GIA | GRUPO DE INTELIGENCIA ARTIFICIAL

isms forum | INTERNATIONAL INFORMATION SECURITY COMMUNITY

Decálogo de Seguridad en la **IA Agéntica**

Fundamentos para un ecosistema de agentes confiable

Índice

Introducción 05

¿Qué es un agente IA y qué es la IA agéntica? 06

- Tipos de agentes 07
- Con qué puede interactuar un agente 08

Aspectos previos de Seguridad 09

- Tendencias en ciberseguridad 10
- Owasp Top 10 for Agentic Applications 11
- AI Agents Attack Matrix 12
- Lethal Trifecta 15

Decálogo 16

- 01. Identificación 16
- 02. Identidad VIP 18
- 03. Datos 19
- 04. Protección 20
- 05. Gestión de la autonomía 21
- 06. Monitorización 22
- 07. Respuesta 23
- 08. Gobierno 24
- 09. Gestión del Riesgo 25
- 10. Cadena de Suministro 26

Decálogo vs OWASP Top 10 for Agentic Applications 28

Decálogo vs AI Agents 30

Referencias 32

Participantes

Autores

Angel Pérez
Eduardo de Prado

Revisor

Fernando Rubio

Editora

Beatriz García

Coordinación del proyecto

Susana Marín

Diseño y Maquetación

Susana Marín

Introducción

La IA Agéntica es un reto emergente de rápida adopción en las empresas. Organizaciones de referencia como Gartner o McKinsey auguran una adopción acelerada a corto plazo, especialmente en áreas como la automatización de procesos, interacción con el cliente y operaciones complejas.

Adicionalmente se observa con preocupación un creciente volumen de amenazas de seguridad que afectan a la IA Agéntica que se suman a las amenazas de ciberseguridad de aplicación para el resto de los sistemas de información.

El Decálogo de Seguridad en la IA Agéntica enumera 10 principios de seguridad específicos de la IA Agéntica, no obstante, es conveniente aclarar que algunos conceptos, debido al estado del arte en esta materia, son de carácter teórico a falta de herramientas específicas que permitan implementarlos. Progresivamente iremos viendo cómo los fabricantes dan respuesta a estas necesidades, así como muy probablemente surjan nuevos principios de seguridad que deban ser atendidos y conlleven la necesidad de revisar este documento.

Los aspectos de ciberseguridad "clásicos" son de plena aplicación a la IA Agéntica y existe bastante documentación de referencia al respecto, este documento evita entrar en los referidos temas de ciberseguridad clásicos y se enfoca en aspectos específicos de seguridad en Agentes IA.

Complementariamente al decálogo, se ha realizado un ejercicio de validación de la adecuación de los principios del decálogo contra dos referentes en amenazas de IA Agéntica:

- OWASP Top 10 for Agentic Applications
- AI Agents Attack Matrix que mantiene una comunidad de expertos

Nota adicional:

La elaboración de este documento ha contado con el apoyo de herramientas de inteligencia artificial (IA). En consonancia con las buenas prácticas de uso responsable de esta tecnología, se ha mantenido una estricta supervisión y control humanos durante todo el proceso.

Todas las decisiones finales sobre el contenido, así como las revisiones y validaciones, han sido tomadas íntegramente por personas. De este modo, la IA actúa únicamente como una herramienta de apoyo para reforzar la capacidad humana, asegurando que el criterio y la responsabilidad últimos recaen siempre en los autores del documento.

Este enfoque garantiza la alineación con los principios de responsabilidad, trazabilidad y control humano exigidos por los marcos regulatorios y normativos vigentes (AI Act / ISO 42001).

Qué es un **agente** y qué es la **IA Agéntica**

Según se cita en el libro *Artificial Intelligence. A Modern Approach:*

> **Un agente** es cualquier cosa que perciba su entorno mediante sensores y actúe sobre él mediante actuadores. Un agente humano tiene ojos, oídos y otros órganos como sensores, y manos, piernas, tracto vocal, etc., como actuadores. Un **agente robótico** puede tener cámaras y telémetros infrarrojos como sensores, y diversos motores como actuadores. Un **agente de software** recibe pulsaciones de teclas, contenido de archivos y paquetes de red como entradas sensoriales y actúa sobre el entorno visualizando en la pantalla, escribiendo archivos y enviando paquetes de red.

En lo relativo a Agentes IA, la clave está en que el agente recibe un objetivo (que puede estar proporcionado por un humano) y elige de forma independiente las acciones más apropiadas para alcanzar el objetivo mediante los recursos que tiene acceso.

Tipos de **agentes**

De forma sencilla se pueden conceptualizar tres tipos de agentes:

Agente de asistencia

- Apoya al usuario en tareas cognitivas
- *Ejemplo:* Copilot, ChatGPT, Gemini, ...

Agente de acción

- Realiza tareas concretas de forma autónoma
- *Ejemplo:* un bot de automatización de procesos que crea tickets en la herramienta de ticketing de la organización

IA Agéntica

- Arquitectura de múltiples agentes coordinados que razonan, planifican, usan herramientas y colaboran para lograr objetivos complejos, mediante el planeamiento y descomposición de tareas dinámico que permite analizar un objetivo complejo y dividirlo en subtareas, actualizando el plan de ejecución en tiempo real.
- *Ejemplo:* un sistema capaz de analizar datos financieros, generar un informe y distribuirlo mediante la plataforma colaborativa de la organización.

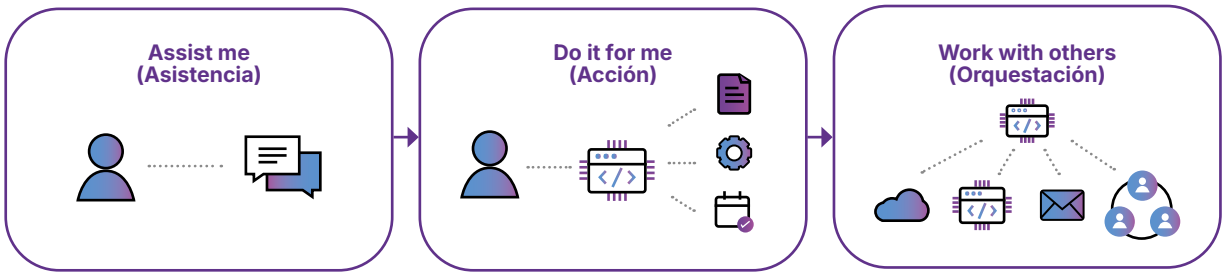

Figura 1: Tipos de Agentes

Con qué puede interactuar un **agente**

Un agente de IA puede interactuar con múltiples tipos de entidades y sistemas:

- **Humanos:** Interacción directa con usuarios para recibir instrucciones, dar respuestas o colaborar en tareas.
- **Otros agentes:** Cooperación con agentes especializados para dividir y resolver problemas complejos.
- **Datos (Web, BBDD):** Acceso y consulta de información estructurada y no estructurada en bases de datos y fuentes externas.
- **Modelos LLM:** Uso de modelos de lenguaje para razonamiento, generación de contenido y comprensión contextual.
- **Herramientas:** mecanismos para interactuar con otros sistemas y entornos
  - ↗ Herramientas conectadas mediante API, MCP, Web Socket, IoT, vehículos, robots, etc. (ERP, CRM, Correo, Chats, etc.)
  - ↗ CUA (Computer User Agent) que permite interactuar como un humano con el teclado, mouse, pantalla, etc.

Figura 2: Interacciones de los Agentes

Aspectos previos de **seguridad**

En el mundo anglosajón el término “Seguridad” se diferencia en dos vertientes que son muy relevantes en el ámbito de la IA Agéntica: Security y Safety.

Safety (seguridad funcional)

Capacidad del sistema para operar de manera segura y fiable, incluso en presencia de fallos internos, errores humanos o condiciones imprevistas del entorno, evitando daños a las personas, al entorno o a otros sistemas.

En IA Agéntica, esto implica diseñar agentes que tomen decisiones previsibles y robustas, minimizando el riesgo de comportamientos peligrosos o inesperados, más allá de los ataques externos.

Security (seguridad informática)

Protección del sistema frente a amenazas deliberadas por parte de actores maliciosos, como ataques de inyección de datos, manipulación de modelos, espionaje o toma de control de agentes.

En el contexto de IA Agéntica, garantiza la confidencialidad, integridad y disponibilidad tanto del propio agente como de los datos y sistemas con los que interactúa.

En este capítulo se enumeran los principales retos y tendencias en materia de seguridad aplicada a los Agentes de IA priorizando los aspectos Safety. Entender estos retos y tendencias permite comprender la necesidad de herramientas como el presente Decálogo.

Tendencias en **ciberseguridad**

Según un informe reciente de Gartner estas son las principales tendencias en ciberseguridad para el presente ejercicio:

1. **La IA Agéntica exige supervisión de ciberseguridad**
2. **La volatilidad regulatoria global impulsa los esfuerzos de resiliencia cibernética**
3. **La computación postcuántica pasa a estar en los planes de acción**
4. **La gestión de identidad y acceso se adapta a los agentes de IA**
5. **Las soluciones SOC impulsadas por IA desestabilizan la operativa habitual**
6. **GenAI rompe con las tácticas tradicionales de concienciación en ciberseguridad**

Tal y como se puede ver las tendencias 1, 4, 5 y 6 impactan directamente en IA y, de forma más específica las tendencias 1 y 4 a los Agentes IA.

TENDENCIA 1: LA IA AGÉNTICA EXIGE SUPERVISIÓN DE CIBERSEGURIDAD

"La IA Agéntica está siendo utilizada rápidamente por empleados y desarrolladores, creando nuevas superficies de ataque. Las plataformas no-code/low-code y el vibe coding amplían aún más esto, impulsando la proliferación de agentes de IA no gestionados, código no seguro y posibles violaciones del cumplimiento normativo."
(Gartner, Top Cybersecurity Trends 2026)

TENDENCIA 4: LA GESTIÓN DE IDENTIDAD Y ACCESO SE ADAPTA A LOS AGENTES DE IA

"El auge de los agentes de IA está introduciendo nuevos desafíos a las estrategias tradicionales de gestión de identidad y acceso (IAM), especialmente en el registro y gobernanza de identidad, la automatización de credenciales y la autorización basada en políticas para actores de máquinas. No abordar estos problemas aumentará el riesgo de incidentes de ciberseguridad relacionados con el acceso a medida que los agentes autónomos se vuelvan más frecuentes."
(Gartner, Top Cybersecurity Trends 2026)

OWASP Top 10 for Agentic Applications

OWASP (Open Web Application Security Project) es una fundación internacional sin ánimo de lucro dedicada a mejorar la seguridad del software. Su proyecto más famoso es el "OWASP Top 10 Web Application Security Risks", un documento de referencia para profesionales de seguridad.

Con el mismo espíritu OWASP ha lanzado la iniciativa "OWASP Gen AI Security Project" con el objetivo de contribuir a estandarizar la seguridad en la IA Generativa.

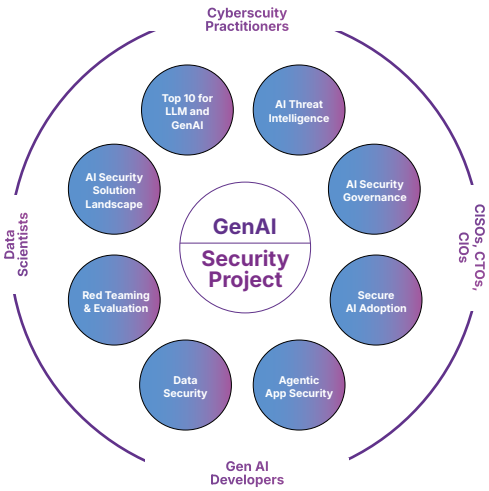

Figura 3: Iniciativa GenAI de OWASP[1]

En el marco de este proyecto y a efectos del presente decálogo, resulta especialmente relevante el reciente informe "OWASP Top 10 for Agentic Applications", que identifica las diez amenazas de mayor impacto en el ámbito de la IA Agéntica.

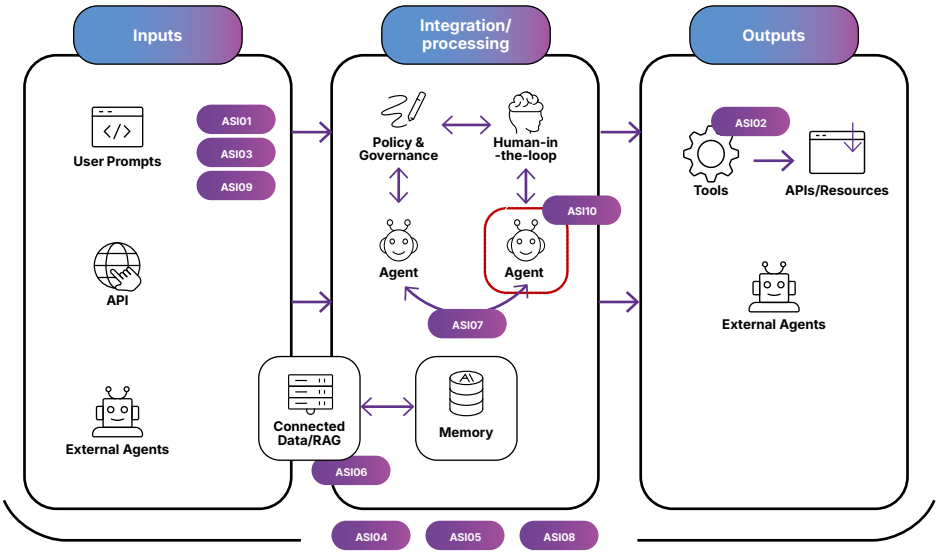

Figura 4: *Mapa de las top 10 amenazas identificadas por OWASP*

[1] *https://genai.owasp.org/*

Las 10 amenazas identificadas en el documento son:

ASI01 - Agent Goal Hijack:

Manipulación directa o indirecta de las instrucciones o inputs de un agente para desviar sus objetivos y decisiones hacia fines no autorizados.

ASI02 - Tool Misuse & Exploitation:

Uso indebido de herramientas legítimas por parte del agente, dentro de sus permisos, que deriva en acciones dañinas, costosas o de exfiltración.

ASI03 - Identity & Privilege Abuse:

Abuso de identidades, delegaciones o privilegios heredados entre agentes para escalar permisos y ejecutar acciones no autorizadas.

ASI04 - Agentic Supply Chain Vulnerabilities:

Introducción de comportamientos maliciosos o inseguros a través de modelos, herramientas, agentes o artefactos de terceros integrados dinámicamente.

ASI05 - Unexpected Code Execution (RCE):

Ejecución de código no previsto generada o activada por el agente que conduce a compromisos del sistema o del entorno de ejecución.

ASI06 - Memory & Context Poisoning:

Corrupción persistente de la memoria, contexto o RAG del agente para sesgar decisiones futuras, facilitar abusos o exfiltrar información.

ASI07 - Insecure Inter-Agent Communication:

Falta de autenticación, integridad o validación semántica en las comunicaciones entre agentes que permite spoofing, manipulación o replay.

ASI08 - Cascading Failures:

Propagación y amplificación de un fallo inicial entre agentes y flujos autónomos hasta provocar impactos sistémicos.

ASI09 - Human-Agent Trust Exploitation:

Explotación del exceso de confianza humana en el agente para inducir aprobaciones o decisiones perjudiciales bajo apariencia legítima.

ASI10 - Rogue Agents:

Agentes que cambian su comportamiento y actúan de forma autónoma, engañosa o dañina dentro del ecosistema.

AI Agents attack **Matrix**

Desde hace años la organización MITRE mantiene el framework MITRE ATT&CK de "tácticas, técnicas y procedimientos" de ataques de ciberseguridad, es un framework de gran prestigio ampliamente adoptado por las organizaciones en su defensa.

Más recientemente, MITRE mantiene el framework MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) que también recoge "tácticas, técnicas y procedimientos" de ataques contra sistemas de IA.

A partir de este framework, una comunidad de expertos mantiene un interesante repositorio de Tácticas y Técnicas específicas sobre Agentes IA[2]. Las 16 tácticas que se definen en este repositorio son:

1. **Reconnaissance (Reconocimiento):** El adversario intenta recopilar información para planificar futuras operaciones, como buscar repositorios de código, bases de datos técnicas o analizar vulnerabilidades de la IA.
2. **Resource Development (Desarrollo de recursos):** Acciones para establecer o adquirir capacidades que apoyen el ataque, como crear prompts específicos, obtener infraestructura o publicar modelos y conjuntos de datos envenenados.
3. **Initial Access (Acceso inicial):** Métodos para entrar en la red o interactuar con el sistema de IA, incluyendo la manipulación del usuario, el envenenamiento de RAG o el compromiso de la cadena de suministro.
4. **AI Model Access (Acceso al modelo de IA):** El atacante busca obtener acceso a la inferencia del modelo (vía API), al producto habilitado con IA o incluso acceso físico o total al modelo.
5. **Execution (Ejecución):** El intento de ejecutar código malicioso o instrucciones no deseadas a través del sistema, destacando técnicas como la inyección de prompts (directa o indirecta) y la invocación de herramientas del agente.
6. **Persistence (Persistencia):** Técnicas para mantener el acceso o control a largo plazo, como el envenenamiento del contexto del agente, la manipulación de la memoria o la autorreplicación de prompts.
7. **Privilege Escalation (Escalada de privilegios):** El adversario intenta ganar permisos de mayor nivel para realizar acciones restringidas, usando métodos como el jailbreak o la explotación de instrucciones del sistema.
8. **Defense Evasion (Evasión de defensas):** Estrategias para evitar ser detectado, como ofuscar prompts, silenciar instrucciones de seguridad o usar caracteres invisibles (ASCII smuggling).
9. **Credential Access (Acceso a credenciales):** El robo de nombres de usuario, contraseñas o tokens, buscándolos en la configuración del agente, en sistemas de almacenamiento (RAG) o mediante recolección directa.
10. **Discovery (Descubrimiento):** Intento de averiguar detalles del entorno, como la ontología del modelo, instrucciones del sistema (system prompts), definiciones de herramientas o limitaciones del sistema de IA.

[2] *https://ttps.ai/*

11. **Lateral Movement (Movimiento lateral):** El atacante se mueve a través del entorno desde el sistema inicial comprometido hacia otros sistemas, por ejemplo, envenenando mensajes o recursos compartidos.
12. **Collection (Recolección):** Recopilación de datos de interés para los objetivos del atacante, incluyendo historial de chats, mensajes de usuarios o datos provenientes de repositorios de información.
13. **AI Attack Staging (Preparación de ataque de IA):** Actividades para verificar el éxito de un ataque o manipular el modelo para que se comporte de forma específica antes del impacto final.
14. **Command and Control (Mando y control):** El establecimiento de comunicación con los sistemas comprometidos para controlarlos, utilizando APIs de servicios de IA o shells inversos.
15. **Exfiltration (Exfiltración):** El robo de datos del sistema, que puede realizarse mediante el renderizado de imágenes, enlaces clicables o abusando de herramientas de inferencia para extraer información sensible.
16. **Impact (Impacto):** Acciones finales para manipular, interrumpir o destruir sistemas y datos, como erosionar la integridad del modelo, causar denegación de servicio (DoS) o generar costes excesivos por uso de recursos.

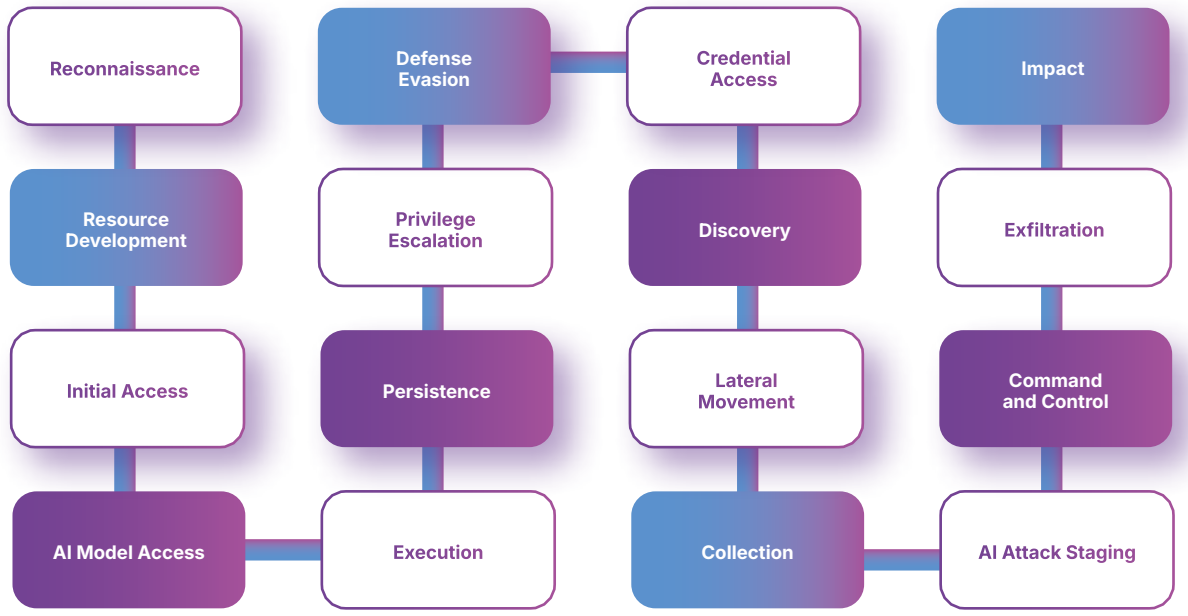

Figura 5: *AI Agents Attack Matrix*

Lethal Trifecta

Lethal Trifecta nos ayuda a definir cuán en riesgo está un agente en función de tres simples preguntas.

1. **¿Tiene acceso a datos sensibles?**
2. **¿Es capaz de consumir contenido no confiable?**
3. **¿Tiene capacidad de comunicarse con el exterior de nuestra compañía?**

Si se responde con un SÍ a estas preguntas, podemos considerar que el agente está en alto riesgo, vulnerable a ataques de tipo "indirect prompt injection" (según OWASP top 10 for Agentic Applications).

Esta variante de la amenaza top 1 "prompt injection" se aprovecha de que, si las manipulaciones no vienen directamente de la interacción con el usuario mal intencionado, sino que provienen de un recurso consumido por el agente, son mucho más difíciles de distinguir por sus mecanismos internos y salvaguardas.

En este escenario, el atacante induce al agente a consumir una página web bajo su control que contiene una inyección de instrucciones a nivel de sistema. El agente, incapaz de distinguir este contenido como malicioso, acata dichas instrucciones accediendo a información sensible de la compañía y exfiltrándola gracias a su capacidad de comunicarse hacia el exterior.

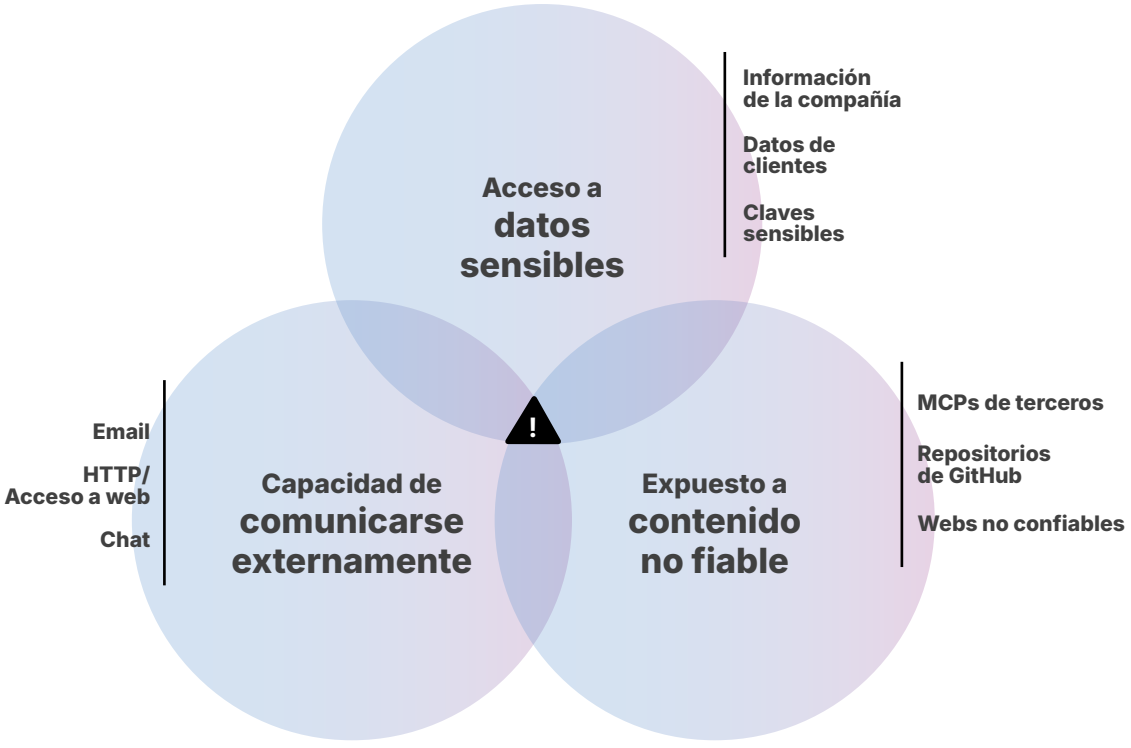

Figura 6: *Lethal Trifecta*

Decálogo

Una vez expuestos los conceptos teóricos de IA Agéntica, así como referencias a las previsiones de despliegue y principales amenazas y retos, se desarrollan en este capítulo los 10 principios que componen el Decálogo de Seguridad en la IA Agéntica.

01. IDENTIFICACIÓN:
CONOCE A TUS AGENTES Y SUS CARACTERÍSTICAS

El primer principio establece la necesidad de disponer de un inventario de agentes, además de mecanismos que permitan el descubrimiento de aquellos que no han pasado por los procesos de gobierno establecidos, es decir, considerados como “Shadow AI”.

El inventario de agentes debe identificar para cada uno de ellos los siguientes aspectos:

- **Infraestructura del agente:** Existe una responsabilidad compartida que afecta a la información que soporta el agente, y a la capacidad de influir en la gestión del cambio asociada, que varía en función del tipo de infraestructura. Esta puede ser:
  - Modo SaaS (Gemini, Manus AI, Copilot Investigador).
  - Low Code (p.ej. N8N, Copilot Studio).
  - PaaS (Azure AI Agents, Langchain).
  - Infraestructura on premise desplegada por la organización.
- **Entornos disponibles:** Desarrollo, Calidad, Producción, etc.
- **Modelos subyacentes, versión y fabricante:** identificar para cada LLM usado qué fabricante es, qué versión se utiliza (p.ej. DeepSeek v3) y sus características de implementación (local, cloud de fabricante, etc.).
- **A qué herramientas tiene acceso** y cuáles son los mecanismos de conexión (MCP/APIs/etc.) así como tipo de permisos (lectura, escritura), es el caso p.ej. de un agente que puede realizar búsquedas en internet, leer/escribir en base de datos, abrir tickets en una plataforma, etc.

- Con qué **otros agentes y cadenas de agentes** se relaciona (A2A), Por cadena de agentes se entienden aquellos agentes que internamente ejecutan otros agentes y que acaba generando el paradigma de la IA Agéntica.
- **Flujos de datos y dependencias:** Para cada actor que interactúa con el agente identificar cuál es el flujo de cada dato que se procesa y qué actor es el propietario de este.
- **Clasificación y categorización de los agentes.**
  - Según el proceso de la organización al que da servicio (p.ej. atención al cliente, fabricación de componentes, etc.).
  - Según la dimensión de compliance afectada (si le afecta AI Act, GDPR, NIS2/DORA, CRA, PCIDSS, etc.).
  - Según sus **características de comportamiento:**
    - *Estático:* cuando su estado y comportamiento no varían durante la ejecución del proceso.
    - *Dinámico:* cuando el agente modifica su comportamiento en función del feedback del usuario, realimentándose y evolucionando con el uso.
  - Según su **ciclo de vida:**
    - *Persistente:* cuando se mantiene operativo de forma indefinida (p. ej., un chatbot con memoria o contexto persistente).
    - *Efímero:* cuando el agente se ejecuta de forma puntual y desaparece tras completar una tarea concreta (p. ej., un agente invocado por otro agente para una acción específica).

- **Grado de Criticidad para el negocio:** atendiendo al impacto potencial del agente en términos operativos, financieros, reputacionales o regulatorios.
- **Ownership:** Para cumplir con los principios de accountability se deben identificar los siguientes roles:
  - *Responsable de Negocio del Agente:* vela para que el agente cumpla con los objetivos del negocio.
  - *Responsable Técnico del Agente:* implementa y modifica la parametrización del agente para cumplir con su objetivo.
  - *Responsable de Seguridad del Agente:* supervisa el correcto funcionamiento de este en términos de riesgo, detecta amenazas y ejecuta tareas de mitigación y respuesta.

Es fundamental que el inventario sea vivo y se actualice dinámicamente conforme evoluciona la organización agéntica. Los inventarios estáticos mediante diagramas tienen un alto riesgo de quedar pronto desfasados.

Adicionalmente al inventario de agentes, es conveniente que las organizaciones adopten herramientas **SBOM (Software Bill Of Materials)** que permite inventariar en detalle todos los componentes software utilizados, con datos clave tales como:

- Componentes
- Detalle de modelos y procedencia
- Dependencias de bibliotecas
- Conjuntos de datos de entrenamiento
- Información de la versión
- Datos de vulnerabilidades asociadas
- Información sobre licencias
- Referencias externas (knowledge base del componente)

En este ámbito es significativo destacar la iniciativa AIBOM de OWASP enfocada a los componentes de un sistema de IA.

El proceso de desarrollo y actualización de un agente es un proceso dinámico en el que es fácil que se requieran iteraciones y cambios continuos (cambios en el agente debidos al feedback, corrección de errores, optimización de proceso, cambio de componentes internos, etc.). Esto conlleva la necesidad de disponer de un proceso de gestión del cambio formal, en el que se reflejen aspectos tales como:

- Cambios en el prompt
- Cambios en las versiones del LLM o del propio LLM
- Otros aspectos relativos al agente

Conforme las organizaciones avancen en el paradigma de la organización agéntica, se prevé la aparición de volúmenes muy elevados de agentes operando, por ello es factible pasar de identificar agentes de forma unívoca, a tratarlos, total o parcialmente, de forma agregada en base a características comunes.

02.IDENTIDAD VIP:
IDENTIFICA, CLASIFICA Y DA IDENTIDAD A TUS AGENTES Y SUS PERMISOS

Un Agente IA es una identidad sintética que tiene características singulares tales como actividad continua (pueden trabajar 7x24), autonomía y múltiples instancias (puede haber miles de instancias de un mismo agente en ejecución de forma simultánea). Por ello, es preciso concienciarnos de que los agentes se deben tratar como una identidad VIP, debiendo considerar aspectos tales como:

- Cada agente debe disponer de un identificador (ID) unívoco que permita su trazabilidad a lo largo de todo su ciclo de vida.
- Disponer de un sistema de gestión de identidades de agentes sólido. Al provisionar un agente se genera su ID correspondiente, y al desprovisionarlo se elimina de forma controlada tanto el ID como las credenciales asociadas.
- Los privilegios deben cumplir con las siguientes características:
  - RBAC (Acceso basado en roles) considerando RBAC dinámico para aquellos agentes que varían sus necesidades de acceso en función de su estado.
  - Mínimo privilegio basado en la necesidad de conocer y hacer.
- Todas las credenciales asociadas a las identidades que intervienen en la IA Agéntica deben tener mecanismos sólidos de protección (API Key, HTTP secrets, etc.)
- Para cada agente y/o herramienta con el que se relaciona debe tener mecanismos sólidos de autenticación y cifrado de comunicaciones.
- Disponer credenciales JIT (Just In Time) para agentes efímeros, que se generen y destruyan en cada invocación.
- Accountability: Se debe diferenciar el responsable del agente bajo estos principios:
  - ¿El agente actúa en nombre de una persona? Accountable=Usuario que lo invoca. Se requiere para casos en que p.ej. una acción malintencionada de un usuario frente un agente provoca un daño específico.
  - ¿El agente actúa en nombre del propio agente? Accountable=Requiere definición específica. Se requiere para casos en que p.ej. una acción errónea de un agente sin influencia humana provoca un daño específico.
- En el caso específico de Agentes de Navegación (p.ej. ChatGPT Atlas o Comet) así como CUA (Computer User Agent) conviene identificar los dos modos posibles de interacción:
  - Modo ***"Logged in"***: el agente accede a sitios donde el usuario que lo invoca ya ha iniciado sesión y mantiene sus credenciales en la interacción.
  - Modo ***"Logged out"***: el agente actúa de forma aislada frente a estos sitios.

03. DATOS:
PROTEGE LOS DATOS EN TODO EL CICLO (INGESTA, USO, SALIDA Y MEMORIA)

Para que el agente IA pueda desempeñar de forma eficaz su tarea, es imprescindible garantizar la seguridad de los datos con con los que interactúa a lo largo de todo su ciclo de vida. Para ello, deben cumplirse los siguientes principios básicos orientados a proteger la ***Confidencialidad, Integridad, Disponibilidad y Autenticidad*** de los datos. Se requiere por tanto cumplir con los siguientes aspectos:

- **Calidad de datos:** es imprescindible verificar que los datos correctos alimentan al agente para evitar sesgos o resultados erróneos:
  - Evaluar la calidad y representatividad de los datos.
  - Asegurar que los Datasets involucrados sean relevantes.
  - Evitar introducir datos sesgados o desactualizados en el contexto.
  - Incluir políticas de retención y eliminación de datos en el ciclo del agente.
- **Gobernanza del contexto:** considerar todo el conjunto de datos que recibe el agente para tomar sus decisiones, aplicar principio de minimización de datos que consuma el agente, así como tener trazabilidad de la cadena de pensamiento. (p.ej. poder detectar si el agente ha accedido a datos clasificados a partir de una orden recibida, o poder detectar la confidencialidad de una cadena de pensamiento)
- **Proteger las bases de datos vectoriales que son consumidas por el agente**, por ejemplo, verificando que el RAG (Retrieval Augmented Generation) consumido por el agente no tiene acceso a información clasificada no autorizada o incluso información alterada intencionadamente.
- **Algunos agentes tendrán memoria persistente**, nuevamente debe protegerse porque puede contener información clasificada que el agente haya decidido almacenar con criterios de mejora de su propio performance.
- **Aplicar en los agentes principios de:**
  - *DLP (Data Loss Prevention):* poder detectar flujos anómalos de datos clasificados que pueden dar indicios de incidente de brecha.
  - *IRM (Information Right Management):* para la información clasificada disponer de mecanismos de cifrado y gestión de permisos.
  - *Minimización de datos:* para reducir al mínimo imprescindible los datos que el agente puede ver, recuperar, procesar o almacenar en cualquier fase de su ciclo de vida.
  - *PET (Privacy-Enhancing Technologies):* para proteger la privacidad de la información mediante técnicas del tipo:
    - Anonimización o Seudonimización, p.ej: Data Masking, K-anonymity, tokenización, etc.
    - Cifrado homomórfico para poder realizar cálculos sobre datos cifrados sin tener que descifrarlos primero.
    - Generación de Datos sintéticos: ideal para entrenar agentes sin datos reales.
    - Etc.

04. PROTECCIÓN:
PROTEGE TUS AGENTES Y SU ENTORNO

Un ataque puede iniciarse en cualquiera de los elementos que forman parte del agente y su ecosistema relacionado, por ello la protección debe ser integral, considerando los siguientes aspectos fundamentales:

- **Los prompts de configuración (system prompt)** del agente son elementos clave para proteger la integridad de sus acciones, por eso deben definirse incorporando mecanismos de protección en el propio prompt. Por ejemplo, si no quieres que el agente pueda dar determinada información clasificada, debe indicarse claramente a nivel de configuración.
- **Evitar inyectar información externa dentro del system prompt** para configurar el comportamiento del agente, si no mediante el user prompt. En la información externa se podrían incluir caracteres especiales u órdenes maliciosas que alteraran el comportamiento del agente a nivel de system prompt. Dicho de otra manera: se debe separar las instrucciones de los datos.
- Minimizar la probabilidad de que el agente ejecute acciones maliciosas inducidas por el atacante mediante el **despliegue de "guardrails"**, como elemento fundamental que limita y supervisa el comportamiento del agente para que opere dentro de límites definidos de seguridad, cumplimiento y propósito.
- Asegurar que el agente tiene solo las capacidades que realmente precisa:
  - **Sobre las herramientas:** acotar las herramientas y acciones que pueda ejecutar al máximo, p.ej. si el agente está concebido para que haga conexiones https no permitas que pueda ejecutar comandos de Linux.
  - **Sobre el modelo usado:** En la medida de lo posible intentar utilizar modelos DSLM (Domain Specific Large Model) en lugar de LLM, dado que el DSLM se concibe como modelo de IA entrenado y acotado para un propósito específico.
  - **Sobre los recursos:** Acotar los permisos de acceso a los recursos, siendo un ejemplo claro el caso cuando el agente debe hacer una consulta a una base de datos, los permisos que tenga sean solo de lectura. Otro ejemplo claro es que se evite ejecutar acciones a nivel de sistema operativo con más privilegios de los precisos.
- Los agentes que ejecuten código deben hacerlo dentro de **Sandbox aisladas.**
- Al igual que en sistemas deterministas **debe separarse la restricción de entornos** en los que opera un agente (Desarrollo, Test, Calidad, Producción).
- Aplica los principios de ZeroTrust (nunca confiar en ningún usuario, dispositivo o sistema y verificar de forma continua toda solicitud de acceso) a todos los niveles:
  - Usuario ↔ Agente
  - Agente ↔ Agente
  - Agente ↔ Herramienta
  - Agente ↔ Modelo.

05. GESTIÓN DE LA AUTONOMÍA:
CONTROLA LA AUTONOMÍA Y LOS LÍMITES DE ACTUACIÓN

Los agentes tienen la característica de poder ejecutarse de forma permanente y sin asistencia humana, es por ello preciso acotar al máximo su autonomía o, dicho de otra forma, qué acciones puede desencadenar de forma proactiva. Para ello deben contemplarse los siguientes aspectos:

- **Establecer controles proporcionales a la criticidad de la acción que va a desempeñar.** Es muy recomendable incorporar controles Human In The Loop (HITL) en aquellas que puedan suponer mayor riesgo. Ejemplos claros serían aquellas acciones que puedan afectar la operativa de la organización o incluso la seguridad de las personas.
  - Estos humanos supervisores deben ser entrenados para reconocer intentos de manipulación de la propia IA o comportamientos inesperados fuera del objetivo.
  - Asimismo, para decisiones de riesgo crítico considerar incluir doble validación humana.
- Complementariamente al punto anterior es muy importante **considerar el riesgo de HITL overwhelming (fatiga de decisión)**, referida a aquellas situaciones en que se generan tantas demandas de decisión que superan las capacidades de la persona, provocando situaciones de cuello de botella o a la toma de decisiones erróneas debido a la fatiga.
- **Gestión dinámica de la autonomía mediante tres principios:**
  - *Desconexión:* debe existir un mecanismo que permita al agente ser desconectado y que el proceso del que formaba parte pueda seguir en modo manual. Complementariamente, para aquellos agentes de mayor riesgo debe existir un "kill switch" (botón rojo) que permita la interrupción inmediata del agente ante la primera señal de alarma. El kill switch requiere que se definan "líneas rojas" en las que se debe activar, así como se deben realizar entrenamientos periódicos para ponerlos a prueba.
  - *Escalado:* debe existir un mecanismo que permita al agente escalar en autonomía de forma ordenada y auditable, el ejemplo más claro es en las fases iniciales del agente, en las que requiere una supervisión estrecha y, conforme gana fiabilidad acumulando un mayor número de decisiones acertadas, aumenta su autonomía.
  - *Reversibilidad:* debe existir un mecanismo para que, en caso de que se detecte un funcionamiento potencialmente erróneo, el agente reduzca su autonomía incluyendo controles intermedios.
- **Mecanismos de fallback:** procurar que el agente tenga mecanismos redundantes para continuar con su operativa en caso de fallo de algún componente. Un ejemplo sería el caso en que un agente se apoya en consulta a un LLM local "A" y puede acceder a un LLM "as a Service" en caso de que falle la disponibilidad del primero.
- **Gestión de costes:** prestar atención a los costes dinámicos derivados del comportamiento autónomo del agente (tokens y llamadas a modelos, uso en cascada de tools/APIs, etc.). Existen casos conocidos en que la toma de decisiones de un agente, al no ser determinista, varía su "cadena de pensamiento" provocando variaciones significativas de coste de ejecución, o casos en los que una petición simple desemboca en la consulta a sitios web que contienen códigos HTML extremadamente largos y que son procesados por el agente.
- **Human manipulation:** existe el riesgo de que un agente comprometido pueda intentar manipular las decisiones del humano con quien actúa, por ello deben establecerse controles sobre decisiones humanas para detectar estas potenciales manipulaciones.

06. MONITORIZACIÓN:
MONITORIZA COMPORTAMIENTO, DECISIONES E INTERACCIONES

Las capacidades de ejecución 7x24 de los agentes y su complejidad provocada por el posible anidamiento de agentes, generan interacciones difícilmente rastreables por el humano de forma explícita. Por ello, es conveniente incorporar mecanismos de detección y análisis, tales como:

- **Observabilidad:** se deben establecer mecanismos para registrar y monitorizar el comportamiento interno y la lógica del agente:
  - Razonamientos o inferencias
  - Cadenas de pensamiento (Chain-of-Thought)
  - Decisiones de ejecución de herramientas
  - Decisiones de ejecución de otros agentes
  - Interacciones con usuarios
  - Detección de anomalías y de comportamiento (drift detection). Parte de la **Observabilidad** tiene como objetivo poder detectar, a nivel operativo, cuando el agente se comporta fuera de los objetivos establecidos, por ejemplo, con inferencias o Chain-of-Thought que se alejan del comportamiento esperado y que pueden generar una amenaza interna a la operativa de la organización, generar sesgos discriminatorios o incumplir la ética de la organización.
- **Detección de anomalías y abusos:** capacidad para detectar situaciones en que un actor desconocido esté intentando enviar prompts con objetivos maliciosos.
- **Monitorización de acceso a datos:** detección de acceso a datos clasificados fuera del propósito, detección de intentos de exfiltración o manipulación.
- **Monitorización de picos de consumo:** para poder detectar posibles abusos o compromisos, con el objetivo de tener un indicador temprano de ataque o pérdida de control.

En general, el objetivo es tener capacidad no solo para detectar anomalías sino también realizar un análisis forense en caso de incidente (análisis de la Chain-of-Thought, e-Discovery para obtener evidencias, etc.).

07. RESPUESTA:
PREPÁRATE PARA INCIDENTES CAUSADOS POR O A TRAVÉS DE AGENTES

Este bloque está intrínsecamente relacionado con la Monitorización de comportamiento, decisiones e interacciones y tiene como objetivo definir, implementar y entrenar capacidades de respuesta frente a incidentes.

En caso de incidentes en IA Agéntica, debido al alto volumen de interacciones y elementos involucrados es conveniente considerar que debemos tener capacidades de respuesta automatizadas desde Agentes IA de defensa.

- **Suspensión de la autonomía:**
  - Activar la suspensión de la autonomía: desarrollado en punto anterior
  - Si hay involucrados más agentes activar la desconexión en cascada de agentes
- **Preservar el contexto cognitivo:** definido en el apartado Observabilidad.
- **Identificar fallos de control:**
  - Fallo del guardrail
  - Fallo del HITL
  - Desviaciones de la lógica del agente versus la política declarada
- **Descontaminación cognitiva:**
  - Eliminar memoria persistente, datos envenenados o instrucciones heredadas
  - Reset parcial o total del estado del agente
- **Reconfiguración segura:**
  - Adaptar system prompt para reforzar control defectuoso
  - Mejorar comportamiento del guardrail
  - etc.
- **Vuelta a la normalidad:**
  - Activar la cadena de ejecución de agentes implicados de forma progresiva
  - Supervisión estrecha de la activación para detectar en fase temprana comportamientos no deseados.
- **Lecciones aprendidas:**
  - En base a toda la información obtenida realizar un análisis detallado de puntos de mejora para poder implementar dichas mejoras posteriormente, del tipo:
    - » Actualizar agent policies, patrones de diseño entre agentes, etc.
    - » Registrar el incidente en el repositorio de incidentes IA
    - » Revisar controles HITL
    - » Evaluar incorporar agentes supervisores que automaticen la detección y respuesta
    - » Definir/actualizar runbooks implicados

08. GOBIERNO:
CUMPLE CON LEGISLACIÓN, ESTÁNDARES Y RESPONSABILIDADES EMERGENTES

Cada organización dispone de un marco legal externo y un marco normativo definido a nivel sectorial o interno que regula sus principios de comportamiento. A nivel de Agentes IA deben considerarse referentes tales como:

- **AI Act como reglamento principal de la Unión Europea**
- **GDPR**
- **NIS2/DORA**

Estándares de referencia a nivel de IA:

- **ISO/IEC 42001:2023** – AI Management System
- **ISO/IEC 23894:2023** – Information technology — Artificial intelligence — Guidance on risk management

Adicionalmente, es conveniente tomar como referencia publicaciones de equipos expertos tales como:

- **Específicas de Agentes IA:**
  - OWASP Top 10 for Agentic Applications
  - CSA MAESTRO
- **Generales sobre IA:**
  - NIST AI Risk Management Framework
  - OWASP Top 10 Risk & Mitigations for LLMs and Gen AI Apps
  - Microsoft Responsible AI Standard
  - Google Secure AI Framework (SAIF)
  - AWS Responsible AI Framework
  - MITRE ATLAS

Para poder afrontar de forma holística el reto que todo ello plantea es muy conveniente definir un Comité de Gobierno de la IA que cubra los siguientes objetivos relativos a IA Agéntica:

- **Obligatorio:** "visión defensiva" orientada a identificar y mitigar riesgos a todos los niveles: ciberseguridad, protección de datos personales y compliance.
- **Opcional:** "visión ofensiva", para aquellas organizaciones que incorporen la IA Agéntica en su propuesta de valor, con la misión principal de definir arquitectura de referencia, infraestructura necesaria, priorización de casos de uso, promoción y acompañamiento de adopción, etc.

El marco de Gobierno debe establecer pilares clave como una política de Agentes IA y planes de formación específicos.

09. GESTIÓN DEL RIESGO:
EVALÚA Y MEJORA CONTINUAMENTE CONFIANZA, RIESGO Y DESEMPEÑO DEL AGENTE

Debe existir un criterio de clasificación que determine la relevancia del agente para la organización considerando aspectos tales como:

- **Cuál es el blast radius (radio de explosión)** referido al alcance máximo del daño potencial que puede causar un fallo del agente:
  - ¿A qué información tiene acceso de lectura y escritura?
  - Agency, ¿qué puede hacer el agente? (no es lo mismo un agente que da recomendaciones que uno que realiza acciones)
  - Con qué sistemas externos está conectado, diferenciando también consultas de acciones.

- **Cómo afectaría un fallo de disponibilidad o integridad de las acciones** del agente.

Nos encontramos ante una encrucijada que consideramos relevante tratar en lo que refiere a la gestión de riesgos. La metodología tradicional de análisis de riesgos periódico pierde sentido frente al reto de la IA Agéntica, en el que el entorno evoluciona de forma dinámica cambiando constantemente, lo que requiere explorar métodos de análisis dinámicos. Asimismo, estos métodos de análisis dinámicos plantean una carga de trabajo superior que, probablemente, solo pueda afrontarse apoyándose en sistemas potenciados por IA.

Conviene medir de forma empírica las afectaciones, utilizando métodos como por ejemplo:

- **Testeo adversarial/Red Team continuo** atendiendo a los siguientes aspectos:
  - No solo considerando las amenazas ciber tradicionales.
  - No solo poniendo a prueba el LLM mediante técnicas como prompt injection, etc.
  - Teniendo en cuenta el Agente o Agentes en su conjunto (poner a prueba herramientas conectadas, el RAG asociado, el propio LLM, etc.) Es decir, analizando todo el ecosistema del agente o agentes.
  - Usando recursos que definen las técnicas y las tácticas aplicables a sistemas agénticos, como por ejemplo el AI Agents Attack Matrix.
  - Utilizando safety datasets que pongan a prueba el agente no solo para evaluar la probabilidad de dar datos erróneos sino también la probabilidad de que realice acciones ofensivas (p.ej.: Nemotron-AIQ Agentic Safety Dataset).

- **Validaciones periódicas:** complementar las pruebas de testing adversario/redteaming con revisiones técnicas periódicas como, p.ej.:
  - Verificar que se cumplen los criterios de gobierno identidad definidos al inicio (p.ej. que el agente mantiene los permisos estrictamente necesarios para su objetivo).
  - Revisar cuándo se renovó por última vez las credenciales de todo el entorno de acción del agente.
  - Evaluar necesidad de actualizar los guardrails y system prompts conforme surjan nuevas amenazas o se requiere acotar más el comportamiento del agente.
  - Verificar que el RAG mantiene su consistencia (no contiene información contaminada).
  - Asegurar que los mecanismos de DLP/IRM se mantienen operativos.
  - Etc.

10. CADENA DE SUMINISTRO: CONTROLA TUS PROVEEDORES

La IA Agéntica depende mucho de la consistencia de la cadena de suministro, requiere prestar atención a la seguridad de distintos elementos:

- **Infraestructura sobre la que corren los LLMs**, que puede ser:
  - Recurso cloud propio del proveedor LLM (OpenAI, Anthropic, etc.)
  - Recurso cloud contratado de forma específica para el LLM por la organización para hospedar LLMs locales (AWS, Azure, etc.)
  - Infraestructura on premise desplegada por la organización.
- **Infraestructura sobre la que corren los agentes** (la infraestructura donde se despliegan los agentes puede ser distinta de donde se ejecutan los LLMs locales), tratado en el capítulo "01 Identificación".
- El propio **modelo LLM** utilizado.
- **Repositorios de modelos open source** (p.ej. Hugging Face).
- Las **librerías de código** implicadas.
- **APIs y MCPs** de herramientas de terceros.

Ante ello debemos tomar acciones tales como:

- **Inventariar toda la cadena de suministro de la IA Agéntica** (todos los elementos enumerados anteriormente).
- Considerar, con espíritu crítico, **en qué caso se deben desplegar agentes locales o mantenerlos en la propia del fabricante.**
- **Verificar la integridad de los modelos** descargados o de las APIs utilizadas en caso de modelos cloud.
- **Aplicar criterios de Third Party Risk Management** para los proveedores implicados
  - Categorizar el nivel de riesgo del proveedor para su elemento proporcionado.
  - Evaluar la madurez de seguridad del proveedor.
  - Cuál es su política de actualización de modelos e infraestructura gestionada.
  - Existencia de un plan de respuesta a incidentes de seguridad.
  - Evaluar la solidez y expectativas de continuidad del proveedor (estamos ante un mercado que se transforma de forma acelerada y conlleva el riesgo de que algunos players ahora relevantes desaparezcan de forma súbita).
  - Disponer de una metodología de revisión periódica del rating del proveedor.

Hay que considerar también que estos procesos potenciados mediante IA Agéntica tienen un sistema de coste variable que, dada la alta transformación que se está experimentando en este mercado, es susceptible de sufrir fluctuaciones de coste relevantes a lo largo del tiempo. Es por ello preciso mantener siempre la atención puesta en la dependencia adquirida por la organización en LLMs de terceros que, de sufrir un incremento de precio, podrían poner en riesgo la viabilidad económica de los procesos soportados por agentes que necesitan de esos LLMs para funcionar adecuadamente.

Decálogo vs OWASP TOP 10 for Agentic Applications

Con objetivo de validar el grado de cobertura del presente decálogo con relación a las amenazas actuales se ha realizado un ejercicio de mapeo, en el que se indica:

- **En rosa:** Aquellos puntos del decálogo que dan cobertura fuerte de las amenazas identificadas en el informe OWASP.
- **En azul:** Aquellos puntos del decálogo que dan una cobertura parcial de la amenaza.

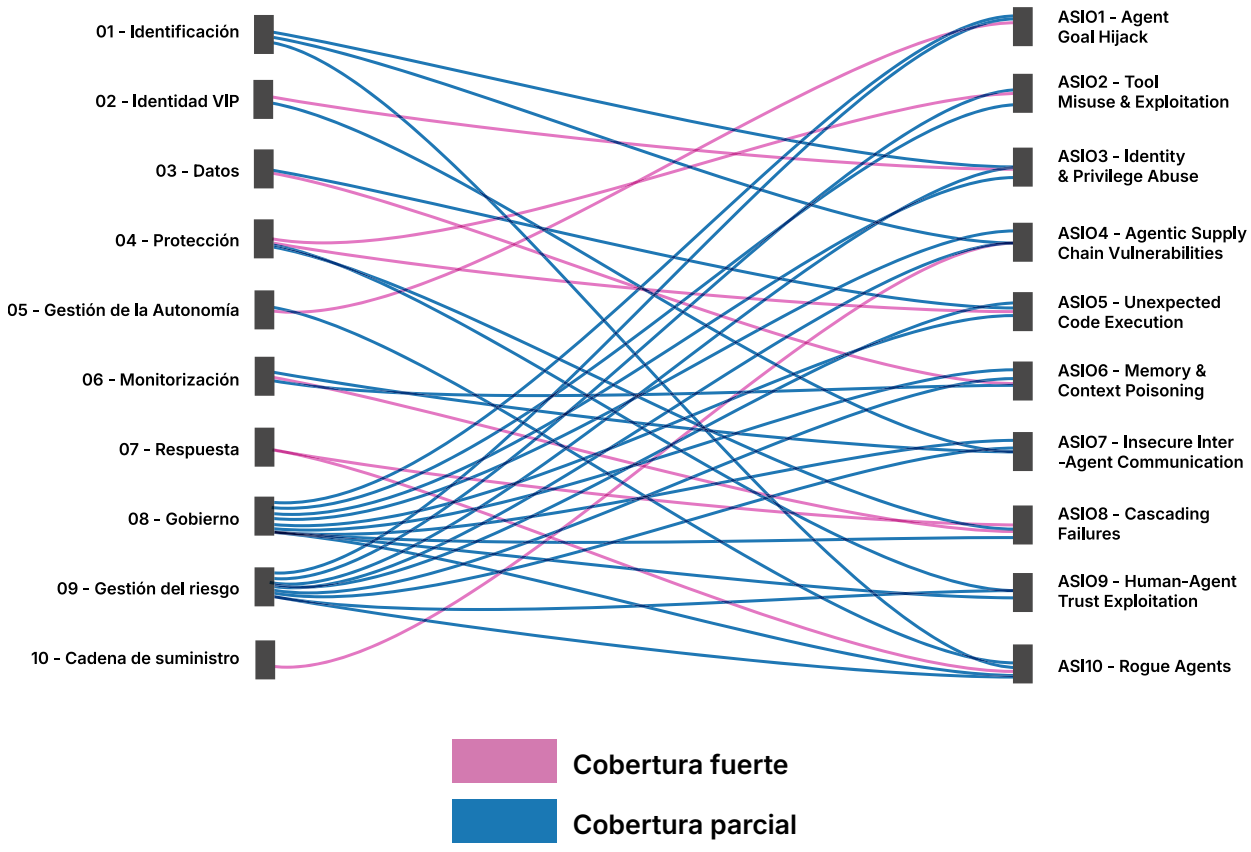

Figura 7: *Decálogo vs. OWASP Top 10 for Agentic Applications*

Decálogo/ OWASP Top 10	Agent Goal Hijack	Tool Misuse & Exploitation	Identity & Privilege Abuse	Agentic Supply Chain Vulnerabilities	Unexpected Code Execution	Memory & Context Poisoning	Insecure Inter-Agent Communication	Cascading Failures	Human-Agent Trust Exploitation	Rogue Agents
01 - Identificación	n/a	n/a	Parcial	Parcial	n/a	n/a	n/a	n/a	n/a	Parcial
02 - Identidad VIP	n/a	n/a	Fuerte	n/a	n/a	n/a	Parcial	n/a	n/a	n/a
03 - Datos	n/a	n/a	n/a	n/a	Parcial	Fuerte	n/a	n/a	n/a	n/a
04 - Protección	n/a	Fuerte	n/a	n/a	Fuerte	n/a	n/a	Parcial	Parcial	n/a
05 - Gestión de la Autonomía	Fuerte	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Parcial
06 - Monitorización	n/a	n/a	n/a	n/a	n/a	Parcial	Parcial	Fuerte	n/a	n/a
07 - Respuesta	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Fuerte	n/a	Fuerte
08 - Gobierno	Parcial	Parcial	Parcial	Parcial	Parcial	Parcial	Parcial	Parcial	Parcial	Parcial
09 - Gestión del riesgo	Parcial	Parcial	Parcial	Parcial	Parcial	Parcial	Parcial	Parcial	Parcial	Parcial
10 - Cadena de suministro	n/a	n/a	n/a	Fuerte	n/a	n/a	n/a	n/a	n/a	n/a

Tabla 1: *Decálogo vs. OWASP Top 10 for Agentic Applications*

Decálogo vs AI Agents Attack Matrix

Se muestra a continuación la cobertura de los 10 principios del catálogo en relación a las 16 tácticas del informe AI Agents Attack Matrix, en rosa se representan las coberturas fuertes y en azul las parciales.

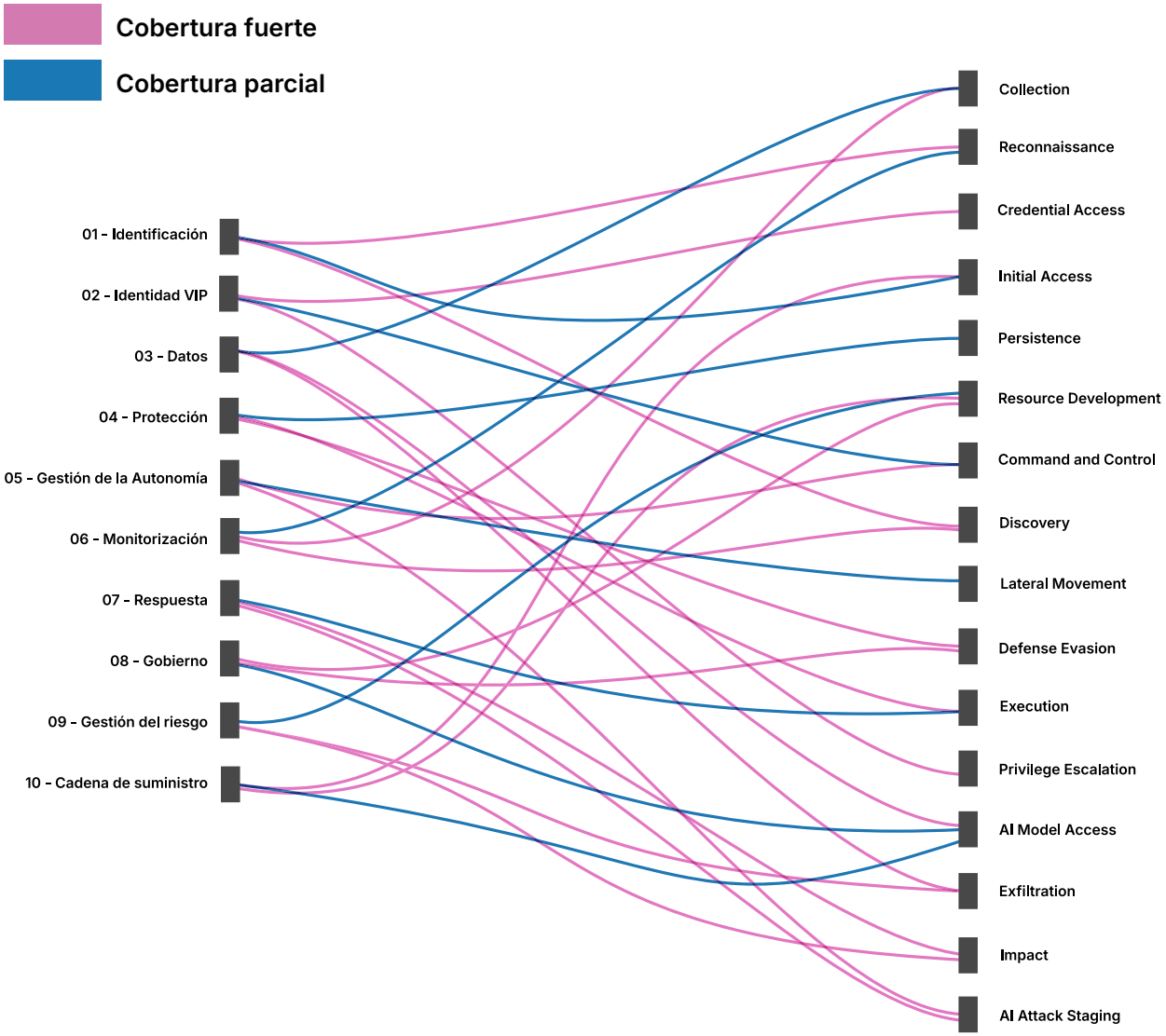

Figura 8: *Decálogo vs. AI Agents Attack Matrix*

AIAgents Attck Matrix/ Decálogo	01 - Identificación	02 - Identidad VIP	03 - Datos	04 - Protección	05 - Gestión de la Autonomía	06 - Monitorización	07 - Respuesta	08 - Gobierno	09 - Gestión del riesgo	10 - Cadena de suministro
Collection	n/a	n/a	Parcial	n/a	n/a	Fuerte	n/a	n/a	n/a	n/a
Reconnaissance	Fuerte	n/a	n/a	n/a	n/a	Parcial	n/a	n/a	n/a	n/a
Credential Access	n/a	Fuerte	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Initial Access	Parcial	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Fuerte
Persistence	n/a	n/a	n/a	Parcial	n/a	n/a	n/a	n/a	n/a	n/a
Resource Development	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Fuerte	Parcial	Fuerte
Command and Control	n/a	Parcial	n/a	n/a	Fuerte	n/a	n/a	n/a	n/a	n/a
Discovery	Fuerte	n/a	n/a	n/a	n/a	Fuerte	n/a	n/a	n/a	n/a
Lateral Movement	n/a	n/a	n/a	n/a	Parcial	n/a	n/a	n/a	n/a	n/a
Defense Evasion	n/a	n/a	n/a	Fuerte	n/a	n/a	n/a	Fuerte	n/a	n/a
Execution	n/a	n/a	n/a	Fuerte	n/a	n/a	Parcial	n/a	n/a	n/a
Privilege Escalation	n/a	Fuerte	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
AI Model Access	n/a	n/a	Fuerte	n/a	n/a	n/a	n/a	Parcial	n/a	Parcial
Exfiltration	n/a	n/a	Fuerte	n/a	n/a	n/a	n/a	n/a	Fuerte	n/a
Impact	n/a	n/a	n/a	n/a	n/a	n/a	Fuerte	n/a	Fuerte	n/a
AI Attack Staging	n/a	n/a	n/a	n/a	Fuerte	n/a	Fuerte	n/a	n/a	n/a

Tabla 2: *AI Agents Attack Matrix vs. Decálogo*

Referencias

AI Agents Attack Matrix Project. (2025). AI Agents Attack Matrix. https://ttps.ai/

Amazon Web Services. (2024). AWS Responsible AI Framework. https://aws.amazon.com/es/ai/responsible-ai/

Cloud Security Alliance. (2025). MAESTRO: Agentic AI Threat Modeling Framework. https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro

European Parliament & Council. (2016). General Data Protection Regulation (GDPR) (Reglamento UE 2016/679). https://www.boe.es/doue/2016/119/L00001-00088.pdf

European Parliament & Council. (2022). Digital Operational Resilience Act (DORA) (Reglamento UE 2022/2554). https://www.boe.es/buscar/doc.php?id=DOUE-L-2022-81962

European Parliament & Council. (2022). NIS2 Directive (Directiva UE 2022/2555). https://www.boe.es/buscar/doc.php?id=DOUE-L-2022-81963

European Union. (2024). AI Act. https://artificialintelligenceact.eu/

Gartner. (2026). Top Cybersecurity Trends for 2026. https://www.gartner.com/en/newsroom/press-releases/2026-02-05-gartner-identifies-the-top-cybersecurity-trends-for-2026

Google. (2024). Secure AI Framework (SAIF). https://safety.google/intl/en_in/safety/saif/

ISMS Forum. (2024). Gobierno de la IA. https://www.ismsforum.es/ficheros/descargas/gobierno-de-la-ia1762540754.pdf

ISO. (2023). ISO/IEC 23894:2023 — Information technology: Artificial intelligence — Guidance on risk management. https://www.iso.org/standard/77304.html

ISO. (2023). ISO/IEC 42001:2023 — Artificial intelligence management system. https://www.iso.org/standard/42001

Microsoft. (2024). Responsible AI Standard. https://www.microsoft.com/en-us/ai/principles-and-approach

Microsoft. (2025). Securing and governing the rise of autonomous agents. https://www.microsoft.com/en-us/security/blog/2025/08/26/securing-and-governing-the-rise-of-autonomous-agents/

MITRE. (2024). MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. https://atlas.mitre.org/

NIST. (2023). AI Risk Management Framework. https://www.nist.gov/itl/ai-risk-management-framework

NVIDIA. (2024). Nemotron-AIQ Agentic Safety Dataset (v1.0). https://huggingface.co/datasets/nvidia/Nemotron-AIQ-Agentic-Safety-Dataset-1.0

OWASP. (2024). AIBOM Generator Initiative. https://genai.owasp.org/ai-sbom-initiative/

OWASP. (2024). GenAI Security Project. https://genai.owasp.org/

OWASP. (2024). Top 10 Risks & Mitigations for LLMs and GenAI Apps. https://genai.owasp.org/llm-top-10/

OWASP. (2025). Generative AI and Agentic Security Solutions Landscape. https://genai.owasp.org/resource/owasp-genai-security-project-solutions-reference-guide-q2_q325/

OWASP. (2025). State of Agentic AI Security and Governance (v1.0). https://genai.owasp.org/resource/state-of-agentic-ai-security-and-governance-1-0/

OWASP. (2026). Top 10 for Agentic Applications. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

Russell, S., & Norvig, P. (2020). Artificial intelligence: A modern approach (3.ª ed.). Pearson. https://people.engr.tamu.edu/guni/csce625/slides/AI.pdf

GIA | GRUPO DE INTELIGENCIA ARTIFICIAL

isms forum | INTERNATIONAL INFORMATION SECURITY COMMUNITY

Decálogo de Seguridad en la **IA Agéntica**

Fundamentos para un ecosistema de agentes confiable