DOCUMENTO 35 · SEGURIDAD DE IA Y AGENTES

# Seguridad de IA y agentes

Amenazas a los sistemas de IA, controles para agentes con capacidad de actuar y exposición a la IA ofensiva

**Fernando García Varela** Versión 0.6 (borrador de trabajo) 19-11-2026 Borrador para revisión

| 4 | 28 | 20 | 9 |
|---|---|---|---|
| niveles de autonomía | controles de seguridad de IA (SEG) | controles de agentes (AG) | requisitos esenciales de un agente |

EN ESTE DOCUMENTO


01 Objeto y alcance

02 Principios de seguridad

03 Amenazas a los sistemas de IA

04 Requisitos esenciales de un agente

05 Niveles de autonomía

06 Catálogo de controles de seguridad de IA (SEG)

07 Catálogo de controles de agentes (AG)

08 Pruebas de seguridad por fase

09 Exposición a la IA ofensiva y uso de la IA en la ciberdefensa

10 Indicadores para el consejo

11 Integración con el ciclo de vida

12 Herramientas y plantillas asociadas

13 Documentos relacionados

14 Control de versiones



**Aviso legal y exención de responsabilidad.** SEVEN-G es un marco metodológico de referencia que se ofrece «tal cual» y con fines exclusivamente informativos. No constituye asesoramiento jurídico, regulatorio, financiero ni profesional, ni garantiza el cumplimiento de ninguna norma. Las referencias a regulación general (como el Reglamento Europeo de IA, el RGPD, DORA o NIS2), a normas técnicas y a regulación específica de cada sector o jurisdicción pueden ser incompletas, no aplicar a un caso concreto o quedar desactualizadas por cambios normativos, interpretaciones o criterios de las autoridades posteriores a su fecha de consulta. **Cada organización que use SEVEN-G es la única responsable de identificar la normativa que le aplica, verificar su vigencia y certificar su propio cumplimiento regulatorio**, con el asesoramiento cualificado que corresponda. Esta metodología es una ayuda genérica y gratuita, compartida con la comunidad para que nadie tenga que empezar desde cero; cada persona u organización puede y debe adaptarla a su propio uso. No debe entenderse que sus partes jurídicamente sensibles hayan sido revisadas por una asesoría jurídica: esas revisiones, para cada empresa o sector, son responsabilidad última de la empresa, el consultor o la organización que la use. Aunque se procura mantenerla al día, alguna norma puede haber cambiado sin que se recoja aquí. En la máxima medida permitida por la ley, el autor no asume responsabilidad alguna por los efectos de su aplicación en ninguna organización ni por su aplicabilidad completa. La metodología no otorga certificación de ningún tipo. El autor no asume responsabilidad alguna por el uso que se haga de este contenido ni por las decisiones que se adopten con él. Los datos, cifras, compañías y casos de los ejemplos son ficticios o ilustrativos.


**LO ESENCIAL** **Condicional: depende de un disparador**

Disparador: IA generativa, agentes (autonomía A1 a A3) o exposición directa a personas externas. El diseño de seguridad (P18) es evidencia de todo caso de uso; los controles de agentes, las pruebas adversarias y los mínimos por nivel de autonomía aplican según el disparador.

Matriz de obligatoriedad y lectura por capas (documento 94)

01

# Objeto y alcance

Este documento establece cómo se protege un sistema de IA frente a ataques y usos indebidos, qué controles adicionales exige un agente según su nivel de autonomía y cómo se gestiona la exposición de la compañía a ataques que usan IA. Desarrolla la sección 10 del documento 01 y los riesgos tipo RT-GEN y RT-SEG del documento 33.

## 1.1 Alcance

| Incluye | No incluye |
|---|---|
| Sistemas de IA propios y de terceros integrados en procesos, incluidos asistentes, sistemas de recuperación de información y agentes. | La seguridad de la información general de la compañía, que se rige por su propio sistema de gestión; este documento la complementa para la IA. |
| Uso corporativo de IA de propósito general en lo que afecta a fuga de información. | La política de uso aceptable (documento 31). |
| Exposición corporativa a la IA ofensiva: suplantación, fraude, explotación acelerada; y uso de la IA en la ciberdefensa de la compañía (sección 9.3). | La gestión de incidentes, que se desarrolla en el documento 37. |

## 1.2 Referencias

Las referencias se han consultado en septiembre de 2026 y deben verificarse en su versión vigente:

- **OWASP Top 10 for LLM Applications**, versión 2025 (riesgos LLM01 a LLM10). El OWASP GenAI Security Project publicó la edición 2026 el 3 de agosto de 2026, con otra numeración: los códigos LLM01 a LLM10 de este documento son los de la edición 2025.
- **OWASP Top 10 for Agentic Applications** (OWASP GenAI Security Project, publicado el 9 de diciembre de 2025; riesgos ASI01 a ASI10) y la guía **Agentic AI – Threats and Mitigations** del mismo proyecto.
- **MITRE ATLAS**, base de conocimiento de tácticas y técnicas adversarias contra sistemas de IA, que se actualiza periódicamente.
- **NIST AI 600-1**, perfil de IA generativa del NIST AI RMF (julio de 2024), en particular los riesgos de seguridad de la información, privacidad de datos e integración de la cadena de valor.
- **NIST CSF 2.0** (NIST CSWP 29, febrero de 2024), para expresar los controles con sus seis funciones (gobernar, identificar, proteger, detectar, responder y recuperar), y su perfil para la IA, el **Cyber AI Profile** (NIST IR 8596), que a fecha de consulta es un **borrador** preliminar (diciembre de 2025) y se usa solo como orientación (34 §5.3).

- **ISO/IEC 42001** e **ISO/IEC 23894** para el encaje en el sistema de gestión y en la gestión de riesgos.
- **Reglamento (UE) 2024/1689**, que exige a los sistemas de alto riesgo un nivel adecuado de precisión, solidez y ciberseguridad, incluida la resistencia a la manipulación de datos y de modelos y a entradas diseñadas para inducir errores.
- **Reglamento (UE) 2024/2847 de ciberresiliencia**, para quien fabrica productos con elementos digitales que incorporan IA: notificación de vulnerabilidades explotadas activamente e incidentes graves desde el 11-09-2026 (art. 14) y obligaciones principales desde el 11-12-2027; se integra con SEG-13, SEG-14 y el documento 37 (34 §7.2).
- **CCN-CERT BP/36**, guía de buenas prácticas frente a la IA ofensiva del Centro Criptológico Nacional (junio de 2026).

Este documento no constituye asesoramiento jurídico.

02

# Principios de seguridad

| # | Principio | Qué implica |
|---|---|---|
| 1 | **El modelo no es una frontera de seguridad** | Las instrucciones al modelo ayudan, pero no protegen. Los límites que importan (permisos, importes, destinatarios, datos) se aplican fuera del modelo. |
| 2 | **Todo contenido externo es no confiable** | Correos, documentos, webs, respuestas de herramientas y de otros agentes son datos, nunca instrucciones. |
| 3 | **Mínimo privilegio y mínima autonomía** | Cada sistema tiene las herramientas, datos y autonomía imprescindibles, y se justifican en la fase 4. |
| 4 | **Cada acción responde a una intención autorizada** | Ningún agente ejecuta una acción que no pueda trazarse a una solicitud o mandato legítimo. |
| 5 | **Una persona valida lo que no se puede deshacer** | Las acciones sensibles requieren validación humana informada. |
| 6 | **Siempre se puede parar** | Todo sistema con capacidad de actuar tiene interruptor de parada probado. |
| 7 | **Se asume el compromiso** | El diseño limita el daño si el modelo es manipulado o una credencial es robada. |
| 8 | **Se prueba como atacaría un adversario** | La eficacia de los controles se demuestra con pruebas adversarias, no con declaraciones. |

03

# Amenazas a los sistemas de IA

## 3.1 Amenazas principales

| Amenaza | Descripción | Referencia | Riesgo tipo (33) | Controles |
|---|---|---|---|---|
| **Inyección directa de instrucciones** | El usuario introduce instrucciones que alteran el comportamiento previsto. | OWASP LLM01; ATLAS | RT-GEN-01 | SEG-02, SEG-03, SEG-11 |
| **Inyección indirecta de instrucciones** | Instrucciones ocultas en contenido que el sistema lee (documentos, correos, webs, respuestas de herramientas, voz) desvían la tarea. | OWASP LLM01; ASI01 | RT-GEN-02 | SEG-02, AG-05, AG-12, AG-18 |
| ***Jailbreak*** | Técnicas para que el modelo ignore sus salvaguardas y produzca contenido o acciones prohibidas. | OWASP LLM01; ATLAS | RT-GEN-01, RT-REP-04 | SEG-03, SEG-11 |
| **Envenenamiento** | Manipulación de datos de entrenamiento, ajuste, evaluación, bases de conocimiento o memoria. | OWASP LLM04, LLM08; ASI06 | RT-SEG-05, RT-GEN-07 | SEG-08, AG-14 |
| **Extracción** | Consultas diseñadas para replicar el modelo, inferir datos de entrenamiento o reconstruir instrucciones del sistema. | OWASP LLM02, LLM07; ATLAS | RT-SEG-06 | SEG-05, SEG-10, SEG-12 |
| **Fuga de información** | Datos personales, confidenciales o secretos expuestos en respuestas, registros, recuperación o herramientas. | OWASP LLM02, LLM07, LLM08; NIST AI 600-1 | RT-GEN-05, RT-DAT-05 | SEG-05, SEG-06, SEG-07 |
| **Tratamiento inseguro de salidas** | La salida se ejecuta o se inserta en otros sistemas sin validar (código, consultas, enlaces). | OWASP LLM05; ASI05 | RT-GEN-06 | SEG-04, AG-11 |
| **Abuso de herramientas** | El sistema usa herramientas legítimas de forma dañina por manipulación o por exceso de permisos. | OWASP LLM06; ASI02, ASI03 | RT-GEN-03, RT-GEN-04 | AG-02, AG-05, AG-07, AG-08 |
| **Cadena de suministro de modelos** | Modelos, pesos, bibliotecas, conectores o servidores de herramientas manipulados o vulnerables. | OWASP LLM03; ASI04; NIST AI 600-1 | RT-SEG-07 | SEG-09, AG-13 |
| **Consumo sin límite** | Uso masivo que agota recursos o dispara costes. | OWASP LLM10 | RT-GEN-08, RT-ECO-02 | SEG-10, AG-16 |

| Amenaza | Descripción | Referencia | Riesgo tipo (33) | Controles |
|---|---|---|---|---|
| **Desinformación del propio sistema** | Contenidos falsos presentados como ciertos que llevan a decisiones erróneas. | OWASP LLM09; NIST AI 600-1 | RT-TEC-04 | SEG-11; controles de fuentes (33) |

## 3.2 Amenazas específicas de agentes

Un agente planifica, usa herramientas, conserva memoria y actúa con autoridad delegada. Eso añade amenazas que no existen en un asistente que solo responde:

| Amenaza | Qué ocurre | Referencia | Controles |
|---|---|---|---|
| **Secuestro del objetivo** | El agente persigue un objetivo distinto del encomendado. | ASI01 | AG-04, AG-05, AG-12 |
| **Abuso de identidad y privilegios** | El agente usa credenciales o permisos heredados más allá de su tarea; un atacante los reutiliza. | ASI03 | AG-01, AG-02, AG-03, AG-20 |
| **Envenenamiento de herramientas y conectores** | La descripción de una herramienta o de un conector lleva instrucciones ocultas que el agente obedece; una herramienta cambia de descripción o de comportamiento después de aprobarse; un conector pide ámbitos o permisos mayores que su función; o una persona instala servidores de herramientas sin control. | ASI02, ASI04 | AG-13, AG-02, AG-12, SEG-27 |
| **Ejecución inesperada de código** | El agente genera o ejecuta código o comandos con efectos no previstos. | ASI05 | AG-11, SEG-04 |
| **Envenenamiento de memoria y contexto** | Contenido manipulado persiste y condiciona tareas posteriores. | ASI06 | AG-14 |
| **Comunicación insegura entre agentes** | Mensajes falsificados o no autenticados entre agentes. | ASI07 | AG-15 |
| **Fallos en cascada** | Un error se propaga entre agentes o se amplifica en bucles. | ASI08 | AG-07, AG-16 |
| **Explotación de la confianza humana** | El agente induce a una persona a aprobar algo inadecuado. | ASI09 | AG-08 |
| **Agentes descontrolados** | Un agente comprometido o desviado sigue operando. | ASI10 | AG-09, AG-17 |

04

# Requisitos esenciales de un agente

Todo sistema clasificado como A2 o A3, y en lo que corresponda A1 (sección 5), **debe** cumplir los nueve requisitos siguientes. Los controles AG de la sección 7 los concretan.

## 4.1 Identidad propia

- El agente **debe** operar con una identidad no humana propia, distinta de la de cualquier empleado y de la de otros agentes (AG-01).
- La identidad se registra en el inventario de identidades no humanas (P54) con su sistema, finalidad y **responsable humano**.
- Cuando actúa en nombre de un usuario, la acción se registra con las dos identidades: la del agente y la del usuario por cuya cuenta actúa. El agente no puede obtener más permisos que el usuario al que representa.

## 4.2 Permisos mínimos

- Lista cerrada de herramientas permitidas y, para cada una, operaciones y datos permitidos (AG-02). Lectura y escritura se autorizan por separado.
- Los permisos se justifican en el diseño de seguridad (P18) y se revisan periódicamente (AG-20).
- Separación de privilegios: el componente que lee contenido no confiable no dispone, sin control intermedio, de herramientas que ejecutan acciones sensibles (AG-12).

## 4.3 Credenciales

- Secretos en un gestor de secretos; nunca en instrucciones, código, memoria ni registros (AG-03).
- Credenciales de corta duración y ámbito limitado; rotación con plazo definido; revocación inmediata vinculada al interruptor de parada.

## 4.4 Control de intención

El control de intención garantiza que **cada acción del agente es trazable a una intención autorizada**. Funciona así:

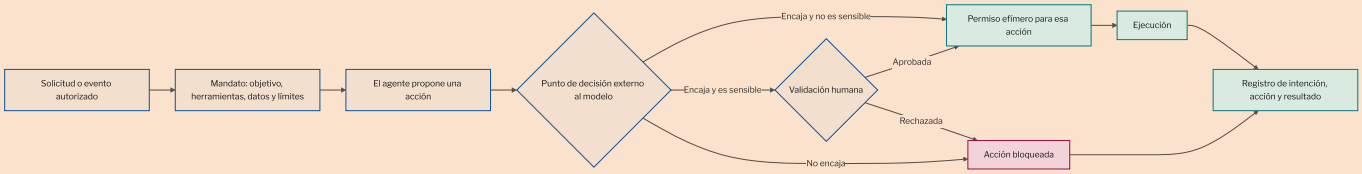
1. **Mandato.** Cada tarea tiene una intención autorizada: quién la solicita o qué evento la origina, objetivo, herramientas, datos y límites (AG-04).
2. **Punto de decisión externo al modelo.** Antes de cada acción, un componente independiente del modelo comprueba que la acción concreta (herramienta, parámetros, datos, destinatario, importe) encaja con la intención y los límites (AG-05).

3. **Permiso efímero.** Si encaja, emite un permiso de corta duración limitado a esa acción. Si no encaja, la bloquea. Si es sensible, la somete a validación humana.
4. **Trazabilidad.** La acción queda registrada con el identificador de la intención que la autorizó (AG-06, AG-10).

---

## Control de intención de un agente

Ninguna acción se ejecuta sin encajar con una intención autorizada




Fuente: SEVEN-G

### 4.5 Límites de actuación

Límites definidos en el diseño y aplicados fuera del modelo (AG-07): importes por operación y acumulados, número de acciones por periodo, destinatarios o dominios permitidos, horarios, volumen de datos, número de iteraciones y presupuesto de consumo. Superar un límite detiene la secuencia y genera alerta.

### 4.6 Validación humana en acciones sensibles

Se consideran **acciones sensibles**, salvo justificación registrada en P17 y P18:

- Pagos, transferencias, devoluciones o cambios de datos bancarios.
- Comunicaciones externas en nombre de la compañía a clientes, proveedores, autoridades o público.
- Decisiones con efectos significativos sobre personas (empleo, crédito, seguros, acceso a servicios, reclamaciones).
- Alta, baja o modificación de datos personales o de permisos.
- Cambios en sistemas de producción, configuraciones de seguridad o código desplegado.
- Borrado de información o acciones no reversibles.
- Cualquier acción fuera de los límites ordinarios.

La validación **debe** ser informada: la persona ve qué acción se propone, con qué datos, por qué y con qué efectos, y puede rechazarla sin coste para ella. Se mide la tasa de rechazo y el tiempo de revisión para detectar aprobación rutinaria (AG-08, RT-ORG-04).

**Derivación inmediata a una persona.** En los agentes y asistentes con exposición directa (01 §9.2), la conversación **debe** pasar de inmediato a una persona cuando quien la usa lo pide expresamente o

cuando aparecen señales de vulnerabilidad: indicios de que es menor o de que tiene dificultades por la edad, confusión sobre lo que se le propone, duelo, angustia o cualquier situación en la que seguir con el sistema pueda perjudicarle. La derivación no exige justificar la petición ni repetirla y no se desactiva por coste (52 §10.3). Las señales y el canal se fijan en P17 y el texto que se muestra, en P49. Cada derivación se registra (motivo, momento y tiempo hasta que atiende una persona) y se mide (derivaciones por motivo, peticiones no atendidas) para revisar en R6 si las señales están bien calibradas. La regulación de cada sector o jurisdicción puede exigir más.

> **Por qué importa.** Un sistema que atiende a personas externas encontrará a alguien que no está en condiciones de tratar con una máquina o que no quiere hacerlo. Retenerle en la conversación para contener costes convierte una eficiencia en un daño a una persona y en un riesgo reputacional. Derivar sin fricción, y medir cuándo se deriva, protege a la persona y dice si el sistema está bien diseñado.

### 4.7 Interruptor de parada

- Mecanismo para detener de inmediato el agente completo o una capacidad concreta (una herramienta, un tipo de acción) **sin desplegar código** (AG-09).
- Activable por el responsable de operación de IA y por seguridad, con procedimiento en el manual de operación (P24).
- Al activarse: revoca credenciales, bloquea nuevas acciones, conserva registros y activa el proceso alternativo.
- Se prueba antes de G5 y periódicamente (sección 5.3). Su tiempo de efecto se mide.

### 4.8 Registro de acciones

Registro íntegro y protegido frente a modificación (AG-10) con, como mínimo: identidad del agente y del usuario representado, intención, acción propuesta, decisión del punto de decisión, validación humana si la hubo, herramienta, parámetros relevantes, resultado, versión de modelo e instrucciones, fecha y hora. La conservación respeta la minimización de datos personales y los plazos regulatorios aplicables; para desplegadores de sistemas de alto riesgo, el Reglamento de IA exige conservar los registros generados automáticamente al menos seis meses, salvo otra disposición.

### 4.9 Entornos aislados

La ejecución de código, la navegación web, la manipulación de ficheros externos y las pruebas se realizan en entornos aislados, con salida de red restringida, sin credenciales de producción y con

destrucción al terminar (AG-11).

05

# Niveles de autonomía

## 5.1 Definición

| Nivel | Nombre | Qué hace el sistema | Papel humano |
|---|---|---|---|
| **A0** | Asistencia | Informa, resume o genera contenido. | La persona decide y ejecuta. |
| **A1** | Recomendación | Propone una decisión o acción concreta. | La persona valida cada acción antes de ejecutarla. |
| **A2** | Actuación supervisada | Ejecuta acciones dentro de límites definidos. | Supervisa, puede interrumpir y revisa a posteriori. |
| **A3** | Actuación autónoma | Ejecuta secuencias de acciones sin revisión individual dentro de límites estrictos. | Fija límites, supervisa agregados y dispone de interruptor de parada. |

## 5.2 Reglas de asignación

1. El nivel se propone en la fase 3, se fija en la fase 4 (P17 y P18) y se registra en el inventario (T02). Se asigna el **nivel más alto** de las acciones que el sistema puede ejecutar, no el de su uso habitual.
2. Un sistema A2 o A3 cuyas acciones tienen efecto sobre terceros, dinero, datos personales o sistemas de producción cumple el criterio Enterprise "agentes con capacidad de actuar" (01 §9.2).
3. **A3** requiere decisión del comité de IA en G4 y en G5, con justificación de por qué A2 no basta.
4. Subir de nivel es un cambio relevante: exige revaloración de riesgos, actualización de P18 y nueva decisión del *gate* correspondiente.
5. En caso de incidente S1 o S2 relacionado con la autonomía, el sistema baja provisionalmente de nivel o se detiene hasta el análisis posterior (documento 37).

## 5.3 Controles mínimos por nivel

**Sí** = obligatorio · **Rec.** = recomendado; su omisión se justifica y registra · **—** = no aplica.

| Control | A0 | A1 | A2 | A3 |
|---|---|---|---|---|
| **AG-01 Identidad propia** | Rec. | Sí | Sí | Sí |

| Control | A0 | A1 | A2 | A3 |
|---|---|---|---|---|
| **AG-02 Mínimo privilegio** | Sí | Sí | Sí | Sí |
| **AG-03 Gestión de credenciales** | Sí | Sí | Sí | Sí |
| **AG-04 Mandato e intención autorizada** | — | Rec. | Sí | Sí |
| **AG-05 Punto de decisión de intención** | — | — | Sí | Sí |
| **AG-06 Trazabilidad acción–intención** | — | Sí | Sí | Sí |
| **AG-07 Límites de actuación** | — | Rec. | Sí | Sí |
| **AG-08 Validación humana** | — | Sí (toda acción) | Sí (sensibles) | Sí (sensibles y fuera de límites) |
| **AG-09 Interruptor de parada** | Rec. | Rec. | Sí | Sí |
| **AG-10 Registro de acciones** | Rec. | Sí | Sí | Sí |
| **AG-11 Entornos aislados** | Rec. | Rec. | Sí | Sí |
| **AG-12 Contenido externo no confiable** | Sí | Sí | Sí | Sí |
| **AG-13 Herramientas y conectores aprobados** | Rec. | Sí | Sí | Sí |
| **AG-14 Integridad de memoria y contexto** | Rec. | Rec. | Sí | Sí |
| **AG-15 Comunicación segura entre agentes** | — | Rec. | Sí | Sí |
| **AG-16 Contención de cascadas y consumo** | — | Rec. | Sí | Sí |
| **AG-17 Monitorización de comportamiento** | Rec. | Rec. | Sí | Sí |
| **AG-18 Pruebas adversarias de agentes** | Rec. | Sí | Sí | Sí |
| **AG-19 Reversibilidad y compensación** | — | Rec. | Sí | Sí |
| **AG-20 Revisión periódica de permisos** | Rec. | Sí | Sí | Sí |

AG-01, AG-03 y AG-13 aplican a A0 como obligatorios si el sistema accede a sistemas o datos corporativos con credenciales propias. AG-11 es obligatorio en cualquier nivel si el sistema ejecuta código o navega. AG-15 solo aplica si intervienen varios agentes.

| Frecuencia mínima | A1 | A2 | A3 |
|---|---|---|---|
| **Prueba del interruptor de parada** | — | Antes de G5 y semestral | Antes de G5 y trimestral |
| **Pruebas adversarias (AG-18)** | Antes de G5 y tras cambios relevantes | Antes de G5, anual y tras cambios relevantes | Antes de G5, semestral y tras cambios relevantes |
| **Revisión de permisos (AG-20)** | Anual | Semestral | Trimestral |
| **Revisión de muestras de acciones** | Semestral | Mensual | Continua con alertas y revisión mensual |
| **Banco de pruebas corporativo (SEG-27)** | Antes de G5 y tras cambios relevantes | En cada despliegue | En cada despliegue |

En A0, el banco de pruebas corporativo se ejecuta antes de G5 y tras cambios relevantes, como en A1.

### 5.4 Controles críticos

En A2 y A3 son **controles críticos** AG-01, AG-02, AG-03, AG-05, AG-08, AG-09, AG-10 y AG-12; en A1, AG-02, AG-03, AG-08 y AG-12. Un control crítico no diseñado bloquea G4; no probado, bloquea G5. No se admite **Continuar con condiciones** sobre ellos (01 §7.3). Su desactivación en producción es no conformidad crítica (01 §12).

06

## Catálogo de controles de seguridad de IA (SEG)

*Aplica* indica el alcance: **IA** (todo sistema de IA), **GEN** (IA generativa y agentes), **EXT** (exposición a personas externas), **CORP** (control corporativo, no de un caso de uso). *Función CSF* indica la función del NIST CSF 2.0 a la que contribuye principalmente el control y, entre paréntesis, su categoría (34 §5.3); con ella se construye el perfil de seguridad de la IA desde los controles existentes.

| Código | Control | Qué exige | Evidencia | Aplica | Fase | Función CSF |
|---|---|---|---|---|---|---|
| **SEG-01** | Modelado de amenazas de IA | Análisis de amenazas con OWASP LLM 2025, OWASP Agentic y MITRE ATLAS, vinculado al registro de riesgos. | Modelo de amenazas en P18. | IA | 3, 4 | ID (ID.RA) |

| Código | Control | Qué exige | Evidencia | Aplica | Fase | Función CSF |
|---|---|---|---|---|---|---|
| **SEG-02** | Separación de instrucciones y datos | El contenido externo se delimita y trata como dato; las instrucciones del sistema no pueden sobrescribirse con él. | Diseño y resultados de pruebas de inyección. | GEN | 4, 5 | PR (PR.DS) |
| **SEG-03** | Filtros de entrada y salida | Detección de intentos de manipulación, contenido prohibido y datos sensibles antes y después del modelo. | Configuración, umbrales y tasa de detección. | GEN | 4–6 | PR (PR.DS) · DE (DE.CM) |
| **SEG-04** | Tratamiento seguro de salidas | Las salidas se validan y codifican antes de ejecutarse o insertarse en otros sistemas. | Revisión de integraciones; pruebas. | GEN | 4, 5 | PR (PR.PS) |
| **SEG-05** | Protección de instrucciones del sistema | Sin secretos, credenciales ni lógica de control en las instrucciones; su divulgación no compromete la seguridad. | Revisión de instrucciones; prueba de extracción. | GEN | 4, 5 | PR (PR.PS) |
| **SEG-06** | Recuperación con permisos | La recuperación de información respeta los permisos del usuario; segmentación de índices por nivel de confidencialidad. | Diseño del índice; pruebas con usuarios de distinto perfil. | GEN | 4, 5 | PR (PR.AA) |
| **SEG-07** | Prevención de fuga de datos | Minimización, enmascarado, clasificación de la información y control de datos en entradas, contexto, salidas y registros. | Reglas de prevención de fuga; muestreo de registros. | IA, CORP | 4, 6 | PR (PR.DS) |
| **SEG-08** | Integridad de datos de entrenamiento y conocimiento | Procedencia verificada, control de cambios y detección de anomalías en datos de entrenamiento, ajuste, evaluación y bases de conocimiento. | Linaje (P16); controles de ingesta. | IA | 4, 6 | PR (PR.DS) |
| **SEG-09** | Cadena de suministro de modelos | Inventario de modelos y componentes con versión y procedencia; fuentes aprobadas; verificación de integridad; análisis de vulnerabilidades. | Inventario de componentes (P54); registro de verificación. | IA | 4, 6 | GV (GV.SC) · ID (ID.AM) |
| **SEG-10** | Límites de uso y consumo | Autenticación, cuotas por usuario y sistema, límites de tamaño y frecuencia, detección de patrones de extracción. | Configuración y alertas. | GEN, EXT | 4, 6 | PR (PR.IR) |
| **SEG-11** | Evaluaciones de seguridad y *red teaming* | Pruebas adversarias antes de G5, tras cambios relevantes y periódicamente (sección 8). | Plan, resultados y acciones cerradas. | GEN | 5, 6 | ID (ID.IM) |

| Código | Control | Qué exige | Evidencia | Aplica | Fase | Función CSF |
|---|---|---|---|---|---|---|
| **SEG-12** | Registro y monitorización de seguridad | Telemetría de entradas, salidas, bloqueos y acciones integrada en la monitorización de seguridad de la compañía, con escenarios de detección específicos. | Escenarios de detección; alertas probadas. | IA | 4, 6 | DE (DE.CM) |
| **SEG-13** | Gestión acelerada de vulnerabilidades | Plazos de corrección reducidos para sistemas expuestos y componentes de IA; priorización por explotabilidad. | Plazos aprobados; cumplimiento. | CORP | 6 | ID (ID.RA) · PR (PR.PS) |
| **SEG-14** | Respuesta a incidentes de IA | Procedimientos para inyección, fuga, acción no autorizada y compromiso de agente, conectados con el documento 37. | Plan (P26); simulacro. | IA, CORP | 5, 6 | RS (RS.MA) · RC (RC.RP) |
| **SEG-15** | Autenticación resistente a *phishing* | Doble factor resistente a suplantación en accesos expuestos, privilegiados y de administración de plataformas de IA. | Cobertura medida. | CORP | C4 | PR (PR.AA) |
| **SEG-16** | Verificación fuera de banda | Confirmación por un canal independiente y preestablecido de órdenes de pago, cambios de cuentas bancarias y peticiones urgentes de directivos. | Procedimiento; pruebas de cumplimiento. | CORP | C4 | PR (PR.AA) |
| **SEG-17** | Concienciación sobre suplantación con IA | Formación y simulacros con mensajes, voz y vídeo sintéticos, dirigidos especialmente a finanzas, dirección, atención al cliente y soporte técnico. | Plan; resultados de simulacros. | CORP | C4 | PR (PR.AT) |
| **SEG-18** | Protocolos frente a contenido sintético | Palabras de verificación o preguntas acordadas en canales de voz y vídeo; herramientas de detección como apoyo, no como única barrera. | Procedimiento publicado. | CORP | C4 | PR (PR.AA) · DE (DE.AE) |
| **SEG-19** | Pruebas continuas de la superficie expuesta | Descubrimiento de activos expuestos y pruebas de intrusión recurrentes que incluyan técnicas automatizadas con IA. | Resultados y plazos de corrección. | CORP | C4, 6 | ID (ID.AM, ID.RA) |

| Código | Control | Qué exige | Evidencia | Aplica | Fase | Función CSF |
|---|---|---|---|---|---|---|
| **SEG-20** | Control del uso corporativo de IA | Herramientas aprobadas, bloqueo o supervisión de las no autorizadas y prevención de fuga hacia servicios externos. Incluye los agentes locales en los equipos (asistentes de programación con credenciales, servidores de herramientas instalados por el usuario) y las extensiones del navegador con IA, que se inventarían y se autorizan como cualquier otra herramienta (31 §4.7). Probar una herramienta externa con datos de la compañía también es usarla (36 §2). | Monitor de uso (T21); inventario de agentes locales y extensiones. | CORP | C4 | GV (GV.PO) · PR (PR.DS) |
| **SEG-21** | Autonomía de la respuesta automatizada | Nivel A0–A3 fijado por tipo de acción de contención; A2 o A3 solo con procedimiento aprobado, límites, reversión probada e interruptor de parada (sección 9.3). | Matriz de acciones y niveles (P18); pruebas de reversión (P19). | CORP | 4–6 | RS (RS.MI) |
| **SEG-22** | Validación humana de la contención | Validación informada antes de contener con efecto en producción, en clientes o sobre personas; revisión a posteriori de esas acciones en A2 y A3. | Registro de validaciones y revisiones; tasa de rechazo. | CORP | 6 | RS (RS.MA) |
| **SEG-23** | Calidad de la detección con IA | Falsos positivos y negativos, tiempo de detección y de respuesta frente a una línea base sin IA, con umbrales; análisis de cada falso negativo en un incidente real. | Indicadores y umbrales (P25); análisis de falsos negativos. | CORP | 5, 6 | DE (DE.AE) |
| **SEG-24** | Dependencia del proveedor de ciberdefensa con IA | Evaluación del proveedor, control de la telemetría que sale de la compañía, plan de salida y operación degradada sin IA. | P14, P55, P57; prueba de operación degradada. | CORP | 3, 6 | GV (GV.SC) |
| **SEG-25** | Pruebas adversarias del sistema defensivo | Evasión, envenenamiento de la telemetría o del entrenamiento e inyección de instrucciones en alertas, tickets o registros que lee el sistema. | Plan e informe de pruebas adversarias (P53). | CORP | 5, 6 | ID (ID.IM) |

| Código | Control | Qué exige | Evidencia | Aplica | Fase | Función CSF |
|---|---|---|---|---|---|---|
| **SEG-26** | Pasarela corporativa de IA | Todo el tráfico hacia modelos, propios o de terceros, pasa por un punto de control común fuera del modelo que aplica, con la misma configuración medida para todos los casos de uso: filtros de entrada, de salida y de contexto (SEG-03), prevención de fuga (SEG-07), cuotas y límites (SEG-10), lista de modelos aprobados, registro con identificador de correlación (SEG-12) y, para agentes, el punto de decisión de intención (AG-05). Falla cerrado en acciones sensibles y degradado en consultas. | Arquitectura (P18); configuración común versionada; prueba de que no hay tráfico a modelos fuera de la pasarela; eficacia medida (52 §5.2). | CORP | C4, 4–6 | PR (PR.PS) · DE (DE.CM) |
| **SEG-27** | Banco de pruebas corporativo de agentes | Biblioteca versionada de escenarios adversarios, derivada del modelo de amenazas (SEG-01) y de los incidentes propios, que cubre todos los canales —texto, documentos, correo, web, respuestas de herramientas, memoria, mensajes entre agentes y voz o audio sintéticos—, las herramientas y conectores, y usuarios simulados. Está integrada en la cadena de despliegue: en A2 y A3 se ejecuta en cada despliegue y un escenario crítico con éxito impide desplegar; en A0 y A1, antes de G5 y tras cada cambio relevante. | Biblioteca de escenarios versionada; resultado por despliegue; informe de pruebas adversarias (P53). | CORP | 5, 6 | ID (ID.IM) |
| **SEG-28** | Canal identificado para agentes de terceros | Con exposición directa a clientes, aplicar en cada recorrido la postura fijada en C2 (13 §5.6). En «atender», un canal con identificación verificable del agente, permisos del cliente concretos, revocables y registrados, límites, registro de cada acción y etiquetado del tráfico; exige a los agentes externos lo que AG-01 (identidad propia) y AG-15 (comunicación entre agentes) exigen a los propios. En «tolerar» o «bloquear», detección y etiquetado. Las reglas de elegibilidad y de concesión se aplican fuera del modelo (sección 9.4). | Postura aprobada (13 §5.6); diseño (P18); pruebas con clientes sintéticos (P53). | EXT | C2, 4–6 | PR (PR.AA) · DE (DE.CM) |

**Por qué importa.** La pasarela corporativa de IA (SEG-26) y el banco de pruebas corporativo de agentes (SEG-27) convierten controles que cada equipo construiría a su manera en una capacidad común, medida y comparable: filtros, prevención de fuga y control de intención se configuran una vez y se miden igual para todos los casos de uso, y cada cambio de un agente se prueba contra los mismos ataques antes de llegar a producción. Sin ellos, la seguridad depende de lo que recuerde cada equipo y nadie puede decir al consejo si los guardarraíles funcionan.

07

## Catálogo de controles de agentes (AG)

La lista **LV-AG** del documento 22 convierte cada control en preguntas binarias verificables en G4 (diseñado), G5 (probado) y R6 (operando). Este documento es la referencia de su contenido. *Función CSF* tiene el mismo significado que en la sección 6.

| Código | Control | Qué exige | Evidencia en G4 · G5 · R6 | Función CSF |
|---|---|---|---|---|
| **AG-01** | Identidad propia | Identidad no humana exclusiva, registrada, con responsable humano; doble identidad cuando actúa por cuenta de un usuario. | Registro de identidad · Prueba de trazas · Inventario vigente | PR (PR.AA) |
| **AG-02** | Mínimo privilegio | Lista cerrada de herramientas, operaciones y datos; lectura y escritura separadas; permisos justificados. | Matriz de permisos · Prueba de acceso denegado · Revisión (AG-20) | PR (PR.AA) |
| **AG-03** | Gestión de credenciales | Gestor de secretos, corta duración, rotación, revocación vinculada al interruptor. | Diseño · Prueba de revocación · Registro de rotaciones | PR (PR.AA) |
| **AG-04** | Mandato e intención autorizada | Cada tarea con solicitante o evento de origen, objetivo, herramientas, datos y límites. | Esquema de mandato · Muestras · Muestras | PR (PR.AA) |
| **AG-05** | Punto de decisión de intención | Componente externo al modelo que valida cada acción contra el mandato y emite permisos efímeros. | Diseño · Pruebas de desvío bloqueado · Tasa de bloqueos | PR (PR.AA) |

| Código | Control | Qué exige | Evidencia en G4 · G5 · R6 | Función CSF |
|---|---|---|---|---|
| **AG-06** | Trazabilidad acción–intención | Toda acción enlaza con el identificador de la intención que la autorizó. | Modelo de datos · Reconstrucción de casos · Muestreo | PR (PR.PS) |
| **AG-07** | Límites de actuación | Importes, volúmenes, destinatarios, horarios, iteraciones y presupuesto aplicados fuera del modelo. | Tabla de límites · Pruebas de superación · Alertas | PR (PR.PS) |
| **AG-08** | Validación humana | Lista de acciones sensibles; validación informada; medición de rechazos y tiempos. | Lista y diseño (P17) · Prueba con usuarios · Tasa de rechazo | PR (PR.AA) |
| **AG-09** | Interruptor de parada | Parada total o por capacidad sin desplegar código; revoca credenciales; activa proceso alternativo. | Procedimiento · Prueba con tiempo de efecto · Pruebas periódicas | RS (RS.MI) |
| **AG-10** | Registro de acciones | Registro íntegro, protegido y con los campos de 4.8; conservación definida. | Especificación · Prueba de integridad · Muestreo | PR (PR.PS) |
| **AG-11** | Entornos aislados | Ejecución de código, navegación y ficheros externos en entornos aislados sin credenciales de producción. | Arquitectura · Prueba de escape · Configuración vigente | PR (PR.IR) |
| **AG-12** | Contenido externo no confiable | Separación de privilegios entre lectura de contenido no confiable y acciones sensibles. | Diseño · Pruebas de inyección indirecta · Resultados periódicos | PR (PR.AA) |
| **AG-13** | Herramientas y conectores aprobados | Inventario de herramientas, conectores y servidores de herramientas —también los locales, instalados en equipos de usuario— con versión fijada, origen verificado y aprobación; revisión de la descripción de cada herramienta antes de aprobarla (sin instrucciones dirigidas al modelo) y de sus ámbitos y permisos frente a su función; nueva aprobación si cambia la descripción o el comportamiento. | Inventario y revisión de descripciones · Verificación y prueba de envenenamiento (SEG-27) · Revisión de cambios con nueva aprobación | ID (ID.AM) · GV (GV.SC) |
| **AG-14** | Integridad de memoria y contexto | Memoria aislada por usuario y tarea, con caducidad, validación de escritura y posibilidad de purga. | Diseño · Prueba de envenenamiento · Purgas registradas | PR (PR.DS) |
| **AG-15** | Comunicación segura entre agentes | Autenticación mutua, mensajes íntegros, sin confianza implícita entre agentes. | Diseño · Prueba de suplantación · Configuración | PR (PR.DS) |
| **AG-16** | Contención de cascadas y consumo | Límites de iteración, reintentos y profundidad; disyuntores; presupuesto por tarea. | Parámetros · Prueba de bucle · Alertas de consumo | PR (PR.IR) |

| Código | Control | Qué exige | Evidencia en G4 · G5 · R6 | Función CSF |
|---|---|---|---|---|
| **AG-17** | Monitorización de comportamiento | Líneas base de acciones y alertas ante desviaciones (volumen, destinos, horarios, herramientas). | Escenarios de detección · Alertas probadas · Revisión de alertas | DE (DE.CM) |
| **AG-18** | Pruebas adversarias de agentes | Inyección directa e indirecta, abuso de herramientas, escalada de privilegios, manipulación del aprobador. | Plan · Resultados · Campañas periódicas | ID (ID.IM) |
| **AG-19** | Reversibilidad y compensación | Preferencia por acciones reversibles; procedimiento para deshacer o compensar las demás. | Diseño · Prueba de deshacer · Casos reales | RC (RC.RP) |
| **AG-20** | Revisión periódica de permisos | Certificación por el responsable humano de identidades, permisos y herramientas con la frecuencia de 5.3. | — · — · Actas de revisión | PR (PR.AA) |

08

# Pruebas de seguridad por fase

| Fase | Pruebas | Quién | Resultado exigido |
|---|---|---|---|
| **3 · Viabilidad** | Modelado de amenazas preliminar; evaluación de seguridad del proveedor; decisión preliminar de nivel de autonomía. | Técnico y seguridad | Riesgos RT-GEN y RT-SEG valorados en P12. |
| **4 · Diseño** | Modelado de amenazas completo (SEG-01); revisión de arquitectura de permisos, control de intención e interruptor. | Seguridad, independiente del equipo | P18 con todos los controles aplicables diseñados. |
| **5 · Entrega** | Evaluaciones adversarias automatizadas; *red teaming* manual; pruebas de inyección directa e indirecta; pruebas de límites, interruptor, revocación y reversión; pruebas de fuga de datos; ejecución del banco de pruebas corporativo de agentes (SEG-27) sobre la versión candidata. | Equipo de pruebas independiente o tercero | Controles críticos probados; hallazgos altos cerrados o aceptados según 33 §7. |
| **6 · Operación** | Evaluaciones continuas de regresión; campañas periódicas (5.3); pruebas tras cambios de modelo, instrucciones, herramientas o proveedor; simulacros de incidente; en A2 y A3, banco de pruebas en cada despliegue y ampliación del banco con cada incidente propio. | Operación y seguridad | Indicadores de la sección 10 dentro de umbral. |

| Fase | Pruebas | Quién | Resultado exigido |
|---|---|---|---|
| **7 · Evolución** | Pruebas del alcance ampliado; en retirada, revocación de identidades y borrado verificado. | Seguridad | Sin identidades ni credenciales huérfanas. |

**Contenido mínimo de una campaña de *red teaming* (plantilla P53):** alcance y reglas de enfrentamiento aprobadas; escenarios derivados del modelo de amenazas; inyección directa e indirecta por todos los canales de entrada; intentos de extracción de instrucciones y datos; abuso de cada herramienta; intentos de superar límites y de inducir aprobaciones; resultados con tasa de éxito por escenario, severidad y acción correctiva; repetición de los escenarios fallidos tras la corrección.

Las evaluaciones adversarias automatizadas se versionan en el banco de pruebas corporativo (SEG-27) y se ejecutan antes de cada cambio relevante y, en A2 y A3, en cada despliegue. Un escenario con éxito en una acción sensible es un hallazgo que bloquea G5 hasta su corrección.

09

# Exposición a la IA ofensiva y uso de la IA en la ciberdefensa

La IA no crea tipos de ataque nuevos en su mayoría, pero reduce su coste, aumenta su verosimilitud y acorta los tiempos. La guía CCN-CERT BP/36 insiste en esta idea: el tiempo entre la existencia de una vulnerabilidad y su explotación se reduce, y las revisiones de seguridad puntuales llegan tarde. La exposición se gestiona como riesgo corporativo (C4), con responsable en seguridad de la información e información al consejo. La sección 9.4 trata otro caso: agentes de terceros que actúan legítimamente por el cliente; su tráfico se parece al de un ataque, pero no lo es.

## 9.1 Amenazas y controles corporativos

| Amenaza | Qué cambia con la IA | Riesgo tipo | Controles corporativos |
|---|---|---|---|
| **Suplantación de identidad** | Voz y vídeo sintéticos convincentes a partir de muestras públicas; suplantación en tiempo real. | RT-SEG-01 | SEG-16, SEG-17, SEG-18; límites de autorización de pagos. |
| ***Phishing* generado** | Mensajes personalizados, sin errores, en cualquier idioma y a gran escala. | RT-SEG-02 | SEG-15, SEG-17; filtrado de correo; reducción de información pública sobre empleados. |

| Amenaza | Qué cambia con la IA | Riesgo tipo | Controles corporativos |
|---|---|---|---|
| **Fraude del CEO con *deepfakes*** | Videollamadas o audios falsos de directivos que ordenan pagos urgentes o confidenciales. | RT-SEG-01 | SEG-16 sin excepciones por urgencia o jerarquía; doble firma; palabra de verificación (SEG-18). |
| **Suplantación del cliente con IA** | Voz sintética o vídeo manipulado para superar la identificación del cliente en canales automatizados o telefónicos; documentos, fotografías o facturas manipulados con IA para respaldar reclamaciones, altas o cambios de datos. | RT-SEG-09 | Detección de voz sintética y prueba de vida solo como apoyo, nunca como única barrera; verificación por un segundo factor independiente del canal antes de operaciones sensibles o de cambios de datos de contacto o bancarios; revisión de documentos e imágenes con comprobación de metadatos y de coherencia, y revisión humana por muestreo y en los casos señalados; SEG-18 en los canales de voz y vídeo con clientes cuando sea posible. |
| **Explotación acelerada de vulnerabilidades** | Descubrimiento y explotación automatizados poco después de publicarse una vulnerabilidad. | RT-SEG-03 | SEG-13, SEG-19; inventario de superficie expuesta; segmentación. |
| **Ataque a identidades y credenciales** | Automatización de pruebas de credenciales y abuso de identidades no humanas. | RT-SEG-04 | SEG-15; AG-01, AG-03, AG-20; detección de tráfico automatizado. |
| **Ataque a los propios agentes** | Inyección indirecta para que los agentes de la compañía actúen contra ella. | RT-GEN-02 | AG-05, AG-08, AG-12, AG-18. |
| **Desinformación sobre la compañía** | Contenidos falsos sobre la compañía o sus directivos difundidos a escala. | RT-REP-03 | Vigilancia de marca; protocolo de comunicación de crisis. |

## 9.2 Medidas de gobierno

- **Responsable:** seguridad de la información, con informe al comité de IA y a la comisión delegada.
- **Evaluación anual** de exposición a la IA ofensiva con las amenazas de 9.1 y, en entidades del ámbito del Esquema Nacional de Seguridad, con los instrumentos de autoevaluación que publique el CCN.
- **Simulacro anual**, como mínimo, de fraude con suplantación sintética que implique a finanzas y dirección.
- **Procesos de pago** revisados para que ninguna orden dependa solo del reconocimiento de voz, imagen o estilo de escritura.
- **Plazos de corrección** de vulnerabilidades en sistemas expuestos revisados y aprobados, con seguimiento de su cumplimiento.

## 9.3 Uso de la IA en la ciberdefensa de la compañía (Defend)

La compañía también puede usar la IA para defenderse: correlación y triaje de alertas, caza de amenazas, análisis de código malicioso y respuesta automatizada. El Cyber AI Profile del NIST, en borrador, lo trata como un área propia, **Defend** (34 §5.3). En SEVEN-G es un **caso de uso** más: se da de alta en el registro, recorre el ciclo de vida con sus puertas y su autonomía se fija con la escala A0–A3 (sección 5). Lo específico es que un error del sistema defensivo puede dejar pasar un ataque (falso negativo) o interrumpir la operación al contener lo que no es un ataque (falso positivo), y que el propio sistema defensivo es un objetivo para el atacante.

| Aspecto | Regla |
|---|---|
| **Autonomía de la respuesta** | Cada tipo de acción de contención (aislar un equipo, bloquear una cuenta, cortar tráfico, revocar credenciales, retirar un mensaje) tiene su nivel A0–A3. Las acciones con efecto en producción o sobre personas son A1, con validación humana previa, salvo que un procedimiento aprobado, con límites, reversión probada e interruptor de parada, permita A2 o A3 (SEG-21). |
| **Supervisión humana** | Antes de una contención con efecto en producción, en clientes o sobre la cuenta de una persona, la valida una persona de seguridad que ve por qué se propone; en A2 y A3, esas acciones se revisan a posteriori (SEG-22). |
| **Calidad de la detección** | Falsos positivos y falsos negativos, tiempo de detección y tiempo de respuesta, medidos frente a una línea base sin IA y con umbrales en P25 (SEG-23). Un falso negativo en un incidente real se analiza como hallazgo (documento 37). |
| **Dependencia del proveedor** | Evaluación del proveedor de ciberdefensa con IA (P14, P55), control de la telemetría que sale de la compañía y de dónde se trata, plan de salida (P57) y operación degradada sin la IA (SEG-24). |
| **Pruebas del propio sistema defensivo** | Pruebas adversarias de evasión, de envenenamiento de la telemetría o de los datos de entrenamiento y de inyección de instrucciones en alertas, tickets o registros que el sistema lee (SEG-25, P53). |
| **Riesgo tipo** | RT-SEG-08 (documento 33 §9.9). |

> **Por qué importa.** Usar IA en la ciberdefensa acelera la detección y la respuesta, pero traslada a un sistema automático decisiones que pueden parar la operación o dejar pasar un ataque. Tratarla como un caso de uso más, con autonomía fijada por tipo de acción, supervisión humana en lo que no se puede deshacer y métricas de acierto frente a una línea base, evita que la defensa se convierta en una nueva fuente de incidentes.

**IA para TI y ciberdefensa: cómo se mide y cómo se financia.** Los casos de uso que usan IA para proteger y operar la tecnología de la propia compañía no se justifican con un VAN: su valor es menos

pérdida esperada y más capacidad de detectar, responder y recuperarse. Se registran y recorren el ciclo de vida como cualquier otro caso de uso, se miden por **resultado operativo** (regla de presentación 11 del documento 40) y se financian desde el sobre de habilitación (documento 13 §12), presentadas aparte del neto de la cartera como las plataformas habilitadoras (40 §7.3).

| Capacidad | Qué es | Cuándo se usa | Cómo se mide |
|---|---|---|---|
| **Detección y respuesta** | Correlación y triaje de alertas, caza de amenazas y contención asistidas por IA (este apartado; SEG-21 a SEG-25). | Cuando el volumen de alertas o el tiempo de respuesta superan lo que el equipo puede atender con calidad. | Tiempo de detección, de respuesta y de recuperación frente a la línea base sin IA; falsos positivos y negativos; contenciones revertidas. |
| **Seguridad de aplicaciones y de interfaces de programación** | Revisión asistida de código, configuraciones e interfaces de programación expuestas, y pruebas continuas de la superficie expuesta (SEG-19). | Cuando la compañía desarrolla o expone servicios y la revisión manual no llega a todo el cambio. | Vulnerabilidades críticas detectadas antes de producción; tiempo de corrección; cobertura de interfaces revisadas. |
| **Banco de pruebas de agentes** | El banco de pruebas corporativo (SEG-27) integrado en la cadena de despliegue. | Desde que hay agentes A1 o superiores, o IA generativa con exposición directa. | Escenarios cubiertos frente al modelo de amenazas; eficacia de los guardarraíles (52 §5.2); escenarios críticos bloqueados antes de producción. |
| **Operación y resiliencia** | IA para anticipar fallos, correlacionar incidentes de operación y acelerar la recuperación de los servicios. | En servicios con niveles de servicio exigentes o funciones críticas. | Disponibilidad; tiempo medio de recuperación; incidentes evitados con causa registrada. |
| **IA en la sombra** | Detección y regularización del uso no autorizado de IA, incluidos agentes locales y extensiones del navegador (SEG-20; 31 §5). | Desde C1, como parte del inventario. | Casos detectados y regularizados; tiempo de regularización; cobertura de equipos y de tráfico inspeccionados. |

## 9.4 Agentes de terceros que actúan por el cliente

La sección 9.1 trata los agentes como amenaza: alguien los usa contra la compañía. Esta sección trata otro caso: **agentes de terceros que actúan legítimamente por el cliente**. Su tráfico se parece al de un ataque (automatizado, repetido, con credenciales delegadas), pero no lo es. Si la compañía lo bloquea

todo, pierde a los clientes que ya delegan en su agente; si lo tolera sin control, acepta acciones que nadie puede atribuir ni limitar.

Un **cliente agéntico** es un cliente que actúa a través de un agente personal de IA: un asistente que investiga, compra, reclama o contrata en su nombre. Llega a la compañía por cuatro caminos, con una base común de pruebas:

## Las cuatro caras del cliente agéntico

Cuando el cliente actúa a través de un agente personal de IA, llega a la compañía por cuatro caminos, con una base común de pruebas.

**1 · INVESTIGA**
**Le investiga con un asistente**
Lo que el asistente sabe y recomienda de la compañía decide si entra en la lista de opciones.

↓

**2 · COMPRA**
**Compra con un agente**
El agente compara y cierra la compra. El recorrido de venta debe poder atenderlo.

↓

**3 · PIDE SERVICIO**
**Reclama y pide concesiones**
Atención, reclamaciones, devoluciones y bajas. El agente pide siempre, y la compañía paga lo que antes nadie reclamaba.

↓

**4 · HABLA CON SU AGENTE**
**Canal entre agentes**
El agente del cliente habla con un agente de la compañía, con identidad, permisos y registro.

**Base común: probar antes y siempre.** Clientes sintéticos, simulación de usuarios, pruebas de intrusión y *red teaming* (SEG-27 y P53).

**QUIÉN RESPONDE DE QUÉ: PROPUESTA DEL SECTOR, NO NORMA**

**El cliente**
De los permisos y las instrucciones que da a su agente.

**El proveedor del servicio**
De lo que ocurra fuera de esos permisos.

**El proveedor del agente**
De lo que ocurra fuera de esas instrucciones.

**Cada uno**
De las brechas de sus propios sistemas.

La responsabilidad sigue al control. SEVEN-G la recoge como criterio de diseño del canal identificado (SEG-28), no como norma.

1. **Investiga.** El cliente consulta a un asistente, y lo que este sabe y recomienda de la compañía decide si entra en la lista de opciones.

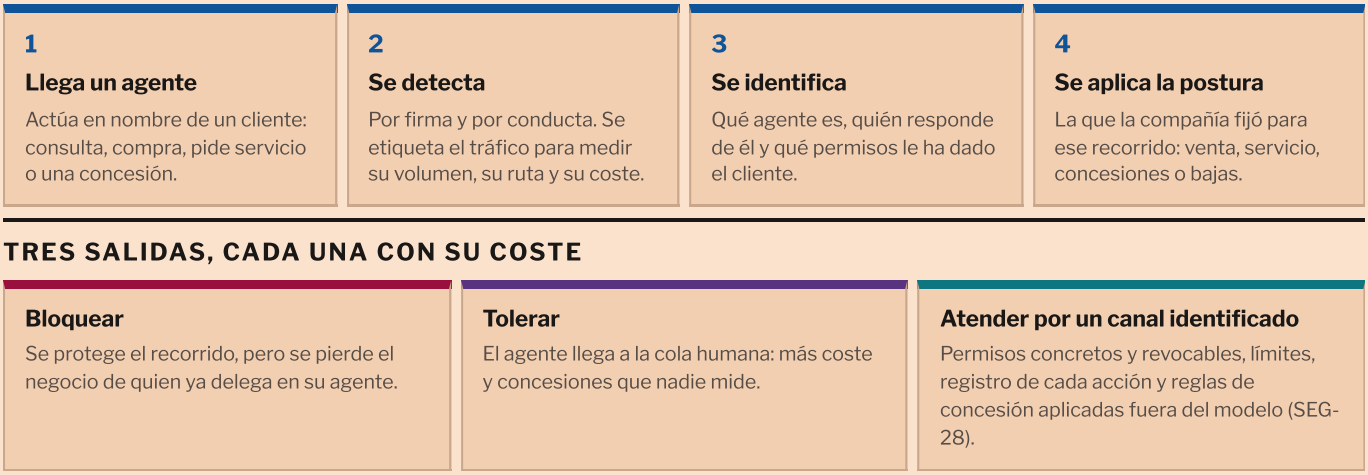
2. **Compra.** Un agente compara y cierra la compra por el cliente.
3. **Pide servicio.** Atención, reclamaciones, devoluciones, concesiones y bajas. El agente pide siempre, y la compañía paga beneficios que antes casi nadie reclamaba.
4. **Su agente habla con el de la compañía.** Un canal entre agentes, con identidad, permisos y registro.

**Base común: probar antes y siempre.** Los cuatro caminos se prueban con clientes sintéticos (agentes de cliente simulados), simulación de usuarios, pruebas de intrusión y *red teaming*. Se hacen con el banco de pruebas corporativo (SEG-27) y se registran en P53, incluida la limitación de que una simulación no sustituye a clientes reales.

**Postura por recorrido.** En C2, la compañía con exposición directa a clientes decide por recorrido (venta, servicio, concesiones, bajas) si **bloquea**, **tolera** o **atiende por un canal identificado** (13 §5.6). Cada opción tiene su coste:

## Postura ante el agente del cliente, por recorrido

En C2 la compañía decide, recorrido a recorrido, si bloquea, tolera o atiende al agente por un canal identificado (documento 13, sección 5.6).




Ninguna postura es gratuita. La decisión se revisa en C5 y se aplica por recorrido, no a toda la compañía a la vez.

**Requisitos del canal identificado (SEG-28).** Atender al agente exige:

| Requisito | Qué exige | Relación |
|---|---|---|
| **Identidad del agente y de quien responde de él (*Know Your Agent*)** | La compañía sabe qué agente es, quién responde de él y qué puede hacer. No acepta un agente anónimo en un canal con efecto. | AG-01 aplicado a agentes externos |
| **Firma de las peticiones** | Las peticiones del agente llevan una firma que la compañía puede verificar. | AG-15 aplicado a agentes externos |
| **Permisos del cliente** | Concretos, revocables y registrados ante la compañía; el cliente puede ver y retirar lo que ha concedido. | 4.2 |

| Requisito | Qué exige | Relación |
|---|---|---|
| **Límites** | Tipo de acción, importe y frecuencia por agente y por cliente. | SEG-10 |
| **Registro de cada acción** | Con el identificador del agente, del cliente y del permiso aplicado. | SEG-12 · 4.8 |
| **Reglas de elegibilidad y de concesión fuera del modelo** | El modelo no decide qué concesión se paga: lo decide la regla de la compañía, igual en todos los canales. | 40 §5.4 |
| **Detección por firma y por conducta** | El tráfico de agentes se distingue del humano y del ataque. | SEG-12 · 52 §4.3 |
| **Etiquetado del tráfico** | Cada contacto de un agente se etiqueta para medir su volumen, su ruta y su coste. | 52 §5.2 |

**Quién responde de qué.** Una propuesta del sector financiero europeo, que SEVEN-G recoge como criterio de diseño del canal y no como norma, liga la responsabilidad al control: el cliente responde de los permisos y las instrucciones que da; el proveedor del servicio, de lo que ocurra fuera de esos permisos; el proveedor del agente, de lo que ocurra fuera de las instrucciones; y cada uno, de las brechas de sus propios sistemas. El reparto se escribe en las condiciones del canal (RT-LEG-09).

**Estado regulatorio** (34 §3.11 y §12). La norma de pagos exige que el tercero pueda identificarse ante la entidad, también en el mecanismo de reserva, y no se afirma aquí que prohíba por sí misma el acceso por pantalla. En pagos, la opinión de la EBA de 2018 sobre la autenticación reforzada (EBA-Op-2018-04) aclara que esa autenticación la aplica, y sus exenciones las decide, el proveedor que emite las credenciales del cliente, y que puede encargarla por contrato a otro proveedor, con reparto de la responsabilidad entre ambos. Para el canal identificado, esto significa que un agente no hereda la autenticación del cliente: la entidad decide cuándo la exige y a quién se la encarga. La opinión no trata los agentes de IA; se usa como orientación. El artículo 50 del Reglamento Europeo de IA obliga a informar a las personas que interactúan con un sistema de IA; no obliga a un agente a identificarse ante un servicio. Que la ley exija a los agentes identificarse es una propuesta, no una obligación vigente. Los protocolos de agente (Web Bot Auth, protocolo de agente personal) son borradores: se citan como orientación y no fundamentan ningún control.

**Riesgos y medición.** Riesgos tipo RT-ECO-06 (subida de demanda y de concesiones) y RT-LEG-09 (disputa de una acción del agente del cliente) en el documento 33 §9; indicadores de la sección 10; exposición económica como escenario aparte en 40 §5.4. La postura no es un criterio nuevo de las puertas: se verifica con los criterios de seguridad existentes (sección 11).

**Por qué importa.** Los agentes de clientes ya llegan, con o sin permiso de la compañía. Decidir la postura por recorrido, en vez de dejar que cada equipo improvise, evita dos fallos opuestos: perder ventas por bloquear a quien delega en su agente y pagar concesiones sin medir por tolerarlo. El canal identificado convierte un tráfico ambiguo en acciones atribuibles, limitadas y registradas.

10

## Indicadores para el consejo

| Indicador | Qué muestra | Lectura |
|---|---|---|
| **Sistemas por nivel de autonomía** | Número de sistemas A0, A1, A2 y A3 en producción. | Crecimiento de A2 y A3 con controles críticos completos. |
| **Controles críticos de agentes** | Porcentaje de controles críticos probados y operando en A2 y A3. | 100 %; cualquier desviación se explica. |
| **Identidades de agentes con permisos excesivos** | Hallazgos de la última revisión (AG-20). | Cero. |
| **Credenciales sin rotar en plazo** | Identidades no humanas con secretos fuera de plazo. | Cero. |
| **Prueba del interruptor de parada** | Sistemas A2 y A3 con prueba vigente y tiempo de efecto medido. | 100 %. |
| **Éxito de inyección de instrucciones** | Porcentaje de escenarios con éxito en la última campaña, y en acciones sensibles. | Tendencia a la baja; cero en acciones sensibles. |
| **Acciones de agentes bloqueadas** | Acciones bloqueadas por el control de intención o por límites. | Explicar picos: ataque, error de diseño o mandato mal definido. |
| **Incidentes de seguridad de IA** | Incidentes S1–S4 por tipo: inyección, fuga, acción no autorizada, compromiso de agente. | Tendencia y tiempo de contención. |
| **Doble factor resistente a *phishing*** | Cobertura en accesos expuestos y privilegiados. | 100 % en privilegiados. |
| **Vulnerabilidades críticas expuestas fuera de plazo** | Número y antigüedad. | Cero. |

| Indicador | Qué muestra | Lectura |
|---|---|---|
| **Simulacros de suplantación** | Resultado del último simulacro de fraude con contenido sintético. | Porcentaje que aplica la verificación fuera de banda. |
| **Uso no autorizado de IA** | Casos detectados y regularizados. | Tendencia a la baja. |
| **Calidad de la ciberdefensa con IA** | Falsos positivos y negativos, tiempo de detección y de respuesta y contenciones automáticas revertidas (sección 9.3). | Mejora frente a la línea base sin IA; cada contención revertida se explica. |
| **Eficacia de los guardarraíles** | Escenarios del banco de pruebas (SEG-27) detectados o bloqueados, falsos positivos y desactivaciones (52 §5.2). | Cero escenarios críticos sin detectar; toda desactivación explicada. |
| **Tráfico por la pasarela corporativa** | Proporción del tráfico hacia modelos que pasa por la pasarela (SEG-26). | 100 %; el tráfico directo se explica y se regulariza. |
| **Derivaciones a una persona** | Derivaciones por motivo y peticiones no atendidas en sistemas con exposición directa (sección 4.6). | Ninguna petición sin atender; tendencia por motivo explicada. |
| **Tráfico de agentes de clientes** | Volumen por recorrido y postura, y proporción etiquetada (sección 9.4; 52 §5.2). | Proporción etiquetada creciente; el tráfico de agentes sin etiquetar se explica. |
| **Concesiones resueltas por agentes** | Número, importe y tasa frente a la línea base humana (40 §5.4). | Variación explicada; no se compara con el escenario de exposición, que no es neto. |

11

# Integración con el ciclo de vida

| Puerta | Qué se exige en seguridad |
|---|---|
| **Viabilidad** (G3) | Riesgos RT-GEN y RT-SEG valorados; nivel de autonomía propuesto; evaluación de seguridad del proveedor (documento 36). |
| **Diseño** (G4) | P18 completo; controles SEG y AG aplicables diseñados; controles críticos diseñados; conformidad de seguridad de la información; tráfico hacia modelos por la pasarela corporativa (SEG-26); con exposición directa a clientes, postura por recorrido aplicada (SEG-28). No es un criterio nuevo del documento 21: se verifica con los criterios de seguridad existentes de la puerta. |
| **Puesta en producción** (G5) | Controles críticos probados; *red teaming* realizado con hallazgos tratados; interruptor probado; firma de seguridad de la información en la firma multinivel (01 §6.7); banco de pruebas corporativo (SEG-27) sin escenarios críticos con éxito. |

| Puerta | Qué se exige en seguridad |
|---|---|
| **Revisión de continuidad** (R6) | Controles operando; campañas y revisiones dentro de frecuencia; incidentes analizados; nivel de autonomía vigente; banco de pruebas ejecutado en cada despliegue (A2 y A3); con exposición directa a clientes, postura por recorrido vigente y tráfico de agentes etiquetado (SEG-28). |
| **Escalado o retirada** (G7) | En escalado, revisión de controles para el nuevo alcance; en retirada, revocación de identidades y credenciales. |

El responsable técnico de IA diseña e implanta los controles; seguridad de la información los revisa y emite conformidad; el responsable de riesgos de IA valora el residual; el auditor de IA verifica las evidencias con la lista LV-AG. Quien construye el agente no realiza su *red teaming* de G5.

12

# Herramientas y plantillas asociadas

| Nombre | Uso |
|---|---|
| **Diseño de seguridad (incluye agentes)** (P18) | Modelo de amenazas, nivel de autonomía, identidades, matriz de permisos, mandato y control de intención, límites, acciones sensibles, interruptor, registro, aislamiento, plan de pruebas y estado de cada control SEG y AG. Fase 4; se actualiza en 5 y 6. |
| **Evaluación de seguridad de agentes** (T10) | A partir del nivel de autonomía, genera los controles SEG y AG exigibles, registra su estado (cumple, no cumple, no aplica, pendiente) con evidencia y bloquea G4 o G5 si un control crítico no está diseñado o probado. Lista de verificación en T03. |
| **Plan e informe de pruebas adversarias** (P53) | Alcance, reglas de enfrentamiento, escenarios y resultados de las campañas de *red teaming* (sección 8). Fases 4 a 6. Incluye los resultados del banco de pruebas corporativo (SEG-27) y las pruebas con clientes sintéticos y simulación de usuarios (sección 9.4). |
| **Inventario de identidades no humanas y componentes** (P54) | Identidades de los agentes con su responsable humano (sección 4) y componentes con versión y procedencia (SEG-09). Fase 4; se revisa en la operación. |
| **Perfil de seguridad de IA (CSF 2.0 / Cyber AI Profile)** (P72) | Perfil actual y objetivo de la seguridad de la IA con las funciones del CSF, a partir de la columna «Función CSF» de las secciones 6 y 7 y del cuestionario del documento 11 (34 §5.5). C1, C2 y C5. |
| **Diseño de supervisión humana · Manual de operación · Plan de respuesta a incidentes** (P17 · P24 · P26) | Acciones sensibles y validación; procedimiento del interruptor; respuesta a incidentes de IA. |

| Nombre | Uso |
|---|---|
| **Inventario · Riesgos · Incidentes · Panel del consejo · Monitor de uso corporativo** (T02 · T06 · T08 · T17 · T21) | Nivel de autonomía, riesgos, incidentes, indicadores y uso no autorizado. |

13

# Documentos relacionados

| Documento | Relación |
|---|---|
| **01 · Metodología fundacional** | Sección 9.2 (agentes con capacidad de actuar), sección 10 y firma multinivel. |
| **13 · Tesis de IA, ambición y apetito de riesgo** | Sección 5.6: postura ante agentes de clientes por recorrido (SEG-28). |
| **22 · Listas de verificación** | Lista LV-AG basada en los controles AG. |
| **31 · Política corporativa y uso aceptable** | Uso corporativo de IA y SEG-20. |
| **33 · Metodología de riesgos de IA** | Escalas y riesgos tipo RT-GEN y RT-SEG. |
| **34 · Mapeo regulatorio** | Secciones 3.11 y 12: lo que dicen y no dicen el artículo 50, la norma de pagos y eIDAS 2 sobre agentes. |
| **36 · Terceros y proveedores de IA** | Seguridad de proveedores, modelos y conectores de terceros. |
| **37 · No conformidades e incidentes** | Respuesta a incidentes de seguridad de IA y desactivación de controles críticos. |
| **40 · Reglas de medición del valor** | Sección 5.4: exposición económica a agentes de clientes. |
| **52 · Manual de operación de IA** | Operación de la monitorización, del interruptor, de las evaluaciones continuas y del tráfico de agentes de clientes. |
| **60 · Paquete para el consejo** | Indicadores de la sección 10. |

14

# Control de versiones

| Versión | Fecha | Cambios |
|---|---|---|
| **0.1** | 16-09-2026 | Primera versión. Define las amenazas a sistemas de IA con referencias a OWASP (LLM 2025 y aplicaciones agénticas), MITRE ATLAS y NIST AI 600-1; los nueve requisitos esenciales de un agente; los niveles de autonomía A0–A3 con controles mínimos y frecuencias; los catálogos SEG-01 a SEG-20 y AG-01 a AG-20; las pruebas por fase; la exposición a la IA ofensiva y los indicadores para el consejo. |
| **0.2** | 25-09-2026 | Añade el NIST CSF 2.0 y el Cyber AI Profile (en borrador) a las referencias (sección 1.2) y la columna «Función CSF» a los catálogos SEG y AG (secciones 6 y 7), con la función y la categoría del CSF a las que contribuye cada control (34 §5.3), y la plantilla P72 en la sección 12. |
| **0.3** | 25-09-2026 | Añade la sección 9.3 sobre el uso de la IA en la ciberdefensa de la compañía (área Defend del Cyber AI Profile): autonomía de la respuesta, supervisión humana de la contención, calidad de la detección, dependencia del proveedor y pruebas del propio sistema defensivo; los controles SEG-21 a SEG-25; un indicador para el consejo y el riesgo tipo RT-SEG-08. |
| **0.4** | 19-10-2026 | Controles SEG-26 (pasarela corporativa de IA) y SEG-27 (banco de pruebas corporativo de agentes), con frecuencias, pruebas por fase, indicadores e integración en las puertas; SEG-20 ampliado a agentes locales y extensiones del navegador; amenaza de envenenamiento de herramientas y conectores (3.2) y AG-13 precisado; derivación inmediata a una persona (4.6); suplantación del cliente con IA (9.1, RT-SEG-09); medición y financiación de la IA para TI y ciberdefensa (9.3) (D159). |
| **0.5** | 16-11-2026 | Sección 9.4 (agentes de terceros que actúan por el cliente: las cuatro caras del cliente agéntico, base común de pruebas, postura por recorrido, requisitos del canal identificado, reparto de responsabilidad y estado regulatorio); control SEG-28; dos indicadores en la sección 10 y la postura por recorrido en G4 y R6 (D185). |
| **0.6** | 19-11-2026 | Sección 9.4, estado regulatorio: quién aplica la autenticación reforzada según la opinión de la EBA de 2018 (EBA-Op-2018-04), como orientación para el canal identificado (D194). |

# Fuentes externas citadas

Enlaces a la fuente oficial o primaria, comprobados antes de publicar. Verifique siempre la versión vigente en la fecha de uso.

**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 (Reglamento de Inteligencia Artificial)** · https://eur-lex.europa.eu/eli/reg/2024/1689/oj/spa · Parlamento Europeo y Consejo de la UE · 13-06-2024

**Reglamento (UE) 2024/2847 del Parlamento Europeo y del Consejo, de 23 de octubre de 2024, relativo a los requisitos horizontales de ciberseguridad para los productos con elementos digitales (Reglamento de Ciberresiliencia)** · https://eur-lex.europa.eu/eli/reg/2024/2847/oj/spa · Parlamento Europeo y Consejo de la UE · 23-10-2024

**Real Decreto 311/2022, de 3 de mayo, por el que se regula el Esquema Nacional de Seguridad** · https://www.boe.es/eli/es/rd/2022/05/03/311/con · Gobierno de España (Ministerio de Asuntos Económicos y Transformación Digital) · 03-05-2022

**Centro Criptológico Nacional (CCN) / CCN-CERT** · https://www.ccn.cni.es/index.php/es/ · Centro Criptológico Nacional (adscrito al Centro Nacional de Inteligencia)

**Agentic AI – Threats and Mitigations (IA agéntica: amenazas y mitigaciones), v1.0** · https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/ · OWASP GenAI Security Project (Agentic Security Initiative), OWASP Foundation · 17-02-2025

**CCN-CERT BP/36 · Guía de buenas prácticas frente al modelo de IA ofensiva** · https://www.ccn-cert.cni.es/es/informes/informes-de-buenas-practicas-bp/7469-ccn-cert-bp-36-buenas-practicas-ia-ofensiva/file.html · Centro Criptológico Nacional (CCN-CERT) · 2026-06

**Marco de gestión de riesgos de IA: perfil de inteligencia artificial generativa (NIST AI 600-1)** · https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence · National Institute of Standards and Technology (NIST) · 26-07-2024

**OWASP Top 10 para aplicaciones agénticas 2026 (OWASP Top 10 for Agentic Applications for 2026)** · https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/ · OWASP GenAI Security Project (Agentic Security Initiative), OWASP Foundation · 09-12-2025

**OWASP Top 10 para aplicaciones con modelos de lenguaje de gran tamaño, edición 2025 (OWASP Top 10 for LLM Applications 2025)** · https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/ · OWASP GenAI Security Project (OWASP Foundation) · 17-11-2024

**Perfil del marco de ciberseguridad para la inteligencia artificial (Cyber AI Profile): perfil comunitario del NIST, NIST IR 8596 (borrador preliminar inicial)** · https://csrc.nist.gov/pubs/ir/8596/iprd · National Institute of Standards and Technology (NIST), NCCoE · 16-12-2025

**ISO/IEC 23894:2023 Tecnología de la información — Inteligencia artificial — Orientación sobre la gestión del riesgo** · https://www.iso.org/standard/77304.html · ISO/IEC (JTC 1/SC 42) · 06-02-2023

**ISO/IEC 42001:2023 Tecnología de la información — Inteligencia artificial — Sistema de gestión** · https://www.iso.org/standard/42001 · ISO/IEC (JTC 1/SC 42) · 18-12-2023

**Marco de ciberseguridad del NIST, versión 2.0 (The NIST Cybersecurity Framework (CSF) 2.0), NIST CSWP 29** · https://doi.org/10.6028/NIST.CSWP.29.spa · National Institute of Standards and Technology (NIST), Departamento de Comercio de EE. UU. · 26-02-2024

**Marco de gestión de riesgos de inteligencia artificial del NIST (AI RMF 1.0), NIST AI 100-1** · https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10 · National Institute of Standards and Technology (NIST), Departamento de Comercio de EE. UU. · 26-01-2023

**MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems), base de conocimiento de tácticas y técnicas adversarias contra sistemas de IA** · https://atlas.mitre.org/ · The MITRE Corporation