KPMG

AI Cyber Frontier

Nuestra vision sobre la ciberseguridad en los entornos de IA



¿Son seguros y están bien gobernados nuestros entornos de Inteligencia Artificial?

KPMG. Make the Difference.

Noviembre 2026



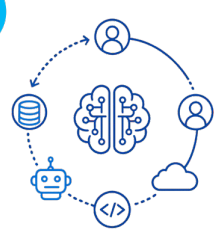
La IA no es solo una herramienta tecnológica, sino que actualmente es uno de los pilares clave dentro del panorama de todas las empresas

1



La IA está dejando de ser un elemento “experimental” para integrarse en **procesos críticos** de negocio, operación, atención al cliente, desarrollo tecnológico y toma de decisiones.

2



Su adopción introduce **nuevos retos** al combinar datos, modelos, identidades humanas y no humanas, agentes, APIs, terceros y decisiones automatizadas en un mismo ecosistema.

3



El **perímetro tradicional** de seguridad **deja de ser suficiente**: ahora también deben protegerse modelos, prompts, datos de contexto, conectores, acciones ejecutadas por agentes y resultados generados.

4



El **riesgo emergente no se limita a ciberseguridad**: impacta en privacidad, cumplimiento normativo, gobierno del dato, continuidad operativa, reputación y responsabilidad corporativa.

5



Las organizaciones necesitan evolucionar desde un enfoque **reactivo** hacia un **modelo de gobierno y control de la IA desde el diseño.**

La IA debe ser entendida como una capa transversal y estratégica para toda la organización, no como una aplicación o un sistema más.

Y a la vez que permite impulsar capacidades de los equipos, trae consigo una serie de retos que hay que abordar de forma ágil y coordinada

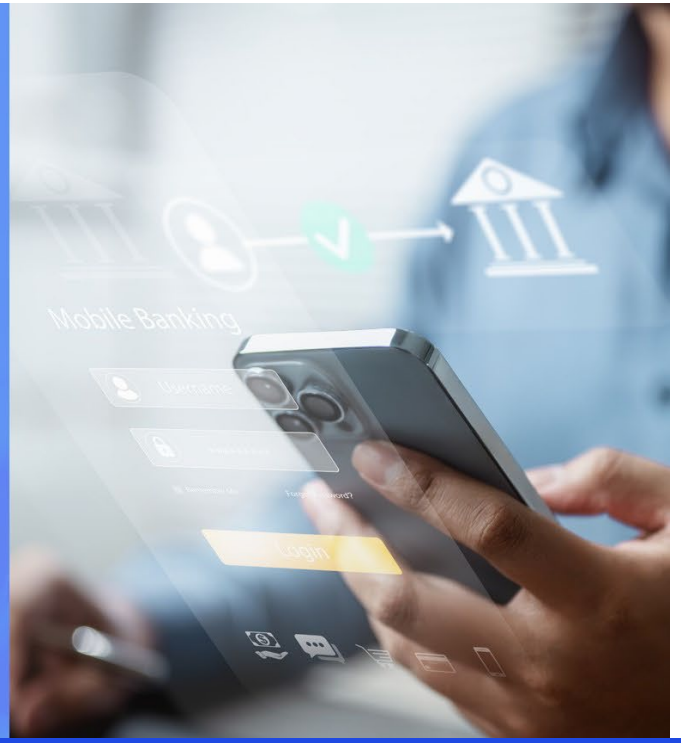
1. **Mayor velocidad y escala:**
   La IA permite ejecutar análisis, decisiones y acciones a una escala que multiplica el impacto potencial de errores, filtraciones o configuraciones inseguras.

2. **Nuevas superficies de ataque:**
   Aparecen vectores específicos como inyección de prompts, extracción de información sensible, manipulación de contexto, abuso de APIs o toma de control de agentes.

3. **Identidades más complejas:**
   Además de usuarios y cuentas técnicas, surgen agentes de IA capaces de interactuar con sistemas, consumir datos, invocar servicios y ejecutar tareas.

4. **Menor trazabilidad tradicional:**
   No siempre es sencillo explicar por qué un modelo ha generado una respuesta, tomado una decisión o activado una acción concreta.

5. **Frontera difusa entre recomendación y ejecución:**
   La IA puede pasar de asistir a operar, lo que complica la asignación de responsabilidad, supervisión y control.



La IA acelera decisiones y operaciones, pero de la misma manera amplía la superficie de exposición, la complejidad de control y la dificultad de supervisión.

Así como un conjunto de riesgos específicos de seguridad, ya que pasamos de proteger sistemas a proteger decisiones autónomas

1



Inyección de prompts y manipulación de contexto

Atacantes o usuarios internos pueden inducir al modelo a **ignorar instrucciones, revelar información, alterar respuestas o activar comportamientos no previstos.**

2



Fuga de información sensible

Datos personales, secretos, código, documentación interna o información de negocio pueden exponerse a través de prompts, respuestas, logs, **integraciones** o sistemas de recuperación de información.

3



Accesos indebidos mediante agentes

Los agentes de IA pueden heredar **permisos excesivos, invocar APIs,** consultar repositorios, operar sobre sistemas o encadenar acciones sin una **validación clara** de identidad, propósito y alcance.

4



Dificultad para validar identidades agenticas

No siempre está claro qué **agente actúa**, en nombre de quién, **con qué permisos**, bajo qué política y con qué **trazabilidad verificable**.

5



Manipulación de datos y conocimiento

Modelos, fuentes documentales, **bases vectoriales** o sistemas de recuperación pueden contaminarse o alterarse, afectando a la calidad, integridad y seguridad de las **decisiones generadas**.

Al hablar de IA, el riesgo ya no reside sólo en la infraestructura: también vive en el comportamiento del modelo, el contexto que recibe y las acciones que puede desencadenar.



La IA también trabaja para el atacante

Las amenazas ya no se reconocen por la mala ortografía: la IA permite fraudes hiperpersonalizados, creíbles y a escala. Reconocerlas es la primera línea de defensa.

1



Phishing generado por IA

Correos y mensajes **sin errores**, personalizados con datos públicos (LinkedIn, web corporativa) y generados en segundos, a escala.

Señales de alerta

- Urgencia o presión inusual
- Enlaces o adjuntos inesperados
- Dominios o remitentes casi idénticos

Qué hacer: No hacer clic: verificar por otro canal y reportar a seguridad.

2



Deepfakes de voz y video

Bastan **unos segundos de audio** para clonar una voz; las videollamadas pueden mostrar rostros sintéticos de ejecutivos reales.

Señales de alerta

- Petición fuera del proceso habitual
- Voz o gestos poco naturales
- Rechazo a cambiar de canal

Qué hacer: Hacer una pregunta de control o devolver la llamada a un número oficial.

3



Suplantación de identidad

Fraude del CEO o del proveedor: se combinan correo, WhatsApp y voz clonada para pedir pagos, cambios de cuenta o accesos.

Señales de alerta

- Cambio de datos bancarios
- “Es confidencial, no lo comentes”
- Saltarse aprobaciones

Qué hacer: Doble validación y regla de “cuatro ojos” en pagos y accesos.

4



Contenidos falsos

Noticias, documentos, capturas o informes **fabricados para manipular decisiones** o reputación; incluye las “alucinaciones” de la propia IA.

Señales de alerta

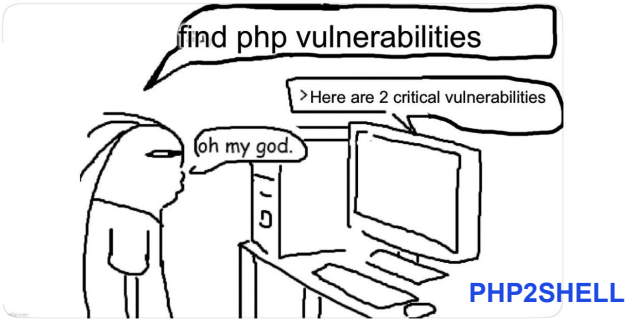
- Fuente no verificable
- Cifras o citas sin respaldo
- Busca una reacción emocional inmediata

Qué hacer: Contrastar con fuentes oficiales antes de compartir o decidir.

Caso real (Hong Kong, 2024): un empleado de Arup transfirió ~US$25 millones tras una videollamada en la que todos los “directivos” eran deepfakes

Y a una velocidad inédita: la IA acelera el ataque más rápido de lo que evoluciona el control

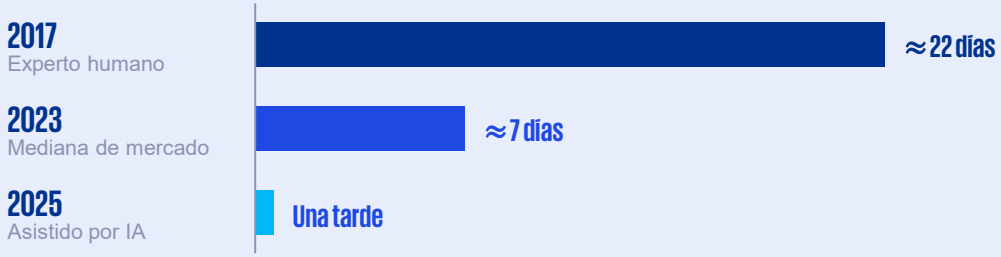
ARTEX



VELOCIDAD DEL ATAQUE

Del parche al exploit: de semanas a horas

Tiempo desde la publicación de un parche hasta un exploit funcional



18 h

tarda hoy un atacante que usa IA en completar un ciberataque típico, desde que entra hasta que causa el daño. Sin IA, ese mismo ataque tomaba **~4 semanas** (Bain)

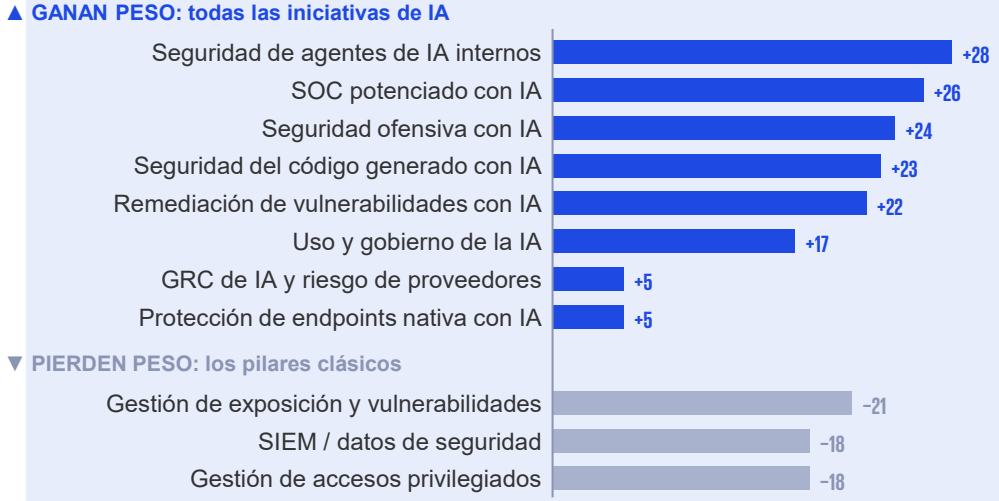
≈37x

más rápido: menos margen para **detectar, decidir y contener** antes del impacto

PRIORIDADES DEL CISO

La agenda del CISO para 2027 gira hacia la IA

Variación 2026 → 2027 en % de CISOs que la sitúan en su top 5 (puntos porcentuales)



Además, las **7 iniciativas más votadas para 2027** involucran IA.

Si el ataque se mide en horas, el control no puede medirse en trimestres: la gobernanza de la IA debe operar a su misma velocidad

Fuentes: Bain & Company; a16z – A. Ezra y S. Zhang, «AI and the patch gap»; RAND (2017); Mandiant (2023); PlatformSecurity (2025); a16z CISO Survey (N=100, empresas de +1.000 empleados, oct. 2026).

De la preocupación al control: hacia un enfoque integral de seguridad en IA

LO QUE DEBE PREOCUPAR

72% de las organizaciones perciben un incremento del riesgo de ciberseguridad

Fuente: World Economic Forum

13% reportaron brechas en modelos o sistemas de IA

97% de ellas carecía de controles adecuados de acceso en IA

Fuente: IBM

Prioridades para equipos de ciberseguridad

- prompts
- conectores
- APIs
- agentes
- datos de contexto
- identidades no humanas

Prioridades equipos de riesgos y cumplimiento

- licitud
- minimización
- trazabilidad
- Inventario
- evidencias de control
- diligencia demostrable

Preocupación



Control

CÓMO RESPONDER



A) Punto de partida pragmático

identificar usos reales de IA, clasificarlos por criticidad, entender qué datos consumen, qué identidades intervienen y qué acciones pueden ejecutar.

B) Identidad como capa de confianza

quién actúa, qué agenta interviene, qué permisos utiliza, que datos consume, que sistemas invoca y qué acciones ejecuta.

C) Controles aplicables y evidenciables

gobierno, diseño seguro, privacidad, IMA, monitorización, pruebas, respuesta ante incidentes y supervisión continua.

La solución no consiste en frenar la IA o paralizar su uso por parte de los negocios, sino en convertir su adopción en un programa gobernable, seguro y trazable.

Principales ámbitos de actuación en ciberseguridad

Nuestro portfolio de capacidades entorno a la seguridad de la IA, se define como un roadmap de soporte para nuestros clientes encaminado a evaluar, establecer y aplicar marcos de gobierno, controles, modelos operativos y tecnológicos para lograr una IA segura y confiable.

Framework de control y operación del modelo

Desarrollar un marco de controles de seguridad IA y operativizarlo para proteger los sistemas de IA durante todo su ciclo de vida y anticipar riesgos desde el diseño.

- AI Secbydesign
- Controles integrales
- Gestión del ciclo de vida

Seguridad en la plataforma e infraestructura

Protección del entorno de IA aplicando controles alineados a riesgos para garantizar integridad, disponibilidad y escalabilidad de la plataforma e infraestructura.

- Escalabilidad segura
- Protección continua
- Alta disponibilidad

Protección de las identidades no humanas

Visión y control sobre identidades no humanas y agentes de IA, identificando relaciones, dependencias y riesgos reales como base para definir accesos basados en contexto.

- Descubrimiento e inventario
- Gobierno del ciclo de vida
- Authority-in-execution

Seguridad en el uso del modelo

Protección y supervisión del comportamiento del modelo mediante guardrails, trazabilidad y detección de anomalías para prevenir abusos y garantizar control en tiempo real.

- Guardrails activos
- Supervisión continua
- Prevención adversaria

Testing y pruebas continuas de modelos

Evaluación de la resiliencia de los modelos frente a pruebas manuales y automatizadas, simulando ataques y condiciones críticas para garantizar su seguridad y robustez.

- Robustez
- Validación automatizada
- Fiabilidad operativa

Formación y concienciación

Sensibilización y capacitación sobre riesgos, oportunidades y buenas prácticas en el uso de la IA, abordando conceptos clave, casos de uso, medidas preventivas y guías de uso responsable.

Un enfoque estructurado para convertir la ambición de IA en una capacidad segura, gobernada y escalable

La adopción de la IA no es solo una cuestión de tecnología. Requiere **gobernanza, seguridad, talento, alianzas y una hoja de ruta clara** para generar valor sostenible y gestionar los riegos de manera proactiva.

Antes de empezar, necesitamos responder las preguntas críticas

- ¿En qué procesos y casos de uso la IA puede **maximizar valor** sin comprometer el perfil de riesgo y la seguridad del negocio?
- ¿Qué **riesgos de seguridad**, **privacidad y cumplimiento** debemos identificar y mitigar desde el diseño?
- ¿Quién define la **gobernanza y ownership** sobre los sistemas de IA, incluyendo decisiones críticas, riesgos y resultados?
- ¿Cómo industrializamos el paso de **piloto a producción** asegurando controles, resiliencia y cumplimiento continuo?

Nuestro punto de partida



La IA no es un proyecto. Es una capacidad que debe de ser segura, trazable y escalable desde el primer día.

Madurez en la gobernanza de la IA: ¿dónde estamos?

Cinco niveles para ubicar a la organización y definir el siguiente paso: de un uso espontáneo a una IA gobernada, medible y auditable.

El punto de partida hoy

13% de las organizaciones reportaron brechas en modelos o sistemas de IA

97% de ellas carecía de controles adecuados de acceso en IA

Fuente: IBM

¿En qué nivel está su organización? Avanzar un nivel cada 6–12 meses es una meta realista.



NIVEL 1 Inicial

Uso espontáneo e individual; Shadow AI; sin inventario, políticas ni responsables definidos.

NIVEL 2 Consciente

Primeras guías de uso, herramientas aprobadas y formación básica; riesgos identificados de forma parcial.

NIVEL 3 Definido

Política formal y comité de IA; inventario de casos de uso clasificados por riesgo y criticidad.

NIVEL 4 Gestionado

Controles técnicos (DLP, guardrails, IAM), monitorización, KPIs/KRIs y auditorías periódicas.

NIVEL 5 Optimizado

Gobierno integrado al negocio, testing adversarial, mejora continua y cumplimiento demostrable (ISO/IEC 42001).

La IA se adopta más rápido de lo que evolucionan los controles: el objetivo es subir de nivel sin frenar al negocio

Políticas a definir y cómo proteger la información

La gobernanza se materializa en políticas claras, controles técnicos y cumplimiento del marco normativo aplicable.

8 políticas que toda organización debería definir

1. **Uso aceptable de IA**
   Qué se puede hacer, con qué herramientas y con qué límites.
2. **Clasificación de la información**
   Qué datos pueden usarse en cada tipo de herramienta.
3. **Catálogo de herramientas aprobadas**
   Proceso para evaluar, aprobar y retirar soluciones de IA.
4. **Supervisión humana**
   Revisión obligatoria y responsables por resultados y decisiones.
5. **Terceros y proveedores de IA**
   Cláusulas de uso de datos, ubicación y no entrenamiento.
6. **Desarrollo seguro de IA**
   Seguridad desde el diseño en modelos, agentes y conectores.
7. **Identidades y accesos**
   Mínimo privilegio para personas y agentes de IA.
8. **Respuesta a incidentes de IA**
   Cómo detectar, reportar y contener fugas o abusos.

¿Cómo proteger la información?



- **Clasificar y etiquetar:** sensibilidad automática en documentos y correos.
- **DLP para IA:** bloqueo de prompts y cargas con información sensible.
- **Control de Shadow AI:** descubrir y bloquear apps de IA no autorizadas.
- **Acceso condicional:** mínimo privilegio para personas y agentes.
- **Trazabilidad:** registro de prompts, respuestas y accesos de riesgo.
- **Cifrado y anonimización** del dato en todo su ciclo de vida.

Marco normativo a considerar

Ley 21.719 · Protección de Datos Personales: vigente desde el 1 de diciembre de 2026, con una nueva Agencia de Protección de Datos.

Ley 21.663 · Marco de Ciberseguridad: gestión de riesgos y reporte de incidentes ante la ANCI.

Referentes internacionales: AI Act (UE), ISO/IEC 42001 y NIST AI RMF.

Gobernar la IA no es frenarla: es convertir su adopción en un programa gobernable, seguro y trazable

KPMG

MAKE AI SECURE



KPMG. Make the Difference.