los retos a los que se enfrenta el diseño y entrenamiento de sistemas de IA con impacto en la protección de datos, es la evitación de sesgo, pero para ello debemos entender qué tipos de sesgo podemos encontrarnos:

- → **Sesgo del conjunto de datos.** Se produce este riesgo cuando los datos con los que se lleva a cabo el entrenamiento de los sistemas de IA no reflejan los datos reales objetivos, como sesgos de muestreo o de medición
- → **Sesgos personales**, en tanto en cuanto los sesgos de los desarrolladores puedan influir en los resultados del modelo mediante la introducción involuntaria de sus propios sesgos y percepciones
- → **Sesgo algorítmico.** Cuando, a través del diseño, se puedan dar prioridad a ciertas características sobre otras, de forma que se generen conclusiones injustas

Sobre el tratamiento y mitigación del sesgo, cabe señalar que el grupo de expertos de la Comisión Europea elaboraban unas Directrices para una IA fiable[18], recogía la necesidad de que los sistemas de IA se fundamenten en el compromiso de utilizarlos al servicio de la humanidad y el bien común, evitando el sesgo injusto, ya que podría tener múltiples consecuencias negativas, desde la marginación de los grupos vulnerables hasta la exacerbación de los prejuicios y la discriminación.

Para evitar la aparición de sesgos deberemos abordar su prevención y deberemos establecer mecanismos desde un enfoque holístico en las distintas fases del proceso:

[18] https://digital-strategy.ec.europa.eu/es/library/ethics-guidelines-trustworthy-ai

Diseño y planificación:

En esta fase deberemos establecer objetivos claros del sistema y los grupos a los que puede afectar, identificando variables sensibles como género o etnia, y analizando posibles impactos discriminatorios.

Se deberán elaborar de las fichas del sistema que recojan información sobre el uso previsto, riesgos y limitaciones. Así como habilitar y diseñar mecanismos que nos permitan gestionar los derechos de los interesados en relación con la protección de datos más allá del derecho de acceso y transparencia, articulando los procedimientos para poder atender derechos como el de supresión, de oposición o no ser objeto de decisiones automatizadas.

01.

Datos

En la fase de tratamiento, se aplicarán técnicas de anonimización y minimización de datos partiendo de los criterios anteriormente señalados y documentando los conjuntos de datos: origen, propósito, limitaciones, etc.

02.

Modelado y entrenamiento

La selección de algoritmos se efectuará atendiendo a criterios que permitan auditoría y explicabilidad para poder cumplir con el principio de transparencia.

Sin perjuicio de ello, se deberán documentar las decisiones de diseño y valoraciones utilizadas para ello que permitan tener una trazabilidad de las decisiones adoptadas y faciliten la identificación de errores.

03.

Evaluación y pruebas

Con carácter previo a la implementación, realizar auditorías internas de sesgos y establecer entornos de pruebas que permitirá evaluar el sistema. Para ello, será necesario el establecimiento de métricas que permitan revisar los resultados. A este respecto, resulta interesante revisar la lectura de la guía de ISMS en las que se recogen las métricas más usadas como técnicas de evaluación y validación[19]. Posteriormente se documentarán los resultados de las auditorías y las limitaciones detectadas.

04.

Implementación y despliegue

En la fase de implementación se procederá al establecimiento de los mecanismos de supervisión humana, incluyendo capacitación y formación a los usuarios del sistema, así como mecanismos y procesos que faciliten corregir o modificar resultados.

05.

Mantenimiento y monitorización

Se incluirán las valoraciones y criterios de los usuarios que permita establecer correcciones, además de la realización de auditorías periódicas de sesgos y rendimiento que nos permitan evaluar el sistema.

06.

[19] ISMS Forum. (s. f.). Introducción a la inteligencia artificial. isms-gt-ia---01-introl-a-la-ia1701173559.pdf

04. **Evaluaciones de impacto** integradas

Dentro de las obligaciones derivadas de la responsabilidad proactiva y gestión de riesgos derivada del uso de sistemas de IA, debemos tener en cuenta las evaluaciones de impacto relativas a la protección de datos **(EIPD o DPIA)** conforme prescribe el **artículo 35 del RGPD** y la **evaluación de impacto relativa a los derechos fundamentales del RIA**. Ambas evaluaciones son piezas distintas que confluyen y complementan en el marco de gobernanza de la IA en las Entidades, como ha reconocido el propio RIA.

Así, **la EIPD se centra en los riesgos relacionados con la protección de datos personales**. Las evaluaciones de impacto relativas a derechos fundamentales (conocidas como FRIA) se centran en analizar la no discriminación, libertad de expresión, seguridad y salud entre otros, cuando se despliegan sistemas de alto riesgo.

Las EIPD del artículo 35 del RGPD deberán efectuarse cuando el tratamiento de datos conlleve un alto riesgo para los derechos y libertades de las personas. Así, se deberán efectuar EIPD cuando se cumplan varios de estos factores:

- → **Procesamiento innovador o uso de nuevas tecnologías**
- → **Decisiones automatizadas con efectos significativos**
- → **Combinación de conjuntos de datos de manera inesperada para el usuario**
- → **Evaluación de personas para valorar comportamientos, productividad o riesgos crediticios**
- → **Tratamiento de datos de personas vulnerables.**

Para la elaboración de las EIPD resulta especialmente relevante la revisión de las Directrices del Grupo 29 de la Comisión Europea como la guía elaborada por la AEPD en la que configura una lista de tipos de tratamientos que requieren EIPD[20].

Contenido mínimo. La EIPD debe incluir:

- → La descripción de operaciones y fines de tratamiento.
- → La evaluación de necesidad y proporcionalidad.
- → Evaluación de los riesgos.
- → Las medidas y salvaguardas previstas destinadas a la prevención de la materialización de los riesgos o las que puedan establecer la mitigación de los mismos. Para ello, se podrá contar con el asesoramiento del DPO (si existe) y consultar previamente a la autoridad si, tras las medidas, persiste un alto riesgo residual.

[20] Agencia Española de Protección de Datos (AEPD). (2019). Evalúa-Riesgo RGPD. https://www.aepd.es/documento/wp248rev01-es.pdf

Sin perjuicio de lo expuesto, en el **art. 27 del RIA** se recoge la obligación de efectuar evaluación de impacto relativa a los derechos fundamentales antes del despliegue de determinados sistemas de IA de alto riesgo definiendo el ámbito subjetivo de aplicación para los responsables del despliegue que sean organismos de Derecho público o entidades privadas que prestan servicios públicos (para sistemas de alto riesgo del **art. 6.2, con la excepción del anexo III, punto 2** para infraestructuras críticas), y los responsables del despliegue de sistemas del **anexo III, punto 5, letras b) y c)** (relativas a solvencia/score crediticio y seguros de vida y enfermedad).

Así mismo, en el citado precepto se establece que cuando ya exista una EIPD la FRIA debe complementarla de manera que pueda utilizarse las sinergias entre ambas evaluaciones.

Por ello, la coexistencia de la Evaluación de Impacto en Protección de Datos (EIPD) detallada en el art. 35 del RPGD y las evaluaciones de impacto sobre derechos fundamentales del RIA implica que las organizaciones deberán desarrollar metodologías conjuntas, que permitan la optimización y eficiencia de los recursos en la identificación, gestión y monitorización del riesgo, estableciendo un proceso de supervisión y control más coherente.

Esta obligación de elaboración de las FRIA, que entrará en vigor el 2 de agosto de 2026, exige que el contenido mínimo que deberá contemplarse en su elaboración contará con:

- → Descripción del proceso en el que se usará el sistema y su alineación con la finalidad prevista.
- → Categorías de personas y grupos afectados debiendo expresar especial consideración con los colectivos vulnerables.
- → Riesgos específicos de daño para los colectivos afectados.
- → Ventana temporal y frecuencia de uso.
- → Medidas de mitigación ante la materialización del riesgo de gobernanza interna y definición de mecanismos de reclamación, que debe incorporar procedimientos que permitan documentar la trazabilidad y monitorización, y recoger los cambios y actualizaciones si se producen. Para ello es recomendable establecer metodologías que nos permitan evaluar los impactos de forma parametrizada.
- → Notificación a la autoridad de control. Tras completar el FRIA se deberá notificar el resultado a la autoridad de control. Es relevante mencionar la metodología promovida por el Consejo de Europa para la realización de las FRIA, llamada "HUDERIA[21]".
- → Medidas de supervisión humana que se van a aplicar: en tanto en cuanto debe definirse cómo se va a producir en los distintos momentos del ciclo de vida del sistema (antes, durante y después del uso) de manera que se integren con medidas técnicas de robustez, calidad del dato y explicabilidad.

Por tanto, a modo de conclusión, cabe destacar que la protección de datos debe entenderse como un valor estratégico que permita su integración en la cultura de las organizaciones, identificando la protección de los datos personales como extensión de la identidad humana y la protección adecuada a la misma.

Proteger los datos no es sólo cumplir con las prescripciones normativas, sino la necesidad de integrar la privacidad por diseño y por defecto para respetar la dignidad de las personas.

Conceptualmente, sobre esta visión se han expresado en marcos éticos internacionales como la **Recomendación de la Unesco sobre la Ética de la IA**[22] o las **Directrices de la OCDE** que destacan la privacidad, la equidad y la rendición de cuentas como pilares de una IA confiable.

Para ello, y para la infiltración de este principio, de forma conceptual y transversal en las Entidades, ***debemos desarrollar programas de formación y concienciación, políticas internas y comités*** que nos permitan facilitar ese proceso de transformación que facilite la protección del individuo y sus datos.

[21] Consejo de Europa. (s.f.). HUDERIA – Risk and impact assessment of AI systems. https://www.coe.int/en/web/artificial-intelligence/huderia-risk-and-impact-assessment-of-ai-systems

[22] Recomendación sobre la ética de la inteligencia artificial | UNESCO https://www.unesco.org/es/articles/recomendacion-sobre-la-etica-de-la-inteligencia-artificial

Cumplimiento normativo en la IA

09.

El contexto regulatorio ha experimentado una transformación radical, pasando de un enfoque inicial centrado en la publicación de directrices éticas (como las de la UNESCO o la OCDE) a la **imposición de normativas con fuerza de ley.**

Los sistemas de IA se han convertido en motores fundamentales de transformación económica, social y política. Desde los algoritmos de recomendación en redes sociales y las aplicaciones de reconocimiento facial hasta los sistemas automatizados de toma de decisiones en sectores como la sanidad, el empleo o las finanzas, la IA está redefiniendo los límites de lo posible. Sin embargo, esta expansión acelerada ha traído consigo una creciente **preocupación pública, institucional y gubernamental sobre los riesgos éticos, legales y sociales asociados a su uso.** En este contexto, el cumplimiento normativo en materia de IA se ha convertido en un pilar esencial para garantizar la confianza, la seguridad y la legitimidad del ecosistema tecnológico contemporáneo. Su objetivo es garantizar que el desarrollo, implementación y uso de los sistemas de IA se realicen conforme a las disposiciones legales vigentes, los principios éticos reconocidos internacionalmente y los estándares técnicos de calidad, seguridad y transparencia.

La primera razón que explica la creciente relevancia del cumplimiento normativo en IA es la **expansión de la presión regulatoria a nivel global.** A medida que los gobiernos reconocen el potencial transformador –y también disruptivo– de la IA, las instituciones nacionales e internacionales están adoptando marcos legales cada vez más estrictos para su desarrollo y uso responsable.

El ejemplo más paradigmático de este nuevo contexto es el ya mencionado RIA. Este instrumento normativo, pionero en el mundo, establece una clasificación de los sistemas de IA en función de su nivel de riesgo (mínimo, limitado, alto o prohibido) e impone obligaciones específicas de transparencia, seguridad, trazabilidad y supervisión humana. De este modo, la UE busca asegurar que los sistemas de alto riesgo – como los utilizados en infraestructuras críticas, selección de personal o diagnóstico médico– sean desarrollados y utilizados cumpliendo con requerimientos específicos.

Pero esta normativa no ha sido la única y son muchas las iniciativas regulatorias y guías relacionadas con la IA.

De este modo, países como Estados Unidos, Canadá, Japón, China, Brasil, Perú, Italia o Reino Unido están elaborando sus propios marcos regulatorios o guías de buenas prácticas para controlar los impactos de la IA. La OCDE, el Consejo de Europa y las Naciones Unidas también promueven principios comunes de gobernanza algorítmica, centrados en la transparencia, la equidad, la responsabilidad y el respeto de los derechos humanos. Este fenómeno ha generado una presión regulatoria creciente, donde las organizaciones – públicas y privadas– deben adaptarse simultáneamente a múltiples marcos legales, estándares técnicos y expectativas sociales.

La situación actual es análoga a la que vivieron las empresas tecnológicas con la entrada en vigor del RGPD en 2018. En aquel momento, la privacidad de los datos personales se convirtió en un requisito estratégico, no solo jurídico. Hoy, con la IA, asistimos a un proceso similar: **las entidades que no integren programas de cumplimiento normativo específicos en materia de IA se exponen a sanciones económicas, pérdida de reputación y restricciones operativas.**

El uso de la IA en la automatización de decisiones que afectan directamente a los derechos y oportunidades de las personas – como la contratación laboral, la concesión de créditos o la vigilancia policial– ha evidenciado la necesidad de mecanismos de control y rendición de cuentas.

Entre las principales **preocupaciones** destacan:

→ La **opacidad de los algoritmos:** a raíz del funcionamiento en modo “caja negra”, con procesos internos difíciles de explicar incluso para sus propios desarrolladores. Esta falta de transparencia socava la confianza pública y dificulta la atribución de responsabilidades en caso de errores o discriminaciones.

→ El **sesgo y la discriminación algorítmica:** los algoritmos aprenden de los datos con los que se entrenan, y si estos contienen prejuicios sociales, históricos o demográficos, la IA puede replicarlos o amplificarlos. Existen numerosos casos documentados de sistemas que han mostrado sesgos de género, raza o edad en procesos de selección o evaluación.

→ La **manipulación informativa y la desinformación:** con el auge de los modelos generativos, como los que producen textos, imágenes o vídeos sintéticos, surgen nuevos riesgos en el terreno político, mediático y cultural. El uso indebido de estas tecnologías puede afectar la integridad de los procesos democráticos o dañar la reputación de personas e instituciones.

→ La **pérdida de control humano y la autonomía de las máquinas:** la creciente complejidad de los sistemas de IA plantea interrogantes sobre la supervisión, corrección y el control humano para revertir decisiones automatizadas.

Estas preocupaciones han generado una demanda social de responsabilidad, exigiendo que los desarrolladores, proveedores y usuarios de IA garanticen el respeto a los valores fundamentales: la dignidad humana, la no discriminación, la justicia, la privacidad y la transparencia. En consecuencia, **los programas de cumplimiento normativo se consolidan como herramientas esenciales para prevenir riesgos, demostrar diligencia debida y construir confianza entre ciudadanos, reguladores e inversores.**

Un programa de cumplimiento normativo en IA no se limita a cumplir formalmente con la ley; implica diseñar una arquitectura organizativa y técnica que integre la ética, la gobernanza y la rendición de cuentas en todo el ciclo de vida de los sistemas de IA.

En la práctica, **estos programas deben incluir**, entre otros:

01. **Mapeo y clasificación de sistemas de IA según su nivel de riesgo**, conforme a las regulaciones vigentes.

02. **Evaluaciones de impacto algorítmico** que identifiquen posibles efectos adversos sobre los derechos fundamentales.

03. **Protocolos de transparencia y explicabilidad** que permitan entender cómo y por qué un sistema toma determinadas decisiones.

04. **Mecanismos de supervisión humana efectiva** para garantizar que la decisión final siga siendo controlable.

05. **Canales internos de reporte y auditoría** que permitan detectar y corregir incumplimientos antes de que generen daños o sanciones.

06. **Formación continua del personal** y creación de comités de ética o gobernanza algorítmica.

El cumplimiento normativo en IA, además, debe entenderse como una ventaja competitiva, permitiendo reducir la exposición a riesgos legales y fortaleciendo la reputación y credibilidad de una organización ante los clientes, inversores y autoridades públicas. En un mercado cada vez más consciente del impacto social de la tecnología, la confianza se convierte en un activo estratégico.

Por eso, el verdadero reto, no radica únicamente en crear documentos de cumplimiento o protocolos aislados, sino en fomentar una cultura organizacional basada en la responsabilidad. La ética y el cumplimiento deben formar parte del ADN de las entidades que desarrollan o utilizan IA, integrándose desde la fase de diseño (lo que se conoce como ethics by design o compliance by design).

En este sentido, el cumplimiento normativo en IA debe concebirse como un proceso dinámico y transversal, que evolucione al mismo ritmo que la tecnología y la regulación. No basta con cumplir las normas vigentes; es necesario anticipar las futuras tendencias regulatorias, participar en la elaboración de estándares sectoriales y promover la transparencia proactiva.

Asimismo, los programas de cumplimiento pueden actuar como un puente entre la innovación y la regulación, demostrando que es posible desarrollar IA avanzada responsable. Frente a la falsa dicotomía entre innovación y control, la experiencia muestra que una innovación ética puede ser, también, sostenible.

9.1. Cumplimiento Normativo **del RIA**

En el actual panorama regulatorio y normativo, resulta imprescindible centrar la atención en el RIA. Este marco normativo destaca no solo por ser el primero de carácter global en establecer un enfoque armonizado sobre el desarrollo y uso de estas tecnologías, sino también por su impacto directo en los Estados miembros de la Unión Europea y, en particular en España.

Su relevancia radica en que está diseñado para **garantizar un equilibrio entre la innovación tecnológica y la protección de los derechos fundamentales, la salud y la seguridad,** aspectos esenciales en un entorno cada vez más digitalizado y automatizado, al tiempo que pretende apoyar la innovación y prevenir la fragmentación del mercado. Además, el RIA se está consolidando como referencia internacional en la regulación de la IA, marcando un estándar que influirá en el desarrollo de políticas tecnológicas más allá de las fronteras europeas.

El Reglamento de Inteligencia Artificial (RIA) o AI Act (Reglamento (UE) 2024/1689, de 13 de junio de 2024) fue publicado en el Diario Oficial de la Unión Europea el 12 de julio de 2024 y entró en vigor el 1 de agosto del mismo año y, como hemos indicado, marcó un hito al ser la **primera norma vinculante** de alcance general sobre IA en el mundo, y que establece un marco jurídico horizontal que regula el desarrollo, la comercialización y el uso de los sistemas de IA dentro del territorio de la Unión, con efectos extraterritoriales sobre actores no europeos cuyos sistemas se utilicen en la UE.

Aplica a cualquier empresa que compre, desarrolle, personalice o utilice sistemas de IA que puedan **afectar a un ciudadano de la UE**, incluyendo proveedores establecidos fuera de la Unión si sus resultados se destinan a ser utilizados en la UE.

Se excluyen de su aplicación los sistemas o modelos de IA utilizados exclusivamente para **fines militares, de defensa o seguridad nacional**, así como la **investigación, prueba o desarrollo científicos** antes de su introducción en el mercado.

La estructura del RIA se fundamenta en un enfoque basado en el riesgo, de modo que las obligaciones de cumplimiento aumentan conforme lo hace el potencial de daño o impacto del sistema de IA sobre los derechos fundamentales, la seguridad o los valores democráticos. El RIA distingue cuatro niveles de riesgo:

01.

Riesgo inaceptable

Abarca usos prohibidos de IA, tales como la manipulación cognitiva de personas vulnerables, los sistemas de puntuación social (social scoring) por parte de gobiernos, la vigilancia biométrica en tiempo real en espacios públicos (salvo excepciones muy restringidas) o la inferencia de emociones en lugares de trabajo y centros educativos.

02.

Riesgo Alto

Comprende sistemas utilizados en ámbitos críticos (salud, empleo, justicia, educación, crédito, infraestructuras esenciales, seguridad, entre otros) y son los que pueden tener un impacto significativo en la salud, la seguridad o los derechos fundamentales.

Ejemplos de alto riesgo incluyen sistemas usados en: empleo y gestión de trabajadores (contratación, evaluación de desempeño), aplicación de la ley (evaluación de riesgo de delitos, fiabilidad de pruebas), migración, asilo y control fronterizo (polígrafos, evaluación de riesgos) y acceso a servicios esenciales (públicos y privados).

03.

Riesgo limitado

Donde se imponen obligaciones de transparencia (por ejemplo, chatbots o sistemas generativos que deben revelar su naturaleza artificial).

04.

Riesgo mínimo o nulo

En el que el cumplimiento se basa en buenas prácticas o códigos voluntarios.

Este marco regulatorio integral, que se implementará de manera escalonada, se traduce en implicaciones prácticas y responsabilidades concretas para todos los operadores de la cadena de valor de la IA, acarreando consecuencias significativas en caso de incumplimiento de sus disposiciones:

→ **Proceso de implementación escalonado:** las diferentes obligaciones entran en vigor en distintos momentos, no todas a la vez, extendiendo su aplicación total hasta el año 2027. Este enfoque por fases da a las empresas y organizaciones un tiempo de adaptación gradual, priorizando los riesgos más altos primero.

→ **Profesionalización de la IA:** las organizaciones están profesionalizando la estrategia de IA, creando roles o departamentos dedicados a definir y gestionar su adopción y despliegue de forma coherente y centralizada.

→ **Transparencia y documentación:** se exige un nivel significativo de inventario de casos de uso, documentación técnica, registros de actividad y transparencia sobre el funcionamiento de los sistemas, así como la identificación clara del contenido generado por IA.

→ **Ciberseguridad:** es esencial fortalecer las medidas de ciberseguridad para proteger los sistemas de IA contra nuevos vectores, técnicas, tácticas y procedimientos de ataque que ya están aflorando, así como brechas de privacidad.

→ **Formación y conocimiento:** dada la complejidad del RIA, es crucial capacitar al personal en sus implicaciones para la seguridad y protección de datos, ya que un gran porcentaje aún no está familiarizado con él.

→ **Evaluación y adaptación:** las empresas deben identificar si utilizan o desarrollan sistemas de IA y, en caso afirmativo, determinar el rol que desempeñan, y su nivel de riesgo.

→ **Gestión de datos rigurosa:** es fundamental establecer políticas robustas de gobernanza de datos para asegurar la calidad, representatividad y minimización de sesgos en los conjuntos de datos de entrenamiento, especialmente si se manejan datos personales.

→ **Supervisión humana y ética:** las organizaciones deben garantizar una supervisión humana efectiva y realizar evaluaciones de impacto ético y de derechos fundamentales, contando con personal capacitado en IA y seguridad.

→ **Copyright y Propiedad Intelectual:** los proveedores de GPAI deben respetar los derechos de autor y proporcionar resúmenes detallados de los datos de entrenamiento.

→ **Riesgos y Responsabilidad:** la ley introduce responsabilidades claras para todos los actores de la cadena de valor de la IA, lo que exige a las empresas una reevaluación de sus procesos, el valor de los recursos humanos y la mitigación de riesgos operativos, regulatorios y de sesgos.

Requisitos de cumplimiento para sistemas de alto riesgo

El núcleo del cumplimiento normativo en la Unión Europea recae sobre los sistemas de IA clasificados como de alto riesgo, regulados en los **artículos 8 a 51 del RIA y en los Anexos II y III** del mismo.

Los proveedores y, en ciertos casos, los desplegadores de estos sistemas deben implementar un sistema de gestión de calidad que abarque todo el ciclo de vida del producto, desde el diseño hasta la vigilancia posterior a su comercialización. Dicho sistema debe documentar, entre otros elementos, los siguientes aspectos:

→ **Gestión del riesgo:** proceso continuo de identificación, análisis y mitigación de riesgos derivados del uso del sistema, que debe actualizarse a lo largo de la vida útil del modelo. Incluye tanto riesgos previstos como emergentes, así como medidas preventivas y correctivas.

→ **Gobernanza de datos:** obligación de garantizar la calidad, pertinencia y representatividad de los datos utilizados para entrenar, validar y probar los modelos. Los datos deben estar libres de sesgos indebidos, debidamente etiquetados y sometidos a mecanismos de control estadístico.

→ **Documentación técnica:** mantenimiento de un expediente técnico exhaustivo que describa el diseño, la arquitectura del sistema, los algoritmos empleados, las fuentes de datos, las métricas de rendimiento, los límites del modelo y las medidas de seguridad implementadas. En el caso de que el sistema sea proporcionado por un proveedor, éste deberá proporcionar la documentación técnica necesaria.

→ **Registro y trazabilidad (logging):** los sistemas deben permitir el registro automático de eventos (archivos de registro) a lo largo de todo su ciclo de vida. Esta obligación consiste en conservar registros automáticos de las operaciones del sistema que permitan reconstruir decisiones o resultados, facilitando auditorías y evaluaciones posteriores.

→ **Transparencia e información al usuario:** los proveedores deben proporcionar instrucciones claras sobre el uso previsto del sistema, su nivel de precisión, limitaciones y las medidas necesarias para su supervisión humana.

→ **Supervisión humana efectiva:** se requiere una intervención y vigilancia humana mediante medidas que permitan a los operadores anular manualmente decisiones. De este modo, el diseño del sistema debe garantizar que las personas responsables puedan comprender, controlar y, en su caso, interrumpir la operación del sistema de IA para prevenir daños o decisiones erróneas.

→ **Robustez, precisión y ciberseguridad:** el sistema debe demostrar resistencia frente a fallos, ataques adversariales, manipulación, envenanimiento de datos o resultados no intencionados.

→ **Vigilancia post-comercialización:** los proveedores deberán establecer un plan de vigilancia continua, recopilar información sobre el comportamiento real del sistema y notificar incidentes graves o anomalías a las autoridades competentes.

El marco legal ya está en vigor (desde agosto de 2024) y la aplicación es escalonada. Las **prohibiciones** entraron en vigor en febrero de 2025, y las normas para GPMs en agosto de 2025, mientras que la mayoría de las obligaciones para SAR están previstas para agosto de 2026. La amenaza de sanciones es severa: **hasta 35 millones de euros o el 7% del volumen anual de negocio global** por incumplimiento. Además, pueden incluir la prohibición de uso o la retirada de certificaciones. Esto convierte el RIA en una de las normativas más estrictas en materia tecnológica, equiparable al RGPD en su régimen sancionador.

Sanciones:

35 millones de euros

7% del volumen anual del negocio global

Evaluación de **conformidad y marcado CE**

Antes de poner en el mercado o en servicio un sistema de alto riesgo, el proveedor debe someterlo a un procedimiento de evaluación de conformidad, que puede ser interno o mediante un organismo notificado independiente, dependiendo de la categoría del sistema y del marco sectorial aplicable (por ejemplo, dispositivos médicos, transporte, o productos sujetos a legislación armonizada).

Una vez verificada la conformidad con los requisitos del RIA, el proveedor debe emitir una declaración UE de conformidad y colocar el marcado CE en el producto o interfaz digital, garantizando que cumple con la legislación de la Unión. Cualquier modificación sustancial del sistema obliga a repetir la evaluación.

Además, los proveedores deberán registrar los sistemas de alto riesgo en una base de datos europea de IA, gestionada por la Comisión Europea, lo que refuerza la transparencia y la trazabilidad de los productos en el mercado interior. También se deberá tener en cuenta que la CE tiene la capacidad de ampliar el listado de casos de uso de riesgo alto del RIA, lo que obliga a todas las entidades a revisar sus inventarios ante la futura publicación de estas potenciales nuevas incorporaciones.

Obligaciones adicionales para **modelos de propósito general (GPAI)**

El RIA también introduce una categoría específica para los modelos de IA de propósito general (General-Purpose AI, GPAI), como los grandes modelos lingüísticos o multimodales que sirven de base a múltiples aplicaciones.

A partir del 2 de agosto de 2025, los proveedores de GPAI deberán:

→ **Documentar** la arquitectura, fuentes de datos y métodos de entrenamiento empleados.

→ **Garantizar la transparencia** en relación con la energía utilizada, el tamaño del modelo y los recursos de cómputo.

→ **Adoptar medidas** para prevenir usos indebidos y facilitar el cumplimiento descendente por parte de quienes integren estos modelos en aplicaciones concretas.

→ En el caso de GPAI con riesgo sistémico, se imponen **obligaciones reforzadas**, como auditorías independientes, pruebas de robustez, reportes técnicos periódicos al AI Office y planes de mitigación de riesgos globales.

Roles en torno a la IA

El RIA define una serie de roles y responsabilidades a lo largo de la cadena de valor de la IA, entre los que se encuentran el rol del Fabricante, el del Proveedor, el del Responsable del Despliegue, el del Importador, el del Distribuidor, o el del Responsable Autorizado.

Es crucial entender estos roles, y cuál de ellos desempeña cada actor en cada caso de uso para conocer las responsabilidades que le corresponden a cada uno, especialmente en el contexto de los anteriormente citados sistemas de alto riesgo.

Mecanismos de supervisión y autoridades competentes

El RIA crea una arquitectura institucional europea de supervisión compuesta por varios organismos:

→ La **Oficina Europea de Inteligencia Artificial (European AI Office)**, encargada de coordinar la aplicación del RIA, supervisar los GPAI y promover la coherencia normativa entre los Estados miembro.

→ Las autoridades nacionales competentes designadas por cada Estado miembro, responsables de la vigilancia del mercado, la imposición de sanciones y la evaluación de conformidad. En España se ha creado para ello la **AESIA, Agencia Española de Supervisión de la Inteligencia Artificial**, si bien se prevé una norma nacional e, incluso, a nivel autonómico, que pudieran derivar en Autoridades de Control adicionales.

→ Los **organismos notificados**, que actúan como entidades independientes de certificación y auditoría técnica.

→ El Comité Europeo de Inteligencia Artificial, de carácter consultivo, que agrupa representantes de los Estados miembros para garantizar la armonización interpretativa y el intercambio de buenas prácticas.

Este entramado institucional persigue un equilibrio entre la seguridad jurídica, la protección de los derechos fundamentales y la innovación responsable.

Cumplimiento organizativo: implicaciones prácticas

Desde la perspectiva del gobierno corporativo, el cumplimiento del RIA exige a las organizaciones desarrollar un marco interno de gestión del cumplimiento de IA (AI Compliance Framework). Dicho marco debe incluir:

→ Inventario y clasificación de los sistemas de IA empleados o desarrollados por la entidad.

→ Evaluaciones de impacto previas al despliegue, que analicen riesgos éticos, sociales y de derechos fundamentales.

→ Políticas de gobernanza de datos y control de calidad de la información utilizada para entrenar modelos.

→ Protocolos de transparencia y documentación técnica, que permitan demostrar la conformidad ante auditorías regulatorias.

→ Mecanismos de supervisión humana y planes de respuesta ante incidentes.

→ Auditorías internas y externas periódicas para verificar la eficacia del sistema de gestión y su alineación con los requisitos regulatorios.

→ Formación continua del personal en materia de IA, ética digital y cumplimiento normativo.

Integración con marcos internacionales y otras normativas

El cumplimiento del RIA puede complementarse con la adopción de estándares internacionales de gestión de la IA, tales como el NIST AI Risk Management Framework (2023), la norma ISO/IEC 42001:2023, la norma ISO/IEC TR 24368:2022, la norma ISO/IEC 23894:2023, Artificial Intelligence - Guidance on risk management, entre otras ya descritas en detalle en el capítulo 4 del presente documento.

Por otra parte, el marco regulatorio actual y de gestión de riesgos (GRC) no se circunscribe al marco establecido en el RIA y el RGPD, que señalábamos, la gobernanza de la IA es una tarea transversal que exige la integración total con normativas convergentes desde diferentes materias:

01.

Protección de datos y privacidad

Respecto a la IA, el RGPD sigue siendo de aplicación crítica en cada etapa de su ciclo de vida (entrenamiento, validación, inferencia, retirada), tal y como se indica en el capítulo específico de la guía, asociado a privacidad y protección de datos.

Por ello, conceptos como privacidad desde el diseño, transparencia y explicabilidad o la garantía de los derechos de los usuarios, así como la realización de evaluaciones de impacto en la protección de datos (EIPD) que de forma integrada con las FRIA resultan esenciales para una gestión proactiva del riesgo y el cumplimiento del principio de privacidad desde el diseño que debe ser intrínseco en las estrategias de implementación de sistemas de IA.

Puede verse el detalle en el capítulo correspondiente a Privacidad y Protección de Datos de la presente guía ya que, dada la importancia de estos aspectos cuando se habla de un sistema de IA y que dicha importancia va más allá del cumplimiento del RGPD o de una normativa de privacidad y protección de datos, estos aspectos tienen un capítulo específico en la presente guía.

02.

Seguridad de la información y ciberseguridad

Los sistemas de IA, además, deben cumplir con los requisitos necesarios de la ciberseguridad.

Así, los productos o servicios pueden estar sujetos al Reglamento (UE) 2024/2847 del Parlamento Europeo y del Consejo, de 23 de octubre de 2024, también conocido como Ley de Ciberresiliencia o CRA, y las Entidades deberán atender a las prescripciones normativas sobre seguridad de la información como Directiva (UE) 2022/2555 del Parlamento Europeo y del Consejo de 14 de diciembre de 2022 relativa a las medidas destinadas a garantizar un elevado nivel común de ciberseguridad en toda la Unión (NIS2), el Reglamento (UE) 2025/38 del Parlamento Europeo y del Consejo, de 19 de diciembre de 2024, (Reglamento de Cibersolidaridad) que pretende fortalecer la resiliencia de la UE frente a incidentes de ciberseguridad, el Reglamento (UE) 2022/2554 de diciembre de 2022 sobre la resiliencia operativa digital del sector financiero (en adelante, DORA) o el propio Esquema Nacional de Seguridad en el ámbito nacional español.

Todo este framework normativo converge y pretende crear un ecosistema que permita que la gobernanza de la IA en las entidades se efectúe también bajo los principios de seguridad desde el diseño y resiliencia.

03.

Datos

Piedra angular básica para el éxito en la implantación de sistemas de IA. Datos y de calidad. Para ello, se aprobó en la Unión Europea el Reglamento 2022/868 (Data Governance Act) también conocido como Data Act cuyo objetivo es promover el acceso y la reutilización de los datos e impulsar la interoperabilidad de los datos entre sectores y proveedores de servicios entre Estados conforme a las prescripciones establecidas en la norma aplicable desde septiembre de 2025.

A este respecto, por su entidad propia, también cabe señalar el reciente Reglamento (UE) 2025/327 del Parlamento Europeo y del Consejo, de 11 de febrero de 2025, relativo al Espacio Europeo de Datos de Salud, como reglamento sectorial que establece medidas de interoperabilidad de datos para su uso primario y eficiencia de recursos sanitarios y un sistema que permite el uso secundario de esos datos.

04.

Propiedad intelectual, secretos empresariales y responsabilidad civil:

Se requiere un enfoque contractual explícito para proteger la Propiedad Intelectual (licenciamiento de software, derechos de autor sobre inputs y outputs de GenAI) y los secretos empresariales, especialmente cuando se introduce información confidencial en sistemas de IA de terceros.

Los LLM plantean enormes retos en cuanto a la autoría y protección de contenidos generados por IA (outputs) y el uso de datos protegidos por derechos de autor para el entrenamiento (inputs), y debemos tener en cuenta también el marco regulatorio desde la perspectiva del daño y la posible responsabilidad extracontractual de los daños provocados por IA.

A tal efecto, cabe recordar que el Tribunal Superior de Justicia de la Unión Europea, Sentencia del Tribunal de Justicia de 20 junio 2024, AT y BT contra PS GbR y otros, C-590/22[23] establece la inaplicación del principio **non bis in idem** del régimen sancionador administrativo y el derecho de daños.

En relación con el derecho de daños, en el transcurso de 2025 hemos visto como la Comisión retiró por falta de consenso entre los distintos Estados miembros y la presión regulatoria de los últimos años, la Directiva sobre responsabilidad extracontractual de daños causados por sistemas de IA. Por tanto, desde la órbita del derecho de daños a nivel europeo nos encontramos con la Directiva (UE) 2024/2853 del Parlamento Europeo y del Consejo, de 23 de octubre de 2024, sobre responsabilidad por los daños causados por productos defectuosos, la Ley de Consumidores y Usuarios, así como el Código Civil para la determinación de las posibles indemnizaciones por daños.

En dicha Sentencia se reconoce el derecho del afectado a percibir una indemnización con la imposición de sanciones al responder a finalidades diferentes. En tanto en cuanto, el TJUE entiende que la indemnización responde al objetivo de indemnizar de forma "total y efectiva" los daños y perjuicios sufridos y las sanciones son una medida disuasoria o punitiva por incumplimiento o inobservancia normativa.

Estas son solo algunas de las regulaciones que han de tenerse también en consideración a la hora de garantizar que el sistema de IA cumple con la normativa vigente ya que no hay que perder de vista que cualquier regulación puede afectar directa o indirectamente a los sistemas de IA, desde el momento en el que se tratan de sistemas que tratan datos de la organización y que pueden utilizarse para múltiples propósitos por lo que para garantizar que el sistema es IA Compliance habrá que analizar tanto el sistema como el caso de uso e identificar las distintas regulaciones que puedan serle de aplicación a fin de garantizar el cumplimiento de todas ellas.

[23] Unión Europea. (2024). Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos y Directivas indicados (Reglamento de Inteligencia Artificial). Diario Oficial de la Unión Europea, L 2024/1689. https://op.europa.eu/es/publication-detail/-/publication/72688445-2ee7-11ef-a61b-01aa75ed71a1/language-es

Cumplir con estas regulaciones no es una mera cuestión de cumplimiento o de evitar sanciones, es una forma de garantizar que los sistemas de IA operen dentro de un marco de confianza, transparencia y supervisión humana efectiva. La implementación temprana de estructuras internas de AI compliance, alineadas con el RIA y con estándares internacionales de gestión, se presenta hoy como una condición imprescindible para la competitividad, la sostenibilidad y la legitimidad de cualquier organización que utilice IA.

05.

Otras normativas de aplicación

Además de las normativas ya mencionadas y el gran framework que supone el RIA, es necesario que las organizaciones tengan presentes tres aspectos adicionales para conocer por completo el marco normativo que les puede afectar:

→ El Gobierno de España presentó a principios de 2025 un **anteproyecto de ley de gobernanza de la Inteligencia Artificial (IA)**, que busca garantizar un uso de la Inteligencia Artificial que sea ético, inclusivo y beneficioso para las personas. Uno de los principales aspectos de su texto es la regulación del régimen sancionador, así como las competencias que diferentes Autoridades de Control tendrán al respecto.

→ La realidad territorial española hace necesario también saber cuándo se publican **leyes autonómicas que regulan de alguna manera el fenómeno de la IA** y quiénes son los sujetos obligados, como por ejemplo la Ley 2/2025, de 2 de abril, para el desarrollo e impulso de la inteligencia artificial en Galicia.

→ Otro tipo de normativas de carácter estatal que regulan en parte o de manera concreta usos de la IA. Cabe mencionar:

- La creación del **Sandbox regulatorio**, por el cual, se crea un entorno controlado de pruebas de sistemas de IA para ver su comportamiento ante los requisitos del RIA **(regulado en el Real Decreto 817/2023, de 8 de noviembre).**
- La **Ley 12/2021 de 28 de septiembre**, por la que se regula el uso de algoritmos en el ámbito laboral, estableciendo el derecho de los trabajadores a ser informados por la empresa de los parámetros, reglas e instrucciones en los que se basan los algoritmos o sistemas de inteligencia artificial que afectan a la toma de decisiones que pueden incidir en las condiciones de trabajo, el acceso y mantenimiento del empleo, incluida la elaboración de perfiles.

Ética en la Inteligencia Artificial

10.

10.1. Contexto actual de la **ética en la IA**

La **ética** como requisito de confianza en la IA

Un sistema puede ser funcionalmente brillante y, aun así, no ser aceptable si erosiona derechos o discrimina. Hablar de ética en IA implica generar confianza. Porque una organización que despliega IA sin un marco ético claro pide a usuarios, clientes y reguladores que confíen a ciegas en sistemas complejos que, además, pueden manejar datos sensibles, tener alucinaciones, sesgos o brechas de seguridad. La confianza se construye con medidas muy concretas.

Citando nuestra anterior Guía del ISMS "Ética y Compliance en el Uso de la Inteligencia Artificial"(2023), estas medidas deben permitir a la persona entender el comportamiento del sistema y el grado de intervención humana en el mismo. Intervención real, contando con mecanismos de información y control, y con capacidad para revisar y corregir. La ética también permite gestionar prioridades cuando hay objetivos divergentes, y reducir el riesgo. Documentar datos y decisiones, auditar sesgos, habilitar canales de reclamación y planificar correcciones rápidas disminuye la probabilidad de incidentes reputacionales, regulatorios o legales. Además, la ética habilita la adopción: usuarios y equipos confían más y usan mejor la tecnología cuando perciben reglas claras, límites bien puestos y vías de recurso. En IA generativa es evidente: con controles y responsabilidades claras, las personas la integran de forma productiva en su día a día. La ética no frena la innovación; la hace sostenible y confiable.

Más allá del **compliance técnico:** Nuevos dilemas de la IA generativa y agéntica

La normativa y el compliance tradicional han permitido ordenar tecnologías razonablemente predecibles –requisitos, certificaciones y auditorías– cada vez ha ido mutando más a un enfoque de riesgos. Y la IA generativa y los agentes autónomos están escalando a otro nivel. Generadores de narrativas, imágenes, vídeos o código, y con capacidad para ejecutar acciones sin intervención de una persona, no limitándose a procesar datos o predecir resultados.

Una de las cuestiones más acuciantes es el impacto en la **dimensión cognitiva de las personas.** Los modelos generativos producen respuestas rápidas, complacientes y convincentes, incluso cuando son falsas, lo cual puede erosionar el pensamiento crítico y fomentar la aceptación pasiva de cualquier respuesta, con el riesgo de perder habilidades de análisis y reflexión porque resulta más cómodo delegarlas.

La transparencia técnica no es suficiente; el reto ético es cómo preservamos la capacidad de las personas para cuestionar, contrastar y decidir por sí mismas en un entorno saturado de textos, imágenes y vídeos generados artificialmente. Para ello, hay que reorientar la capacitación y la formación de las personas a este nuevo entorno, así como introducir guardarraíles cognitivos en las herramientas de IA.

Así mismo, si bien un modelo que responde a preguntas puede parecer inofensivo, el mismo sistema integrado con herramientas externas –capaz de comprar, reservar, enviar mensajes o programar tareas– la reflexión ya no es teórica: **¿Qué tareas son legítimas para entregar a un agente? ¿Qué significa "supervisión humana" cuando el sistema opera a la velocidad de una máquina?**

Otro conflicto de exponencial expansión es la alteración de la fe y autenticidad pública: deepfakes hiperrealistas o textos persuasivos a gran escala amenaza las bases democráticas, las pruebas de verdad y la reputación de personas e instituciones.

Los estándares de cumplimiento técnico hablan de marcas de agua o de avisos al usuario, pero la ética va más allá: ¿en qué contextos es legítimo usar contenido sintético? ¿qué deberes de diligencia tenemos antes de difundir una imagen o un audio generado? La cuestión ya no es solo técnica, sino cultural: preservar la confianza en lo que vemos y escuchamos.

El compliance aporta reglas, pero la ética debe aportar sentido y proteger a las personas. Y esta reflexión no es baladí: como ejemplo extremo e ilustrativo, a mediados de 2025[24] se conoció el caso de un adolescente californiano que se quitó la vida tras mantener durante meses conversaciones intensas con ChatGPT, y según la demanda presentada por sus padres contra el fabricante no solo validaba sus pensamientos de desesperanza, sino que incluso llegó a sugerir métodos de suicidio, revisar fotografías de autolesión y reforzar la idea de mantener en secreto sus intenciones. Un episodio que ilustra que no hablamos de abstracciones teóricas, sino de impactos reales sobre vidas humanas

[24] Reuters. (2025, 26 de agosto). OpenAI, Altman sued over ChatGPT's role in California teen's suicide. https://www.reuters.com/sustainability/boards-policy-regulation/openai-altman-sued-over-chatgpts-role-california-teens-suicide-2025-08-26

10.2. **Contexto y directrices éticas** aplicables

Marco regulatorio de referencia y estándares

A lo largo de los últimos años se han ido configurando marcos normativos y de gobernanza que trasladan principios, directrices y recomendaciones éticas a un plan más operativo. Entre los más relevantes se encuentran:

- **Reglamento de IA, RIA (Unión Europea, 2024)**: marco basado en el riesgo. Convierte en obligaciones legales principios como la transparencia, la supervisión humana, la no discriminación y la robustez.
- **RGPD (Unión Europea, 2016):** regula transversalmente la privacidad, la minimización de datos y el derecho a la protección frente a decisiones automatizadas.
- **NIST AI Risk Management Framework (EE. UU., 2023)[25]:** no es vinculante. Estándar de facto para la gestión de riesgos de IA. Recoge principios como explicabilidad, fiabilidad, trazabilidad y alineamiento con valores sociales.
- **Recomendación de la UNESCO sobre la Ética de la IA (2021)[26]:** primer instrumento global adoptado por 193 Estados Miembros que fija principios éticos universales: derechos humanos, inclusión, diversidad cultural, sostenibilidad ambiental.
- **Principios de la OCDE sobre IA (2019)[27]:** marco de referencia para países de la OCDE y del G20, que incorpora valores como crecimiento inclusivo, transparencia, robustez y rendición de cuentas.

Este contexto comprende normativas vinculantes (RIA, RGPD), y guías de gobernanza adoptadas por consenso internacional (OCDE, UNESCO, NIST). En conjunto, podemos decir que engloban los ejes éticos aplicables que toda organización debe considerar al diseñar, implantar y supervisar sistemas de IA.

[25] National Institute of Standards and Technology (NIST). (2023). AI Risk Management Framework (AI RMF 1.0). https://www.nist.gov/itl/ai-risk-management-framework

[26] UNESCO. (2021). Recomendación sobre la ética de la inteligencia artificial. https://unesdoc.unesco.org/ark:/48223/pf0000381137

[27] Organisation for Economic Co-operation and Development (OECD). (2019). OECD AI Principles. https://oecd.ai/en/ai-principles

Entendiendo la tradicional **orientación al riesgo** para sistemas tasados

El RIA materializa el enfoque orientado al riesgo de Europa, al clasificar los sistemas en categorías según su peligrosidad potencial. Pero va más allá, estableciendo, en su artículo 5, una lista de prácticas inaceptables, como la manipulación subliminal o la explotación de vulnerabilidades de menores (enraizado en la Carta de Derechos Fundamentales de la Unión Europea y en la Recomendación de la UNESCO sobre la ética de la IA, que remarca que la innovación tecnológica no puede erosionar la dignidad humana).

No todo lo que es técnicamente posible es jurídicamente aceptable ni éticamente legítimo. El respeto a la ley y a los derechos humanos marca una frontera: descartar desde el inicio usos que atenten contra la libertad y la autonomía de las personas.

Por otra parte, el terreno de la **privacidad desde el diseño y la gobernanza del dato** se articula principalmente a través del RGPD, pero también se integra en el RIA (artículo 10). Los principios de minimización, limitación de finalidad y control por parte del interesado son tan relevantes en la IA como en cualquier otro tratamiento de datos, y cobran más interés que nunca.

Otro eje fundamental es la exigencia de **transparencia y explicabilidad**. Europa lo recoge como obligación imponiendo a los sistemas de alto riesgo deberes de documentación y, en el caso de los GPAI, la obligación de informar sobre el entrenamiento y etiquetar los contenidos generados artificialmente. En paralelo, el **NIST AI Risk Management Framework** en Estados Unidos resalta que las características de una IA confiable incluyen que los sistemas sean "responsables y transparentes, explicables e interpretables", lo que implica ofrecer explicaciones ajustadas a cada tipo de público y no meramente técnicas. En la misma línea, la **OCDE** incorpora como principio rector la "transparencia y divulgación responsable en torno a los sistemas de IA para garantizar que las personas comprendan cuándo están interactuando con ellos y puedan cuestionar los resultados".

La justicia, equidad y no discriminación constituyen otro eje fundamental. El RIA exige gobernanza de datos y medidas para mitigar sesgos en sistemas de alto riesgo. La UNESCO y la OCDE, por su parte, insisten en que la IA debe servir al bien común y a la inclusión social.

No se trata solo de revisar conjuntos de datos en busca de errores, sino de evaluar de forma continua si los resultados generan desigualdades injustas.

Las organizaciones deben diseñar sistemas con equipos diversos, testear con datos de grupos representativos y garantizar que los algoritmos no reproducen discriminaciones. El RIA también dedica el artículo 14 a la **supervisión humana en los sistemas de alto riesgo,** que el NIST denomina human-in-the-loop o human-on-the-loop: la supervisión no es un trámite simbólico, y debe garantizar un poder real de revisión o cancelación. En la práctica, esto debe implicar que en ámbitos sensibles como la justicia o la sanidad nunca se pueda sustituir por completo la decisión humana.

Más recientemente se han incorporado al debate cuestiones como la **sostenibilidad y proporcionalidad,** que merece la pena seguir. El desarrollo de modelos cada vez más grande exige enormes cantidades de energía y recursos: por ejemplo, según la Agencia Internacional de la Energía (IEA), el consumo eléctrico de los centros de datos vinculados a IA podría doblarse hacia 2030, alcanzando 945 TWh anuales, equivalente al consumo total de Japón. La proporcionalidad exige calibrar el valor social de cada aplicación frente a su huella ecológica.

GPAI y agentes: límites del marco regulatorio actual y necesidad de nuevos enfoques

A diferencia de los sistemas tradicionales, donde era posible tasar la finalidad y asignar obligaciones proporcionales, los GPAI desbordan esa lógica de clasificación. Su carácter abierto y multifuncional hace inviable encasillarlos en un riesgo concreto: lo mismo pueden impulsar aplicaciones médicas que generar campañas de desinformación, redactar un informe jurídico o programar un malware.

En los GPAI no existe, por tanto, una finalidad única que permita anticipar riesgos y regularlos de manera proporcional. Lo que antes se preveía como un marco de finalidades tasadas, se convierte en un escenario desdibujado de usos legítimos, abusos y amenazas emergentes.

¿Qué finalidad puede asignarse a un modelo que lo mismo redacta un ensayo que escribe código, genera imágenes, traduce un texto jurídico o simula una voz humana? El RIA impone obligaciones adicionales a los GPAI –documentación técnica, resúmenes sobre datos de entrenamiento, etiquetado de contenidos sintéticos–. Medidas importantes, pero los riesgos permanecen: modelos liberados al mercado y que se integran en productos de terceros, se encadenan con otras aplicaciones y se reutilizan en contextos muy alejados de los objetivos de sus desarrolladores, llegando a construirse ecosistemas enteros sobre un mismo núcleo tecnológico, donde los límites entre innovación, abuso y riesgo inesperado se difuminan.

Es la naturaleza de un ecosistema de IA agéntica: sistemas que no solo generan contenido, sino que también actúan, planifican, interactúan con herramientas externas, ejecutan tareas y toman decisiones de manera autónoma.

Es necesaria la conceptualización de nuevos enfoques, y su traslado al orden jurídico. En primer lugar, la **protección y autonomía cognitiva**. Porque más allá de la transparencia técnica, las organizaciones deben proteger la capacidad de las personas para pensar y decidir por sí mismas. Esto implica diseñar interfaces que fomenten la verificación crítica y ofrecer avisos claros sobre la naturaleza generada de los contenidos.

También la r**esponsabilidad distribuida y la trazabilidad compartida.** La ética exige aquí un reparto claro de roles y obligaciones, evitando tanto el vacío de responsabilidad como la dilución que deja al usuario final desprotegido. Pero esta responsabilidad compartida solo es efectiva si se acompaña de mecanismos de trazabilidad que permitan auditar el recorrido del modelo: qué datos se usaron, qué modificaciones se introdujeron y en qué contextos se aplicó. Es decir, rendición de cuentas verificable.

[25] National Institute of Standards and Technology (NIST). (2023). AI Risk Management Framework (AI RMF 1.0). https://www.nist.gov/itl/ai-risk-management-framework

[26] UNESCO. (2021). Recomendación sobre la ética de la inteligencia artificial. https://unesdoc.unesco.org/ark:/48223/pf0000381137

[27] Organisation for Economic Co-operation and Development (OECD). ([2019]). OECD AI Principles. https://oecd.ai/en/ai-principles

Propiedad intelectual y vulnerabilidad de la verdad

La tensión entre la propiedad intelectual y la capacidad de la IA generativa para replicar, recombinar y recrear contenidos protegidos es uno de los focos éticos actuales, ya que la confianza en la autenticidad de lo que vemos, oímos o leemos se ve amenazada cuando los resultados de la IA pueden mezclar obras protegidas con contenidos reales de forma indistinguible. El entrenamiento de grandes modelos con enormes volúmenes de datos plantea dudas sobre el uso de materiales sujetos a copyright sin autorización, y aunque existen argumentos sobre el uso legítimo o excepciones de minería de datos, muchos autores sienten que sus obras se convierten en materia prima gratuita para sistemas que luego generan contenidos que compiten con ellos, lo que añade además el dilema de la autoría:

¿Quién es el titular de una canción compuesta por un algoritmo entrenado con miles de piezas musicales humanas? ¿El programador, la empresa, el usuario que pidió la obra, o nadie en absoluto? La ausencia de respuestas claras amenaza con desincentivar la creación, pues la apropiación indebida y el plagio inadvertido se vuelven riesgos constantes.

Este conflicto no se detiene en lo jurídico. La capacidad de la IA generativa para producir imágenes hiperrealistas, audios convincentes o textos persuasivos a gran escala coloca en jaque la **fe pública en la verdad**, haciendo que dejemos de confiar en lo que vemos o escuchamos: antes, una fotografía servía como prueba, una grabación como testimonio, un documento como registro fiable. Hoy, esa confianza se resquebraja a golpe de deepfake, textos generados en masa alimentando la desinformación o vídeos manipulados.

Lo que está en juego no es solo la protección de derechos de autor, sino la infraestructura sobre la que se sostiene la convivencia y la confianza en que existe una verdad accesible y contrastable.

Desde la perspectiva ética, ambos dilemas están profundamente entrelazados, erosionando la base misma de la cultura e incluso de la democracia. Las soluciones técnicas –como marcas de agua digitales o etiquetas de contenido generado– son necesarias, pero no suficientes. El verdadero reto es normativo y cultural, redefiniendo las reglas de la propiedad intelectual y reforzando la alfabetización digital de la ciudadanía para que pueda distinguir, contrastar y no caer en la manipulación.

Más allá de estas reflexiones, ya se están librando batallas judiciales que muestran lo relevante y urgente que es este frente. desde la demanda de un editor digital contra OpenAI en Delaware por usar sin permiso su contenido (Reuters[28]), hasta la acción encabezada por The New York Times contra OpenAI y Microsoft, que se suma a más de una docena de casos de autores por el uso no autorizado de sus obras (The Guardian[29]) .

[28] Reuters. (2025, 24 de abril). Publisher Ziff Davis sues OpenAI for copyright infringement. https://www.reuters.com/business/publisher-ziff-davis-sues-openai-copyright-infringement-2025-04-24/

[29] The Guardian. (2025, 4 de abril). US authors' copyright lawsuits against OpenAI and Microsoft combined in New York with newspaper actions. https://www.theguardian.com/books/2025/apr/04/us-authors-copyright-lawsuits-against-openai-and-microsoft-combined-in-new-york-with-newspaper-actions?utm_source=chatgpt.com

Ética de la IA y neuroderechos: ¿el futuro?

Pronto tendremos tecnologías capaces de interactuar directamente con la actividad cerebral, lo cual abre una cuestión ética sin precedentes: **los neuroderechos**. Y es que la mente humana y sus procesos más íntimos merecen una protección específica frente al avance de las neurotecnologías y de las aplicaciones de IA que, combinadas con sensores y modelos predictivos, pueden acceder, inferir o incluso modificar estados mentales.

La ética aquí ya no se limita a garantizar un uso responsable de datos personales o de algoritmos en la toma de decisiones: hablamos de la defensa de la **libertad cognitiva**, del derecho a mantener nuestros pensamientos en la esfera privada y a no ser manipulados en nuestras emociones o conductas mediante estímulos dirigidos por sistemas inteligentes.

Algunos países han comenzado a dar pasos concretos. Chile, por ejemplo, reformó su Constitución en 2021 para reconocer los neuroderechos como parte de los derechos humanos básicos, marcando un precedente internacional (Norma 1166983[30]). También organismos como la UNESCO o el Consejo de Europa han advertido también sobre la necesidad de extender la protección de la dignidad humana a este nuevo ámbito que plantea dilemas sin precedentes en términos de consentimiento informado, manipulación subliminal y control social (UNESCO NEUROTECH[31])

¿Qué ocurre si una empresa es capaz de inferir el estado emocional de una persona en tiempo real y adaptar su publicidad a ello? ¿Qué implicaciones tiene un sistema que, conectado a un casco de realidad virtual, pueda alterar o inducir experiencias sensoriales que el usuario perciba como auténticas?

[30] Chile. (2023). Ley N.º 21.595 sobre delitos informáticos. Biblioteca del Congreso Nacional de Chile. https://www.bcn.cl/leychile/navegar?idNorma=1166983
[31] UNESCO. ([2023]). Recomendación sobre la ética de la neurotecnología. https://www.unesco.org/en/ethics-neurotech/recommendation

Pautas prácticas y mecanismos de gobernanza para empresas que desarrollan, usan o adoptan IA

10.3.

La gobernanza ética de la IA en las empresas requiere integrar prácticas y responsabilidades, no limitarse a cumplir con la regulación.

Para ello, esta guía distingue tres ejes complementarios:

01. Empresas que desarrollan IA, responsables de introducir salvaguardas desde el diseño, el entrenamiento y el despliegue de los modelos.

02. Usuarios de IA –empleados, clientes o consumidores de GPAI–, quienes necesitan pautas claras de uso crítico, verificación y formación continua.

03. Líderes con capacidad de decisión y responsables en la adopción de IA en sus organizaciones. Deben garantizar la coherencia entre ética y estrategia, dotar de recursos al gobierno y asumir la rendición de cuentas.

Pautas de referencia **para empresas que desarrollan IA**

→ Anonimizar y minimizar datos desde el inicio, sistemas entrenados con la información estrictamente necesaria, eliminando o transformando identificadores personales o sensibles desde el diseño. Por ejemplo, una empresa que desarrolla un modelo para analizar patrones de absentismo laboral sustituirá por identificadores anónimos, nombres de empleados, direcciones, etc., conservando solo lo relevante, como edad, categoría profesional y tipo de contrato.

→ Documentar y trazar datos y modelos, esto es, identificar, inventariar y documentar el origen de los datos, su tipo de licencia/permisos/derechos, así como las modificaciones que han sufrido. Trazabilidad, diligencia en materia de propiedad intelectual y protección de datos y facilidad para auditorías regulatorias. Así, una startup que entrena un modelo de generación de imágenes puede documentar que el 60% de su dataset proviene de repositorios con licencias Creative Commons, un 30% de contratos con fotógrafos y un 10 % de material propio, conservando evidencias en caso de disputa judicial.

→ Tomar decisiones conscientes sobre la finalidad del sistema, y sobre qué datos excluir, qué objetivos priorizar y qué límites establecer antes de comenzar el entrenamiento. Las empresas tienen la responsabilidad de definir explícitamente los escenarios en los que sus sistemas no deben usarse. Un ejemplo: al diseñar un algoritmo de recomendación de empleo, la empresa decide desde el inicio no incluir variables como género o estado civil, que podrían sesgar la asignación de oportunidades. Otro ejemplo es el de una compañía que comercializa un modelo de generación de voz, que puede prohibir expresamente su uso para imitar a menores de edad o a figuras políticas, y documentar esta exclusión como parte de su política de ética por diseño.

→ Garantizar supervisión humana con poder de acción real. En base a tres elementos básicos: acceso a explicaciones comprensibles de las decisiones del sistema, formación específica para interpretar esas explicaciones y botones de acción claros de “confirmación”, “modificación”, “responsabilidad” etc. dentro de las propias herramientas. Por ejemplo, al usar un sistema de riesgo crediticio, el analista vea en pantalla no solo la puntuación propuesta por el algoritmo, sino también los principales factores que la motivan (por ejemplo, nivel de ingresos o historial de pagos). Además, la interfaz debe incluir opciones directas para aprobar, rechazar o modificar la decisión automática, de modo que la persona conserve el control.

→ Establecer mecanismos ágiles y responsables de interacción con el usuario: escuchar a los usuarios e integradores para identificar sesgos, errores o daños no previstos. Para ello, las empresas deben ofrecer canales simples y accesibles de reclamación, así como compromisos de respuesta. Una compañía que ofrece una API de IA para clasificación de imágenes incorpora un botón en el panel de control que permite reportar “clasificación errónea” o “contenido inapropiado”, comprometiéndose a revisar y responder en un plazo máximo.

→ Atacar y forzar el sistema, haciéndolo capaz de no colapsar frente a usos maliciosos o entradas inesperadas. Es clave simular ataques de inyección de prompts y escenarios de manipulación deliberada antes de su despliegue. Por ejemplo, una compañía lanza un chatbot de atención médica donde el usuario intenta forzar recetas ilegales o solicitar instrucciones de autolesión, simula cómo el sistema detecta y bloquea estas conductas.

→ Documentar decisiones y procesos de entrenamiento: qué decisiones se tomaron, por qué y con qué criterios, documentando claramente los datos empleados, versiones del modelo, objetivos de optimización, y dilemas éticos que hayan podido surgir. Un ejemplo es una empresa que desarrolla un sistema de selección de personal, que archiva los criterios por los que decidió dar más peso a la experiencia laboral que a la titulación académica, para poder justificar esta elección si en el futuro se detecta un sesgo.

→ Realizar evaluaciones periódicas de impacto ético y social, y planificar ciclos de monitorización continua. Cuando sea relevante, es importante anticipar el impacto social, cognitivo y ambiental de los sistemas de IA. Estas evaluaciones ayudan a identificar riesgos menos evidentes, como la pérdida de autonomía del usuario o el consumo excesivo de energía. Por ejemplo, en educación, un modelo generativo puede afectar la capacidad crítica de los estudiantes si se abusa de la herramienta. La monitorización continua permite detectar desviaciones, sesgos o vulnerabilidades, y adaptar el sistema a cambios normativos o sociales, como ajustar un motor de recomendación que favorece contenidos sensacionalistas.

Pautas de referencia **para usuarios de IA (empleados, clientes, consumidores)**

01 Formarse en uso crítico de la IA: capacitarse en cómo funcionan los modelos, sus limitaciones y su confiabilidad, reduciendo el riesgo de aceptar salidas erróneas, alucinaciones o manipulaciones.

02 Evitar introducir datos sensibles en sistemas externos, ya que prompts que contienen información sensible pueden ser almacenados o reutilizados sin control.

03 Establecer límites de delegación en la IA, decidiendo de antemano qué tareas pueden ser asistidas por IA y cuáles no, con criterio profesional y previniendo que decisiones críticas recaigan por completo en un sistema automático.

04 Desarrollar hábitos de contraste cognitivo, fomentando cuestionar lo que devuelve la herramienta, comparar alternativas y no dar por buenas las respuestas más rápidas. Preservando el pensamiento crítico y evitando la “comodidad intelectual” por outputs convincentes.

05 Utilizar la IA como apoyo y no como sustituto, combinando el criterio humano con la capacidad del modelo, en lugar de delegar totalmente. Así, se reduce el riesgo de dependencia excesiva y se refuerza la calidad del resultado.

06 Ser transparentes en el uso de IA, comunicando a clientes o interlocutores cuándo un texto, imagen o informe ha sido generado con ayuda de un modelo. Esta práctica refuerza la confianza y evita malentendidos.

07 Mantener copias y respaldos de los resultados, de forma que no se dependa exclusivamente de la herramienta ni se pierda trazabilidad de los outputs. Esto permite demostrar qué se generó, cuándo y con qué uso.

Pautas de referencia para líderes (CEOs, directivos, decisores) que adoptan IA en sus organizaciones

01

Definir una estrategia de IA alineada con los valores de la organización, para que su adopción no solo esté basada en eficiencia o reducción de costes.

02

Evaluar el impacto de la IA en la fuerza laboral, anticipando qué tareas se automatizarán y qué funciones se transformarán, con el objetivo de planificar la transición y evitar despidos masivos no gestionados.

03

Invertir en reskilling y upskilling de empleados, ofreciendo programas de capacitación para que los trabajadores adquieran nuevas competencias digitales y puedan adaptarse al cambio, e incluso incrementar su productividad personal.

04

Incorporar la perspectiva de colectivos vulnerables en la toma de decisiones.

05

Establecer un comité de ética y gobernanza de IA diverso y con capacidad de decisión real.

06

Asumir la rendición de cuentas personalmente como líderes, reconociendo que la responsabilidad última del uso de IA en la organización no recae en el algoritmo ni en el proveedor, sino en la dirección.

Auditoría y Trazabilidad de Sistemas de IA

11.

La gobernanza de la IA dentro de una organización debe asegurar que todos los sistemas de IA, propios o de terceros, cuenten con mecanismos de trazabilidad que respalden la transparencia en sus procesos. Además, resulta esencial que estos sistemas sean auditables para prevenir errores, sesgos o usos inadecuados de la tecnología.

11.1. **¿Por qué auditar** la IA?

Auditar un sistema de IA, ya sea interno o como parte de un servicio ofrecido por terceros, es un paso imprescindible para ofrecer garantías acerca de su funcionamiento. Dada la adopción masiva de la IA en entornos donde las decisiones automatizadas pueden afectar derechos fundamentales, procesos críticos productivos o la reputación, la auditoría es una herramienta fundamental para identificar riesgos, corregir desviaciones y fortalecer la confianza de clientes, reguladores y negocio.

Sin ser exhaustivos, dentro de las actividades de la auditoría de sistemas de IA habría que considerar, al menos, los siguientes objetivos:

→ Verificar desde el diseño el cumplimiento de requisitos técnicos, éticos y legales.
→ Evaluar el impacto sobre los derechos fundamentales, en especial lo que pueda afectar a poblaciones vulnerables.
→ Detectar sesgos, errores o vulnerabilidades en los modelos.
→ Asegurar trazabilidad en la toma de decisiones y en todos los procesos algorítmicos.
→ Documentar la plataforma, necesario para fortalecer la gobernanza organizativa y la supervisión humana mediante su conocimiento.
→ Colaborar en la promoción de buenas prácticas sobre la sostenibilidad y eficiencia energética del sistema.

11.2. **Trazabilidad:** seguir el rastro de cada decisión

La trazabilidad, generalmente referida a los aspectos relacionados con la trazabilidad algorítmica y técnica, es requisito para saber qué datos se usaron, qué modelo se aplicó, qué parámetros estaban activos y/o en qué contexto se tomó cierta decisión o un conjunto de decisiones. **Una trazabilidad completa y efectiva que contemple no sólo mecanismos técnicos orientados a los sistemas internos de IA o los organizativos en el marco de la organización.** Deben contemplarse, además, elementos que formarían parte también de la "trazabilidad organizativa", y que abarca procesos, decisiones y responsabilidades dentro de la organización que afectan al desarrollo y uso del sistema de IA, incluyendo documentación de reuniones, aprobaciones, roles, coordinación entre equipos internos y externos, o el clausurado de la contratación de servicios en el ámbito de IA.

La trazabilidad de una plataforma o sistema de IA **debe contar, al menos, con los siguientes mecanismos:**

→ **Registro** de todas las actividades del ciclo de vida
→ **Control de versiones** de los modelos e historial de actualizaciones, uso de los datos y control del código productivo.
→ **Registro de decisiones** clave relacionadas con la plataforma, conectando la parte organizativa con la implantación
→ Mecanismos de trazabilidad técnica confiables en todos los sistemas con **identificadores únicos**, que contemplen las acciones realizadas por cada componente, inferencias, errores y decisiones dadas por el sistema y/o fruto de la conexión o el uso de terceros.
→ **Repositorio documental** orientado a la trazabilidad organizativa y contractual: roles, aprobaciones y responsabilidades internas.
→ **Registro de proveedores y explotación** proactiva de Acuerdos de nivel de servicio (SLA), cláusulas de trazabilidad con terceros, KPIs sobre el cumplimiento de políticas de gobernanza compartida o la gestión de cambios.
→ La **integración con sistemas de gestión documental y GRC** facilita dotar de elementos de supervisión y verificación cruzada, diseñar pruebas de consistencia y validación y facilitaría la realización de auditorías técnicas.

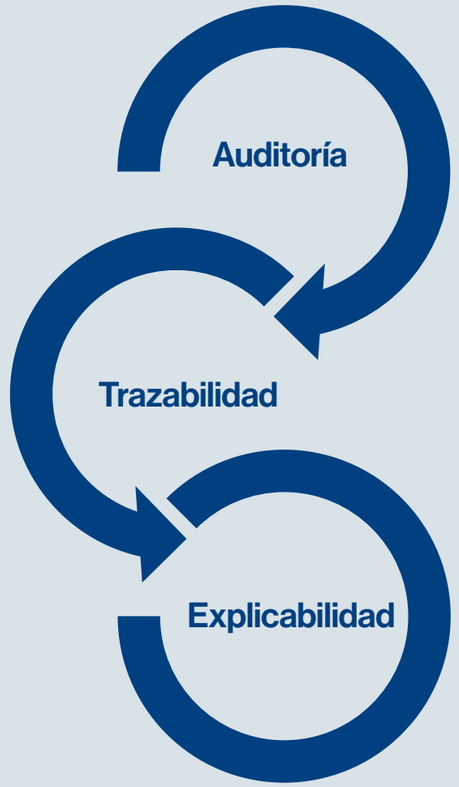
11.3. Dependencias entre **Auditoría y Trazabilidad**

La auditoría debe valorar los procesos de trazabilidad y su tratamiento por parte del resto de sistemas, para asegurar que la información dada es correcta y suficiente.

La ausencia de trazabilidad en la ejecución de alguna función crítica, el uso de formatos inadecuados o deficiencias en el seguimiento de los registros dificulta significativamente el proceso de auditoría, ya que impide una revisión estructurada y eficiente del sistema.

	Trazabilidad	Auditoría	Relación
Datos	Registra origen, transformación y uso	Evalúa calidad, sesgo y legalidad	La auditoría usa la trazabilidad para verificar el uso correcto
Modelos	Versiona y documenta cambios	Revisa rendimiento y equidad	Permite auditar qué versión se usó y cómo se entrenó
Decisiones	Guarda contexto y lógica	Evalúa impacto y justificación	Facilita auditar decisiones automatizadas
Cumplimiento	Documenta procesos	Verifica alineación con normas	La trazabilidad debe ser evidencia para la auditoría

Además de los conceptos de auditoría y trazabilidad, un tercero, la explicabilidad – capacidad de un sistema de IA para explicar a usuarios y cualquiera sus decisiones y resultados -, es requisito para aquellos sistemas clasificados como "Riesgo alto" por el RIA, junto con la supervisión humana.
La forma en la que se relacionan los conceptos de auditoría, trazabilidad y explicabilidad es esencial para abordar correctamente los tres:



Requiere trazabilidad para reconstruir decisiones y explicabilidad para entenderlas.

Base de conocimiento que permite auditar y explicar el comportamiento del sistema.

Depende de la trazabilidad para acceder a los datos y lógica, y podría ser un objetivo clave de la auditoría.

11.4. **Metodología de Auditoría** en sistemas IA

Las tareas necesarias para la realización de una auditoría de sistemas de IA deben tener como objetivo evaluar su funcionamiento desde múltiples dimensiones: técnica, funcional, ética, organizativa y sostenibilidad. Esto hace necesario contar con una metodología estructurada que ofrezca una revisión rigurosa, proporcional al riesgo, y basada en criterios normativos. Este capítulo incluye un anexo 2, donde se recogen las preguntas clave para la auditoría, organizadas según las dimensiones mencionadas. Dicho anexo constituye una herramienta práctica para orientar el proceso de evaluación.

Dimensiones de Auditoría

Dimensión Técnica

Objetivo: garantizar precisión, robustez, explicabilidad y seguridad del sistema.

Acciones clave:
- Evaluación de precisión, robustez, explicabilidad y seguridad.
- Validación de datos: calidad, representatividad, sesgo.
- Revisión del ciclo de vida del modelo: diseño, entrenamiento, validación, despliegue.
- Verificación de cumplimiento de requisitos técnicos.

Referencias: ISO/IEC 42001 (controles técnicos), RIA Art. 9 a15 (requisitos de sistemas de alto riesgo), RGPD Art. 5 y 25 (minimización y privacidad desde el diseño).

Dimensión Funcional

Objetivo: verificar que el sistema cumple su propósito operativo y se adapta al contexto de uso.

Acciones clave:
- Evaluación de impacto en derechos fundamentales.
- Equidad y no discriminación.
- Transparencia y explicabilidad.
- Revisión de mecanismos de supervisión humana y gobernanza ética.

Referencias: RIA Art. 11 (documentación técnica), ISO/IEC 42001 (gestión del ciclo de vida), RGPD Art. 22.

Dimensión Ética

Objetivo: asegurar respeto a los derechos fundamentales, equidad y supervisión humana.

Acciones clave:
- Evaluación de impacto ético.
- Revisión de sesgos.
- Mecanismos de supervisión.

Referencias: RIA Art. 14 (supervisión humana), RGPD Art. 22 (decisiones automatizadas), ISO/IEC 42001 (valores organizativos).

Dimensión Organizativa

Objetivo: auditar la gobernanza interna, roles y procesos asociados al sistema.

Acciones clave:
- Documentación de roles, responsabilidades y procesos.
- Coordinación entre equipos y proveedores.
- Revisión de políticas internas y cultura organizativa.

Referencias: RIA Art. 16–29 (obligaciones del proveedor), RGPD Art. 24–32 (responsabilidad organizativa), ISO/IEC 42001 (estructura de gobernanza).

Dimensión de Sostenibilidad y Eficiencia Energética

Objetivo: evaluar el impacto energético y la sostenibilidad del sistema.

Acciones clave:
- Medición de consumo, incluyendo huella de proveedores.
- Coordinación entre equipos y proveedores.
- Revisión de políticas internas y cultura organizativa.

Referencias: RIA Art. 16–29 (obligaciones del proveedor), RGPD Art. 24–32 (responsabilidad organizativa), ISO/IEC 42001 (estructura de gobernanza).

Evidencias **requeridas**

La auditoría debe valorar los procesos de trazabilidad y su tratamiento por parte del resto de sistemas, para asegurar que la información dada es correcta y suficiente.

La ausencia de trazabilidad en la ejecución de alguna función crítica, el uso de formatos inadecuados o deficiencias en el seguimiento de los registros dificulta significativamente el proceso de auditoría, ya que impide una revisión estructurada y eficiente del sistema.

Dimensión	Trazabilidad	Auditoría	Relación
TÉCNICA	Precisión, robustez, explicabilidad	Informes de validación, métricas, documentación técnica	ISO/IEC 42001, RIA Art. 9–15
TÉCNICA	Seguridad algorítmica	Pruebas de ciberseguridad, análisis de vulnerabilidades	ISO/IEC 42001
TÉCNICA	Minimización de datos	Revisión de datos recolectados, justificación de uso	RGPD Art. 5, 25
FUNCIONAL	Adecuación al contexto de uso	Requisitos funcionales, pruebas de usuario	RIA Art. 11
ÉTICA	Supervisión humana, equidad, transparencia	Evaluación de impacto ético, registros de decisiones	RIA Art. 14, RGPD Art. 22
ORGANIZATIVA	Gobernanza, roles y responsabilidades	Políticas internas, organigramas, actas de reuniones	ISO/IEC 42001, RIA Art. 16–29, RGPD Art. 24–32
AMBIENTAL	Eficiencia energética y sostenibilidad	Informes de consumo, políticas ambientales	ISO/IEC 42001, políticas internas

A partir de esta base metodológica, se propone un procedimiento estructurado que permite aplicar la auditoría de forma secuencial, rigurosa y adaptada al nivel de riesgo del sistema evaluado.

Procedimiento de auditoría de sistemas de IA

Se requiere de un procedimiento claro y estructurado que permita evaluar su funcionamiento desde múltiples dimensiones, agregando también de manera transversal el tratamiento de la normativa y regulación aplicable. Como en cualquier proceso de auditoría, es necesario contar con un contexto adecuado en tiempo y forma que permita que sus resultados no estén desvirtuados por posibles tratamientos realizados en paralelo o entre la finalización y la entrega de hallazgos. Las fases son las habituales en cualquier auditoría.

Fase 1: Planificación

Se observan los principios de auditoría establecidos por normas como la ISO 19011 y el enfoque de gestión de riesgos de ISO/IEC 31000. Dentro del alcance se considerarán aspectos como el tipo de sistema (predictivo, generativo, autónomo), su nivel de riesgo según lo recogido en el RIA, y una valoración sobre el impacto potencial sobre derechos fundamentales.

→ Definir el alcance, objetivos y criterios de auditoría.
→ Identificar los sistemas de IA a auditar (idealmente, seleccionados y segregados por riesgo, impacto o criticidad).
→ Identificar y asignar roles, recursos y abrir calendario.
→ Revisar el marco normativo aplicable (entre otras, RIA, ISO/IEC 42001, RGPD y políticas de la organización).

Fase 2: Recolección de Evidencias

Es recomendable aplicar técnicas de verificación cruzada de fuentes (documental, testimonial, observacional...) para garantizar la fiabilidad de las evidencias. El uso de herramientas como model cards, data sheets for datasets[32] y logs del sistema facilita la trazabilidad.

→ Solicitar documentación técnica, funcional, ética y organizativa.
→ Realizar entrevistas con responsables de desarrollo, cumplimiento y usuarios.
→ Recopilar bitácoras, trazabilidad, informes de validación y registros normativos.

[32] "El Foro Económico Mundial sugiere que todas las entidades deben documentar la procedencia, la creación y el uso de los conjuntos de datos de aprendizaje automático con el fin de evitar resultados erróneos o discriminatorios. Este marco de trabajo propone que todo conjunto de datos debe ser acompañado de una "ficha de datos", llamada datasheet, que consiste en un cuestionario que guía en la documentación de los datos y la reflexión a lo largo del ciclo de vida de los datos." - Documentación de datos: Datasheets for datasets | datos.gob.es

Fase 3: Evaluación por Dimensiones

Apoyada en matrices de cumplimiento normativo, análisis de brechas (gap analysis) y mapas de calor de riesgos, bien empleando Frameworks como AI Risk Assessment Framework de NIST o herramientas como AI Fairness 360 o LIME para explicabilidad o cualquier otra herramienta propia o de terceros.

→ Diseñar una plantilla de auditoría por dimensión.
→ Analizar el cumplimiento normativo y los indicadores clave.
→ Identificar brechas, riesgos y oportunidades de mejora.

Fase 4: Informes de Auditoría

Deberán estructurarse según criterios de materialidad, criticidad y priorización de riesgos dados en la evaluación inicial. Siempre es importante incluir anexos técnicos, mapas de riesgos y recomendaciones clasificadas por impacto y urgencia. El almacenamiento del bruto de la captación de las evidencias y la metodología para su obtención también debería ser tenido en cuenta.

→ Redactar el informe con hallazgos, evidencias y recomendaciones.
→ Clasificar los riesgos (críticos, moderados, menores).
→ Proponer medidas correctivas y de mitigación.
→ Promover la presentación del informe en los comités responsables de IA, así como al Comité de seguridad de la Información u órgano similar.

Fase 5: Seguimiento

Se deben incorporar indicadores que puedan usarse en procesos de mejora continua (KPI), revisiones periódicas y el calendario para auditorías planificadas. Establecer un ciclo de retroalimentación con los equipos técnicos y de cumplimiento, y documentar las acciones correctivas en sistemas de gestión facilitará enormemente la realización del seguimiento y la detección de desviaciones de manera más eficiente.

→ Verificar la implementación de las recomendaciones.
→ Actualizar registros y documentación.
→ Programar auditorías recurrentes o revisiones periódicas

Enfoque Prioritario para la Auditoría de Sistemas de IA

Una forma de priorizar la auditoría es aplicar un enfoque basado en riesgo y criticidad, alineado con prácticas de auditoría interna (COSO, ISO 31000) y con el espíritu del RIA, que clasifica los sistemas según su nivel de riesgo. Este enfoque permite asignar recursos de forma eficiente y centrar la evaluación en los elementos que pueden generar mayor impacto ético, legal, técnico u organizativo:

1. Clasificación del sistema

- ¿Es un sistema de alto riesgo según el RIA (por ejemplo: biometría, decisiones judiciales, selección de RRHH)?
- ¿Afecta derechos fundamentales, procesos críticos o poblaciones vulnerables?

2. Evaluación preliminar de impacto

- ¿Qué consecuencias tendría un fallo técnico, ético o legal?
- ¿Qué nivel de autonomía tiene el sistema en la toma de decisiones?

3. Asignación de profundidad de auditoría

NIVEL DE RIESGO	ENFOQUE DE AUDITORÍA	FRECUENCIA ORIENTATIVA
ALTO	Auditoría completa por dimensiones + trazabilidad extendida	Trimestral / antes de cada despliegue
MEDIO	Auditoría técnica y ética + revisión organizativa	Semestral / tras actualizaciones
BAJO	Revisión funcional + trazabilidad básica	Anual / bajo demanda

4. Criterios de priorización

→ **Impacto sobre personas:** ¿Puede generar discriminación, exclusión o daño?
→ **Complejidad técnica:** ¿Utiliza modelos opacos o difícilmente explicables?
→ **Madurez organizativa:** ¿La organización tiene procesos sólidos de gobernanza?

11.5. **Metodología de implantación o verificación** de la Trazabilidad de sistemas de IA

Es fundamental implantar y gestionar adecuadamente los sistemas de trazabilidad, abordando aspectos técnicos, organizativos, éticos, funcionales y ambientales para facilitar auditorías rigurosas, fomentando los procesos asociados a la mejora continua.

Implantación de la Trazabilidad

Trazabilidad técnica

Objetivo: documentar y seguir el ciclo de vida técnico del sistema de IA, desde el diseño hasta el despliegue.

Acciones clave:
- Registro de versiones de modelos, datasets y código fuente.
- Bitácoras de entrenamiento, validación y ajustes.
- Logs estructurados de inferencias, errores y decisiones automatizadas.
- Uso de herramientas como Model Cards, Data Sheets, Lineage Trackers.

Referencias:
→ ISO/IEC 42001 (controles técnicos y ciclo de vida)
→ RIA Art. 9–15 (requisitos técnicos de sistemas de alto riesgo)
→ RGPD Art. 5, 25 (minimización de datos y privacidad desde el diseño)

Trazabilidad técnica

Objetivo: documentar y seguir el ciclo de vida técnico del sistema de IA, desde el diseño hasta el despliegue.

Acciones clave:
- Registro de versiones de modelos, datasets y código fuente.
- Bitácoras de entrenamiento, validación y ajustes.
- Logs estructurados de inferencias, errores y decisiones automatizadas.
- Uso de herramientas como Model Cards, Data Sheets, Lineage Trackers.

Referencias:
→ ISO/IEC 42001 (controles técnicos y ciclo de vida)
→ RIA Art. 9–15 (requisitos técnicos de sistemas de alto riesgo)
→ RGPD Art. 5, 25 (minimización de datos y privacidad desde el diseño)

Dimensión Ética

Objetivo: asegurar respeto a los derechos fundamentales, equidad y supervisión humana.

Acciones clave:
- Evaluación de impacto ético.
- Revisión de sesgos.
- Mecanismos de supervisión.

Referencias: RIA Art. 14 (supervisión humana), RGPD Art. 22 (decisiones automatizadas), ISO/IEC 42001 (valores organizativos).

Dimensión Organizativa

Objetivo: auditar la gobernanza interna, roles y procesos asociados al sistema.

Acciones clave:
- Documentación de roles, responsabilidades y procesos.
- Coordinación entre equipos y proveedores.
- Revisión de políticas internas y cultura organizativa.

Referencias: RIA Art. 16–29 (obligaciones del proveedor), RGPD Art. 24–32 (responsabilidad organizativa), ISO/IEC 42001 (estructura de gobernanza).

Dimensión de Sostenibilidad y Eficiencia Energética

Objetivo: evaluar el impacto energético y la sostenibilidad del sistema.

Acciones clave:
- Medición de consumo, incluyendo huella de proveedores.
- Coordinación entre equipos y proveedores.
- Revisión de políticas internas y cultura organizativa.

Referencias: RIA Art. 16–29 (obligaciones del proveedor), RGPD Art. 24–32 (responsabilidad organizativa), ISO/IEC 42001 (estructura de gobernanza).

11.6. Procedimiento para aplicar o verificar **la trazabilidad**

La utilización de la trazabilidad en un sistema IA o su valoración en sistemas de terceros, debe verificarse durante la auditoría, al depender de esta varias de las evidencias más importantes. El alineamiento con la metodología descrita en este documento permite integrar la trazabilidad en el ciclo de vida del sistema de IA. Tal y como se ha hecho en el procedimiento de auditoría, se ha previsto su ejecución tanto en entornos propios como híbridos con terceros. Al igual que este, está alineado con los principios del RIA, ISO/IEC 42001 y el RGPD.

Fase 1: Diseño del marco de trazabilidad

- **Objetivo:** Definir la estructura, dimensiones y mecanismos que se aplicarán.
- **Acciones clave:**
  → Seleccionar las dimensiones relevantes (técnica, organizativa, ética, funcional, ambiental).
  → Establecer los tipos de evidencias requeridas por dimensión.
  → Identificar herramientas de registro (registro de eventos, control de versiones, logs, matrices).
  → Alinear el marco con los requisitos normativos aplicables.
- **Referencias:** ISO/IEC 42001 (cláusulas de gobernanza y ciclo de vida), RIA Art. 9–29, RGPD Art. 5, 30.

Fase 2: Integración en el ciclo de vida del sistema

- **Objetivo:** Incorporar los mecanismos de trazabilidad desde el diseño hasta el despliegue.
- **Acciones clave:**
  → Implementar registros estructurados en cada etapa (diseño, entrenamiento, validación, operación).
  → Documentar decisiones técnicas, éticas y organizativas.
  → Establecer puntos de control y validación por dimensión.
  → Coordinar con proveedores externos para asegurar trazabilidad compartida.
- **Referencias:** “Audit-by-design” (AI Now Institute), ISO/IEC 42001, RGPD Art. 25.

Fase 3: Verificación y auditoría de trazabilidad

- **Objetivo:** Evaluar la calidad, completitud y consistencia de los registros generados.
- **Acciones clave:**
  → Aplicar plantillas de auditoría por dimensión.
  → Revisar logs, evidencias y documentación técnica y organizativa.
  → Identificar brechas, incoherencias o riesgos de opacidad.
  → Clasificar el nivel de trazabilidad alcanzado (básico, intermedio, completo).
- **Referencias:** NIST AI Risk Management Framework, RIA Art. 11, 14, 16.

Fase 4: Informe de trazabilidad (Complementario al de auditoría y revisable durante esta)

- **Objetivo:** Consolidar los hallazgos y proponer mejoras.
- **Acciones clave:**
  → Redactar un informe por dimensión con evidencias y recomendaciones.
  → Incluir mapas de trazabilidad, matrices de cumplimiento y anexos técnicos.
  → Presentar el informe a los comités de gobernanza y responsables de IA.
- **Referencias:** ISO 19011 (auditoría), RIA Art. 29 (documentación), RGPD Art. 30.

Fase 5: Seguimiento y mejora continua

- **Objetivo:** Asegurar que la trazabilidad evolucione con el sistema.
- **Acciones clave:**
- Verificar la implementación de recomendaciones.
  → Actualizar registros y mecanismos de trazabilidad.
  → Programar revisiones periódicas y auditorías recurrentes.
  → Establecer indicadores de mejora continua (KPI de trazabilidad).
- **Referencias:** ISO/IEC 42001 (ciclo de mejora), políticas internas ESG.

Marco Normativo 11.7.

La auditoría y trazabilidad de sistemas de IA puede enmarcarse en el conjunto de normas y principios que garanticen su desarrollo y uso responsable.

En el conjunto de países dentro de la Unión Europea, se podrán tomar como referencia el RIA y el RGPD, como bases mínimas ya desarrolladas en esta guía. También hay otras guías de referencia, como la Guía de la Agencia Española de Protección de Datos, AEPD, sobre Auditorías de Tratamientos con IA, que ya es ampliamente utilizada en España y otros países, como México.

Además, existen diversas normas técnicas, con mayor o menos recorrido, que pueden complementar el marco legal, ofreciendo metodologías para implementar auditoría y trazabilidad de forma estructurada, como son la ISO/IEC 42001- Sistema de gestión de IA (AIMS)o la ISO/IEC 23894 - Gestión de riesgos en IA.

Otras normas trasversales también pueden ser utilizadas para fortalecer la auditoría y trazabilidad de sistemas de IA, especialmente cuando se aplican en contextos complejos o regulados. Además, cuentan con la ventaja de una mayor madurez en el mercado. Es el caso de la ISO/IEC 27001 (Seguridad de la Información), la ISO 9001 (Gestión de la Calidad), el COSO (Marco de Control Interno), la IEEE 7000 Series (Estándares Éticos para Sistemas Autónomos) o la ISAE 3000 (Auditoría de Información No Financiera), entre otras.

Gobernanza de modelos generativos y LLMs

12.

La IA se ha consolidado como una tecnología revolucionaria, y la irrupción de la IA generativa y los grandes modelos de lenguaje (LLM) como GPT, Claude AI y Gemini ha marcado un punto de inflexión, ofreciendo una versatilidad y potencia extraordinarias. No obstante, su creciente complejidad y adopción masiva plantean desafíos importantes en cuanto a la interpretabilidad de sus decisiones, la privacidad de los datos, la seguridad, la opacidad y la necesidad de marcos éticos y regulatorios sólidos. Una gobernanza adecuada es crucial para garantizar un desarrollo y uso responsable y fiable de estas tecnologías. Este capítulo ahonda en lo ya visto en el capítulo 5, Gobernanza de la IA, pero profundizando en las especificidades de ese tipo de sistemas

12.1. **Desafíos y Riesgos Específicos** de los Modelos Generativos y LLM

Tal y como se desarrolla en el capítulo sobre La IA como Riesgo Estratégico y Operativo, así como en el capítulo sobre Gestión y Evaluación de Riesgos en IA, los modelos generativos y LLM presentan una serie de riesgos específicos que deben considerarse dentro del marco global de gestión del riesgo de IA.

Entre los riesgos más reconocidos destacan los deepfakes, prompt injection, data poisoning, la pérdida de información sensible, robo de modelos, discriminación algorítmica...

La identificación, análisis y tratamiento de estos riesgos se aborda con detalle en los capítulos dedicados al riesgo en esta guía y en los marcos de referencia y normativos correspondientes (ISO/IEC 27005, NIST AI RMF, y RIA).

Marco Regulatorio y la Gobernanza de LLM 12.2.

Desde la irrupción masiva de las LLM, se han definido diferentes sensibilidades con respecto a la necesidad de regulación. En Europa, destaca el RIA, que cataloga el uso de algunas LLM como de Riesgo Inaceptable (Prácticas de IA Prohibidas) o como Sistemas de IA de Alto Riesgo (se ve más adelante, en esta guía) y establece obligaciones específicas para los GPAI, entendidos como modelos entrenados con un gran volumen de datos utilizando autosupervisión a gran escala, que demuestran un alto grado de generalidad y son capaces de realizar una amplia variedad de tareas. Para aquellos GPAI que la Comisión Europea califique como de riesgo sistémico (en función de criterios tasados, como el número de parámetros, la calidad de los datos, la cantidad de cálculo utilizada para entrenar el modelo, modalidades de entrada/salida, y el número de usuarios finales registrados), se aplicarán requisitos adicionales.

Las obligaciones específicas para los proveedores de GPAI incluyen:

→ Garantizar la **transparencia** de los modelos.
→ Asegurar la **verificabilidad de los datos** utilizados para el entrenamiento.
→ Proporcionar la **explicabilidad** de sus operaciones y decisiones.
→ **Respetar los derechos de autor** y proporcionar resúmenes detallados de los datos de entrenamiento.

La aplicación de estas normas se realizará de manera escalonada, con las reglas de gobernanza y obligaciones para los GPAI aplicables a los doce meses de la entrada en vigor del Reglamento (agosto de 2025). El incumplimiento de estas obligaciones puede acarrear multas significativas en cuantía y/o en porcentaje de facturación global anual.

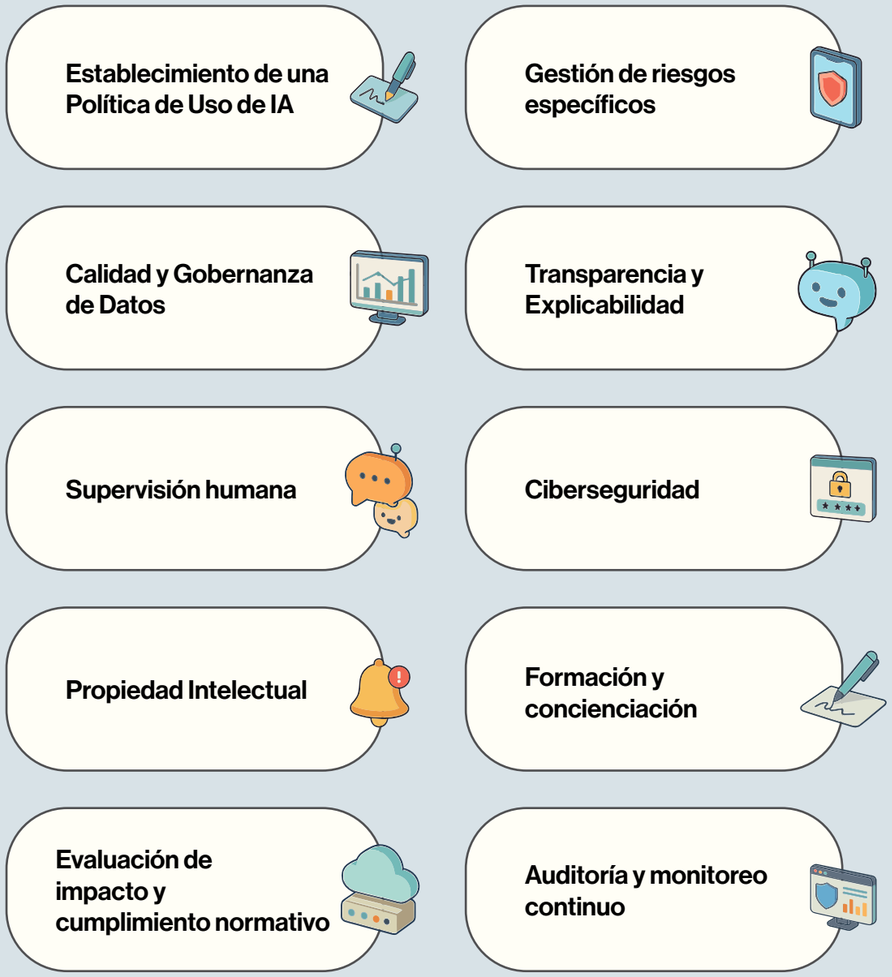
Tal y como se indica en el capítulo 9, la figura que regula y supervisa la implementación del RIA en España es la AESIA.

En este contexto, la gobernanza de los LLM debe ser un proceso integral que involucre a diferentes roles dentro de la organización, como el Chief AI Officer (CAIO), Chief Information Security Officer (CISO), Chief Data Officer (CDO), Data Protection Officer (DPO), y el equipo Legal y de Compliance. La colaboración y coordinación entre estos perfiles es esencial para asegurar el cumplimiento normativo, la gestión de riesgos y el uso ético de la IA.

12.3.

Recomendaciones y Medidas Prácticas para la Gobernanza de LLM

La implementación de un marco de gobernanza sólido para los modelos generativos y LLM es fundamental para mitigar riesgos y aprovechar sus beneficios de manera responsable.



01.

Establecimiento de una Política de Uso de IA:

- La organización debe definir una política interna que regule el uso de sistemas de IA, incluyendo LLM, alineada con la estrategia de la compañía y sus valores éticos. Se deben diferenciar los usos aceptables de los prohibidos.
- Es crucial llevar un registro interno de las soluciones de IA utilizadas, tanto externas como de desarrollo propio.

02.

Gestión de Riesgos Específicos:

- Implementar un sistema de gestión de riesgos para gestionar los riesgos que puedan afectar la seguridad, los derechos fundamentales y el cumplimiento regulatorio durante todo el ciclo de vida del LLM.
- Realizar evaluaciones de riesgo continuas para identificar y tratar riesgos conocidos y previsibles, así como aquellos que puedan surgir del uso indebido razonablemente previsible del sistema.
- Utilizar algún marco de gobierno para comprender y gestionar las amenazas, incluyendo la validez, fiabilidad, seguridad y resiliencia de los modelos.

03.

Calidad y Gobernanza de Datos:

- Garantizar que los conjuntos de datos de entrenamiento, validación y prueba se sometan a prácticas de gobernanza y gestión de datos adecuadas.
- Establecer procedimientos internos para asegurar la exactitud, integridad, fiabilidad, veracidad, actualización y adecuación de los datos, así como mecanismos para analizar, medir y detectar posibles desequilibrios y sesgos.
- Definir el origen de las fuentes de datos, justificar su elección y establecer la base legitimadora para el uso de datos personales, especialmente si se utilizan categorías especiales.
- Implementar técnicas de minimización, seudoanonimización y protección de datos sensibles en todas las fases.

04.

Transparencia y Explicabilidad:

- Los sistemas de LLM deben diseñarse para funcionar con un nivel de transparencia suficiente que permita a los responsables del despliegue interpretar y usar correctamente sus resultados.
- El proveedor debe proporcionar documentación técnica clara y soporte para comprender la lógica, funcionamiento y limitaciones del LLM, así como los principios y métodos utilizados para la toma de decisiones.
- La información dirigida a los usuarios finales debe ser inteligible y de fácil acceso, empleando un lenguaje claro y sencillo, adaptándose a circunstancias particulares, especialmente para colectivos vulnerables.

05.

Supervisión Humana

- El sistema debe permitir una supervisión humana efectiva, lo que implica incluir una interfaz humano-máquina adecuada y mecanismos de intervención que permitan a los operadores detectar y mitigar posibles riesgos o anomalías.
- La supervisión debe ser realizada por personas con la competencia, formación y autoridad necesarias, garantizando una acción significativa y no simbólica en la decisión.

06.

Ciberseguridad:

- Fortalecer las medidas de ciberseguridad para proteger los LLM contra ataques específicos como la inyección de prompt, el envenenamiento de datos, y el robo de modelos.
- Implementar sistemas de registro de eventos (logs) para asegurar la trazabilidad y auditar el funcionamiento de la IA, detectando situaciones de riesgo.
- Realizar auditorías periódicas de seguridad que incluyan el análisis de vulnerabilidades, cumplimiento normativo y la resiliencia del sistema.

07.

Propiedad Intelectual:

- Desarrollar políticas internas sobre el respeto de los derechos de autor para los datos utilizados en el entrenamiento y los contenidos generados por los LLM.
- Definir claramente los derechos sobre los conjuntos de datos del proveedor y de terceros, así como los derechos sobre las creaciones o innovaciones resultantes de la IA.

08.

Formación y Concienciación:

- Capacitar al personal en el uso responsable de los LLM, los principios éticos, la seguridad de la información y la protección de datos.
- Concienciar sobre los riesgos de fuga de información confidencial o personal al interactuar con sistemas de IA públicos.

09.

Evaluación de Impacto y Cumplimiento Normativo:

- Realizar evaluaciones de impacto para los derechos y libertades (EIPD/DPIA) y evaluaciones de impacto del RIA, especialmente para sistemas de LLM de alto riesgo, para identificar y mitigar riesgos potenciales.
- Asegurar que la contratación de LLM de terceros cumpla con un marco contractual robusto que regule condiciones y responsabilidades de las partes, abordando aspectos de privacidad, seguridad, calidad de datos y propiedad intelectual.

10.

Auditoría y Monitoreo Continuo:

- Establecer un sistema de monitorización continua para detectar vulnerabilidades o incidentes de seguridad y asegurar la conformidad con los estándares de privacidad y seguridad.
- Realizar auditorías periódicas de los LLM para verificar su funcionamiento, precisión, sesgos y cumplimiento de políticas.

La gobernanza de los modelos generativos y LLM es un proceso dinámico que exige una adaptación constante a los avances tecnológicos y a la evolución del marco regulatorio. Un enfoque bien planificado y ejecutado proporcionará mayor seguridad jurídica, transparencia y confianza a todas las partes involucradas, promoviendo un uso beneficioso y ético de la IA.

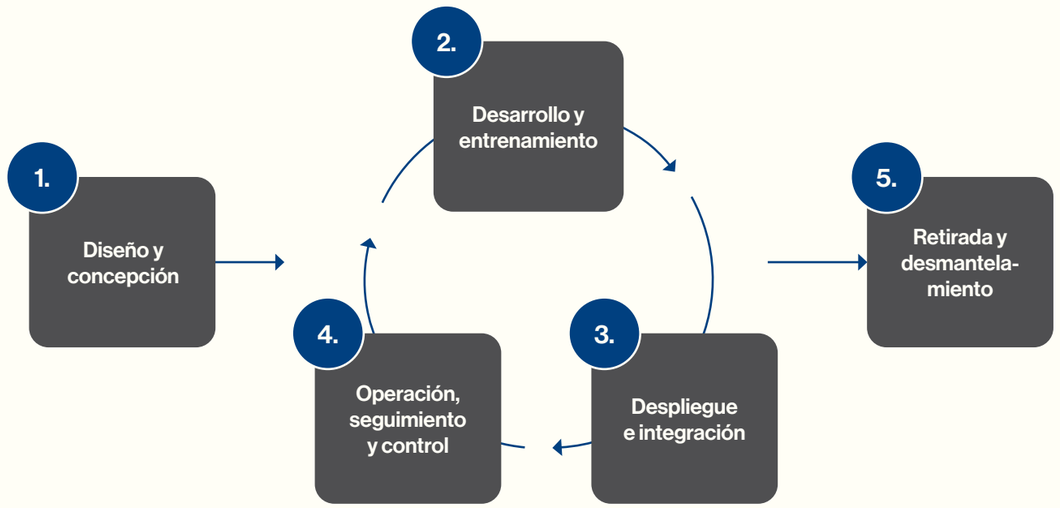
Aquellas empresas que lo deseen también pueden certificarse en la norma internacional ISO/IEC 42001:2023, “Information technology -Artificial intelligence- Management system” que especifica los requisitos para la gestión de IA en las organizaciones a lo largo de todo el ciclo de vida.

Seguridad desde el diseño en sistemas de IA

13.

La gobernanza efectiva de la IA exige que la seguridad no sea un añadido, sino un pilar fundamental integrado en cada etapa del ciclo de vida de los sistemas. Un enfoque de Seguridad desde el Diseño (Security by Design) es indispensable para gestionar proactivamente los riesgos, garantizar la resiliencia y construir una IA confiable. Muchas implementaciones de IA fracasan debido a deficiencias en la gobernanza, la calidad de los datos o la falta de una base tecnológica segura.

A continuación, se detalla el Ciclo de Vida Seguro de Soluciones basadas en IA, detallándose los controles y consideraciones de seguridad críticos a lo largo de las cinco fases clave del ciclo de vida de una solución de IA, desde su concepción hasta su retirada.



Fase 1: Diseño y Concepción

En esta fase se sientan las bases de la seguridad del sistema. Las decisiones tomadas aquí determinarán la robustez y resiliencia de la solución final. El objetivo es anticipar y planificar la mitigación de riesgos antes de escribir una sola línea de código.

→ **Evaluación de Riesgos y Modelado de Amenazas:** identificar los posibles vectores de ataque específicos de la IA, como el envenenamiento de datos, la evasión de modelos o los ataques de inferencia. Utilizar marcos como MITRE ATLAS para estructurar este análisis, como ya expuesto en anteriores capítulos.

→ **Definición de Requisitos de Seguridad y Privacidad:** incorporar la seguridad como un requisito no funcional desde el inicio. Esto incluye aplicar los principios de Privacidad desde el Diseño y por Defecto, como la minimización de datos y la seudonimización, tal como exige el RGPD.

→ **Análisis de Impacto Ético y en Derechos Fundamentales:** evaluar cómo el sistema puede afectar a los derechos de las personas, considerando la equidad, la no discriminación y la autonomía humana, tal y como prevé el RIA, tantas veces mencionado en esta guía.

→ **Selección de Arquitectura Segura:** diseñar una arquitectura que incorpore controles de seguridad nativos, como la segmentación de redes, el cifrado de datos en reposo y en tránsito, y una gestión robusta de identidades y accesos.

Fase 2: Desarrollo y Entrenamiento

La fase de desarrollo es donde los datos se convierten en el combustible del modelo. La seguridad en esta etapa se centra en proteger la integridad de los datos y del propio modelo durante su construcción.

→ **Gobernanza y Seguridad de la Cadena de Suministro de Datos (DataOps):** asegurar la calidad, integridad y procedencia de los datos de entrenamiento, validación y prueba. Es crucial documentar el linaje de los datos para garantizar la trazabilidad y verificar que su uso es lícito y respeta la propiedad intelectual.

→ **Protección contra el Envenenamiento de Datos:** implementar controles para detectar y prevenir la inyección de datos maliciosos en los conjuntos de entrenamiento, lo que podría corromper el comportamiento del modelo.

→ **Entorno de Desarrollo Seguro**[33]**:** utilizar entornos de desarrollo y experimentación aislados para evitar que las pruebas afecten a los sistemas de producción. Gestionar de forma segura las dependencias de software y las librerías de código abierto para evitar la introducción de vulnerabilidades.

→ **Documentación Continua:** mantener una documentación técnica exhaustiva, incluyendo datasheets para los conjuntos de datos y model cards (anexo 3) para los modelos, que describan su funcionamiento, limitaciones y métricas de rendimiento.

[33] ISMS Forum. (2024). II Edición de la Guía de DevSecOps. https://www.ismsforum.es/ficheros/descargas/iiedicionguiadevsecopsv1746557372.pdf

Fase 3: Despliegue e Integración

El despliegue es el proceso de integrar el modelo de IA en un entorno de producción para su uso práctico. La seguridad en esta fase es importante para proteger el sistema en un entorno operativo real.

→ **Infraestructura Segura:** asegurar que la infraestructura subyacente (servidores, contenedores, plataformas en la nube) esté correctamente configurada y securizada. Siguiendo guías de hardening, por ejemplo, las publicadas en NIST para cada sistema y aplicando principios de mínimo privilegio.

→ **Protección de APIs y Endpoints:** las interfaces de programación de aplicaciones (API) a través de las cuales se consume el modelo son un vector de ataque principal. Deben protegerse contra ataques de inyección (prompt injection), abuso y acceso no autorizado.

→ **Gestión Segura de Salidas (Insecure Output Handling):** validar y sanear siempre las salidas del modelo antes de que sean procesadas por otros sistemas o mostradas a los usuarios, para prevenir ataques como Cross-Site Scripting (XSS) si el modelo genera código.

→ **Pruebas de Seguridad Pre-Producción:** realizar pruebas de penetración y auditorías de seguridad específicas para la solución de IA antes de su lanzamiento, simulando ataques adversarios para identificar debilidades.

Fase 4: Operación, Seguimiento y Control

Una vez desplegado, un sistema de IA requiere una supervisión continua para garantizar que su comportamiento sigue siendo seguro, fiable y alineado con los objetivos previstos.

→ **Monitorización Continua del Modelo:** supervisar métricas de rendimiento, precisión y equidad para detectar degradaciones del modelo (model drift o concept drift) que puedan introducir riesgos o sesgos.

→ **Registro y Trazabilidad (Logging):** implementar un sistema de registro robusto que capture eventos clave como las decisiones tomadas, las intervenciones humanas y las alertas de seguridad. Estos registros son fundamentales para la auditoría, la rendición de cuentas y el análisis forense.

→ **Supervisión Humana Efectiva:** habilitar mecanismos que permitan una intervención humana significativa, especialmente en sistemas de alto riesgo. Los operadores deben tener la capacidad y la autoridad para anular o corregir las decisiones del sistema cuando sea necesario. (anexo 4)

→ **Plan de Respuesta a Incidentes:** disponer de un protocolo específico para gestionar incidentes de seguridad relacionados con la IA, que defina los pasos para la detección, contención, erradicación y recuperación.

Fase 5: Retirada y Desmantelamiento

El ciclo de vida de un sistema de IA concluye con su retirada. Esta fase debe gestionarse de forma segura para evitar la exposición de datos sensibles y propiedad intelectual.

→ **Archivo Seguro del Modelo y los Datos:** archivar de forma segura los modelos, los datos de entrenamiento y la documentación asociada, de acuerdo con las políticas de retención de la organización y los requisitos regulatorios.

→ **Eliminación Segura de Datos:** eliminar de forma segura los datos personales y confidenciales de los sistemas de producción, de acuerdo con el principio de limitación del plazo de conservación del RGPD.

→ **Revocación de Accesos:** deshabilitar todas las API, cuentas de servicio y credenciales asociadas al sistema retirado para prevenir accesos no autorizados.

→ **Documentación del Proceso:** registrar todas las acciones realizadas durante la fase de retirada para garantizar una trazabilidad completa del fin de vida del sistema. (anexo 5)

Conclusiones y recomendaciones finales

14.

La Inteligencia Artificial (IA) se ha consolidado como un motor de transformación profunda en las organizaciones, pero también como un vector de riesgo estratégico, operativo, ético y regulatorio. El recorrido de esta guía evidencia que la gobernanza de la IA ya no es una opción, sino un imperativo legal, competitivo y social. El Reglamento Europeo de IA (RIA) marca un antes y un después, exigiendo pasar de los principios y la ética declarativa a la operacionalización efectiva y auditable de la IA en todas sus fases.

De la ética a la obligación legal: **el RIA como catalizador**

La entrada en vigor del RIA convierte la gestión de riesgos y la gobernanza de la IA en una obligación estructurada, con plazos y sanciones equiparables al RGPD. Las organizaciones deben identificar, clasificar y gestionar sus sistemas de IA según el nivel de riesgo, anticipando las exigencias regulatorias y evitando sanciones de cuantía muy elevada. El cumplimiento ya no es solo una cuestión de reputación, sino de viabilidad y sostenibilidad del negocio.

Como muestra de que todavía queda terreno por recorrer en este aspecto, según el último AI Readiness Index 2025 de Cisco , solo el:

23%

de las empresas encuestadas tienen sus procesos de Gobernanza preparados para hacer frente a los nuevos retos de la IA.

Nuevos vectores de riesgo y la necesidad de **marcos y contramedidas especializadas**

La irrupción de la IA generativa y los grandes modelos de lenguaje (LLM) han multiplicado los riesgos: ataques novedosos (prompt injection, data poisoning, model theft), generación de desinformación, pérdida de control sobre los datos y desafíos inéditos en propiedad intelectual.

Para hacer frente a estos riesgos no podemos contar con las herramientas que teníamos hasta ahora: necesitamos un marco gobernanza específico; así como contramedidas y modelos de seguridad especializados siguiendo una adaptación constante a la evolución tecnológica y regulatoria.

Gobernanza por diseño: de la reacción a la proactividad

La gestión del riesgo de la IA debe ser proactiva e integrada. El concepto de AI Governance by Design (AIGD) implica incorporar desde el inicio consideraciones éticas, legales, de seguridad y sociales en el ciclo de vida de la IA. Esto exige madurez organizativa, estructuras de GRC robustas, y la definición clara de roles como CAIO, CISO, CDO y DPO, así como la creación de comités multidisciplinares de IA.

Operacionalización: de la estrategia a la práctica

La clave del éxito reside en traducir los marcos y principios en políticas, procesos y controles efectivos. Esto implica:

- → Formalizar una estructura de gobierno de IA, con inventario centralizado de sistemas, políticas internas, responsabilidades definidas y procedimientos claros.
- → Desarrollar una estrategia de seguridad de IA robusta, basada en Zero Trust, segmentación, monitorización continua y respuesta ágil ante incidentes.
- → Implementar trazabilidad y auditoría en todos los sistemas, garantizando la rendición de cuentas y la capacidad de demostrar cumplimiento.
- → Robustecer la gestión de terceros, exigiendo garantías contractuales de privacidad, seguridad, explicabilidad y cumplimiento normativo.
- → Fomentar una cultura organizacional consciente, con formación específica y transversal en riesgos, ética y buenas prácticas de IA.

Solamente si somos capaces de establecer estas políticas, procesos y controles, podremos medir de forma efectiva el impacto de la IA en nuestras organizaciones, algo que, según el citado estudio AI Readiness Index de Cisco, sólo están en disposición de hacer el 32% de las compañías.

Protección de **datos, derechos y sostenibilidad**

La protección de datos personales y la privacidad deben integrarse desde el diseño, alineando el cumplimiento del RGPD con las obligaciones del RIA. La evaluación de impacto (EIPD/DPIA y FRIA) y la minimización de datos son esenciales para anticipar y mitigar riesgos. Además, la sostenibilidad y el impacto ambiental de la IA deben ser considerados en la toma de decisiones, alineando la innovación con los objetivos de desarrollo sostenible.

Siguiendo con los datos presentados en Cisco AI Readiness Index, sólo el 34% de las organizaciones tienen integrados los nuevos riesgos de la IA dentro sus políticas de protección de datos; siendo este el campo donde más se nota el grado de madurez en la adopción de la IA.

93%

de las empresas que presentan un grado avanzado de adopción de la IA en su negocio, tienen ya integrados los nuevos riesgos de la IA dentro sus políticas de protección de datos

Supervisión humana, explicabilidad y confianza

La supervisión humana efectiva, la transparencia y la explicabilidad son pilares para generar confianza y garantizar el control sobre los sistemas de IA, especialmente en contextos de alto riesgo. La trazabilidad organizativa y técnica, junto con la documentación exhaustiva, permiten auditar y justificar decisiones, facilitando la rendición de cuentas ante usuarios, reguladores y la sociedad.

Adaptación sectorial y tamaño organizativo

Las conclusiones y recomendaciones deben adaptarse al sector y tamaño de cada organización. Sectores críticos como salud, finanzas o infraestructuras requieren especial atención, mientras que las pymes deben priorizar la visibilidad y el control sobre el uso de IA, especialmente en la gestión de proveedores y herramientas externas.

Hacia una IA confiable, ética y competitiva

La IA solo generará valor sostenible si se desarrolla y utiliza dentro de un marco de confianza, legalidad y responsabilidad. La gobernanza efectiva de la IA es la mejor garantía para aprovechar su potencial, mitigar sus riesgos y fortalecer la competitividad y la reputación de las organizaciones.

Recomendaciones finales

Integrar la gobernanza desde el diseño

Adoptar el enfoque AI Governance by Design, incorporando la ética, la seguridad y la protección de datos desde la concepción de cada sistema.

Anticipar y cumplir el RIA

Iniciar cuanto antes la adaptación al Reglamento Europeo de IA, identificando sistemas, roles y obligaciones

Fortalecer la ciberseguridad y la gestión de riesgos

Aplicando marcos especializados y controles específicos para IA generativa, LLM e IA Agéntica; avanzando en la madurez de la automatización sin perder la supervisión humana. No en vano, hoy día sólo el 34% de las organizaciones confía en la resiliencia de su infraestructura actual de ciberseguridad para hacer frente a los riesgos de la IA

Garantizar la trazabilidad y la auditoría

Mantener registros completos y accesibles que permitan demostrar el cumplimiento y facilitar la supervisión.

Impulsar la formación y la cultura organizativa

Capacitar a todos los perfiles implicados en IA, desde la dirección hasta los usuarios finales, para reducir la brecha de confianza y mejorar la gestión de riesgos.

Revisar y adaptar contratos con terceros

Exigir garantías de cumplimiento, seguridad y explicabilidad a proveedores de IA.

Evaluar el impacto social, ético y ambiental

Incorporar la sostenibilidad y la protección de derechos en la toma de decisiones sobre IA.

En definitiva, la gobernanza de la IA es un proceso dinámico y transversal, que exige liderazgo, visión estratégica y capacidad de adaptación. Las organizaciones que integren estos principios estarán mejor preparadas para afrontar los retos y aprovechar las oportunidades de la nueva era digital.

En este contexto, resulta imprescindible reforzar los planes de formación continua y reskilling del talento interno, dotando a los equipos de competencias avanzadas en el uso responsable, seguro y estratégico de herramientas de IA. Esta capacitación no solo reduce la brecha tecnológica, sino que actúa como catalizador de una adopción ética, eficiente y alineada con los objetivos de negocio.

De hecho, terminando con las lecciones clave del Cisco AI Readiness Index, aquellas empresas que presentan mayor madurez en la gobernanza de la IA son más susceptibles de aprovechar el potencial de la misma en estos tres frentes:

1. **Identificando los casos de uso** en su negocio para la aplicación de la IA.
2. **Midiendo el impacto económico** de sus inversiones en IA.
3. **Reportando beneficios en rentabilidad, productividad e innovación** como resultado de la aplicación de la IA a sus negocios.

ANEXOS

Guías y recursos prácticos

Las siguientes guías y recursos prácticos se encuentran disponibles para su descarga en formato editable en el siguiente enlace: https://www.ismsforum.es/ficheros/descargas/descargables-gobierno-de-ia1762519105.docx

Anexo 1

Casos de Uso y Desafíos Comunes de Coordinación

Caso de uso 1:
Implementación de un modelo generativo para atención al cliente

Desafío: Validar el cumplimiento normativo, proteger datos personales y garantizar seguridad.

Coordinación:

- CAIO lidera; CISO valida arquitectura segura
- DPO revisa implicaciones de privacidad
- Legal revisa licencias de uso
- CDO asegura calidad del dato.

Caso de uso 2:
Automatización de decisiones de crédito

Desafío: Asegurar la trazabilidad de decisiones automatizadas y evitar sesgos algorítmicos.

Coordinación:

- **CAIO** y **CDO** colaboran en la transparencia de modelos
- **Legal** y **DPO** evalúan cumplimiento de requisitos de explicabilidad

La explicabilidad (o explicabilidad algorítmica) es la capacidad de un sistema de inteligencia artificial para ofrecer razones comprensibles sobre cómo ha llegado a una decisión. Esto es especialmente importante cuando la IA toma decisiones que afectan a personas, como:

- **Aprobar o denegar un crédito**
- **Seleccionar candidatos para un empleo**
- **Determinar el precio de un seguro**

Caso de uso 3:
Integración de IA en procesos internos de RRHH

Desafío: Riesgos reputacionales por sesgos o decisiones discriminatorias.

Coordinación:

- CAIO y CIO supervisan arquitectura
- DPO y Legal evalúan compatibilidad con principios de igualdad y no discriminación

Anexo 2

Preguntas clave para la Auditoría

Seguridad en el ciclo del sistema

- **Observación del principio de "Seguridad en el diseño":**
  → ¿Existe un procedimiento formal que incluya la trazabilidad dentro de los requisitos de seguridad en el diseño de todos los elementos que conforman el sistema?
- **Control de versiones y cambios:**
  → ¿Se documentan y registran las actualizaciones del modelo?
- **Gestión de dependencias y librerías:**
  → ¿Se auditan los componentes externos que pueden introducir vulnerabilidades?
- **Registro de decisiones humanas:**
  → ¿Se trazan las intervenciones humanas en el diseño y ajuste del sistema?

Seguridad de los Datos

- **Origen y calidad de los datos:**
  → ¿Provienen estos de fuentes confiables?
  → ¿Se han anonimizado o tokenizado correctamente?
- **Protección de datos sensibles:**
  → ¿Se aplican medidas para el cifrado o enmascaramiento?
  → ¿Se ha valorado el uso de datos sintéticos en entornos no productivos?
  → ¿Cuentan con un correcto control de acceso?
- **Trazabilidad de datos:**
  → ¿Se puede reconstruir el flujo de datos desde su obtención en repositorios, BBDD y otros sistemas de IA hasta su consumo por parte del área usuaria?

Robustez del Modelo

- **Resistencia a entradas adversarias:**
  → ¿El modelo puede ser manipulado con de forma maliciosa?

- **Evaluación de ataques tipo "model extraction" o "data poisoning":**
  → ¿Hay mecanismos para detectar intentos de replicar o robar el modelo de IA o manipular maliciosamente los datos de entrenamiento de este con el objetivo de introducir errores, sesgos o vulnerabilidades?
- **Pruebas de estrés y simulaciones:**
  → ¿Cómo responde el sistema ante condiciones extremas o inesperadas?
  → ¿Las excepciones y errores cuentan con un manejo que permita elevarlos a eventos de seguridad si fuera necesario?

Seguridad de plataformas hardware y software

- **Infraestructura física:**
  → ¿Qué medidas de seguridad se han considerado para la protección de servidores, centros de datos, dispositivos Edge, etc.?
  → ¿Cómo se entrega la información relativa a estas medidas?
- **Entorno de ejecución:**
  → ¿Qué estrategia de reporte se ha considerado para recoger toda la información relativa a la protección la seguridad de la plataforma base, el sistema operativo, contenedores, virtualización, comunicaciones, etc.?
- **Dependencias y librerías:**
  → ¿Están activos los mecanismos de auditoría de componentes internos y externos?
  → ¿Se monitorizan los mecanismos empleados para verificar la integridad de los componentes?
  → ¿Cómo se gestiona la configuración de los sistemas?
- **Gestión de parches y actualizaciones:**
  → ¿Existe y se gestiona correctamente el control de versiones?
  → ¿Hay un plan para la gestión y corrección de vulnerabilidades?
- **Seguridad en APIs y canales de comunicación:**
  → ¿Existen mecanismos para la protección contra inyecciones, accesos no autorizados y fugas de información?
  → ¿Se utilizan medios seguros y cifrados para el acceso a los sistemas y a las aplicaciones?
  → ¿Ambos están cubiertos por el nivel suficiente de traza que permita depurar cualquier evento?

Supervisión y Monitoreo en Producción

- **Auditoría en tiempo real:**
  → ¿Se registran las interacciones y decisiones del sistema en producción?
- **Alertas y mecanismos de respuesta:**
  → ¿Existe un protocolo ante comportamientos anómalos o violaciones de seguridad?
- **Evaluación de uso indebido:**
  → ¿Se detectan intentos de manipulación por parte de usuarios o agentes externos?

Gestión de autenticación y autorización de identidades y accesos

- **Autenticación de usuarios, agentes o entidades que interactúan con el sistema de IA:**
  → ¿Hay mecanismos robustos de autenticación (MFA, certificados, tokens)?
  → ¿La autorización está basada en roles (RBAC), atributos (ABAC) o políticas dinámicas?
- **Controles para prevenir, detectar y responder ante el uso indebido de identidades privilegiadas o impersonación dentro del sistema.**
  → ¿Se audita el uso de cuentas de servicio, identidades de máquinas?
  → ¿Existe un proceso de trazabilidad que asegure la identificación del responsable en los accesos administrativos o automatizados?
- **Control de acceso a modelos y datos:**
  → ¿Quién puede modificar, consultar o entrenar el sistema?
- **Seguridad en interfaces de interacción (APIs, prompts, etc.):**
  → ¿Se protegen contra inyecciones maliciosas o abusos?
- **Registro de accesos y acciones:**
  → ¿Se auditan los accesos administrativos y técnicos?
  → ¿Las trazas y registros pueden incorporarse a una investigación forense en caso necesario?

Anexo 3

Ejemplo de "Model Card" para un sistema de IA ficticio de evaluación de solvencia crediticia

Información general	
Campo	**Descripción**
Nombre del Modelo	**Scoring de Solvencia Crediticia (SC-2025-v1)**
Versión del Modelo	**1.0 (Lanzamiento: Octubre 2025)**
Tipo de Modelo	**Aprendizaje Supervisado (Clasificación binaria)**
Algoritmo	**Gradient Boosting (XGBoost)**
Propietario Interno	**Departamento de Riesgos**
Contacto Técnico	**Ejemplo@ISMSForum.es**
Clasificación de Riesgo	**Alto Riesgo (según el Reglamento de IA de la UE, Anexo III)**

01

Descripción General y Finalidad del Modelo

Este modelo está diseñado para evaluar la probabilidad de que un solicitante de crédito (persona física) incurra en impago en los próximos 24 meses. El sistema genera una puntuación de solvencia (scoring) que sirve como un elemento de apoyo en el proceso de toma de decisiones para la concesión de créditos al consumo.

Uso Previsto:

→ Evaluar solicitudes de nuevos créditos al consumo.

→ Servir como una de las herramientas de apoyo para los analistas de riesgo en la toma de decisiones. No está diseñado para tomar decisiones de forma 100% autónoma.

→ Segmentar a los solicitantes en perfiles de riesgo (bajo, medio, alto) para determinar las condiciones del crédito.

Usos No Previstos (Prohibidos):

→ Tomar decisiones de denegación de crédito de manera totalmente automatizada sin posibilidad de revisión humana.

→ Utilizar la puntuación para cualquier otra finalidad distinta a la evaluación de solvencia (ej. marketing, evaluación de empleados).

→ Aplicar el modelo a solicitantes que no sean personas físicas.

02

Datos Utilizados

La calidad, gobernanza e integridad de los datos son críticas para el rendimiento y la equidad del modelo, como se subraya en los principios de gobernanza de datos y privacidad.

Datos de Entrenamiento:

- **Fuente:** Datos históricos internos anonimizados de operaciones de crédito entre 2018 y 2023. No se utilizan fuentes de datos externas para el entrenamiento.
- **Volumen:** 150.000 registros de clientes.
- **Variables Utilizadas (Features)**:
  - → **Información Sociodemográfica:** rango de edad, situación laboral (asalariado, autónomo, etc.), antigüedad laboral. Nota: No se utilizan datos de categoría especial como origen étnico o género para evitar sesgos discriminatorios.
  - → **Información Financiera:** nivel de ingresos netos mensuales, nivel de endeudamiento actual, historial de pagos de créditos anteriores, tipo de contrato laboral.
  - → **Información de la Operación: i**mporte del crédito solicitado, plazo de devolución.

- **Preprocesamiento de Datos:**
  - → Se aplicaron técnicas de **seudonimización** para proteger la identidad de los individuos.
  - → Se realizó un análisis y mitigación de sesgos en los datos de entrenamiento para asegurar la representatividad de diferentes grupos demográficos.

Datos de Evaluación (Test):

- **Fuente:** conjunto de datos independiente (20% del total) extraído de la misma fuente histórica (hold-out dataset), no utilizado durante el entrenamiento para garantizar una evaluación objetiva del rendimiento.

03 Métricas de Rendimiento

Las métricas se han seleccionado para evaluar no solo la precisión global del modelo, sino también su fiabilidad y equidad.

Métricas de Precisión:

Métrica	Valor	Descripción
Accuracy (Precisión Global)	**92.5%**	**Porcentaje de predicciones correctas sobre el total.**
Precision (Precisión)	**88.0%**	**De todos los casos que el modelo predijo como "impago", el 88% realmente lo fueron.**
Recall (Sensibilidad)	**85.0%**	**El modelo identificó correctamente al 85% de todos los casos de "impago" reales.**
AUC-ROC	**0.94**	**Mide la capacidad del modelo para distinguir entre las clases (solvente vs. impago). Un valor cercano a 1 indica un excelente rendimiento.**

Métricas de Equidad y Sesgo: se realizaron análisis para medir el impacto dispar entre diferentes grupos demográficos (segmentados por rango de edad y situación laboral).

Métrica de Equidad	Resultado	Conclusión
Paridad Demográfica	**Diferencia < 2% en la tasa de predicción de "impago" entre grupos.**	**No se detectó un impacto adverso significativo.**
Igualdad de Oportunidades	**Diferencia < 3% en la tasa de verdaderos positivos entre grupos.**	**El modelo identifica a los solicitantes con riesgo de impago de manera equitativa entre los grupos analizados.**

04 Limitaciones y Consideraciones Éticas

La transparencia sobre las limitaciones del modelo es un requisito fundamental para un uso responsable.

- **Correlación no implica Causalidad:** el modelo identifica patrones y correlaciones en los datos históricos, pero no explica las causas subyacentes del comportamiento financiero de una persona. Las decisiones finales deben ser contextualizadas por un analista humano.
- **Riesgo de "Drift":** el rendimiento del modelo puede degradarse con el tiempo si las condiciones económicas y sociales cambian (concept drift). Se ha implementado un plan de monitorización continua para detectar estas desviaciones y planificar el reentrenamiento del modelo cuando sea necesario.
- **Supervisión Humana:** este sistema está clasificado como de Alto Riesgo y, por tanto, todas las decisiones automatizadas que tengan un impacto significativo (como una pre-denegación) deben ser revisadas por un analista cualificado antes de ser comunicadas al cliente. Se han habilitado mecanismos para que los analistas puedan anular o modificar la decisión del sistema.
- **Explicabilidad:** aunque el modelo XGBoost es complejo, se han implementado técnicas de explicabilidad (como SHAP - SHapley Additive exPlanations) para generar una justificación comprensible de cada predicción individual, que estará disponible para los analistas internos y para el cliente si lo solicita, en cumplimiento del derecho a obtener una explicación sobre las decisiones automatizadas.

05 Consideraciones de Seguridad y Robustez

El sistema ha sido desarrollado siguiendo principios de seguridad desde el diseño para garantizar su robustez técnica y resiliencia frente a ataques.

- → **Robustez:** se han realizado pruebas de estrés con datos anómalos y escenarios adversarios para evaluar la estabilidad del modelo.
- → **Seguridad:** la infraestructura que aloja el modelo cumple con las políticas internas de ciberseguridad, incluyendo controles de acceso estrictos, cifrado de datos y monitorización de la actividad de la API.
- → **Trazabilidad:** todas las predicciones generadas por el modelo son registradas en un sistema de logs inmutable, que incluye la versión del modelo utilizada, los datos de entrada (seudonimizados) y la puntuación resultante. Esto garantiza la auditabilidad y la rendición de cuentas.

Anexo 4

Formulario de Registro de Supervisión Humana Efectiva
sobre un sistema ficticio

Este formulario debe ser completado por un analista cada vez que se revisa una decisión generada por un sistema de IA de Alto Riesgo. Su propósito es crear un registro auditable que demuestre una intervención humana significativa y no meramente simbólica.

Registro de Supervisión Humana Efectiva	
Sección 1: Identificación del Caso	
ID del Caso:	**SOL-20251027-8A34F**
ID del Modelo IA:	**IA-002 (SC-2025-v1)**
Fecha/Hora de Decisión del Modelo:	**27/10/2025 - 11:34 AM**
Analista Revisor (Nombre e ID):	**XXXX**
Fecha/Hora de Revisión Humana:	**27/10/2025 - 02:15 PM**
Motivo de la Revisión:	☑ **Protocolo Estándar (Decisión de Alto Impacto)** ☐ **Solicitud de Impugnación del Cliente** ☐ **Alerta de Monitorización del Modelo** ☐ **Otro (especificar):**

Registro de Supervisión Humana Efectiva	
Sección 2: Resumen de la Decisión Automatizada	
Resultado del Modelo	**Justificación (Factores Principales reportados por la IA)**
Decisión Sugerida: **Denegar Crédito Puntuación: 350 / 1000 Confianza del Modelo: 91%**	**1. (Impacto Alto) Antigüedad laboral: 8 meses.** **2. (Impacto Medio) Ratio de endeudamiento: 45%.** **3. (Impacto Bajo) Ausencia de historial crediticio interno.**

Registro de Supervisión Humana Efectiva		
Sección 3: Proceso de Verificación Humana		
Acción	**Verificado**	**Observaciones del Analista**
1. Verificación de Datos de Entrada	☑	**Los datos en el sistema coinciden con la documentación del solicitante.**
2. Análisis de la Lógica del Modelo	☑	**La justificación del modelo es coherente con las políticas de riesgo.**
3. Recopilación de Contexto Adicional	☑	**Se ha contactado al solicitante y revisado documentación complementaria.**
4. Hallazgos y Contexto Adicional (información no considerada por el modelo):	• **El solicitante ha presentado un contrato de trabajo indefinido en un sector de alta empleabilidad (Ingeniería de Software), lo que reduce el riesgo asociado a la "baja antigüedad".** • **El endeudamiento preexistente corresponde a un préstamo estudiantil, que tiene un perfil de riesgo diferente a un crédito al consumo.**	

Registro de Supervisión Humana Efectiva	
Sección 4: Decisión Final del Analista y Justificación	
Decisión Final:	☐ **Mantener la decisión del modelo** ☑ **Anular (Override) la decisión del modelo** ☐ **Modificar la decisión del modelo**
Acción Específica Tomada	**Aprobar el crédito con modificación de condiciones: se aprueba un 80% del importe solicitado, con un incremento de 0.5 puntos en el tipo de interés para reflejar el riesgo residual**
Justificación Obligatoria de la Decisión Humana	**La decisión del modelo, aunque técnicamente correcta con los datos disponibles, no considera factores cualitativos clave. La estabilidad laboral confirmada por el contrato indefinido y la naturaleza del endeudamiento previo (educativo vs. consumo) reducen sustancialmente el perfil de riesgo real del solicitante. La anulación de la recomendación automática y la consiguiente aprobación parcial se basan en un juicio profesional que integra este contexto adicional, alineándose con una gestión prudente y personalizada del riesgo.**

Anexo 5

Formulario de Registro de Retirada de un Sistema de IA
sobre un sistema ficticio

Este formulario se utiliza para documentar y verificar todas las acciones ejecutadas durante el proceso de retirada (desmantelamiento o decommissioning) de un sistema de IA. Su objetivo es asegurar que la retirada se realiza de forma segura, controlada y cumpliendo con las políticas internas y la normativa aplicable, garantizando una trazabilidad completa del fin de vida del sistema.

Registro de Retirada de Sistema de IA			
Control de documento			
ID de Registro:	**RET-20260115-IA002**	**Fecha de Inicio:**	**15/01/2026**
Responsable del Proceso:	**xxx**	**Fecha de Finalización:**	**20/01/2026**
Estado:	**COMPLETADO**		

Identificación del Sistema y Motivo de la Retirada	
Campo	**Descripción**
ID del Modelo en Inventario:	**IA-002**
Nombre del Modelo:	**Scoring de Solvencia Crediticia (SC-2025-v1)**
Fecha de Puesta en Producción:	
Justificación de la Retirada:	**Sustitución por una nueva versión. El modelo SC-2025-v1 será reemplazado por la versión SC-2026-v2, que ha demostrado un rendimiento superior (mejora del 4% en AUC-ROC) y una mayor equidad en las pruebas de validación. La retirada del modelo v1 es necesaria para evitar la coexistencia de versiones y asegurar la consistencia en las decisiones de riesgo.**
Autorización de Retirada:	**Aprobado por el Comité de Gobernanza de IA en la sesión del 10/01/2026 (Acta CGIA-2026-01).**

Registro de Retirada de Sistema de IA					
Categoría	**Acción Específica**	**Responsable**	**Fecha**	**Verificado**	**Observaciones / Ruta de Evidencia**
1. Desactivación del Servicio	Deshabilitar el endpoint de la API del modelo en el entorno de producción.	Equipo de MLOps		Sí	El endpoint api.ejemplo.com/scoring/v1 ya no está activo. Verificado mediante pruebas de conexión.
	Redirigir el tráfico del antiguo endpoint a un mensaje de "servicio obsoleto".	Equipo de Infraestructura		Sí	Configuración actualizada en el balanceador de carga.
2. Archivo de Activos	Archivar el código fuente final del modelo (versión 1.0) en el repositorio de archivado.	Equipo de Ciencia de Datos		Sí	Repositorio: git-archive/models/SC-2025-v1. Commit final: f4a2b1c.
	Archivar los datasets de entrenamiento y validación (anonimizados) utilizados.	Equipo de Datos		Sí	Ubicación: data-archive/datasets/credit-scoring/2025_v1. Acceso restringido.
	Archivar la Ficha de Modelo (Model Card) y toda la documentación técnica asociada.	Equipo de Gobernanza		Sí	Documentación archivada en el gestor documental, ref: DOC-IA-002-FINAL.
3. Eliminación de Datos	Eliminar de forma segura los datos de entrenamiento y el modelo de los servidores de producción.	Equipo de MLOps		Sí	Proceso de borrado seguro ejecutado en los servidores de producción prod-ml-01 y prod-ml-02. Certificado de borrado adjunto.
	Purgar los logs de ejecución del modelo que contengan datos personales, de acuerdo con la política de retención (conservar únicamente metadatos agregados para análisis históricos).	Equipo de Seguridad		Sí	Script de purga ejecutado. Los logs de más de 30 días han sido eliminados.
4. Gestión de Accesos	Revocar todas las claves de API y credenciales de servicio asociadas al modelo v1.	Equipo de Seguridad		Sí	Claves de API API_KEY_SC_V1_... revocadas.
	Eliminar los roles de acceso específicos para la administración del modelo v1 en el sistema de gestión de identidades.	Equipo de TI		Sí	Roles Admin-SC-v1 y User-SC-v1 eliminados.
5. Comunicación y Cierre	Notificar a todos los equipos de desarrollo y negocio que utilizaban el modelo v1 sobre su retirada definitiva.	Propietario del Modelo		Sí	Correo electrónico enviado a las listas de distribución dev-teams y risk-analysts.
	Actualizar el estado del modelo a "Retirado" en el Inventario de Sistemas de IA.	Equipo de Gobernanza		Sí	Inventario actualizado. Ref: IA-002.

Verificación Final y Aprobación de la Retirada

Mediante la firma de este documento, los responsables abajo firmantes certifican que todas las acciones descritas en la Sección 2 han sido ejecutadas de acuerdo con los procedimientos establecidos, y que el sistema de IA SC-2025-v1 ha sido retirado de forma segura y completa.

Responsable del Proceso de Retirada:

- Nombre: XXXX
- Cargo: Jefe de MLOps
- Firma: XXXX
- Fecha: XXXX

Aprobación Final de Gobernanza:

- Nombre: XXX
- Cargo: Directora de Gobernanza de Datos y IA
- Firma: XXXX
- Fecha: XXXX

Anexo 6

Preguntas clave para la **evaluación de proveedores**

La dependencia de proveedores externos para soluciones de IA (más del 65% de las organizaciones según la encuesta de 2023) exige un proceso de due diligence robusto. Este checklist ayuda a evaluar y gestionar el riesgo asociado a terceros.

Registro de Retirada de Sistema de IA		
Categoría	**Aspecto a Evaluar**	**Preguntas Clave y Evidencias Requeridas**
1. Cumplimiento Regulatorio (RIA & RGPD)	Clasificación y Documentación del RIA	• ¿Cómo clasifican su sistema de IA según el RIA (alto riesgo, etc.)? • Evidencia: solicitar la Declaración de Conformidad de la UE y la documentación técnica completa exigida para sistemas de alto riesgo
	Gobernanza de Datos y Privacidad	• ¿Cómo garantizan la licitud de los datos de entrenamiento y el respeto a la propiedad intelectual? • ¿Nuestros datos serán utilizados para reentrenar sus modelos? ¿Existe una opción de opt-out? • ¿Dónde se alojan los datos? ¿Qué garantías ofrecen para transferencias internacionales de datos según el RGPD? • Evidencia: cláusulas contractuales específicas, certificaciones (ej., ISO 27701), resultados de auditorías de privacidad.
2. Seguridad de la Información (Ciberseguridad)	Robustez del Modelo	• ¿Qué medidas aplican para mitigar riesgos específicos de IA como prompt injection, envenenamiento de datos o extracción de modelos? • Evidencia: resultados de pruebas de red teaming, informes de pentesting, certificaciones de seguridad (ISO 27001, SOC 2).
	Gestión de Incidentes	• ¿Cuál es su protocolo de respuesta a incidentes de seguridad? ¿En qué plazo notifican las brechas? • Evidencia: Plan de Respuesta a Incidentes, cláusulas contractuales con los apropiados SLA de notificación definidos (ej. 48 horas).
3. Transparencia y Rendición de Cuentas	Explicabilidad	• ¿Qué nivel de explicabilidad ofrecen sobre las decisiones del modelo? ¿Es suficiente para que podamos cumplir con nuestras propias obligaciones de transparencia? • Evidencia: model cards, datasheets, documentación técnica sobre la lógica del modelo.
	Trazabilidad y Auditoría	• ¿El sistema genera logs auditables para todas las decisiones y operaciones críticas? • ¿Permiten la realización de auditorías por nuestra parte o por un tercero independiente? • Evidencia: muestras de logs, derecho a auditoría explícito en el contrato
4. Aspectos Contractuales y Operativos	Responsabilidad y SLA	• ¿Cómo se delimita la responsabilidad por daños causados por resultados erróneos o perjudiciales? • ¿Qué SLA garantizan en cuanto a disponibilidad, latencia y precisión del modelo? • Evidencia: cláusulas de limitación de responsabilidad, acuerdos de nivel de servicio detallados
	Control sobre IA embebida	• ¿Su servicio integra componentes de IA de terceros? ¿Cómo gestionan el riesgo de esa cadena de suministro? • -Evidencia: declaración sobre uso de IA embebida, cláusulas que exijan autorización previa para integrar sistemas de IA con riesgo.

Anexo 7

Plantillas de **evaluación de riesgos y cumplimiento**

Plantilla 1: Evaluación de Impacto y Riesgos de la IA

Registro			
ID de Evaluación:	EIR-IA-[YYYYMMD-D]-[ID_SISTEMA]	**Fecha:**	[Fecha de elaboración]
Autor(es):	[Nombres y Departamentos]	**Revisado por:**	[Nombre, Comité de Gobernanza de IA]
Estado:	EN BORRADOR / EN REVISIÓN / APROBADO	**Versión:**	1.0

Descripción y Clasificación del sistema	
Campo	**Descripción**
ID del Sistema en Inventario:	**[ID asignado en el inventario de IA]**
Nombre del Sistema:	**[Nombre descriptivo del sistema de IA]**
Propietario del Negocio:	**[Departamento responsable de la iniciativa]**
Rol de la Organización:	☐ **Proveedor** ☐ **Responsable del Despliegue**
Descripción y Finalidad:	**Describir en detalle el propósito del sistema, el problema de negocio que resuelve, su funcionamiento general y los objetivos explícitos para los que ha sido diseñado.**
Clasificación de Riesgo (RIA):	☐ **Riesgo Inaceptable (Prohibido)** ☐ **Alto Riesgo** ☐ **Riesgo Limitado (Obligaciones de Transparencia)** ☐ **Riesgo Mínimo** **Justificación: adjuntar el resultado del "Cuestionario de Clasificación de Riesgo" y justificar la clasificación final.**

Evaluación de Impacto en la Protección de Datos (EIPD)

(Obligatoria si el sistema trata datos personales y es probable que entrañe un alto riesgo para los derechos y libertades de las personas físicas)

Naturaleza, Alcance y Contexto del Tratamiento:

- **Datos Personales Tratados:** listar las categorías de datos (identificativos, financieros, etc.) y si se incluyen categorías especiales.
- **Ciclo de Vida del Tratamiento:** describir cómo se tratan los datos en cada fase: entrenamiento, validación, inferencia y retirada.
- **Base Jurídica (Licitud):** indicar la base legal del tratamiento para cada finalidad (consentimiento, interés legítimo, obligación legal, etc.)

Evaluación de Necesidad y Proporcionalidad:

- **Minimización de Datos:** ¿Se recogen y tratan únicamente los datos estrictamente necesarios para la finalidad? Justificar la necesidad de cada categoría de datos.
- **Limitación de la Finalidad:** ¿Existen medidas técnicas y organizativas para asegurar que los datos no se utilicen para fines incompatibles con los iniciales?

Identificación y Evaluación de Riesgos para los Derechos y Libertades:

ID	Descripción del Riesgo Potencial	Derecho(s) Fundamental(es) Afectado(s)	Probabilidad (Baja/Media/Alta)	Impacto (Severidad) (Bajo/Medio/Alto/Crítico)
R-001	Ej: Discriminación algorítmica en la selección de candidatos.	Ej: Derecho a la no discriminación, Derecho al trabajo.	Ej: Media	Ej: Crítico
R-002	Ej: Reidentificación de individuos a partir de datos supuestamente anonimizados.	Ej: Derecho a la privacidad, Protección de datos.	Ej: Baja	Ej: Alto
R-003	Ej: Decisiones automatizadas incorrectas (falsos positivos/negativos) que afectan legalmente al individuo.	Ej: Derecho a la tutela judicial efectiva, Derecho a la protección de datos (exactitud).	Ej: Media	Ej: Alto
R-004	Ej: Acceso no autorizado a categorías especiales de datos (salud, biométricos).	Ej: Derecho a la privacidad, Protección de datos especiales.	Ej: Baja	Ej: Crítico

Medidas de Mitigación Previstas

- **Técnicas:** listar medidas como seudonimización, cifrado, técnicas de privacidad diferencial, etc.
- **Organizativas:** controles de acceso, políticas de retención, formación del personal, etc.

Evaluación de Impacto en los Derechos Fundamentales (FRIA)

(Obligatoria para sistemas de Alto Riesgo desplegados por organismos públicos o para casos de uso específicos como la evaluación crediticia)

Aspecto a Evaluar	Descripción y Análisis
Grupos Afectados:	Identificar las categorías de personas que pueden verse afectadas, prestando especial atención a colectivos vulnerables.
Riesgos Específicos de Daño:	Analizar los riesgos potenciales para derechos fundamentales como la no discriminación, la dignidad humana, la libertad de expresión, el derecho a la tutela judicial efectiva, etc.
Medidas de Supervisión Humana:	Describir los mecanismos de intervención humana implementados: ¿son preventivos o a posteriori? ¿Qué autoridad tiene el supervisor para anular la decisión? ¿Es una supervisión significativa y no simbólica?
Mecanismos de Reclamación:	Describir los canales y procedimientos para que las personas afectadas puedan expresar su punto de vista, impugnar una decisión y solicitar una reparación.

Evaluación de Riesgos de Seguridad y Operacionales (ejemplo genérico)				
Categoría de Riesgo	**Riesgo Específico**	**Probabilidad**	**Impacto**	**Medidas de Mitigación Propuestas**
Seguridad del Modelo	**Envenenamiento de Datos (Data Poisoning): manipulación de los datos de entrenamiento para corromper el modelo.**			
	Ataques de Evasión: entradas adversarias diseñadas para engañar al modelo en la fase de inferencia.			• **Entrenamiento adversario (Adversarial Training).** • **Validación de entradas (Input Sanitization)**
	Inyección de Prompts (Prompt Injection): (Para LLM) Manipulación de las entradas para eludir las restricciones del modelo.			
Riesgo Operacional	**Degradación del Modelo (Model Drift): pérdida de rendimiento debido a cambios en el entorno de datos.**			
Cadena de Suministro	**Vulnerabilidades en Proveedor Externo: dependencia de un servicio de IA de terceros que sufre una brecha de seguridad.**			• **Due diligence exhaustiva del proveedor (ver checklist 13.4).** • **Cláusulas contractuales de seguridad y notificación**

Conclusión y Plan de acción

1. Resumen del Riesgo Residual: tras aplicar las medidas de mitigación, ¿cuál es el nivel de riesgo residual aceptado?
2. Declaración de Conformidad: ¿El sistema, con las medidas propuestas, cumple con los requisitos legales y las políticas internas?
3. Decisión Final del Comité de Gobernanza de IA:

   APROBAR DESPLIEGUE ☐

   APROBAR CON CONDICIONES (detallar abajo) ☐

   RECHAZAR (justificar abajo) ☐

Plantilla 2: Checklist de cumplimiento y Auditoría de IA

Registro General	
Campo	**Descripción**
ID del Sistema en Inventario:	**[ID del sistema auditado]**
Nombre del Sistema:	**[Nombre del sistema de IA]**
Fecha de la Auditoría:	**[Fecha de realización]**
Auditor(es):	**[Nombre(s) y Departamento(s)]**
Alcance de la Auditoría:	**Revisión periódica de cumplimiento (Anual / Semestral / Trimestral).**

Hallazgos y Plan de Acción (Ejemplos)			
ID Hallazgo	**Descripción de la No Conformidad**	**Riesgo Asociado**	**Severidad (Crítica/Alta/Media/Baja)**
CCA-01	**Los registros de supervisión humana en 3 de los 10 casos revisados carecen de una justificación detallada para la anulación de la decisión del modelo.**	**Incumplimiento del requisito de "intervención significativa" del RIA y falta de trazabilidad.**	**Alta**
CCA-02	**La información al usuario en la web no se ha actualizado para reflejar la última versión del modelo.**	**Falta de transparencia con los usuarios.**	**Media**
ID Hallazgo	**Acción Correctiva**	**Responsable**	**Fecha Límite**
CCA-01	**1. Revisar y completar los registros deficientes. 2. Realizar una sesión de refresco formativo para todos los analistas de riesgo sobre la importancia de la documentación detallada.**		
CCA-02	**Actualizar el texto informativo en la interfaz web del simulador de crédito.**		

Checklist ejemplo básico de verificación de Cumplimiento y Gobernanza			
	Punto de Control	**Estado (Conforme / No Conforme / N/A)**	**Evidencia / Observaciones**
CUMPLIMIENTO NORMATIVO			
2.1	**RIA: a documentación técnica del sistema está actualizada y refleja el estado actual del modelo.**		**Revisión de la "Model Card" v1.1, fecha 25/09/2025.**
2.2	**RIA: los registros de trazabilidad (logs) de las decisiones del sistema son accesibles y completos para el período auditado.**		
2.3	**RGPD: El Registro de Actividades de Tratamiento (RAT) está actualizado para este sistema.**		
2.4	**RGPD: Las solicitudes de ejercicio de derechos (acceso, rectificación, oposición a decisiones automatizadas) han sido gestionadas en plazo.**		
GOBERNANZA INTERNA Y ÉTICA			
2.5	**El uso actual del sistema se alinea con la finalidad aprobada en la EIR-IA. No se han detectado usos no previstos.**		
2.6	**(Para Alto Riesgo) Los registros de supervisión humana demuestran una intervención efectiva y no simbólica.**		
2.7	**Se han realizado nuevas pruebas de sesgo y equidad. Los resultados están dentro de los umbrales aceptados.**		**Informe de re-evaluación de sesgo adjunto.**
2.8	**La información proporcionada a los usuarios sobre la interacción con la IA es clara, visible y está actualizada.**		
RENDIMIENTO DEL MODELO			
3.1	**Las métricas clave de rendimiento (ej. Accuracy, AUC) se mantienen por encima de los umbrales definidos. No se ha detectado model drift significativo.**		**Dashboard de monitorización, período [fecha] a [fecha].**
3.2	**El sistema de monitorización y alertas de rendimiento está operativo y ha funcionado correctamente.**		
SEGURIDAD			
3.3	**La infraestructura que soporta el sistema está actualizada y parcheada según la política de gestión de vulnerabilidades.**		**Informe de escaneo de vulnerabilidades.**
3.4	**Se han revisado los controles de acceso. No se han detectado accesos no autorizados a los datos ni al modelo.**		**Revisión de logs de acceso.**
3.5	**No se han registrado incidentes de seguridad relacionados con vectores de ataque específicos de IA (ej. prompt injection, evasión).**		**Registro de incidentes del SOC.**

Contacta con nosotros

Si estás interesado en colaborar con nosotros o necesitas más información sobre nuestros proyectos, escríbenos a: **proyectos@ismsforum.es**