

Gobernar la ejecución de agentes

Plan para crear una arquitectura que proteja a las empresas que usan agentes

¿Dónde están mis agentes?	¿Qué pueden hacer?	¿Qué están haciendo?	¿Cómo respondo?

Elaborado por los miembros fundadores de Blueprint Alliance

Una arquitectura de referencia intersectorial para CISO, CIO, CTO y arquitectos de seguridad empresarial

Tabla de contenido

1	Resumen ejecutivo
2	Introducción
3	Principios del plan
4	Arquitectura de referencia del plan
5	Puesta en marcha con integradores de sistemas globales (GSI)
6	Conclusión: es hora de moldear el futuro de la seguridad de la IA
7	El plan en acción
8	Prácticas recomendadas para empezar

Resumen ejecutivo

Los agentes de IA son un desafío para el perímetro de seguridad tradicional de las empresas. En apenas dos años, los chatbots experimentales de LLM han dado paso a flujos de trabajo con agentes autónomos que ejecutan varios pasos. Antes, el software hacía lo que se le ordenaba; ahora, decide, actúa y se conecta por cuenta propia. Si bien esa autonomía ofrece un valor empresarial real, también plantea un desafío de gobernanza para el que la mayoría de los equipos de seguridad aún no están preparados.

[Gartner predice](#) que, para 2028, una empresa promedio de la lista Fortune 500 global tendrá más de 150 000 agentes en uso, frente a menos de 15 en 2025, lo que generará una gran proliferación de agentes, complejidad de TI y desafíos de gestión. Estos agentes no solo muestran datos, sino que también razonan, enrutan, llaman a las API de destino, crean subagentes y ejecutan operaciones de varios pasos en numerosos sistemas empresariales.

Sin un estándar arquitectónico unificado, las organizaciones se enfrentan a una brecha cada vez mayor en términos de identidad, seguridad y gobernanza: los agentes en la sombra se multiplican, las credenciales se heredan saltando límites de confianza y la ejecución se descontrola. La arquitectura de referencia Blueprint Alliance es un marco abierto y multiproveedor —publicado para el uso de toda la comunidad— que traslada la seguridad del perímetro de red a un plano de control unificado, Zero Trust y basado en agentes, lo que les brinda a los líderes tecnológicos una forma concreta de escalar las implementaciones con agentes sin perder el control.

Introducción

Durante décadas, la seguridad empresarial se basó en una clara separación entre identidades humanas (gestionadas mediante IAM, MFA y SSO) e identidades de máquinas (gestionadas mediante cuentas de servicio estáticas, claves de API o servicios de informática basada en la nube). Los agentes de IA autónomos eliminan esta dicotomía. Cuando un agente de IA actúa en nombre de un empleado humano, hereda privilegios de acceso, consulta bases de datos vectoriales y activa transacciones financieras, se convierte en un actor operativo y una entidad de seguridad activa que requiere controles explícitos de identidad, acceso y gobernanza.

La seguridad no se puede implementar posteriormente en la capa del modelo o la red. Debe estar anclada en un plano de control con agentes que gobierne continuamente la visibilidad, la conectividad y la ejecución.

Software tradicional (determinista)	IA con agentes (no determinista)
Ejecuta rutas de código explícitas, paso a paso.	Recibe intenciones humanas de alto nivel y formula planes de ejecución de varios pasos.
Se basa en permisos y cuentas de servicio estáticos y predefinidos.	Selecciona dinámicamente API y aplicaciones, aprovechando los almacenes de tokens, los intermediarios de credenciales y las sesiones delegadas.
Limitado estrictamente a rutas de integración insertadas en el código.	Actúa de forma autónoma, utiliza herramientas a través del Protocolo de Contexto de Modelo (MCP) y genera subagentes.
Funciona vinculado dentro de sesiones de usuario autenticadas.	Funciona de forma asíncrona, a menudo mucho después de que la persona que lo delegó haya cerrado sesión.

La brecha de gobernanza:

Cinco preguntas que se hacen todos los consejos directivos

Los agentes de IA se multiplican mucho más rápido que la capacidad de seguimiento de la mayoría de los equipos de seguridad. [Gartner predice](#) que, para 2028, una empresa global promedio de la lista Fortune 500 tendrá más de 150 000 agentes en uso, frente a menos de 15 en 2025. Esto generará importantes niveles de proliferación de agentes, complejidad de TI y desafíos de gestión. Solo el 13 % de las organizaciones cree tener implementada la gobernanza adecuada para los agentes de IA. Esa brecha es la razón por la que los comités de ciberseguridad y cumplimiento normativo regresan siempre a las mismas cinco preguntas.

1. ¿Dónde están nuestros agentes de IA y a quién pertenecen?
2. ¿Cómo podemos restringir lo que los agentes pueden hacer y mantenerlos enfocados en sus tareas?
3. ¿Podemos realizar una auditoría y explicar la toma de decisiones autónoma una vez realizadas las acciones?
4. ¿Cuál es nuestro nivel de exposición de datos y riesgo de terceros en toda la cadena de suministro de agentes?
5. ¿Tenemos un interruptor de desconexión y un plan de reversión verificados en caso de que un agente pierda el control?

Estas cinco preguntas se corresponden con los cuatro pilares de la arquitectura de referencia del plan en la Sección 3. La detección y el registro de identidad responden a la primera pregunta; la gestión de acceso responde a la segunda; la autorización y supervisión en tiempo de ejecución responden a la tercera y la cuarta; y la contención activa responde a la quinta. Responder a estas cinco preguntas con evidencia, y no solo con promesas, permite a un CISO acelerar de forma segura la adopción de agentes, al tiempo que adapta el nivel de control a la exposición normativa y operativa. Aquellos que no pueden responderlas permanecen vulnerables a incidentes de seguridad importantes y cada vez más graves.

Estas preguntas sirven como herramienta de diagnóstico para evaluar el riesgo, más que como una lista de verificación unificada. Naturalmente, los requisitos de seguridad varían según el ámbito operativo, ya que un asistente de productividad de solo lectura no necesita los mismos controles de contención que un agente que ejecuta transacciones financieras. Diagnosticar cuáles son los riesgos reales permite a los equipos de seguridad adaptar los controles a la exposición en el mundo real. Esto evita que las organizaciones complejicen excesivamente las implementaciones de bajo riesgo y sufran retrasos de aprobación de varios meses, al tiempo que garantiza que los agentes de alto riesgo se sometan a los límites operativos adecuados.

Principios del plan

El marco Blueprint Alliance extiende el principio de Zero Trust a los agentes para tratarlos como identidades de primera clase, no como código, credenciales o una incorporación posterior a la pila de identidad humana. Los siguientes seis principios sustentan todo lo demás.

Cada agente alojado es una identidad de seguridad distinta, no una ocurrencia tardía. Los agentes se aprovisionan, autentican y desaprovisionan con el mismo rigor que los empleados y las cuentas de servicio. Todos los agentes alojados en la empresa, incluidos los pilotos y las herramientas internas, deben operar dentro de un marco de gobernanza adecuado, con un nivel de control proporcional a su riesgo y alcance operativo. Los entornos de ejecución locales operan en un sistema especializado de contención de extremos, en lugar de un registro formal entre directorio e identidades.

El acceso se limita a cada tarea y no es permanente. Si bien las concesiones de acceso dinámicas con ámbito definido son lo ideal, las implementaciones prácticas combinan concesiones de tareas de corta duración con permisos básicos resilientes para equilibrar la sobrecarga.

La delegación es rastreable de principio a fin. Cuando una persona delega en un agente y este, a su vez, genera un subagente, la cadena de responsabilidad debe mantenerse en cada etapa. Cada acción debe remontarse a la persona o al proceso que la autorizó. Si bien la delegación de extremo a extremo en múltiples etapas sigue siendo un objetivo ambicioso con las herramientas actuales, este modelo define la trayectoria de referencia a medida que los estándares maduran.

Además de aprovisionarse, el entorno de ejecución del agente se aísla y se supervisa. Si bien las decisiones de acceso tomadas durante la configuración son necesarias, no alcanzan. El plan requiere una observación constante de todos los ciclos de ejecución para establecer referencias de comportamiento, más allá de los permisos concedidos.

La contención es instantánea y reversible. Todo agente, independientemente del proveedor o la plataforma, necesita un interruptor de desconexión específico que pueda suspenderlo, ponerlo en cuarentena o finalizarlo en tiempo real, con una estrategia clara para revertir la operación. La combinación de la supervisión constante del comportamiento con un límite de ejecución estructural cierra la brecha entre lo “autorizado” y lo “seguro”, pues contiene el radio de explosión en el momento en que un agente se desvía del estándar de referencia.

La gobernanza se adapta a la velocidad de la IA. Las políticas de seguridad estáticas no pueden seguir el ritmo de las capacidades de los agentes, que evolucionan rápidamente. La gobernanza debe abordar lo que las políticas por sí solas no pueden anticipar. Este plan funciona como una arquitectura adaptable que se ajusta a la evolución de los comportamientos de los agentes y combina límites de permisos estáticos con una definición de ámbito dinámica en tiempo real para permitir a las organizaciones mantener estrictas medidas de seguridad mientras adaptan el acceso al contexto.

En conjunto, estos seis principios extienden el concepto de Zero Trust a una nueva clase de identidades: agentes que operan de forma constante, persistente y autónoma en toda la empresa. Si se implementan correctamente, se cerrará la brecha en materia de identidad, seguridad y gobernanza.

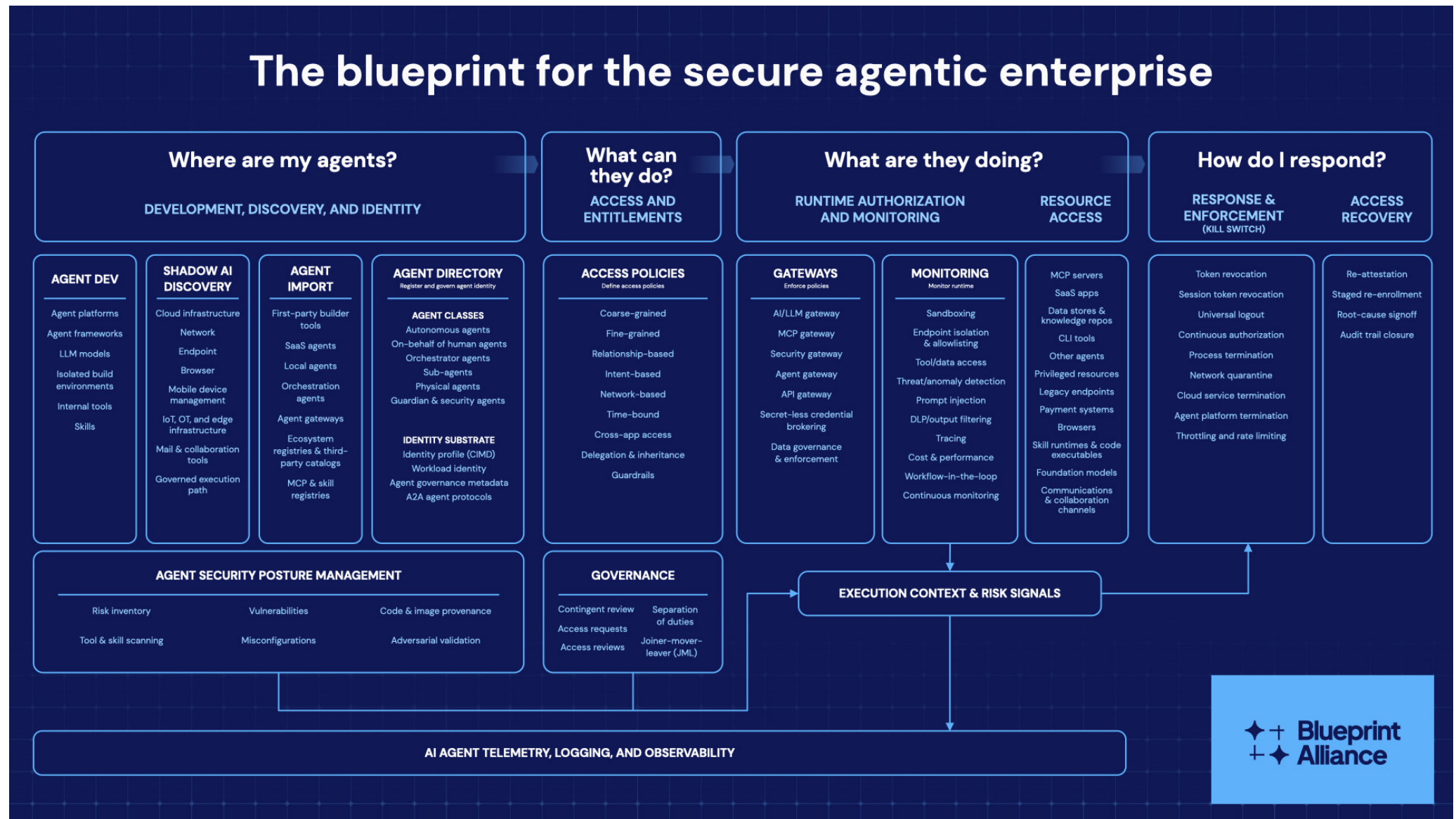
Arquitectura de referencia del plan

La arquitectura de referencia del plan basa la seguridad empresarial en cuatro preguntas operativas necesarias para gobernar la ejecución autónoma. Cada una se corresponde con un pilar de la arquitectura de referencia:

1. ¿Dónde están mis agentes?
2. ¿Qué pueden hacer?
3. ¿Qué están haciendo?
4. ¿Cómo respondo?

Todas estas cuestiones responden a una necesidad de una base sólida de contexto de ejecución y señales de riesgo, que se alimenta continuamente con telemetría en tiempo de ejecución, registros de eventos y capacidad de observación. En conjunto, estos elementos actúan como el tejido conectivo que permite al sistema detectar patrones de riesgo y responder de forma coordinada.

Arquitectura de referencia del plan



3.1 Pilar 1:

¿Dónde están mis agentes? (Desarrollo, detección e identidad)

El desafío fundamental para garantizar la seguridad de la empresa autónoma es la visibilidad. Las organizaciones no pueden gobernar ni proteger aquello cuya existencia desconocen. Este pilar establece una capacidad de seguimiento integral en todos los entornos corporativos para identificar, catalogar y evaluar a cada agente de IA que pase por la infraestructura corporativa, autorizado o no. Para terminar, este pilar registra a los agentes validados como identidades gobernadas con propietarios responsables, lo que genera el inventario fidedigno al que el pilar 2 vincula las políticas de acceso.

3.1.1 Detección de agentes de IA

El sistema de detección de agentes de IA funciona como un mecanismo de descubrimiento continuo de la huella que dejan los agentes empresariales. Este sistema asigna el ciclo de vida completo de los agentes, desde los pipelines de ingeniería internos hasta las integraciones externas con terceros, para garantizar que ningún modelo activo o proceso autónomo opere en un punto ciego.

3.1.1.1 Desarrollo de agentes

Este componente se centra en los entornos de ingeniería internos donde se crean, entrenan y optimizan de manera activa los agentes propios. Proteger el pipeline de desarrollo ayuda a garantizar la seguridad desde el diseño, incluso antes de que un agente interactúe con los datos de producción.

- **Plataformas de agentes:** entornos de orquestación gestionados por la empresa donde los agentes se alojan, se entrenan y se implementan de manera sistemática.
- **Marcos de agentes:** bibliotecas y arquitecturas de software subyacentes (como LangChain, LlamaIndex o AutoGen) que utilizan los desarrolladores para crear la lógica de los agentes.
- **Modelos LLM:** modelos de lenguaje grande fundacionales y perfeccionados que sirven como motores cognitivos que impulsan a los agentes.
- **Entornos de compilación y ejecución aislados:** límites de CI/CD seguros y aislados a nivel de cómputo que ejecutan el código del agente durante las pruebas y la evaluación, lo que garantiza que los equipos de seguridad detecten las ejecuciones de pruebas efímeras, en lugar de depender únicamente de las declaraciones estáticas del registro.
- **Herramientas internas adicionales:** repositorios de código personalizado, asistentes de código y registros de modelos internos integrados en el pipeline de desarrollo.
- **Habilidades:** módulos de procedimiento definidos, fragmentos de código o herramientas desarrolladas para ampliar las capacidades de un agente.

3.1.1.2 Importación de agentes

Las empresas modernas rara vez dependen exclusivamente de una única fuente de inteligencia; en cambio, incorporan diversas arquitecturas de agentes que varían en origen, marco de trabajo y límites de cómputo como medida de protección contra las capacidades de los agentes, que cambian rápidamente. La función de importación de agentes supervisa y gobierna la incorporación de estas distintas clases de agentes al ecosistema corporativo, y se centra principalmente en cómo fluyen los datos en toda la organización según el modelo de implementación subyacente.

El tipo de agente, la categoría y el canal son tres preguntas distintas que se pueden formular sobre cualquier agente importado, y un único agente normalmente responderá las tres a la vez. Un ejemplo sería un agente SaaS (creado por un proveedor) utilizado para flujos de trabajo internos de RR. HH. (B2E) e importado a través de una puerta de enlace de agente.

Tipo de agente: ¿quién lo creó?

- **Agentes propios:** capacidades de IA desarrolladas internamente, que van desde implementaciones totalmente personalizadas creadas con código fuente (como Python y LangChain) que se ejecutan en sistemas de propósito general (por ejemplo, Kubernetes o Serverless Functions) hasta agentes creados con marcos de agentes específicos del proveedor que gestionan de forma nativa el tiempo de ejecución, la memoria y la orquestación del flujo de trabajo.
- **Agentes SaaS:** productos de IA de terceros, diseñados específicamente para funcionar como entidades totalmente independientes dentro del ecosistema de software empresarial.
- **Agente local:** una arquitectura de cliente instalada directamente en el dispositivo local del empleado, donde cada paso de razonamiento se envía a un proveedor remoto de LLM en la nube. Se comporta como una capa no determinista entre la intención humana y la acción de la herramienta, con datos que cruzan el límite de la organización en cada llamada de inferencia.
- **Agentes de orquestación:** agentes creados en plataformas de orquestación visual que sirven para “conectar” diferentes API empresariales.

Categoría de agente: ¿para quién es?

- **Agentes B2E (de empresa a empleado):** agentes de IA internos orientados al empleado, implementados para automatizar los flujos de trabajo corporativos, las operaciones de TI/RR. HH. y la productividad diaria de los empleados dentro de los controles internos de identidad y acceso basado en roles de la empresa.
- **Agentes B2B:** agentes de IA orientados a la empresa que funcionan más allá de los límites corporativos para automatizar flujos de trabajo entre empresas, transacciones de la cadena de suministro e integraciones de API entre inquilinos.
- **Agentes B2C:** agentes de IA orientados al consumidor, implementados para interactuar directamente con usuarios finales externos para tareas de servicio, comercio o soporte, lo que exige estrictas medidas de seguridad contra la inyección de solicitudes de fuentes públicas, la exposición de la marca y las filtraciones de información personal del consumidor.
- **Agentes B2B2C:** agentes de IA de marca blanca o proporcionados por el proveedor, entregados a través de un cliente corporativo para atender a los consumidores finales de ese socio, que operan bajo un aislamiento de múltiples niveles y límites de confianza transitivos.
- **Agentes de asistentes personales:** asistentes de IA dirigidos por el usuario, orientados a consumidores, que operan en nombre de un individuo en contextos personales o empresariales, y que requieren controles de límites de acceso a datos personales y corporativos.

Canal de incorporación: ¿cómo se detecta?

- **Puertas de enlace de agentes:** puntos de incorporación y proxies a través de los cuales estos tipos de agentes externos o importados interactúan con las API internas de la empresa.
- **Registros del ecosistema y catálogos de terceros:** pipelines de conectores automatizados y adaptadores de API que supervisan, extraen y asignan continuamente definiciones de agentes desde registros externos nativos del proveedor, lo que garantiza que los agentes se cataloguen automáticamente sin intervención.
- **Registros de MCP y habilidades:** puntos de incorporación a través de los cuales los agentes obtienen esquemas de herramientas, habilidades y definiciones de contexto de registros de MCP y habilidades públicos o que aloja el proveedor. Los registros son en sí mismos un vector de inyección; una entrada dañina o disfrazada puede extraerse y consumirse automáticamente, por lo que cada herramienta o habilidad proveniente de un registro se trata como una entrada no confiable y se somete al análisis de la cadena de suministro detallado en la sección 3.1.2 antes de que un agente pueda invocarla.

3.1.1.3. Detección de IA en la sombra

Los empleados recurren de forma natural a herramientas de IA no autorizadas para optimizar sus flujos de trabajo, lo que plantea riesgos de cumplimiento normativo, filtración de datos y seguridad operativa. La detección de la IA en la sombra analiza activamente los perímetros corporativos para detectar implementaciones y usos no autorizados de agentes. Las capacidades de detección y supervisión deben implementarse de acuerdo con los requisitos aplicables en materia de privacidad, empleo, relaciones laborales, comités de empresa y protección de datos, con la transparencia, proporcionalidad, minimización de datos, retención y controles de acceso adecuados.

- **Infraestructura en la nube:** instancias de computación en la nube, entornos sin servidor y extremos de API no administrados detectados mediante inspección activa de red/API.
- **Red:** análisis de tráfico en tiempo real y monitoreo de salida para identificar llamadas a la API no reconocidas que se realizan a proveedores de LLM externos y extremos de IA.
- **Extremo:** supervisión a nivel de dispositivo para dar seguimiento a la ejecución no autorizada de binarios de IA, secuencias de comandos o pesos de modelos locales.
- **Navegador:** seguimiento de extensiones y supervisión de la sesión del navegador para detectar cuándo los usuarios inician sesión en plataformas web de IA o sistemas de chat no autorizados.

- **Gestión de dispositivos móviles (MDM) y gestión de extremos:** auditoría continua a nivel de dispositivo para detectar archivos ejecutables de IA locales no autorizados, pesos de modelos locales, herramientas de agentes para la CLI y extensiones para el IDE, al tiempo que se aplica una postura de seguridad básica y se bloquea la ejecución de agentes no gestionados en la capa del sistema operativo.
- **Infraestructura de IoT, OT y perímetro:** supervisión de seguridad industrial e inspección de protocolos de red para detectar modelos de lenguaje pequeños (SLM) y agentes autónomos en equipos de fabricación, PLC y hardware integrado donde no se pueda implementar software tradicional.
- **Flujo de correo y colaboración:** análisis del tráfico de mensajes, chats y calendarios para detectar agentes no autorizados que envían mensajes, responden automáticamente o gestionan calendarios en nombre de un usuario antes de que la actividad sea visible en los registros de salida de la red o el extremo.
- **Ruta de ejecución gobernada como superficie de detección:** límite de la arquitectura donde todos los agentes autorizados se ejecutan mediante un plano de control gestionado de forma centralizada. Cualquier proceso de agente que opere fuera de este límite se señala automáticamente como un indicador de IA en la sombra, sin necesidad de reglas de detección separadas específicas para el entorno.

3.1.2. Gestión de postura de seguridad de los agentes de IA (AIASPM)

Una vez que se detectan los agentes, hay que evaluarlos de manera constante. La AIASPM proporciona la capa de diagnóstico que evalúa los perfiles de riesgo, el mantenimiento de la configuración y las debilidades de software conocidas en todo el inventario de agentes.

- **Inventario de riesgos:** registro centralizado y dinámico que realiza un seguimiento de la postura de seguridad, las vulnerabilidades, las dependencias de software y las configuraciones incorrectas en todos los agentes detectados, independientemente de su estado respecto del ciclo de vida. Proporciona la perspectiva de riesgo de seguridad para complementar el Directorio de agentes centrado en IAM.
- **Vulnerabilidades:** gestión estándar de vulnerabilidades de software extendida a las dependencias de los agentes y a los extremos de los modelos, que abarca las CVE en el código base del agente, la cadena de suministro y los modelos fundacionales, asignadas a marcos estandarizados para una puntuación de riesgo uniforme.
- **Procedencia del código y las imágenes:** verificación criptográfica del origen, la integridad de la compilación y las firmas del código del agente, las imágenes de los contenedores, las dependencias y los pesos del modelo (por ejemplo, las certificaciones del marco SLSA) antes de la ejecución en producción. El uso de imágenes base mínimas y precertificadas, desprovistas de paquetes innecesarios y reconstruidas con base en los datos de las CVE actuales reduce la superficie de ataque de la cadena de suministro y simplifica la verificación de procedencia en el momento de la compilación.
- **Análisis de la cadena de suministro de herramientas y habilidades:** inspección estática y dinámica automatizada de conjuntos de instrucciones de agentes locales, habilidades de ejecución y esquemas remotos de herramientas de Model Context Protocol (MCP) —incluidos los datos extraídos de los registros de MCP y de habilidades— para detectar código malicioso insertado, operaciones ocultas y rutas de salida no autorizadas antes del registro.
- **Errores de configuración:** análisis continuo para detectar configuraciones de parámetros débiles, solicitudes inseguras del sistema, credenciales de larga duración en el código, vectores de suplantación de identidad y desviaciones de configuración que se introduzcan después de aprobar la revisión inicial de un agente.
- **Pruebas adversarias:** evaluaciones continuas y automatizadas para los agentes y sus medidas de seguridad contra ataques simulados y conocidos de inyección de solicitudes, jailbreak y rutas de exfiltración de datos, dentro de límites de ejecución aislados y desechables, para evaluar de forma segura la facilidad de explotación en el mundo real sin poner en riesgo la producción.

3.1.3 Directorio de agentes (registrar y gobernar la identidad del agente)

Mientras la importación de agentes (sección 3.1.1.2) captura la procedencia y los metadatos durante la detección, el Directorio de agentes es el repositorio de identidad autorizado donde las credenciales criptográficas, la propiedad y los derechos de acceso basados en políticas se registran y se mantienen formalmente para todos los agentes validados alojados en la empresa. El registro constituye el límite entre un agente detectado y uno gobernado, y es el requisito previo para cada decisión de acceso en el pilar 2.

Todo agente alojado y todo subagente con permisos distintos deben registrarse aquí con un perfil verificado, una identidad criptográfica, un propietario responsable y relaciones estructurales claras antes de poder solicitar acceso a los sistemas empresariales. El proceso de registro debe ser lo suficientemente sencillo como para que los equipos no se vean tentados a eludirlo utilizando credenciales humanas compartidas. Los agentes locales y los subagentes efímeros que heredan el ámbito de un agente principal se rigen por los controles de contención y delegación que se describen a continuación, en lugar de por entradas individuales del directorio. Una vez que se importa un agente, el directorio responde a dos preguntas más sobre el funcionamiento del agente en la empresa: qué función desempeña y cómo demuestra su identidad.

Clases de agentes: ¿qué papel desempeñan?

- **Agentes autónomos:** agentes totalmente independientes que operan de forma asíncrona para lograr objetivos comerciales a largo plazo sin supervisión humana inmediata.
- **Agentes humanos OBO (On-Behalf-Of):** agentes que ejecutan tareas directamente vinculadas a la sesión de un usuario humano específico, actuando como un intermediario y sujetos a los permisos que tenga dicho usuario.
- **Agentes de orquestación:** agentes de nivel superior que dividen un objetivo en subtarear y generan, coordinan y supervisan subagentes de etapa final; el directorio debe dar seguimiento a las cadenas de delegación y las relaciones de generación como atributos de primera clase.
- **Subagentes:** agentes secundarios especializados, generados o invocados de forma dinámica por un orquestador principal para ejecutar subtarear específicas, que se rigen por ámbitos de sesión heredados en lugar de por el registro individual en directorios. Dado que los límites de tareas dentro del modelo que se aplican en un LLM pueden fallar, los subagentes deben tener límites de ejecución aislados que apliquen estructuralmente el ámbito de los permisos heredados, a fin de evitar que los errores de razonamiento escalen los privilegios.

- **Agentes físicos/corporizados:** agentes que controlan dispositivos robóticos, actuadores u otros sistemas físicos. Esta arquitectura rige la identidad y el acceso, pero complementa, en lugar de reemplazar, los controles existentes de seguridad de planta, seguridad de OT y gestión de cambios.
- **Agentes de seguridad y protección:** agentes de IA administrativa especializados que operan dentro del plano de control de seguridad (por ejemplo, agentes de clasificación del SOC, motores de políticas dinámicas, escáneres AIASPM continuos). Dadas sus elevadas facultades para inspeccionar la telemetría y ejecutar acciones de contención, las identidades de los agentes guardianes requieren una certificación criptográfica obligatoria de la carga de trabajo, una estricta separación de funciones y la aprobación continua del propietario.

Sustrato de identidad: ¿cómo probar que realmente se trata de un agente y no de otro?

- **Perfil de metadatos de identidad del agente:** perfiles de metadatos enriquecidos (CIMD) que definen la intención, el origen, el linaje y el contexto operativo del agente para fundamentar las decisiones de acceso basadas en el riesgo.
- **Identidad de la carga de trabajo:** marcos de producción de identidad criptográficos, independientes de plataformas y basados en estándares, utilizados para emitir identificaciones seguras y verificables a los agentes en entornos de nube dinámicos.

- **Metadatos de gobernanza del agente:** metadatos estructurados vinculados al registro del directorio del agente, que incluyen su propietario humano responsable, el equipo operativo designado o el personal de guardia, el estado del ciclo de vida (por ejemplo, activo, suspendido o retirado), el nivel de riesgo y el ámbito operativo, que los motores de políticas utilizarán continuamente para la creación de controles de acceso y la evaluación durante la ejecución.
- **Protocolos de agentes A2A:** agentes de IA que aprovechan marcos de interoperabilidad estandarizados para detectar capacidades de forma dinámica, delegar tareas y colaborar entre pares superando los límites de la empresa. Al interactuar entre organizaciones sin un directorio compartido o un origen de confianza, los agentes utilizan estándares de acreditación criptográfica e intercambio de tokens (por ejemplo, WIMSE/AIMS, OAuth Token Exchange, HTTP Message Signatures) para preservar las cadenas de delegación y verificar la identidad principal entre intermediarios.

3.2 Pilar 2:

¿Qué pueden hacer? (Acceso y derechos)

En esencia, los agentes autónomos están diseñados para actuar, lo que requiere que se conecten a datos, API y sistemas. A partir de las identidades verificadas establecidas en el pilar 1, este pilar define los límites y permisos que rigen la forma en que los agentes interactúan con los recursos corporativos tanto en sesiones directas delegadas por humanos como en flujos de trabajo complejos de múltiples agentes. Esto añade la gobernanza de la IA entre máquinas a los modelos heredados centrados en el ser humano.

3.2.1 Políticas de acceso (definición de políticas de acceso)

Las Políticas de acceso especifican los límites operativos de lo que un agente con identidad verificada puede leer, escribir o ejecutar. Dada la naturaleza impredecible de los resultados de la IA generativa, estas políticas deben ser extremadamente contextuales y adaptativas.

- **Generales:** controles amplios, basados en roles, que definen a qué entornos o macroservicios generales puede conectarse un agente.
- **Detalladas:** permisos basados en atributos y a nivel de objeto que limitan a un agente a filas de la base de datos, archivos o métodos de API específicos, teniendo en cuenta explícitamente las etiquetas de clasificación de datos corporativos (por ejemplo, “confidencial” o “información personal”) para garantizar el cumplimiento de los límites de datos. Las estructuras de las políticas pueden abarcar la clasificación de datos, la identidad del agente, el ámbito del riesgo, el contexto ambiental y el tipo de acción.

- **Basadas en relaciones:** acceso derivado de las relaciones en el grafo de un agente (quién lo generó, en nombre de quién actúa y dónde está anidado).
- **Basadas en intenciones:** motores de políticas avanzados que analizan la intención semántica de la solicitud o el plan del agente antes de otorgar acceso a los recursos. Inspeccionar un plan es posible con las herramientas actuales; tratar una solicitud literal del usuario como único desencadenante es poco realista, ya que los agentes no siempre tienen en cuenta solo las solicitudes cuando actúan. Los privilegios efímeros y por tiempo limitado se crean estrictamente para la duración de una tarea específica y se revocan inmediatamente al finalizarla.
- **Basadas en la red:** la macrosegmentación restringe el tráfico de agentes a través de VPC empresariales y se combina con controles de microsegmentación de nivel 3 y 4 que restringen a rangos de IP o mallas de servicios autorizados la comunicación entre agentes y objetivos.

- **Por tiempo limitado:** las concesiones de tiempo limitado se restringen a una sola tarea o sesión y se revocan automáticamente al finalizar o vencer el plazo. Dado que las concesiones de aprobación automática y vencimiento automático aportan una capacidad limitada por sí solas, se recomienda combinar la definición de ámbito con límite de tiempo y la detección de anomalías de la sección 3.3.1.2 en lugar de considerar la fecha de vencimiento como única medida de seguridad.
- **Acceso entre aplicaciones:** marcos de políticas y controles de límites que rigen la capacidad de un agente para recorrer distintos ecosistemas de aplicaciones, lo que permite la traducción segura de datos, el intercambio de tokens y la ejecución de funciones a medida que orquesta flujos de trabajo en diferentes plataformas empresariales.
- **Delegación y herencia:** el límite efectivo de permisos para un subagente se calcula como la intersección de sus propios derechos y los derechos del agente o persona que delega, para evitar un escalamiento por confusión. Las políticas también deberían restringir el alcance y la profundidad de la delegación para evitar una expansión ilimitada.
- **Protecciones:** restricciones de seguridad y límites de comportamiento estrictos y deterministas que anulan los bucles de razonamiento autónomo para evitar globalmente acciones que no cumplan con las normas, resultados tóxicos o pasos operativos de alto riesgo, independientemente de la intención del agente o de sus privilegios detallados.
- **Revisión contingente:** activa las aprobaciones de flujo de trabajo automatizadas o con intervención humana cuando la acción de un agente supera los umbrales financieros u operativos definidos (por ejemplo, una compra o reservación que supere los USD 1000), más allá de que la acción se encuentre dentro de los permisos técnicos otorgados.
- **Revisiones de acceso:** análisis automatizado y continuo de permisos que ayuda a los propietarios del sistema a validar los permisos que no cumplen con los requisitos.
- **Solicitudes de acceso:** flujos de trabajo formales y auditados para aprovisionar nuevas capacidades, modelos o accesos al sistema a un agente actual. Si bien las solicitudes pueden originarse desde propietarios humanos o de forma dinámica desde agentes en ejecución, las solicitudes iniciadas por agentes deben evaluarse estrictamente en función de los límites de permisos estáticos y requieren una verificación obligatoria del propietario antes de ampliar el ámbito, a fin de evitar que los agentes manipulen los flujos de trabajo automatizados para otorgar acceso excesivo.
- **Separación de funciones (SoD):** reglas algorítmicas que impiden que un mismo agente tenga privilegios contrapuestos, como generar y aprobar transacciones financieras al mismo tiempo.
- **Altas, cambios y bajas (JML):** aplique la misma disciplina de activación, gestión de cambios y desaprovechamiento de los ciclos de vida de la identidad humana al ámbito, la propiedad o la evolución de la versión del modelo de un agente.

3.2.2. Gobernanza

La gobernanza brinda una referencia normativa y un ciclo de supervisión continua, y usa herramientas automatizadas en lugar de revisiones manuales superficiales para alinear —a la velocidad de los agentes— los privilegios de los agentes con el principio de privilegio mínimo.

3.3 Pilar 3:

¿Qué están haciendo? (Autorización y control durante la ejecución)

La visibilidad y la identidad no tienen sentido sin una supervisión y aplicación en línea durante la ejecución. Este pilar se centra en interceptar, inspeccionar y autorizar las acciones de los agentes en tiempo real mientras ejecutan tareas, lo que asegura que no participen en comportamientos maliciosos, no manipulen indebidamente datos confidenciales ni pongan en riesgo objetivos empresariales críticos.

3.3.1. Autorización y supervisión durante la ejecución

La autorización y supervisión durante la ejecución actúa como una capa de inspección activa para las sesiones de agentes en vivo. Se ubica directamente en la ruta de ejecución y analiza tanto las entradas que llegan al agente como las acciones o salidas que se generan en consecuencia.

3.3.1.1. Puertas de enlace (aplicación de políticas de acceso)

Las puertas de enlace actúan como puntos de aplicación de políticas (PEP) en una cadena de defensa profunda. Interceptan el tráfico entre la capa de agentes y el resto de la pila empresarial, analizan los protocolos y bloquean las llamadas no autorizadas antes de que lleguen a destino. Las API Gateways de etapa final son la autoridad definitiva de autorización, pues realizan su propia validación de cada llamada, independientemente de lo que las puertas de enlace de IA o agentes de etapa inicial ya hayan inspeccionado. Las puertas de enlace de IA/LLM, MCP, seguridad, agente y las API Gateways suelen implementarse en conjunto, pero cada una posee un punto de control distinto:

las puertas de enlace de IA/LLM gestionan el tráfico de la API del modelo; las puertas de enlace de MCP gestionan el tráfico del protocolo de intercambio entre herramientas y contexto; las puertas de enlace de seguridad brindan inspección de amenazas independiente del protocolo; las puertas de enlace de agente gestionan las cuestiones de identidad y delegación específicas del agente; y las API Gateways aplican los controles clásicos de la capa de API a las herramientas y servicios que convocan los agentes.

- **Puerta de enlace de IA/LLM:** intercepta las llamadas a la API de bajo nivel a los modelos fundacionales, y gestiona la limitación de velocidad, el recuento de tokens y las anulaciones de las indicaciones del sistema.
- **Puerta de enlace de MCP (Model Context Protocol):** un proxy especializado optimizado para gestionar protocolos de intercambio de contexto e integraciones de estándar abierto entre agentes y aplicaciones, con especial atención a los servidores de MCP.
- **Puerta de enlace de seguridad:** una capa de mitigación de amenazas integrada que proporciona seguridad de contenido dedicada y verificación de protocolos para las interacciones de los agentes. Se sitúa directamente en el flujo de tráfico para realizar una inspección exhaustiva de la carga útil en tiempo real con terminación TLS, e intercepta y filtra activamente inyecciones complejas de solicitudes, jailbreaks, cadenas de texto de adversarios ocultas y cargas útiles de código malicioso antes de que lleguen al agente o a los recursos de etapa final. También se encarga del análisis de clasificación de datos para garantizar la redacción de los datos salientes, las comprobaciones de integridad estructural de los resultados del modelo y la inspección criptográfica de las cargas útiles encriptadas del agente.

- **Puerta de enlace del agente:** sirve como un plano de control durante la ejecución específico del agente; vincula la identidad del agente a cada solicitud, mantiene el registro de herramientas/MCP del que se nutren los agentes y gobierna las acciones del agente (no solo el acceso) en función de la identidad, el contexto y la acción específica que se solicita. Detiene las interacciones inseguras en tiempo real, bloquea o desinfecta el tráfico riesgoso de los modelos y rechaza las acciones de herramientas confidenciales o exige su autorización mediante políticas integradas y personalizadas. Además, inyecta credenciales de etapa final sin secretos y de corta duración en las llamadas a la API en el momento de la ejecución, de modo que los agentes nunca gestionen, almacenen ni expongan claves de API empresariales de acceso permanente ni secretos de cuentas de servicio. Puede proporcionar telemetría gobernada en interacciones multiagente compatibles y detectar o redactar datos confidenciales, incluso información personal, de acuerdo con las políticas y capacidades configuradas, al tiempo que defiende contra comportamientos no seguros y ataques maliciosos basados en solicitudes.
- **API Gateway:** aplica controles clásicos de la capa de API, limitación de velocidad, cuotas, validación de esquemas, mTLS, WAF, modelado de tráfico y revocación en la puerta de enlace para las herramientas y servicios que utilizan los agentes y los servidores de MCP, y realiza su propia comprobación de autorización final en cada llamada, en lugar de dar por sentado que las puertas de enlace de IA/agente de etapa inicial ya detectaron todas las anomalías.

3.3.1.2 Supervisión (monitoreo durante la ejecución)

La supervisión ofrece niveles exhaustivos de capacidad de observación, detección de amenazas y cumplimiento activo. Analiza la telemetría durante la ejecución en busca de comportamientos anómalos, filtraciones de datos e intentos de explotación externa, y aplica distintos controles, como el uso de entornos de prueba y el aislamiento de extremos en tiempo real.

- **Uso de entornos de prueba:** consiste en aislar la ejecución de los agentes (especialmente agentes no confiables, externos o recién actualizados) dentro de entornos de ejecución seguros y restringidos para evitar movimientos laterales y contener los riesgos de ejecución mediante límites estructurales.
- **Aislamiento de extremos y listas de permitidos:** controles locales durante la ejecución para agentes alojados en extremos, que incluyen entornos de prueba a nivel del sistema operativo, creación de listas de permitidos para aplicaciones de extremos y comprobaciones locales de la postura del dispositivo a fin de restringir la ejecución de comandos del agente local y el acceso a archivos.
- **Acceso a herramientas y datos:** inspección estructural y auditoría en tiempo real de argumentos, cargas útiles, esquemas y consultas que se transmiten a herramientas y repositorios de etapa final. Realizar inspecciones dentro del límite de ejecución o de forma directamente adyacente ayuda a garantizar que los planes de ejecución dinámicos no se desvíen de los parámetros autorizados ni introduzcan mutaciones no deseadas a mitad de la sesión.
- **Detección de amenazas o anomalías:** los motores de aprendizaje automático rastrean los patrones de comportamiento habituales de los agentes para detectar picos inusuales en el acceso a los datos, bucles rápidos de API o patrones de ejecución extraños.

- **Inyección de solicitudes:** escáneres integrados diseñados para captar cadenas de texto maliciosas, tanto directas como indirectas, destinadas a secuestrar la lógica central del agente o anular las instrucciones del sistema.
- **Filtrado de salida/DLP:** las herramientas de prevención de pérdida de datos (DLP) analizan las salidas del agente para mitigar el riesgo de filtración de código fuente propio, información personal o secretos comerciales a proveedores externos o usuarios no autorizados.
- **Seguimiento:** análisis integral de transacciones que traza la ruta de ejecución exacta, las pilas de llamadas y las invocaciones de subagentes desencadenadas por una solicitud principal del usuario.
- **Costo y rendimiento:** seguimiento del uso de tokens, costos operativos, latencia y consumo de recursos para prevenir bucles descontrolados de agentes o ataques de agotamiento de recursos.
- **Flujo de trabajo en bucle:** activadores programáticos que pausan la ejecución del agente para acciones de alto riesgo y las redirigen a comprobaciones cruzadas y automatizadas de políticas o a la aprobación humana explícita antes de continuar.
- **Supervisión continua:** supervisión constante y capacidad de exportación de la actividad del agente, que incluye el rastreo detallado, los registros de auditoría y la imposibilidad de rechazo de las acciones del agente.

3.3.1.3. Cumplimiento continuo de intención y alineación

El cumplimiento continuo de la intención y la alineación determina si un agente se mantiene dentro de los límites de su tarea delegada durante toda la ejecución o si se desvía hacia una expansión de propósito no autorizada. Complementa las políticas de acceso inicial: la política basada en intenciones determina a qué puede acceder inicialmente un agente, mientras que el cumplimiento continuo de la alineación evalúa si los comportamientos evolutivos de varios pasos siguen siendo apropiados con el paso del tiempo.

A continuación, se enumeran algunas de las capacidades principales.

- **Vinculación de propósito:** constatación de un vínculo estricto entre el usuario que delega, la identidad del agente, el flujo de trabajo aprobado y la tarea solicitada.
- **Evaluación de planes y acciones:** comparación, durante la ejecución, entre las llamadas planificadas a herramientas y las acciones observadas, a fin de detectar desviaciones de tareas, desviaciones impulsadas por inyección de solicitudes y comportamientos de agentes confundidos.
- **Cumplimiento según el contexto:** ejecución de acciones en tiempo real, que incluyen la aprobación, el rechazo, la censura, la aprobación escalonada, la reducción del alcance, la suspensión o la contención.
- **Evidencia rastreable de decisiones:** mantenimiento de registros de auditoría irrefutables que expliquen por qué se permitió, se modificó o se bloqueó cada acción material.

3.3.2. Acceso a recursos

El acceso a los recursos delimita el radio de explosión real de un agente. Divide en categorías los objetivos de etapa final con los que un agente puede interactuar, lo que ayuda a garantizar que cada interfaz cuente con las protecciones de seguridad adecuadas.

3.3.2.1. Recursos de destino (componentes a los que se accede)

Los recursos de destino representan los destinos de las acciones de un agente. Ya sea para obtener contexto o enviar cambios, estos extremos deben catalogarse y defenderse de manera continua.

- **Servidores de MCP:** repositorios de datos y servidores secundarios configurados explícitamente para compartir información contextual (datos estructurados, documentos, inserciones, herramientas y estados) con los agentes mediante el Model Context Protocol.
- **Aplicaciones SaaS:** aplicaciones comerciales principales y de productividad empresarial que contienen grandes cantidades de datos corporativos confidenciales.
- **Almacenes de datos y repositorios de conocimiento:** repositorios estructurados y no estructurados, incluidas bases de datos vectoriales, lagos de datos, almacenes y sistemas de almacenamiento transaccional, que los agentes usan para extraer el contexto empresarial o confirmar los resultados operativos. Estos repositorios se rigen por prácticas de gestión de postura de seguridad de los datos (DSPM), que dan seguimiento al linaje y la clasificación a medida que los agentes acceden a cada almacén.

- **Herramientas de CLI:** interfaces de línea de comandos que permiten a los agentes ejecutar secuencias de comandos, configuraciones o comandos operativos a nivel del sistema.
- **Otros agentes:** subagentes de etapa final o agentes especializados encargados de ejecutar componentes parciales de una carga de trabajo más amplia.
- **Recursos privilegiados:** entornos de destino altamente sensibles, incluidos directorios raíz, almacenes de identidades y consolas de configuración maestra.
- **Extremos estándar y heredados:** API REST/GraphQL empresariales, bases de datos y aplicaciones heredadas que no admiten de forma nativa protocolos de contexto de agente, como MCP.
- **Sistemas de pago:** las redes financieras, las puertas de enlace de transacciones y los libros de contabilidad requieren el máximo grado de validación criptográfica y supervisión humana.
- **Navegadores:** sesiones web automatizadas que los agentes usan para recopilar información pública o interactuar con consolas web externas.
- **Entornos de ejecución de habilidades y archivos ejecutables de código:** intérpretes de código, entornos de prueba de ejecución de secuencias de comandos y módulos procedimentales que los agentes invocan de forma dinámica para ejecutar código generado y que requieren límites de cómputo aislados para reducir los riesgos de fuga durante la ejecución.
- **Modelos fundacionales:** proveedores de inteligencia sin procesar donde se alojan los pesos del modelo y que exigen claves de API del modelo seguras y validadas.
- **Canales de comunicación y colaboración:** correo electrónico, chat y sistemas de tickets mediante los que interactúan los agentes; representan vías de entrada de solicitudes no confiables y deben someterse a niveles exhaustivos de inspección y análisis de la carga útil.

3.3.2.2 Seguridad de datos según el propósito

La seguridad de datos según el propósito evalúa si un agente determinado debe acceder a datos específicos, procesarlos, transformarlos, resumirlos, divulgarlos o transmitirlos en función del propósito delegado, la autoridad del usuario, la identidad del agente, el estado del flujo de trabajo, la clasificación de datos, el destinatario, el destino y la acción de etapa final. Aplica controles de datos en el momento en que se usan, no solo cuando se accede a ellos. Esto es importante porque los agentes combinan datos de diferentes sistemas, infieren información confidencial a partir de datos no confidenciales y divulgan los resultados a través de resultados, API, mensajes, archivos y llamadas a subagentes. La clasificación y la DLP determinan qué son los datos y si son confidenciales; la seguridad de datos según el propósito determina si este uso de los datos es apropiado en el contexto. Entre las capacidades principales se incluyen la vinculación de propósitos entre la solicitud del usuario, la tarea del agente, la fuente de datos y el uso permitido; la evaluación durante la ejecución para determinar si el acceso y la divulgación se ajustan al paso actual del flujo de trabajo; los controles para la síntesis, la transformación, la exportación, el movimiento entre aplicaciones y la divulgación a agentes de terceros; la detección de alcance excesivo de los datos, recuperación excesiva, inferencia no autorizada e intercambio inapropiado; y la aplicación de medidas que abarcan la censura, el enmascaramiento, la respuesta parcial, el flujo de trabajo de aprobación, el registro, el bloqueo o la contención.

3.4 Pilar 4:

¿Cómo respondo? (Contención activa)

El último pilar del plan transforma la telemetría en contención inmediata. Cuando los sistemas de supervisión señalan comportamientos anómalos, credenciales vulneradas o un ataque de inyección de solicitudes, la empresa debe contar con mecanismos de mitigación altamente automatizados y precisos que prioricen las respuestas proporcionales y menos perjudiciales (limitación de velocidad, cuarentena) sobre las acciones destructivas que podrían autoprovocar una interrupción del servicio.

3.4.1. Respuesta y cumplimiento (limitación y contención)

El bucle de control operativo automatizado está diseñado para responder a las señales de riesgo en tiempo real. Actúa como una capa de control coordinada, no como un único punto de cierre centralizado, y ejecuta respuestas de seguridad específicas y proporcionales en toda la pila técnica.

El cumplimiento de las medidas de seguridad resulta más eficaz cuando se combina con un protocolo de notificación estandarizado: el sistema que detecta el riesgo publica un evento de desactivación, los sistemas suscritos actúan y el resultado se publica como una señal de confirmación.

3.4.1.1 Medidas de cumplimiento (mecanismos de mitigación)

Las medidas de cumplimiento representan los instrumentos concretos y multicapa que los equipos de seguridad y los orquestadores automatizados pueden activar para aislar, contener o neutralizar por completo a un agente problemático o en riesgo. Estas medidas forman una biblioteca de opciones de respuesta que es propiedad del SOC y no son interruptores aislados que se pueden usar una sola vez: los responsables de la respuesta seleccionan la acción menos perjudicial que sea proporcional a la señal de riesgo.

Las acciones de contención que activan los agentes guardianes (suspender, aislar, finalizar) requieren que el agente de destino opere dentro de un límite de ejecución que admita un control granular a nivel de sesión, ya sea que se ejecute en un servicio en la nube administrado, una aplicación SaaS o un entorno de ejecución en contenedores. Si no existe una plataforma subyacente o soporte durante la ejecución para pausar a un agente sin impactar en los flujos de trabajo adyacentes, la contención basada en API genera una falsa sensación de seguridad. Al contener agentes vinculados a equipos físicos u OT, los disparadores automatizados deben calibrarse para garantizar que un aislamiento abrupto no genere riesgos para la seguridad física ni peligros operativos en la planta.

- **Revocación de tokens:** se invalidan instantáneamente los tokens de acceso a la API o tokens OAuth activos, lo que bloquea de inmediato la capacidad del agente para llamar a las aplicaciones conectadas.
- **Revocación de tokens de sesión:** se invalidan los tokens OAuth activos y se revocan los contextos de sesión posteriores en los servicios empresariales conectados, o se interrumpe la conexión durante la ejecución activa de una única cadena de interacción específica, sin necesidad de eliminar la identidad general del agente.
- **Autorización continua:** reevaluación dinámica y en tiempo real de los contextos de seguridad que reduce instantáneamente los permisos en el momento exacto en que cambia la postura de riesgo.

- **Finalización del proceso:** cierre forzoso de los entornos de cómputo durante el tiempo de ejecución como último recurso cuando falla la cuarentena; tenga en cuenta que la finalización sacrifica la telemetría forense en memoria, por lo que se recomienda, siempre que sea posible, la cuarentena de la red, la suspensión de la sesión o la suspensión del entorno de ejecución.
- **Cuarentena de red:** se usan reglas de microsegmentación para aislar el entorno donde se aloja el agente, a fin de interrumpir su comunicación tanto con la red interna como con la internet abierta.
- **Finalización del servicio en la nube:** apagado instantáneo de las instancias en la nube subyacentes, las funciones sin servidor o los clústeres de máquinas virtuales que alojan al agente que actúa por cuenta propia.
- **Finalización de la plataforma del agente:** un comando administrativo de alto nivel que instruye a la plataforma centralizada de orquestación de agentes para que suspenda la cuenta operativa del agente y detenga todas las tareas programadas.
- **Limitación de velocidad y restricción del tráfico:** restricción progresiva de la tasa de solicitudes, el presupuesto de tokens o la concurrencia de un agente como primera respuesta proporcional, lo que degrada la capacidad sin interrumpir completamente la sesión.

3.4.2. Recuperación del acceso

Un interruptor de desconexión constituye solamente la mitad del bucle de control: cada acción de contención necesita una ruta correspondiente que permita regresar a un último estado válido verificado. La recuperación de acceso gobierna la forma en que se restablece el servicio de un agente contenido, en cuarentena o dado de baja, y confirma que la reincorporación sea deliberada, esté documentada y pueda rastrearse, en lugar de que se revierta automáticamente la medida de cumplimiento. Los flujos de trabajo de recuperación totalmente automatizados y estandarizados siguen siendo un objetivo por lograr; la mayoría de las organizaciones aún requieren la aprobación manual en uno o más de los pasos que se detallan a continuación.

- **Recertificación:** volver a verificar la identidad, el código y la configuración del agente comparándolos con una referencia válida antes de restablecer el acceso en una instancia de ejecución nueva y aislada.
- **Reinscripción gradual:** restablecer el acceso de forma paulatina (por ejemplo, solo lectura antes de escritura, entorno aislado antes de producción) en lugar de restablecer todos los privilegios de una sola vez.
- **Aprobación de la causa raíz:** se requiere que el propietario humano o el equipo de seguridad documenten la causa raíz y el saneamiento antes de reactivar al agente.
- **Cierre del registro de auditoría:** vincular el evento original del interruptor de desconexión, las acciones correctivas y la decisión de reinscripción en un único registro de incidentes cerrado y que pueda auditarse.

3.5

Componentes fundamentales transversales

Estas capas de la arquitectura no existen de forma aislada; de hecho, son la base y el tejido que vincula los cuatro pilares, y la estructura que conecta la visibilidad, la identidad, la supervisión y la respuesta en una estructura unificada de seguridad.

3.5.1. Contexto de ejecución y señales de riesgo

El contexto de ejecución y las señales de riesgo hacen las veces del tejido analítico conectivo del plan. Este motor de orquestación —alimentado continuamente por telemetría en tiempo real, evaluaciones de la postura de seguridad y señales de identidad de usuario delegadas (como indicadores de riesgo de cuenta o riesgo elevado del usuario principal)— evalúa dinámicamente las amenazas al correlacionar condiciones individuales de bajo riesgo en combinaciones tóxicas. También publica un contexto de ejecución positivo, es decir, una confirmación estructural de lo que ya se sabe que es válido, y no solo alertas sobre lo que salió mal. Cuando se supera un umbral (por ejemplo, una inyección de solicitudes repentina combinada con un intento de extraer datos financieros), las señales de riesgo activan los mecanismos de respuesta o cumplimiento correspondientes de forma directa y programática. Para ayudar a garantizar la interoperabilidad en tiempo real y las respuestas coordinadas en entornos de proveedores diversos, este sistema puede aprovechar estándares abiertos de ingeniería de seguridad como **Shared Signals Framework (SSF)** y **Continuous Access Evaluation Profile (CAEP)** para publicar y consumir de forma asíncrona modificaciones de estado de Zero Trust.

3.5.2. Telemetría, registro y capacidad de observación de agentes de IA

Esta es la capa fundamental que sustenta todos los pilares de la arquitectura de referencia. Funciona como un plano de datos inmutable, que captura metadatos de ejecución en toda la pila empresarial, respaldado por una censura obligatoria de la carga útil en línea para evitar que la información de identificación personal, los secretos y las credenciales ingresen al repositorio de registros. Sin este repositorio de registros estructurado y de alta fidelidad, la detección automatizada de amenazas es imposible, las evaluaciones de la postura de seguridad carecen de datos y los responsables de la respuesta a incidentes forenses quedan completamente a ciegas. Para ayudar a garantizar la uniformidad y reducir los silos de proveedores en interacciones complejas entre múltiples agentes, esta capa puede aprovechar estándares abiertos como **Open Cybersecurity Schema Framework (OCSF)** para normalizar diversos flujos de registros en una sola taxonomía de seguridad unificada.

- **Origen de la telemetría a prueba de manipulaciones:** capturar la telemetría en el límite de ejecución —en lugar de depender de los registros de agentes autonotificados— ayuda a garantizar la imposibilidad de rechazo e impide que un agente en riesgo suprima, modifique o falsifique su registro de ejecución.

4

Puesta en marcha con integradores de sistemas globales (GSI)

Un plan de arquitectura requiere un modelo de implementación operativa para lograr un valor empresarial duradero. En las siguientes secciones, se define cómo los equipos de seguridad y los socios consultores ponen en práctica estos cuatro pilares dentro de los flujos de trabajo de empresas existentes. Las organizaciones pueden poner en práctica la arquitectura a través de equipos empresariales internos, integradores de sistemas globales (GSI), proveedores de servicios gestionados u otros socios calificados, en función de las capacidades y los requisitos de la organización. Una arquitectura de referencia publicada sin un modelo operativo se convierte en un producto obsoleto en el plazo de un año. En las siguientes secciones, se define cómo implementar dicha arquitectura, cómo dotarla de personal y cómo mantenerla dentro de una empresa.

4.1. Asesoramiento para la implementación y habilitación de la arquitectura

Los GSI traducen los cuatro pilares en una implementación secuenciada y específica para cada organización, en lugar de un despliegue simultáneo y de una sola vez. Estos marcos incluyen lo siguiente:

- **Evaluación del estado actual:** se compara la infraestructura vigente de IAM, red y supervisión de una empresa con la arquitectura de referencia del plan para identificar las brechas reales frente a las capacidades que ya existen con otro nombre.
- **Diseño del modelo operativo objetivo:** se define claramente la responsabilidad operativa en todos los pilares para evitar brechas en la rendición de cuentas entre los equipos. Este modelo suele asignar las puertas de enlace y la supervisión al Departamento de Ingeniería de Seguridad; el Directorio de agentes, las Políticas de acceso y la gobernanza a IAM; la respuesta y la aplicación de la normativa al SOC; y los límites de ejecución (entornos de prueba, ejecución en aislamiento, gestión del ciclo de vida) al Departamento de Ingeniería Informática y de Plataformas.
- **Planificación de la implementación por fases:** se organiza la implementación según las inversiones en herramientas actuales del cliente y el calendario normativo.
- **Integración de la interoperabilidad de proveedores:** se configuran las puertas de enlace específicas de IA/LLM, MCP, seguridad y agentes (así como las API Gateways) para las que una empresa ya tenga licencia a fin de que interactúen en toda la capa compartida de telemetría y señales de riesgo, en lugar de exigir que se reemplace por completo la infraestructura actual.

4.2 Matrices de puntuación de riesgo estandarizadas

No todos los agentes requieren el mismo nivel de control, y los GSI proporcionan marcos de puntuación calibrados que impiden que las organizaciones sobredimensionen las implementaciones de bajo riesgo o que no protejan adecuadamente las de alto riesgo. Estos marcos incluyen lo siguiente:

- **Modelos de clasificación de riesgos** que evalúan a los agentes según su alcance operativo (solo lectura frente a transaccional), clasificaciones de datos, nivel de autonomía (OBO frente a totalmente autónomo) y radio de impacto de los recursos objetivo.
- **Umbral calibrado por sector:** un agente de servicios financieros que ejecuta flujos de trabajo de pago y un agente de fabricación que lee datos de sensores de tecnología operativa requieren bases de riesgo muy diferentes. Los GSI aportan la biblioteca de patrones intersectoriales para calibrar estos umbrales, en lugar de empezar de cero.
- **Evaluación del impacto empresarial para la gobernanza semántica:** definición de los umbrales monetarios, normativos o de reputación que activan la aprobación obligatoria con intervención humana, adaptados al nivel de riesgo real del cliente y no a un valor predeterminado genérico.
- **Profundidad de aislamiento ajustada al nivel de riesgo:** la clasificación por niveles de riesgo debe escalar la resistencia del límite de ejecución estructural (por ejemplo, los entornos de prueba a nivel de kernel, los ciclos de vida de ejecución efímeros) junto con los mecanismos de revisión, para garantizar que un agente de alto riesgo no tenga el mismo radio físico de explosión que una herramienta de bajo riesgo si se ve vulnerado.

4.3 Gestión continua del ciclo de vida

La gobernanza de agentes no es una certificación única. Los GSI gestionan los procesos continuos del ciclo de vida que mantienen actualizados el Directorio de agentes y las Políticas de acceso a medida que cambia el entorno:

- **Ciclos automatizados de certificación de acceso:** el servicio gestionado extrae el estado del Directorio de agentes y de las Políticas de acceso, y señala a los agentes cuyos propietarios humanos han abandonado la organización o cuyos permisos se alejaron del principio de privilegio mínimo.
- **Equidad de altas, cambios y bajas para los agentes:** se aplica la misma disciplina de desaprovechamiento utilizada para la desvinculación de seres humanos a los cambios de propiedad de los agentes, las actualizaciones de la versión del modelo y el retiro de los subagentes, de modo que las credenciales huérfanas no se acumulen como la principal fuente de la brecha de identidad.
- **Enrutamiento de excepciones para separación de funciones:** se gestionan los flujos de trabajo que enrutan las excepciones de privilegios en conflicto (por ejemplo, un agente que puede generar y aprobar una transacción financiera) al propietario del sistema correcto para su resolución, con la misma periodicidad que la gobernanza de la identidad humana.
- **Gestión de la desviación del entorno de ejecución:** aplicación programada de parches, actualización de dependencias y recertificación de los entornos de ejecución de agentes y las configuraciones de la plataforma base para ayudar a eliminar la desviación de vulnerabilidades en implementaciones alojadas, en la nube y en contenedores.

4.4 Adhesión normativa y de cumplimiento

Los GSI conectan la arquitectura con los marcos regulatorios específicos bajo los que opera una empresa y traducen los controles técnicos en evidencia diseñada para cumplir con los requisitos regulatorios, de cumplimiento y de auditoría:

- **Asignación de controles a marcos normativos:** alinear los resultados de la telemetría, los registros de auditoría y la revisión de accesos con los requisitos específicos del sector (por ejemplo, gestión de riesgos de modelos de servicios financieros, manejo de datos de atención médica, protección de infraestructuras críticas) a medida que evolucionan, junto con indicaciones específicas para la IA con agentes.
- **Paquetes de evidencia listos para auditorías:** estructurar los datos de la capa de telemetría y registro de manera que las revisiones de acceso, los eventos de contención y limitación, y los registros de recuperación de acceso (recertificación, aprobación de la causa raíz, cierre del registro de auditoría) cumplan con los estándares de documentación de los examinadores y auditores, y no solo con las necesidades internas del SOC.
- **Orientación sobre transferencias transfronterizas y residencia de datos:** asesoramiento sobre cómo los canales de importación de agentes (agentes SaaS, agentes propios que cruzan los límites de la organización) se relacionan con los requisitos de residencia de datos y transferencia transfronteriza específicos de cada jurisdicción en la que opera la empresa.
- **Procedencia de auditoría a prueba de manipulaciones:** estructurar los datos de telemetría recopilados de manera que los reguladores y examinadores reciban registros de auditoría irrefutables generados directamente en el límite de ejecución, lo que reduce la necesidad de reconstruir registros secundarios a partir de la infraestructura de red circundante.

4.5 Manuales de estrategias específicos para cada sector

Los requisitos de control varían según el sector, lo que obliga a los GSI a desarrollar manuales de estrategias adaptados a los distintos entornos operativos:

- **Servicios financieros:** prioriza el aislamiento estricto de las ejecuciones, los umbrales de alta seguridad para las transacciones y los controles de datos normativos.
- **Fabricación e industria:** hace hincapié en los sistemas de OT, IoT y la detección de agentes en el perímetro dentro de hardware heredado.
- **Venta minorista y comercio:** se centra en la defensa contra la inyección de solicitudes B2C, la protección de la marca y la filtración de información personal de los clientes.

Los miembros fundadores actualizan continuamente estos manuales de estrategias a medida que se publican los resultados conjuntos de interoperabilidad, verificando que el modelo operativo se vaya adaptando de acuerdo con las nuevas capacidades de los agentes.

Conclusión: es hora de moldear el futuro de la seguridad de la IA

Blueprint Alliance rige el comportamiento de los agentes en ejecución durante la etapa de producción. Esta arquitectura, integrada directamente en la pila empresarial en lugar de ser una capa externa que se agrega después, proporciona las protecciones estructurales necesarias para implementar la automatización con agentes de forma segura y a gran escala.

Para lograr una verdadera interoperabilidad entre múltiples proveedores, los miembros fundadores validan activamente sus plataformas con respecto a estándares abiertos esenciales, entre ellos, Model Context Protocol (MCP) para la interacción entre herramientas, Open Cybersecurity Schema Framework (OCSF) para el registro unificado, Shared Signals and Events (SSF/CAEP) para el intercambio de riesgos en tiempo real, y HTTP Message Signatures (RFC 9421) para la autenticación de solicitudes. Esta referencia compartida de telemetría ayuda a garantizar que una señal de amenaza o una infracción de los límites detectada en un entorno de ejecución pueda procesarse de forma nativa y aplicarse en todo el plano de control. Las capacidades variarán según el proveedor y la implementación; las integraciones de referencia y los resultados de interoperabilidad validados se publicarán una vez finalizadas las pruebas.

Mediante la publicación periódica de resultados conjuntos de interoperabilidad e integraciones de referencia, Blueprint Alliance ofrece un ecosistema para múltiples proveedores y en constante evolución que avanza al ritmo de la innovación de los agentes.

El plan en acción

A continuación, se describe la aplicación práctica y conjunta de los pilares de la sección 3 en cinco situaciones diferentes. Cada ejemplo es una suma creada a partir de patrones observados en las primeras colaboraciones de Blueprint Alliance, en lugar de basarse en un solo cliente. Estos ejemplos ilustran el comportamiento previsto de la arquitectura, pero no constituyen evidencia de los resultados que obtienen los integrantes de Blueprint Alliance en la actualidad. Lo aconsejable es interpretarlos como ejemplos de la arquitectura en movimiento, y no como estudios de caso de una única implementación.

De agente en la sombra a identidad registrada. Un analista financiero conecta un asistente de IA basado en el navegador a su sesión SSO corporativa para resumir los contratos con proveedores. La detección de IA en la sombra descubre el tráfico de salida desconocido desde su primera conexión. La gestión de postura de seguridad del agente de IA comprueba si la herramienta presenta vulnerabilidades conocidas y la dirige al Directorio de agentes. Antes de que se acceda a datos confidenciales, el asistente queda registrado como una identidad controlada con un perfil de metadatos verificado y un ámbito de acceso definido, o bien como un asistente bloqueado en la capa de red. Lo que antes permanecía invisible durante meses ahora es visible y se revisa, y se puede decidir qué hacer antes de que entre en contacto con datos confidenciales.

Preservar la delegación en un flujo de trabajo multiagente. Un responsable de operaciones de ventas solicita a un agente de orquestación que actualice un conjunto de registros de CRM y notifique a una lista de distribución. El agente de orquestación crea un subagente para gestionar la escritura y otro para redactar la notificación. Cada salto devuelve una notificación de actor anidada a la sesión humana original, de modo que el motor de Directorio de agentes y Políticas de acceso pueda aplicar restricciones OBO en cada paso. Si un subagente intenta acceder a un sistema al que el usuario humano original no pudo, lo bloquean los límites principales de las políticas y la IAM, incluido Cross-App Access. Además, dado que cada subagente se ejecuta en su propio contexto aislado, un ataque bloqueado no puede afectar a la sesión del agente de orquestación ni a las tareas relacionadas, mientras que todo el flujo de trabajo permanece unificado en un único registro de ejecución.

Ajustar el acceso a lo que el agente realmente intenta hacer. Un agente de soporte recibe acceso de lectura general a una plataforma de datos de clientes. Cuando intenta emitir un reembolso, el motor de políticas basado en intenciones detecta el cambio de búsqueda a transacción y requiere la aprobación humana antes de otorgar un privilegio efímero y detallado, limitado a ese único pedido. La ejecución del reembolso dentro de un entorno de ejecución efímero, limitado a esa única tarea aprobada, ayuda a garantizar que los privilegios elevados y el entorno de ejecución se desactiven una vez finalizada la tarea, lo que reduce los permisos permanentes y las superficies de ataque persistentes. El privilegio elevado caduca en el instante en que se completa la tarea.

De anomalía a contención automatizada. La supervisión durante la ejecución, que suele ser un modelo tradicional de detección de anomalías de ML (en lugar de un LLM), detecta que un agente realiza un volumen inusualmente alto de llamadas a una API interna que rara vez utiliza. Un guardián de IA integrado señala por separado una firma de inyección de solicitudes en su entrada más reciente. El motor de señales de riesgo correlaciona ambos eventos a través de Shared Signals Framework, supera un umbral de riesgo definido y activa automáticamente la revocación del token y la cuarentena de la red, así como la suspensión del entorno de ejecución activo del agente. De ese modo, preserva el estado en la memoria para una revisión forense, en lugar de finalizarlo directamente, y todo sin esperar a que un humano detecte la alerta. El equipo de seguridad recibe la información completa como un registro de confirmación de circuito cerrado, lo que reduce la contención de una investigación manual de varias horas a una respuesta automatizada y casi instantánea.

Convertir las revisiones de acceso en un proceso continuo. Durante cada ciclo continuo de certificación de acceso, el servicio gestionado de un integrador de sistemas global obtiene el estado completo del Directorio de agentes y las Políticas de acceso mediante la capa de telemetría abierta del plan, identifica a los agentes cuyos propietarios abandonaron la empresa y dirige las excepciones de separación de funciones al propietario del sistema correspondiente. Lo que era una revisión manual con hojas de cálculo se convierte en un flujo de trabajo continuo y preparado para auditorías.

Prácticas recomendadas para empezar

No necesita todas las funcionalidades de este plan desde el primer día. Este es el orden en que los integrantes de Alliance realizaron sus implementaciones.

Implemente primero un sistema unificado de registro y telemetría.

No se puede depurar la aplicación de políticas ni el comportamiento de referencia sin un único repositorio de registros normalizado; considérela como un requisito previo para cada paso que se describe a continuación.

Luego, comience con la detección, no con las políticas. No se puede gobernar lo que no se puede ver. Implemente la detección de agentes de IA en la sombra y un inventario inicial de agentes antes de redactar una sola política de acceso. La mayoría de los equipos se sorprenden de la cantidad de agentes que ya están en funcionamiento.

Registre antes de restringir. Reúna a todos los agentes detectados en un único directorio de agentes con identidades y propietarios verificados antes de aplicar políticas más detalladas o basadas en intenciones. Una política sobre un agente no registrado es una política sobre nada.

Autorización por capas. Los controles detallados (FGA) y basados en relaciones (ReBAC) deben estar implementados antes de que la autorización basada en intenciones pueda definir correctamente el alcance de los permisos del agente.

Instrumente la ejecución antes de automatizar la respuesta.

Implemente sistemas de supervisión, rastreo y filtrado de salida/DLP con la suficiente amplitud como para establecer una referencia de comportamiento real antes de conectar el cumplimiento automatizado a las señales de riesgo. Si automatiza la contención con telemetría ruidosa o incompleta, los falsos positivos erosionarán la confianza en todo el sistema.

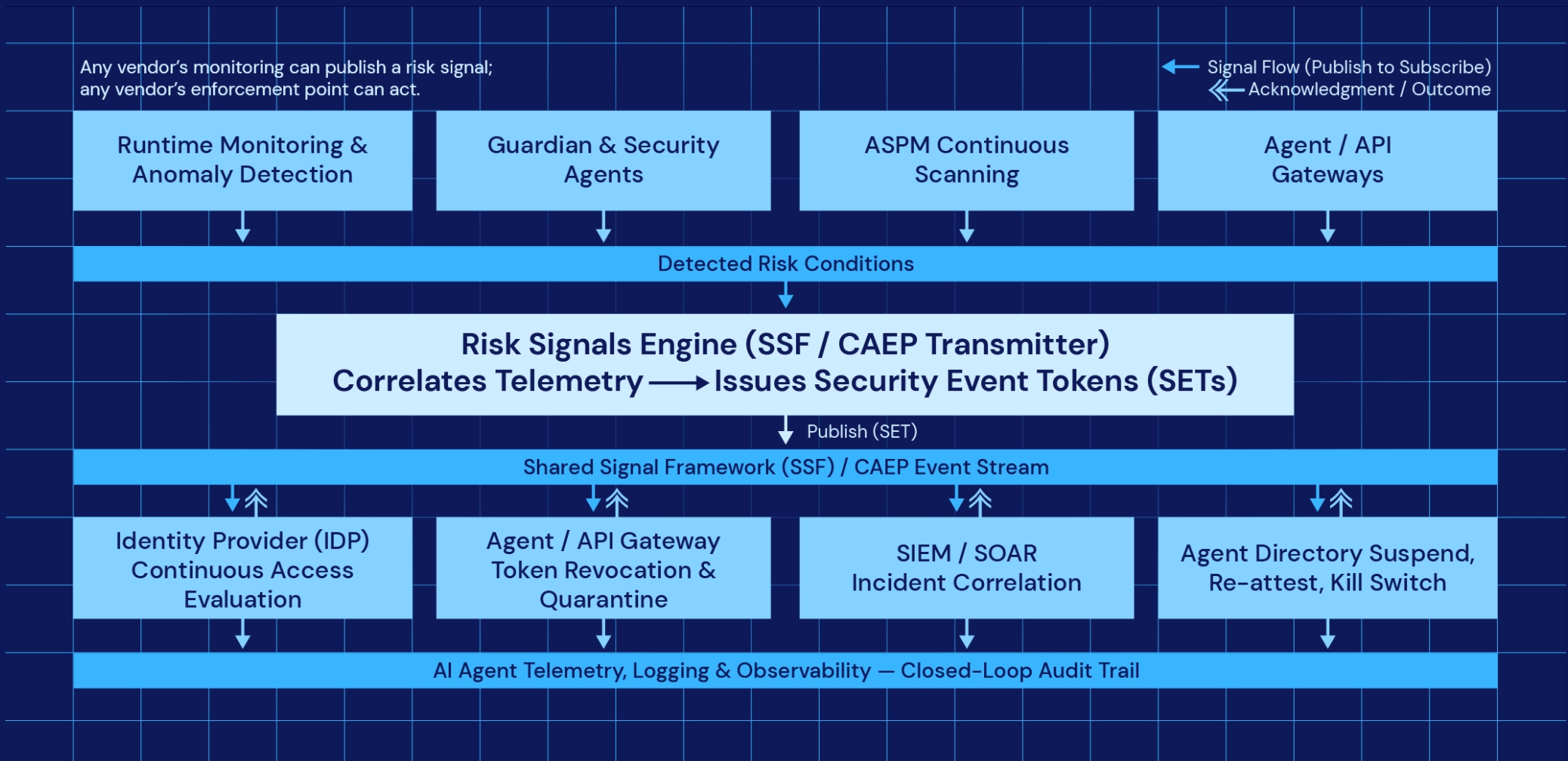
Pruebe el interruptor de desconexión antes de necesitarlo. Valide periódicamente la revocación de tokens, la finalización de sesiones y la cuarentena de la red con un agente que no sea de producción. Un interruptor de desconexión sin probar es una hipótesis, no un mecanismo de control.

Considere la gobernanza un proceso continuo, no un proyecto. Las comprobaciones de revisiones de acceso, solicitudes de acceso y separación de funciones deben realizarse con la misma frecuencia que la gobernanza de la identidad humana, no como una iniciativa puntual de la IA con agentes. Un mecanismo de integración de sistemas global puede ayudar a realizar esa transición entre modelo de proyecto y modelo operativo.

Estos materiales tienen fines informativos generales y no constituyen asesoramiento legal, de privacidad, de seguridad, de cumplimiento ni empresarial. El contenido puede no reflejar los desarrollos más recientes en materia de seguridad, aspectos legales o privacidad. Usted es el único responsable de obtener asesoramiento de su propio asesor legal o profesional y no debe basarse únicamente en estos materiales como sustitutos del asesoramiento profesional. Los autores no hacen declaraciones ni ofrecen garantías respecto a este contenido y no se hacen responsables de pérdidas o daños que puedan derivarse de la implementación de estas recomendaciones.

Flujo de datos en tiempo de ejecución:

Cómo se propagan las señales de riesgo (SSF/CAEDP)



Vista unificada:

Flujo de datos + flujo de señales en el plano de control

