Telefónica Tech

Secure Journey to AI

La hoja de ruta de Telefónica Tech para adoptar
la IA con seguridad, gobernanza y cumplimiento

telefonicatech.com

Índice

Introducción

La Inteligencia Artificial (IA) es un motor clave de transformación para empresas de todos los tamaños, impulsando la eficiencia, personalización y competitividad en numerosos sectores. Su capacidad para analizar datos, automatizar procesos y potenciar la innovación la convierte en una tecnología transversal con impacto directo en las operaciones empresariales.

En este contexto, la IA Generativa representa un salto cualitativo: amplía el potencial creativo de la tecnología y democratiza el uso de la IA en toda la empresa. Sin embargo, la adopción de la IA aumenta riesgos conocidos en el ámbito de la privacidad, la seguridad, el cumplimiento normativo y la gobernanza; e introduce otros inéditos, como los relacionados con sesgos y envenenamiento del modelo.

Para responder a estos retos, en Telefónica Tech proponemos el marco **Secure Journey to AI**, una estrategia integral que permite a las empresas aprovechar el potencial de la IA de forma segura, responsable y alineada con sus objetivos de negocio.

Nuestra propuesta se apoya en tres pilares clave:

- **Identificación de riesgos:** detección temprana de vulnerabilidades en datos, identidades, infraestructuras y modelos.
- **Protección frente a amenazas:** implementación de medidas robustas para asegurar infraestructuras resilientes, aplicaciones seguras y un control efectivo de accesos y datos.
- **Respuesta 360º:** monitorización continua, integración con un SOC especializado y un marco de mejora continua ante incidentes.

En este documento analizamos el contexto empresarial de la IA y los retos y oportunidades de su adopción abordados con la seguridad como eje central: nuestras recomendaciones estratégicas y principios de responsabilidad están diseñados para que la IA se implemente de forma segura, gobernada y en cumplimiento normativo, siempre alineada con los objetivos de negocio.

Con Telefónica Tech como socio tecnológico, las empresas pueden convertir la IA en un motor de transformación fiable, capaz de impulsar la innovación sin poner en riesgo ni la seguridad ni la continuidad del negocio. Nuestro marco Secure Journey to AI asegura que su adopción genere valor real, resiliencia operativa y una ventaja competitiva sostenible.

Las empresas que integren la IA de manera segura transformarán la regulación en confianza y la confianza en ventaja competitiva.

Contexto de la IA en entornos empresariales

La presencia de la IA en las empresas es incuestionable, y su evolución hacia la IA Generativa ha supuesto un cambio significativo. Esta nueva tecnología potencia la eficiencia operativa e introduce capacidades creativas y de automatización avanzada. Permite que personas de diferentes áreas accedan a herramientas innovadoras y participen en la transformación digital de sus empresas.

Este salto cualitativo trae una paradoja difícil de ignorar. Por un lado, las empresas ven en la IA una oportunidad inédita para innovar, acelerar su transformación digital y abrir nuevos modelos de negocio. Por otro, enfrentan riesgos: pérdida de trazabilidad en los resultados, sesgos amplificados en las decisiones, exposición accidental de información sensible o ataques diseñados para manipular los datos y modelos de IA.

En este escenario, la cuestión ya no es si las empresas incorporarán la IA en sus operaciones, sino cómo recorrer este viaje de manera segura y responsable. Gobernar la tecnología, anticipar riesgos y alinear su uso con los valores y objetivos del negocio será lo que marque la diferencia entre aprovechar su potencial o sucumbir a sus amenazas. Para lograrlo, las empresas necesitan un marco integral que proporcione seguridad, cumplimiento normativo y gobernanza desde el primer paso.

> *La IA ha dejado de ser una promesa futura: hoy es el motor que transforma sectores y redefine cómo las empresas compiten, innovan y se relacionan con la sociedad.*

Consideraciones regulatorias y responsabilidad en IA

La adopción de la IA en un entorno empresarial no puede entenderse al margen de la seguridad y la confianza. Cada vez que un modelo de IA procesa información sensible, como datos personales, historiales clínicos o transacciones financieras, entra en juego un marco regulatorio y de responsabilidad que condiciona su diseño y gestión.

En Europa, el Reglamento de Inteligencia Artificial (AI Act) establece un marco integral para clasificar los sistemas de IA en función de su riesgo y exigir transparencia, trazabilidad y medidas de gobernanza específicas. Junto a él, el Reglamento General de Protección de Datos (GDPR) salvaguarda la privacidad de los individuos.

También normativas sectoriales como HIPAA en el ámbito sanitario o PCI-DSS en el financiero imponen obligaciones estrictas para proteger la confidencialidad de la información médica y de las operaciones de pago y con tarjetas bancarias. Estos marcos subrayan que el valor de la IA solo es sostenible si se basa en datos protegidos, tratados con rigor y responsabilidad.

Además, seguridad de la infraestructura es igual de crítica que la privacidad. Estándares internacionales como ISO/IEC 27001 ayudan a establecer sistemas de gestión de la seguridad de la información,

reforzando controles en todo el ciclo de vida de la IA. En el plano regulatorio, la directiva NIS2 amplía las exigencias de ciberseguridad y resiliencia digital en sectores esenciales, mientras que DORA (Digital Operational Resilience Act) establece requisitos específicos para que el sector financiero resista y se recupere de incidentes tecnológicos y ciberataques.

Cumplir con estas normativas no es solo una obligación legal, sino una condición para generar confianza en clientes, reguladores y en la sociedad. Esto asegura que la IA se utiliza de forma alineada con valores sociales y objetivos de negocio, evitando impactos negativos en derechos fundamentales y en las operaciones y los resultados empresariales.

La gobernanza y el cumplimiento no son un componente añadido al final, sino la base que permite que las empresas adopten la IA de manera segura, responsable, adaptada a la normativa y alineada con sus objetivos estratégicos.

Retos principales en la adopción de la IA

El camino hacia la adopción exitosa de la IA no está exento de obstáculos. Las empresas se enfrentan a una serie de desafíos recurrentes que, si no se gestionan con visión estratégica, pueden comprometer tanto la eficacia de los proyectos como la confianza de clientes, socios, empleados y reguladores.

Estos retos abarcan dimensiones técnicas, organizativas y éticas, y requieren un enfoque proactivo que combine gobernanza, seguridad y capacidad de adaptación continua.

Gobierno y cumplimiento

La gobernanza de la IA sigue siendo uno de los grandes retos para las empresas. No basta con desplegar soluciones avanzadas: es necesario establecer marcos claros que definan responsabilidades, asignen roles específicos y garanticen la trazabilidad de los procesos.

La capacitación de los empleados en el uso responsable de la IA es igualmente crítica, ya que el desconocimiento o la falta de formación pueden abrir la puerta a errores, sesgos o usos indebidos. Un buen modelo de gobernanza asegura que la IA cumple con las normativas vigentes y, además, está alineada con los valores y objetivos estratégicos de la empresa.

Gestión de redes y seguridad

La seguridad se ha convertido en una preocupación central a medida que los modelos de IA e IA Generativa se conectan con más datos, aplicaciones e infraestructuras. Las amenazas evolucionan a gran velocidad: desde ataques como el *prompt injection* hasta fugas de información o intentos de *model poisoning* capaces de manipular la integridad de los sistemas.

Ante este panorama, proteger los modelos y los datos asociados exige rediseñar arquitecturas, reforzar controles de acceso y aplicar soluciones avanzadas de ciberseguridad. Solo así las empresas pueden reducir la superficie de ataque y salvaguardar la confianza en los resultados que generan sus sistemas de IA.

Supervisión y mejora continua

La IA no es estática. Los marcos regulatorios, las tecnologías y los riesgos evolucionan de forma constante, lo que obliga a las empresas a mantener un enfoque de supervisión permanente y de mejora continua. Esto implica auditar regularmente políticas, revisar procesos y actualizar controles en función de los cambios tecnológicos y normativos.

La supervisión no debe interpretarse como una carga burocrática, sino como una oportunidad para fortalecer la resiliencia de los sistemas y adaptarlos a nuevas realidades. Un ciclo de mejora continua convierte la seguridad en un activo estratégico, capaz de anticipar amenazas y asegurar la sostenibilidad de los proyectos a largo plazo.

Estos retos no deben verse como obstáculos a la innovación, sino como condiciones necesarias para que la IA contribuya de manera sostenida a la competitividad. Superarlos permite a las empresas asegurar la resiliencia, mitigar riesgos y reforzar la confianza de clientes, socios estratégicos y reguladores, consolidando un entorno en el que la IA pueda desplegar todo su potencial de forma responsable.

Ataques frecuentes a la IA

Pérdida de información	• **Privacy attacks.** Incluyen la extracción de información sensible de los datos de entrenamiento o del contexto de uso.
	  ○ ***Data reconstruction*:** reconstrucción de información privada a partir de respuestas.
	  ○ ***Inferencia de pertenencia*:** determinar si un individuo formó parte del conjunto de entrenamiento.
	  ○ ***Model extraction*:** replicar arquitectura y parámetros del modelo mediante consultas.
	  ○ ***Prompt/context stealing*:** recuperar instrucciones de sistema o contexto confidencial.

Inyección de prompts	• **Prompt injection.** Instrucciones ocultas en el prompt para alterar el comportamiento del modelo, filtrar información sensible o eludir controles de seguridad. • **Jailbreaks.** Manipulación de un modelo de IA para ignorar sus restricciones de seguridad y generar respuestas indebidas o prohibidas
Respuestas de riesgo	• **Generación de información errónea o inapropiada.** Incluye respuestas falsas (*alucinaciones*) y contenido sesgado o malicioso que puede afectar la reputación o la toma de decisiones.
Envenenamiento	• **Data poisoning.** Manipulación de los datos de entrenamiento para introducir sesgos, puertas traseras o vulnerabilidades que comprometan la seguridad o fiabilidad del modelo. • **Model poisoning.** Alteración directa de los parámetros del modelo (por ejemplo en *federated learning*), degradando su rendimiento o forzando salidas manipuladas.

Riesgo de acceso	• **Accesos no autorizados.** Uso indebido para extraer, modificar o eliminar información crítica.
Otros	• **Supply chain attacks.** Ataques a la cadena de suministro de IA (datasets, librerías, modelos preentrenados) para introducir componentes maliciosos en aplicaciones empresariales.

El viaje hacia una IA Segura

Identificación de los riesgos de la IA

Como hemos visto, la adopción de la IA plantea riesgos específicos que, si no se gestionan adecuadamente, pueden comprometer la seguridad, la privacidad, la resiliencia operativa y el cumplimiento normativo de las empresas.

Este apartado aborda el primer pilar de nuestra estrategia Secure Journey to AI: la identificación de riesgos.

El objetivo es detectar de manera temprana vulnerabilidades y amenazas que puedan afectar tanto al desarrollo como al uso de la IA, posibilitando un despliegue seguro y responsable. Para ello, se requiere una visión integral que abarque desde la infraestructura tecnológica hasta los datos, identidades, modelos y aplicaciones.

> *Identificar de forma temprana los riesgos de la IA permite construir modelos de confianza y resiliencia.*

Consultoría para el despliegue de la IA

Antes de implementar soluciones de IA, es imprescindible diseñar arquitecturas seguras desde la fase de planificación. La consultoría especializada permite incorporar la seguridad como un habilitador estratégico y no como un añadido.

Esto implica:

- Evaluar la infraestructura tecnológica para asegurar que soporte la IA de manera segura y escalable.
- Diseñar y analizar las arquitecturas de referencia que integren controles de ciberseguridad desde el inicio.

- Analizar los riesgos relacionados con la cadena de suministro de modelos y herramientas.
- Aplicar el enfoque de *security by design*, fundamental para anticipar las amenazas antes de que surjan, reduciendo costes y vulnerabilidades.

Por ejemplo, un despliegue que omite la revisión de dependencias de software de terceros puede abrir la puerta a ataques de cadena de suministro, impactando la integridad de los modelos.

Integrar la seguridad en estas fases tempranas evita sobrecostes, asegura la confianza y facilita el cumplimiento de normativas como el AI Act europeo o la directiva NIS2, reduciendo la exposición a sanciones, brechas y crisis reputacionales.

Auditoría de identidad, datos y seguridad

Los datos, las identidades y la seguridad de los entornos son el núcleo de la IA. Si no se gestionan correctamente los accesos, permisos y configuraciones, las fugas de información o los accesos indebidos pueden comprometer tanto los modelos como la información sensible de los usuarios.

- La auditoría debe incluir la evaluación de permisos y la segmentación de roles aplicando el principio de mínimo privilegio, junto con mecanismos como la autenticación multifactor (MFA) o el acceso condicional. Esto garantiza que cada usuario solo acceda a la información estrictamente necesaria, reduciendo la superficie de exposición.
- Es importante clasificar y etiquetar los datos sensibles, aplicar políticas de prevención de fuga de información (DLP) y establecer una revisión continua de registros para identificar usos indebidos o no autorizados. Así se protege tanto la información personal como la propiedad intelectual de la empresa.
- Resulta esencial realizar una auditoría de seguridad integral como diagnóstico inicial, que proporcione a las organizaciones un mapa claro y priorizado de riesgos antes de escalar la adopción de la IA.

Un mal diseño de accesos en un hospital, por ejemplo, podría exponer historiales médicos a usuarios no autorizados, con consecuencias legales y reputacionales graves.

Por eso, un control robusto de identidad, datos y seguridad contribuye a que los sistemas de IA se utilicen de manera adecuada, protegiendo tanto la propiedad intelectual de la empresa como la privacidad de los usuarios. Esto en línea con normativas como el RGPD en materia de datos personales, HIPAA en el ámbito sanitario y NIS2 en lo relativo a la protección de infraestructuras críticas y resiliencia digital.

Descubrimiento de riesgos

La identificación temprana de riesgos requiere aplicar metodologías y herramientas avanzadas para detectar vulnerabilidades antes de que sean explotadas.

- La gestión de vulnerabilidades (VRM) es un proceso continuo de identificación, evaluación, priorización y remediación de vulnerabilidades en los sistemas, aplicaciones e infraestructuras de la empresa.

- El análisis de superficies de ataque revela puntos vulnerables en modelos y aplicaciones.
- Las evaluaciones de *shadow AI* ayudan a descubrir usos no autorizados de herramientas de IA dentro de la empresa.
- La priorización de riesgos mediante sistemas de criticidad permite enfocar los recursos en las amenazas con mayor impacto potencial.

En sectores como la banca, identificar el uso no autorizado de aplicaciones de IA por parte de empleados, fuera de los entornos validados por la empresa, puede marcar la diferencia entre un incidente menor y una filtración masiva de datos sensibles.

Adoptar un enfoque de descubrimiento proactivo convierte la seguridad en un proceso continuo, capaz de anticipar incidentes y garantizar una adopción de la IA con confianza y control.

Seguridad de modelos de IA

Los modelos de IA son el núcleo de las aplicaciones generativas, pero también uno de sus vectores de riesgo más críticos. Protegerlos requiere pruebas y técnicas específicas que garanticen su resiliencia.

- Las pruebas adversarias (*adversarial testing*) permiten comprobar cómo reaccionan los modelos frente a entradas maliciosas diseñadas para manipular sus resultados.
- El análisis de las API y puntos de integración es esencial para prevenir ataques como *prompt injection*, *model extraction* o *data poisoning*.
- Las metodologías de *offensive security* orientadas a IA generativa y el endurecimiento de modelos (*hardening*) contribuyen a reforzar su resistencia frente a intentos de manipulación.

Por ejemplo, un modelo de atención al cliente sin controles robustos puede ser manipulado para entregar información sensible o dar respuestas fraudulentas.

La seguridad de los modelos protege el rendimiento técnico, la reputación y la continuidad del negocio: un fallo en esta capa puede derivar en pérdida de confianza de clientes, incumplimiento regulatorio y riesgos operativos significativos.

Principales riesgos de la IA Generativa

Shadow AI.
Uso no autorizado por empleados de herramientas de IA sin control corporativo, lo que aumenta el riesgo de fuga de información sensible y dificulta la trazabilidad de actividades.

Fuga de datos sensibles
Prompts mal diseñados o configuraciones inseguras pueden provocar que asistentes generativos conectados a sistemas SaaS o repositorios internos expongan información confidencial de la empresa.

Alucinaciones y desinformación.
La IA puede generar respuestas falsas, incoherentes o engañosas, dañando la reputación corporativa, la confianza del cliente y afectando al cumplimiento normativo.

Sesgos algorítmicos.
La falta de diversidad en los conjuntos de entrenamiento puede introducir sesgos que conduzcan a decisiones discriminatorias, poco fiables o contrarias a los objetivos de negocio.

Manipulación de pipelines de RAG (Retrieval Poisoning).
Los atacantes pueden introducir datos alterados en las fuentes de información corporativas que alimentan dinámicamente al modelo, comprometiendo la fiabilidad del sistema y exponiendo información sensible de la empresa.

Ataques de privacidad.
Técnicas como la extracción de información de entrenamiento, inferencia de pertenencia (descubrir un registro concreto) o robo de prompts pueden revelar datos personales, propiedad intelectual o secretos corporativos.

> Estos riesgos explican la necesidad de contar con un marco sólido de seguridad y gobernanza para identificar y gestionar vulnerabilidades desde el inicio.

Protección frente a amenazas

Una vez identificados los riesgos, el siguiente paso en nuestro Secure Journey to AI es implementar medidas de protección que salvaguarden los modelos, aplicaciones y datos frente a ciberamenazas.

Este segundo pilar se centra en fortalecer la resiliencia de la infraestructura de IA, corregir vulnerabilidades y asegurar que los entornos tecnológicos operen bajo un marco de seguridad integral.

La protección frente a amenazas no debe limitarse a la corrección puntual de fallos, sino que ha de concebirse como un proceso continuo y dinámico en el que se combinan buenas prácticas de desarrollo seguro, controles avanzados de ciberseguridad e integración con la gestión de identidades y accesos.

La verdadera innovación solo prospera cuando está protegida: blindar datos, modelos e infraestructuras es la base para aprovechar con confianza todo el potencial de la IA.

Creación de aplicaciones resilientes

El desarrollo de aplicaciones de IA debe incorporar la seguridad desde el diseño para prevenir vulnerabilidades que puedan ser explotadas más adelante. La metodología DevSecOps, que integra seguridad en cada fase del ciclo de desarrollo, permite detectar y corregir fallos en etapas tempranas, reduciendo el coste y el impacto de los incidentes.

- Esto implica contar con procesos sólidos de gestión de vulnerabilidades, capaces de identificar y resolver errores antes de la puesta en producción.
- También supone aplicar técnicas de análisis de código, tanto estático como dinámico e interactivo (AST, SAST, DAST, IAST), que aseguren que el software esté libre de vulnerabilidades críticas.
- A ello se suman soluciones como WAAP *(Web Application and API Protection)* y CNAPP *(Cloud-Native Application Protection Platform)*, diseñadas para proteger aplicaciones web y entornos nativos de la nube frente a ataques sofisticados.
- Además, el enfoque de *shift-left security* refuerza la idea de mover los controles de seguridad lo más cerca posible del inicio del ciclo de desarrollo, reduciendo riesgos antes de que la aplicación entre en operación.

En conjunto, estas prácticas permiten crear aplicaciones resilientes que reducen la superficie de ataque y refuerzan la confianza en la IA, evitando que se convierta en un punto débil dentro del ecosistema empresarial.

Protección de datos e identidad (IAM)

Los datos y las identidades son activos críticos para la IA. La protección frente a amenazas pasa por reforzar el perímetro lógico y asegurar que solo las personas autorizadas accedan a la información sensible.

- Un esquema de gestión de identidades y accesos (IAM) avanzado incluye mecanismos como la autenticación multifactor (MFA), el acceso condicional y el control granular de roles, asegurando que cada usuario solo acceda a la información estrictamente necesaria.
- A ello se suman auditorías automáticas y sistemas de alerta temprana que permiten detectar comportamientos anómalos en tiempo real, así como políticas de prevención de fuga de datos (DLP) para evitar que información sensible salga de los sistemas de forma accidental o intencionada.

La dimensión humana también juega un papel clave: la protección frente a ataques de ingeniería social, como el phishing o el robo de credenciales, es imprescindible para reducir el riesgo de accesos indebidos.

En la práctica, un simple clic en un correo fraudulento puede abrir la puerta a intrusiones que comprometan modelos y datos críticos.

> ***Una estrategia sólida de protección de datos e identidad evita la sobreexposición, reduce la probabilidad de fugas y asegura la integridad de los modelos, reforzando así la confianza de clientes, socios y usuarios.***

Corrección de riesgos en la arquitectura de la IA

La arquitectura que sustenta la IA requiere una protección por capas que permita prevenir configuraciones inseguras o exposiciones accidentales que comprometan la seguridad global.

Esto requiere:

- Desplegar infraestructuras seguras, aplicando técnicas como la microsegmentación y utilizando herramientas de seguridad específicas para entornos híbridos y en la nube. Entre ellas destacan:
    - CSPM *(Cloud Security Posture Management)*, que monitoriza la configuración de la nube para detectar desviaciones.
    - CWPP *(Cloud Workload Protection Platform),* que protege cargas de trabajo en la nube.
    - CIEM *(Cloud Infrastructure Entitlement Management),* que controla permisos excesivos en entornos cloud.
- Revisar y corregir de manera continua configuraciones erróneas en pipelines y modelos de IA, así como el uso de *AI Gateways* y políticas de seguridad que regulen cómo se exponen y utilizan los modelos generativos.
- Proteger las API y *plugins* utilizados en la integración para reducir los riesgos de manipulación, extracción de datos o abusos en el consumo de los servicios de IA.

Seguridad en la arquitectura de alto nivel de la IA generativa

La seguridad de una aplicación de IA generativa debe entenderse de manera integral, aplicando controles específicos en cada nivel de su arquitectura. Estos niveles abarcan desde los datos que alimentan los modelos hasta la interacción final con el usuario.

- Entrenamiento del modelo. La protección de los datos de entrenamiento es esencial para garantizar la fiabilidad de los resultados. Requiere una sólida gobernanza del dato, clasificación y etiquetado de información sensible, mecanismos DLP y auditorías periódicas que eviten la exposición o el uso indebido de la información.
- Runtime del LLM. El entorno en el que se procesan los prompts y se generan respuestas necesita una protección continua. Aquí destacan las capacidades de AI-SPM (AI Security Posture Management), que permiten monitorizar desviaciones y aplicar medidas correctivas de forma

proactiva, junto con evaluaciones ofensivas específicas (como pruebas de *prompt injection*, extracción de datos o manipulación de salidas) que validan la resiliencia del modelo frente a ataques avanzados.

- Capa de aplicación. En el punto de contacto con el usuario es crítico reforzar la seguridad de las API, plugins y extensiones. Para ello se aplican soluciones WAAP o WAF de nueva generación, combinadas con pruebas ofensivas recurrentes que aseguren la resiliencia de la interacción frente a abusos o fugas de información.

En conjunto, esta visión de tres bloques garantiza que la seguridad acompañe todo el ciclo, desde los datos de entrenamiento hasta la experiencia del usuario final, consolidando la arquitectura de IA generativa como un entorno confiable, resiliente y en cumplimiento normativo.

Buenas prácticas para proteger datos en IA

Clasificación y etiquetado de datos
Identificar el nivel de sensibilidad de la información y aplicar políticas específicas de protección según su criticidad.

Prevención de fuga de datos (DLP)
Implementando controles que impidan el envío no autorizado de información sensible, tanto en entornos cloud como en dispositivos corporativos.

Gestión de identidades y accesos (IAM)
Aplicando autenticación multifactor (MFA), controles granulares de roles y el principio de mínimo privilegio para limitar la exposición.

Cifrado extremo a extremo
Protege datos tanto en reposo como en tránsito para evitar accesos indebidos.

Políticas de uso responsable
Capacitación a empleados y colaboradores sobre buenas prácticas al interactuar con modelos de IA, para reducir errores humanos y malas configuraciones.

> Estas medidas convierten la **seguridad en un habilitador** que blinda la información crítica y refuerza la confianza de clientes, reguladores y socios.

Respuesta 360º

El tercer pilar de nuestro Secure Journey to AI se centra en una respuesta completa, de 360º. Un enfoque integral que asegura visibilidad continua, capacidad de reacción ante incidentes y una mejora continua frente a amenazas cambian con rapidez. La IA introduce riesgos nuevos y dinámicos y las empresas necesitan un marco de acción que no solo detecte vulnerabilidades, sino que también permita responder de inmediato y aprender de cada evento.

Este pilar combina la monitorización avanzada, la integración con un SOC *(Security Operations Center)* especializado y la gestión sistemática de incidentes con el objetivo de asegurar que las empresas operan con confianza y transparencia en un entorno tecnológico complejo y cambiante.

En un entorno donde las amenazas evolucionan a la velocidad de la tecnología, la capacidad de responder con rapidez y transparencia será el mayor factor de confianza.

Visibilidad end-to-end y gobernanza de la seguridad

Para que las empresas puedan confiar en sus sistemas de IA necesitan visibilidad total sobre lo que ocurre en cada capa: desde la infraestructura hasta los modelos y sus datos. Sin trazabilidad ni auditoría, los riesgos se vuelven invisibles y la confianza se resquebraja.

- La visibilidad end-to-end no es solo un registro técnico de actividad: es la garantía de que cada interacción puede ser explicada, revisada y corregida si es necesario.
- Contar con logs detallados de usuarios y prompts, aplicar monitorización simultánea de infraestructuras y aplicaciones.
- Aprovechar la telemetría avanzada basada en IA permite anticipar anomalías antes de que se conviertan en incidentes.

Un banco que monitoriza en tiempo real las interacciones de sus asistentes virtuales, por ejemplo, no solo protege la integridad de sus modelos, sino que puede demostrar ante auditores y clientes que cada recomendación se genera bajo criterios de seguridad y cumplimiento normativo.

En este sentido, la visibilidad se convierte en un activo de gobernanza y transparencia, esencial para cumplir con normativas como la AI Act europeo y reforzar la confianza de clientes y reguladores.

Integración con SOC especializado de Telefónica Tech

El SOC *(Security Operations Center)* especializado en IA de Telefónica Tech actúa como una torre de control que vigila, anticipa y neutraliza amenazas en tiempo real.
Esta capacidad permite a las empresas contar con la seguridad de que sus entornos de IA se encuentran monitorizados de forma continua, asegurando la detección y gestión inmediata de cualquier incidente.

Más allá de la vigilancia, el SOC ofrece:

- Automatización de respuestas, integrando *playbooks* que activan defensas en segundos.
- Inteligencia frente a nuevas amenazas, desde ataques de *prompt injection* hasta intentos de *data poisoning* o fugas de contexto.
- Intervención humana cuando el incidente lo requiere, coordinando la respuesta con equipos multidisciplinares capaces de escalar y contener situaciones críticas.

El valor del SOC no está solo en la tecnología, sino en la confianza que genera un servicio gestionado y experto. De este modo, el SOC convierte la ciberseguridad en un elemento estratégico, asegurando la continuidad del negocio, la reputación y la tranquilidad de las empresas que adoptan la IA.

Gestión de incidentes y mejora continua

Responder a incidentes no es suficiente: es esencial convertir cada situación en una oportunidad para reforzar la resiliencia organizativa. La gestión de incidentes en IA implica identificar causas raíz, ajustar políticas y mejorar los procesos de forma que cada ataque frustrado fortalezca la defensa global.

Este enfoque se traduce en:

- Un ciclo posincidente, donde cada evento genera aprendizajes que se incorporan al marco de seguridad.
- Políticas adaptativas se ajustan en función de nuevas amenazas, las métricas específicas de IA ayudan a evaluar la eficacia de los controles.
- Entornos de simulación o sandbox AI permiten probar los modelos frente a escenarios adversos antes de ponerlos en producción.

Pensemos en una empresa de salud que detecta un intento de manipulación de su modelo de diagnóstico. Gracias a un proceso estructurado de gestión de incidentes detiene la amenaza y, además, adapta sus controles para evitar que un ataque así pueda repetirse. Así, esa empresa de salud no reacciona, sino que evoluciona.

Un marco de gestión de incidentes y mejora continua asegura que la IA no quede atrapada en un ciclo de vulnerabilidad y reacción, sino que avance hacia un uso más robusto y fiable, reduciendo la recurrencia de ataques y consolidando la confianza organizacional en su uso.

La Respuesta 360º no solo refuerza la resiliencia, también asegura el cumplimiento normativo.

La seguridad como un proceso vivo

Monitorizar de extremo a extremo la actividad	Modelos, datos, identidades y aplicaciones para asegurar trazabilidad y gobernanza.
Identificar anomalías	Intentos de manipulación o fugas de datos antes de que escalen a incidentes críticos.
Integrar con SOC especializado	Usar *playbooks* automáticos para contener y mitigar incidentes de manera ágil.
Aprender de cada incidente	Ajustar políticas y reforzar controles, para asegurar que la protección evoluciona al mismo ritmo que las amenazas.
Auditorías continuas	Validan periódicamente la eficacia de los controles y midan la resiliencia frente a nuevas amenazas y requisitos normativos.

> La seguridad en la IA no es un objetivo final, sino un ciclo permanente de anticipación, adaptación y aprendizaje que convierte la confianza en una ventaja competitiva sostenible. Este ciclo continuo de monitorización, respuesta y auditoría contribuye de forma decisiva al cumplimiento normativo. Marcos como el AI Act, NIS2 o DORA exigen trazabilidad, gobernanza y validación periódica de los sistemas de IA, requisitos que se integran de manera natural en un enfoque de Respuesta 360º

Conclusiones y próximas acciones

Conclusiones

La IA se ha consolidado como un catalizador de transformación, pero introduce riesgos inéditos que obligan a las empresas a replantearse cómo la adoptan. Nuestra hoja de ruta Secure Journey to AI aborda esos nuevos riesgos mediante tres pilares: identificación y detección temprana de riesgos, protección ante amenazas y respuesta 360° a incidentes.

Es importante tener en cuenta que estos pilares no funcionan de manera aislada: constituyen un marco cohesionado que habilita un uso de la IA seguro, responsable y alineado con los objetivos de negocio.

La conclusión más relevante es que la seguridad y la confianza no son un obstáculo para la innovación, sino su habilitador. Solo las empresas que integren el cumplimiento normativo, la gobernanza y la protección de datos en su estrategia podrán desplegar la IA de forma sostenible y obtener un valor real y duradero.

Solo las empresas que integren la IA en su ADN de forma segura podrán gobernarla con eficacia, cumplir con la regulación y convertir la confianza en ventaja competitiva.

Acciones inmediatas

Las conclusiones nos marcan un camino claro: la adopción de la IA exige situar la seguridad como eje central y abordarla con un enfoque proactivo. No basta con reaccionar cuando aparecen riesgos: hay que anticiparse y convertir la seguridad en una ventaja competitiva.

Para avanzar, las empresas deben iniciar estas acciones de inmediato:

- Auditar el uso actual de IA para identificar y clasificar aplicaciones en uso (incluido *Shadow AI*) para obtener una visión precisa de la situación.
- Evaluar la madurez en IA para diagnosticar brechas en gobernanza, seguridad y competencias internas, y diseñar una hoja de ruta factible.
- Establecer un modelo de gobernanza transversal que permita definir responsabilidades claras, alinear áreas y ofrecer formación a todos los niveles para construir una cultura de IA responsable.
- Integrar la seguridad desde el diseño incorporando controles en datos, modelos y aplicaciones desde el inicio de los proyectos, evitando así sobrecostes y generando confianza desde el principio.

Adoptar un enfoque de mejora continua revisando y actualizando políticas y procesos para adaptarlos a nuevas amenazas y marcos regulatorios, como el AI Act, el RGPD o la directiva NIS2.

Estas acciones inmediatas son la base para que la adopción de la IA maximice la rentabilidad de las inversiones, minimice riesgos y acelere la transformación digital de manera sostenible y fiable.

La adopción de la IA solo es sostenible cuando la seguridad es el eje central y se aborda de forma proactiva.

Consolidar una IA segura y responsable

Tras recorrer el camino de la identificación de riesgos, la protección frente a amenazas y la respuesta 360º, el gran desafío para las organizaciones es consolidar un modelo de IA que mantenga la seguridad como eje central y que inspire confianza en el largo plazo. No se trata solo de iniciar proyectos, sino de garantizar que su evolución se mantenga segura, estable y resiliente, capaz de adaptarse a necesidades cambiantes y a nuevas exigencias regulatorias.

El futuro de la IA no depende solo de lo que es capaz de hacer, sino de cómo decidimos implementarla, consolidarla y gobernarla.

La sostenibilidad, la resiliencia y la transparencia serán los factores que marcarán la diferencia, permitiendo a la IA convertirse en un activo estable que acompañe la estrategia de negocio. Alcanzar este objetivo exige un enfoque que combine innovación con seguridad, responsabilidad con resiliencia y transparencia con cumplimiento normativo, para que la tecnología tenga propósito y aporte valor.

En Telefónica Tech acompañamos a empresas de cualquier tamaño y sector en este proceso con Secure Journey to AI, un enfoque integral que asegura que la adopción de la IA se traduzca en impacto positivo, valor sostenible y confianza a largo plazo.

Por qué Telefónica Tech

En el viaje hacia la adopción de una IA segura y responsable, en Telefónica Tech no actuamos solo como un proveedor: somos el socio tecnológico y estratégico que acompaña a las empresas en cada paso y en cada hito de este recorrido.

En Telefónica Tech convertimos la seguridad en el lenguaje de la innovación: acompañamos a las empresas para que recorran el camino hacia una IA segura y con impacto positivo en su negocio.

Nuestro enfoque integral protege los activos clave (datos, aplicaciones e infraestructuras) mientras habilita un uso innovador, responsable y alineado con los objetivos de negocio. Aportamos un valor diferencial a través de capacidades únicas que convierten la seguridad en motor de confianza:

- Protección integral de activos. Aseguramos que los datos, las aplicaciones y las infraestructuras de IA estén siempre protegidos sin comprometer su valor estratégico, aplicando una visión de seguridad transversal que se extiende de extremo a extremo.
- Hoja de ruta en tres pilares. Identificación de riesgos, protección frente a amenazas y respuesta 360° forman nuestro marco integral Secure Journey to AI que permite a las empresas recorrer con confianza su viaje hacia la IA.
- Confianza regulatoria garantizada. Aseguramos que cada proyecto cumpla con marcos como AI Act, NIS2 o DORA y anticipe sus exigencias, ofreciendo transparencia y tranquilidad a clientes, empleados y reguladores.
- Capacidades reconocidas en ciberseguridad. Contamos con una infraestructura robusta, servicios avanzados de monitorización 24x7 y certificaciones de terceros para una defensa proactiva y gestión de identidades y datos sensibles.
- Talento y especialización en IA. Más de 400 expertos y 10 centros especializados que combinan conocimiento en IA y seguridad para crear entornos fiables y resilientes.

Con Telefónica Tech como partner, las empresas pueden integrar la IA de forma segura, eficiente y alineada con sus objetivos de negocio, avanzando en su transformación digital segura.

Nuestra visión integral asegura tanto la protección como la generación de valor, fortaleciendo la confianza de clientes y la resiliencia operativa en un entorno competitivo, cambiante y regulado.

El viaje hacia una IA segura ya ha comenzado: recorrerlo con confianza marcará la diferencia entre experimentar y transformar.

Telefónica Tech