

Governança para a execução agêntica

Blueprint de arquitetura para uma empresa agêntica segura

Onde estão os meus agentes?

O que eles podem fazer?

O que eles estão fazendo?

Como responder?

Elaborado pelos membros fundadores da Blueprint Alliance
Uma arquitetura de referência intersetorial para CISOs, CIOs, CTOs e arquitetos de segurança corporativa

Índice

1	Resumo executivo
2	Introdução
3	Princípios do blueprint
4	Arquitetura de referência em formato blueprint
5	Operacionalização com Global System Integrators (GSIs)
6	Conclusão: definindo o futuro da segurança em IA
7	O blueprint em ação
8	Boas práticas para começar

Resumo executivo

O avanço dos agentes de IA representa um desafio para o modelo de segurança empresarial tradicional. Em apenas dois anos, os chatbots experimentais baseados em LLM se transformaram em fluxos de trabalho autônomos com agentes capazes de executar várias etapas. O software antes fazia o que lhe era dito; agora ele decide, age e se conecta por conta própria. Essa autonomia gera valor comercial real, mas introduz um desafio de governança que a maioria das equipes de segurança ainda não está preparada para gerenciar.

[A Gartner prevê](#) que, até 2028, uma média global da lista Fortune 500 Enterprise terá mais de 150 mil agentes em uso, um aumento significativo em relação aos menos de 15 em 2025, gerando uma proliferação considerável de agentes, complexidade de TI e desafios de gerenciamento. Esses agentes não se limitam a exibir dados. Eles raciocinam, roteiam, chamam APIs downstream, criam subagentes e executam operações de várias etapas em diversos sistemas corporativos.

Sem um padrão arquitetônico unificado, as organizações enfrentam uma lacuna crescente em termos de identidade, segurança e governança: agentes paralelos se multiplicam, credenciais são herdadas fora das fronteiras de confiança e a execução sai do controle. A arquitetura de referência da Blueprint Alliance é uma estrutura aberta com vários fornecedores, publicada para uso pela comunidade aberta, que transfere a segurança do perímetro de rede para um plano de controle agêntico, unificado e Zero Trust. Com ela, os líderes de tecnologia passam a ter uma forma concreta de expandir as implantações agênticas sem perder o controle.

Introdução

Durante décadas, a segurança empresarial baseou-se em uma separação clara entre identidades humanas (gerenciadas via IAM, MFA, SSO) e identidades de máquina (gerenciadas via contas de serviço estáticas, chaves de API ou serviços de computação em nuvem). Os agentes autônomos de IA acabam com essa dicotomia. Quando um agente de IA age em nome de um funcionário humano, herda privilégios de acesso, consulta bancos de dados vetoriais e aciona transações financeiras, ele se torna um principal de segurança ativo e um agente operacional que exige controles explícitos de identidade, acesso e governança.

A segurança não pode ser implementada posteriormente apenas na camada de modelo ou de rede. Ela precisa estar ancorada em um plano de controle agêntico que governe a visibilidade, a conectividade e a execução de maneira contínua.

Software tradicional (determinístico)	IA agêntica (não determinística)
Executa caminhos de código explícitos, passo a passo.	Recebe intenções humanas de alto nível e formula planos de execução com várias etapas.
Depende de permissões estáticas e predefinidas e de contas de serviço.	Seleciona APIs e aplicativos dinamicamente, utilizando token vaults, corretores de credenciais e sessões delegadas.
Limitado estritamente a caminhos de integração embutidos no código.	Atua de forma autônoma, utiliza ferramentas por meio do Model Context Protocol (MCP) e gera subagentes.
Opera vinculado a sessões de usuários autenticados.	Opera de forma assíncrona, frequentemente muito tempo depois de o delegador humano ter encerrado a sessão.

A lacuna de governança:

Cinco perguntas que todo conselho está fazendo

Os agentes de IA estão se multiplicando mais rápido do que a maioria das equipes de segurança consegue acompanhar. [A Gartner prevê](#) que, até 2028, uma média global da Fortune 500 Enterprise terá mais de 150 mil agentes em uso, um aumento significativo em relação aos menos de 15 em 2025, gerando uma proliferação considerável de agentes, complexidade de TI e desafios de gerenciamento. Apenas 13% das organizações acreditam ter a governança adequada para agentes de IA. Essa lacuna é o motivo pelo qual os comitês de segurança cibernética e conformidade continuam voltando às mesmas cinco perguntas.

1. Onde estão nossos agentes de IA e quem é responsável por eles?
2. Como restringir o que os agentes podem executar e mantê-los focados na tarefa?
3. É possível auditar e explicar a tomada de decisão autônoma posteriormente?
4. Qual é a nossa exposição de dados e o risco de terceiros em toda a cadeia de suprimentos de agentes?
5. Temos um mecanismo de desativação e um plano de reversão verificado caso um agente saia do controle?

Essas cinco perguntas formam os quatro pilares da arquitetura de referência do blueprint na Seção 3. “Descoberta e inscrição de identidades” responde à primeira pergunta; “gerenciamento de acesso” responde à segunda pergunta e “autorização e monitoramento durante a execução” respondem às perguntas três e quatro. Já a resposta da quinta pergunta é “contenção ativa”. Respondendo a essas cinco perguntas com evidências, e não apenas com garantias, um CISO pode acelerar com segurança a adoção de agentes e, ao mesmo tempo, adequar a profundidade do controle à exposição regulatória e operacional. Aqueles que não conseguem responder permanecem vulneráveis a incidentes de segurança significativos e crescentes.

Essas perguntas não são uma lista de verificação uniforme. Na verdade, funcionam como uma ferramenta de diagnóstico para avaliar o risco. Os requisitos de segurança variam naturalmente de acordo com o escopo operacional: um assistente de produtividade somente leitura não precisa dos mesmos controles de contenção que um agente que executa transações financeiras. As equipes de segurança capazes de diagnosticar quais riscos realmente se aplicam podem adequar os controles à exposição no mundo real. Isso impede que as organizações compliquem demais as implantações de baixo risco e enfrentem atrasos de aprovação de vários meses, além de garantir que os agentes de alto risco recebam limites operacionais adequados.

Princípios do blueprint

A estrutura da Blueprint Alliance leva o princípio do Zero Trust aos agentes: é preciso tratá-los como identidades de primeira classe, não como código, credenciais ou um mero complemento à estrutura de identidade humana. Esses seis princípios sustentam tudo o que se segue.

Cada agente hospedado é uma identidade de segurança distinta, não uma reflexão tardia. Os agentes são provisionados, autenticados e desprovisionados com o mesmo rigor que os funcionários e as contas de serviço. Todos os agentes corporativos hospedados, incluindo pilotos e ferramentas internas, devem operar dentro de uma estrutura de governança adequada, com profundidade de controle proporcional ao seu risco e escopo operacional. Os runtimes locais operam mecanismos especializados de contenção de endpoints, em vez de inscrição formal em diretórios de identidade.

O acesso é limitado a cada tarefa, não sendo permanente. Embora as concessões dinâmicas com escopo de tarefa representem o objetivo ideal, as implementações práticas combinam concessões de curta duração para cada tarefa com permissões básicas resilientes para equilibrar a sobrecarga.

A delegação é rastreável de ponta a ponta. Quando um humano delega a um agente e esse agente gera um subagente, a cadeia de responsabilidade deve sobreviver a cada salto. Toda ação deve ser rastreada até o ser humano ou processo que a autorizou. Embora a delegação de múltiplos saltos de ponta a ponta continue sendo uma meta ambiciosa, considerando as ferramentas disponíveis atualmente, este blueprint define a trajetória de referência à medida que os padrões amadurecem.

O runtime do agente é isolado e monitorado, não apenas provisionado. As decisões de acesso tomadas no momento da configuração são necessárias, mas não suficientes. O modelo exige observação contínua durante as execuções para estabelecer parâmetros comportamentais, e não apenas o que foi concedido.

A contenção é instantâneo e reversível. Cada agente, independentemente do fornecedor ou plataforma, precisa de um mecanismo de desativação direcionado que possa suspendê-lo, colocá-lo em quarentena ou encerrá-lo em tempo real, com um caminho claro de retorno. A combinação do monitoramento comportamental contínuo com um limite de execução estrutural reduz a lacuna entre “autorizado” e “seguro”, contendo o raio de impacto no momento em que um agente se desvia de seu parâmetro.

A governança se adapta na velocidade da IA. Políticas de segurança estáticas não conseguem acompanhar a rápida evolução das capacidades agênticas. É a governança que precisa abranger aquilo que as políticas sozinhas não conseguem prever. Este blueprint funciona como uma arquitetura adaptável que acompanha os comportamentos emergentes dos agentes, combinando limites de permissão estáticos com definição do escopo dinâmica em tempo real, permitindo que as organizações mantenham limites operacionais rígidos enquanto adaptam o acesso ao contexto.

Juntos, esses seis princípios levam o Zero Trust a uma nova classe de identidades: agentes que operam de forma constante, persistente e autônoma em toda a empresa. Se forem aplicados corretamente, a lacuna de identidade, segurança e governança se fecha.

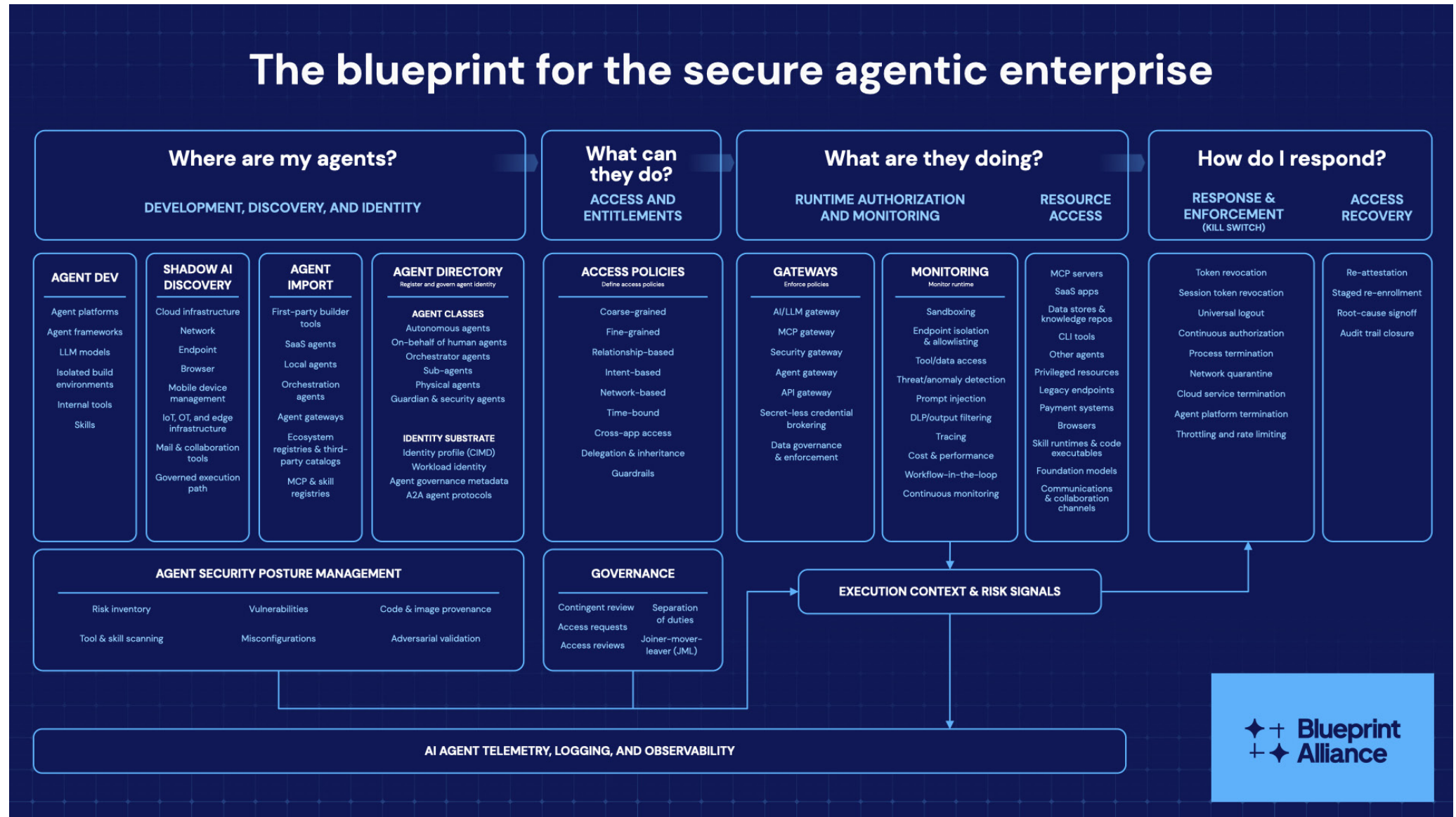
Arquitetura de referência em formato blueprint

A arquitetura de referência em formato blueprint fundamenta a segurança empresarial agêntica em quatro questões operacionais necessárias para governar a execução autônoma, cada uma correspondendo a um pilar da arquitetura de referência:

1. Onde estão os meus agentes?
2. O que eles podem fazer?
3. O que eles estão fazendo?
4. Como responder?

O que permeia todas essas questões é a necessidade de uma base sólida de contexto de execução e sinais de risco, alimentada continuamente por telemetria em runtime, registros e observabilidade. Em conjunto, funcionam como o tecido conjuntivo que permite ao sistema identificar padrões de risco e responder de forma coordenada.

Arquitetura de referência em formato blueprint



3.1 Pilar 1:

Onde estão os meus agentes? (desenvolvimento, descoberta e identidade)

O maior desafio para garantir a segurança da empresa agêntica é a visibilidade. As organizações não podem governar ou proteger aquilo que não sabem que existe. Esse pilar estabelece uma capacidade de rastreamento abrangente em todos os ambientes corporativos para identificar, catalogar e avaliar cada agente de IA que transita pela infraestrutura corporativa, seja ele autorizado ou não gerenciado. Ele conclui com o registro de agentes validados como identidades governadas com proprietários responsáveis, produzindo o inventário autorizado ao qual o Pilar 2 vincula as políticas de acesso.

3.1.1 AI Agent Discovery

O AI Agent Discovery funciona como um mecanismo de detecção contínua da presença de agentes na empresa. Ele mapeia todo o ciclo de vida dos agentes, desde os pipelines de engenharia internos até as integrações externas de terceiros, garantindo que nenhum modelo ativo ou processo autônomo opere em um ponto cego.

3.1.1.1 Desenvolvimento de agentes

Esse componente tem como foco os ambientes de engenharia internos nos quais agentes proprietários são construídos, treinados e ajustados ativamente. A proteção do pipeline de desenvolvimento ajuda a garantir a segurança desde o design, antes mesmo que um agente interaja com os dados de produção.

- **Plataformas de agentes:** ambientes de orquestração gerenciados pela empresa, onde os agentes são hospedados, treinados e implantados sistematicamente.
- **Estruturas de agentes:** as bibliotecas de software e arquiteturas subjacentes (como LangChain, LlamaIndex ou AutoGen) utilizadas pelos desenvolvedores para construir a lógica dos agentes.
- **Modelos LLM:** os modelos de linguagem de grande escala, fundamentais e refinados, que funcionam como os motores cognitivos que impulsionam os agentes.
- **Ambientes isolados de compilação e execução:** limites de CI/CD seguros e isolados computacionalmente que executam o código dos agentes durante os testes e avaliações, garantindo que as equipes de segurança descubram execuções de teste efêmeras em vez de depender exclusivamente de declarações estáticas de registro.
- **Ferramentas internas adicionais:** repositórios de código personalizados, assistentes de código e registros de modelos internos integrados ao pipeline de desenvolvimento.
- **Habilidades:** módulos procedurais definidos, trechos de código ou ferramentas desenvolvidas para estender as capacidades de um agente.

3.1.1.2 Agent Import

As empresas modernas raramente dependem exclusivamente de uma única fonte de inteligência. Em vez disso, como forma de se protegerem contra as capacidades dos agentes em rápida transformação, incorporam diversas arquiteturas de agentes que variam em origem, estrutura e limites computacionais. O Agent Import monitora e controla a introdução dessas diferentes classes de agentes no ecossistema corporativo, concentrando-se principalmente em como os dados fluem pela organização com base no modelo de implantação subjacente.

Tipo de agente, categoria e canal são três perguntas distintas que você pode fazer sobre qualquer agente importado. Em geral, um único agente terá uma resposta para todas as três simultaneamente — por exemplo, um agente SaaS (desenvolvido por um fornecedor) usado para fluxos de trabalho internos de RH (B2E), importado por meio de um gateway de agentes.

Tipo de agente: quem construiu?

- **Agentes próprios:** recursos de IA desenvolvidos internamente, que variam desde implementações totalmente personalizadas, construídas com código puro (como Python e LangChain) executadas em computação de propósito geral (por exemplo, Kubernetes ou Serverless Functions), até agentes construídos usando estruturas agênticas específicas do provedor que cuidam nativamente do runtime, da memória e da orquestração do fluxo de trabalho.
- **Agentes de SaaS:** produtos de IA externos, desenvolvidos especificamente para esse fim, que operam como entidades totalmente independentes dentro do ecossistema de software empresarial.
- **Agente local:** uma arquitetura cliente instalada diretamente no dispositivo local do funcionário, onde cada etapa de raciocínio é enviada para um provedor de LLM remoto na nuvem. Ela se comporta como uma camada não determinística entre a intenção humana e a ação da ferramenta, com dados cruzando a fronteira organizacional a cada chamada de inferência.
- **Agentes de orquestração:** agentes construídos em plataformas de orquestração visual que servem para unir APIs empresariais distintas.

Categoria do agente: para quem é?

- **Agentes B2E (Business-to-Employee):** agentes de IA internos, voltados para o funcionário, implantados para automatizar fluxos de trabalho corporativos, operações de TI/RH e produtividade diária do funcionário dentro de controles internos de identidade empresarial e acesso baseados em funções.
- **Agentes B2B:** agentes de IA voltados para a empresa que atravessam as fronteiras corporativas para automatizar fluxos de trabalho entre empresas, transações da cadeia de suprimentos e integrações de API entre locatários.
- **Agentes B2C:** agentes de IA voltados para o consumidor, implantados para interagir diretamente com usuários finais externos para serviços, comércio ou suporte, exigindo defesas rigorosas contra injeções públicas de prompts, exposição da marca e vazamentos de PII de consumidores.
- **Agentes B2B2C:** agentes de IA personalizados com marca própria ou fornecidos pelo fornecedor, disponibilizados por meio de um cliente corporativo para atender os consumidores finais desse parceiro, operando sob isolamento de vários níveis e limites de confiança transitivos.
- **Agentes assistentes pessoais:** assistentes de IA de nível do consumidor, direcionados pelo usuário, que operam em nome de um indivíduo em contextos pessoais e empresariais, exigindo controles de limites entre o acesso a dados pessoais e corporativos.

Canal de ingestão: como encontrá-lo?

- **Gateways de agentes:** pontos de ingestão e proxies através dos quais esses tipos de agentes externos ou importados interagem com as APIs internas da empresa.
- **Registros de ecossistemas e catálogos de terceiros:** pipelines de conectores automatizados e adaptadores de API que monitoram, extraem e mapeiam continuamente definições de agentes de registros nativos de fornecedores externos, garantindo que os agentes sejam catalogados de forma automática, sem intervenção.
- **Registros de MCP e Skills:** pontos de ingestão pelos quais os agentes extraem esquemas de ferramentas, habilidades e definições de contexto de registros de MCP e Skills públicos ou hospedados por fornecedores. Os próprios registros são um vetor de injeção. Como uma entrada maliciosa ou disfarçada pode ser captada e consumida automaticamente, toda ferramenta ou habilidade proveniente de registros é tratada como entrada não confiável, sujeita à verificação da cadeia de suprimentos na Seção 3.1.2 antes que um agente possa invocá-la..

3.1.1.3 Shadow AI Discovery

Os funcionários tendem a recorrer naturalmente a ferramentas de IA não aprovadas para otimizar seus fluxos de trabalho, o que introduz riscos de conformidade, exfiltração de dados e segurança operacional. O Shadow AI Discovery analisa ativamente os perímetros corporativos para detectar implantações e usos não autorizados de agentes. As funcionalidades de descoberta e monitoramento devem ser implementadas em conformidade com os requisitos aplicáveis de privacidade, emprego, trabalho, comissões de trabalhadores e proteção de dados, com transparência, proporcionalidade, minimização de dados, retenção e controles de acesso adequados.

- **Infraestrutura em nuvem:** instâncias de computação em nuvem, ambientes sem servidor e endpoints de API não gerenciados descobertos por meio de inspeção ativa de rede/API.
- **Rede:** análise de tráfego em tempo real e monitoramento de saída para identificar chamadas de API não reconhecidas para provedores de LLM externos e endpoints de IA.
- **Endpoint:** monitoramento em nível de dispositivo para rastrear a execução não autorizada de binários de IA, scripts ou pesos de modelos locais.
- **Navegador:** rastreamento de extensões e monitoramento de sessão do navegador para detectar quando os usuários fazem login em plataformas de IA baseadas na web ou sistemas de chat não autorizados.

- **Gerenciamento de dispositivos móveis (MDM) e gerenciamento de endpoints:** auditoria contínua em nível de dispositivo para detectar executáveis de IA locais não aprovados, pesos de modelos locais, ferramentas de agentes de CLI e extensões de IDE, ao mesmo tempo que uma postura de segurança básica é imposta e runtimes de agentes não gerenciados são bloqueados na camada do sistema operacional.
- **IoT, OT e infraestrutura de borda:** monitoramento de segurança industrial e inspeção de protocolos de rede para detectar modelos de linguagem de pequeno porte (SLMs) e agentes autônomos em equipamentos de manufatura, PLCs e hardware embarcado em que o software de endpoint tradicional não pode ser implementado.
- **Fluxo de e-mail e colaboração:** análise do tráfego de mensagens, chats e calendários para detectar agentes não autorizados enviando mensagens, respondendo automaticamente ou gerenciando calendários em nome de um usuário antes que a atividade se torne visível nos registros de endpoint ou de saída de rede.
- **Caminho de execução governado como superfície de detecção:** um limite arquitetônico onde todos os agentes autorizados são executados por um plano de controle gerenciado centralmente. Qualquer processo agêntico que opere fora desse limite é sinalizado de forma automática como um sinal de IA paralela, sem a necessidade de regras de detecção específicas do ambiente.

3.1.2 Gerenciamento da postura de segurança de agentes de IA (AIASPM)

Uma vez descobertos, os agentes devem ser avaliados de forma contínua. O AIASPM fornece a camada de diagnóstico que avalia perfis de risco, integridade de configuração e vulnerabilidades de software conhecidas em todo o inventário de agentes.

- **Inventário de riscos:** um registro centralizado e dinâmico que, sem considerar a etapa do ciclo de vida, rastreia a postura de segurança, as vulnerabilidades, as dependências de software e as configurações incorretas em todos os agentes descobertos, oferecendo uma visão dos riscos de segurança para complementar o Agent Directory focado em IAM.
- **Vulnerabilidades:** o gerenciamento padrão de vulnerabilidades de software foi estendido às dependências dos agentes e aos endpoints do modelo, abrangendo CVEs em toda a base de código do agente, cadeia de suprimentos e modelos fundamentais, mapeados para estruturas padronizadas para uma pontuação de risco consistente.
- **Proveniência de código e imagem:** verificação criptográfica da origem, integridade da compilação e assinaturas do código do agente, imagens de contêiner, dependências e pesos do modelo (por exemplo, atestados da estrutura SLSA) antes da execução no ambiente de produção. A utilização de imagens base mínimas e pré-atestadas, desprovidas de pacotes desnecessários e reconstruídas com base em dados CVE atuais, reduz a superfície de ataque da cadeia de suprimentos e simplifica a verificação de procedência no momento da compilação.
- **Análise da cadeia de suprimentos de ferramentas e habilidades:** inspeção estática e dinâmica automatizada de conjuntos de instruções de agentes locais, habilidades de execução e esquemas de ferramentas do Model Context Protocol (MCP) remoto, incluindo entradas extraídas de registros de MCP e de habilidades, para detectar código malicioso incorporado, operações ocultas e caminhos de saída não autorizados antes da inscrição.
- **Configurações incorretas:** varredura contínua para detectar configurações de parâmetros fracas, prompts de sistema inseguros, credenciais de longa duração embutidas no código, vetores de personificação e desvios de configuração introduzidos após a aprovação da revisão inicial do agente.
- **Validação adversarial:** testes adversariais contínuos e automatizados de agentes e seus limites operacionais contra métodos conhecidos de injeção de prompts, jailbreak e exfiltração de dados dentro de limites de execução descartáveis e isolados. O objetivo é avaliar com segurança a explorabilidade no mundo real sem risco de impacto na produção.

3.1.3 Agent Directory (registrar e governar a identidade do agente)

Enquanto o Agent Import (Seção 3.1.1.2) captura a proveniência e os metadados durante a descoberta, o Agent Directory é o repositório de identidade autorizado onde as credenciais criptográficas, a propriedade e os direitos de política são formalmente registrados e mantidos para todos os agentes validados e hospedados na empresa. O registro é a fronteira entre um agente descoberto e um agente governado, sendo o pré-requisito para todas as decisões de acesso no Pilar 2.

Cada agente hospedado e cada subagente com direitos distintos deve ser registrado aqui com um perfil verificado, identidade criptográfica, proprietário responsável e relações estruturais claras antes de poder solicitar acesso aos sistemas Enterprise. Mantenha a inscrição simplificado o suficiente para que as equipes não sejam levadas a contorná-lo usando credenciais humanas compartilhadas. Agentes locais e subagentes efêmeros que herdaram o escopo de um agente principal são regidos pelos controles de contenção e delegação descritos abaixo, em vez de por entradas de diretório individuais. Após a importação de um agente, o Diretório responde a mais duas perguntas sobre como ele opera em sua Enterprise: qual o papel que desempenha e como comprova sua identidade.

Classes de agentes: qual é a função exercida?

- **Agentes autônomos:** agentes totalmente independentes que operam de forma assíncrona para atingir objetivos comerciais de longo prazo sem supervisão humana imediata.
- **Agentes humanos OBO (On-Behalf-Of):** agentes que executam tarefas diretamente vinculadas à sessão de um usuário humano específico. Eles atuam como um proxy e estão sujeitos às restrições das permissões subjacentes desse usuário.
- **Agentes orquestradores:** agentes de nível superior que decompõem um objetivo em subtarefas e criam, coordenam e supervisionam subagentes downstream. O diretório deve rastrear cadeias de delegação e relações de geração como atributos de primeira classe.
- **Subagentes:** agentes filhos especializados, gerados ou invocados dinamicamente por um orquestrador pai para executar subtarefas discretas. São regidos por escopos de sessão herdados em vez de inscrição de diretório individual. Como os limites de tarefas no modelo, impostos dentro de um LLM, podem falhar, os subagentes devem ser envolvidos em limites de execução isolados que imponham estruturalmente o escopo de permissão herdado, impedindo que erros de raciocínio resultem na escalada de privilégios.

- **Agentes físicos/corporificados:** agentes que controlam robôs, atuadores ou outros sistemas físicos. Essa arquitetura rege a identidade e o acesso, mas complementa (em vez de substituir) os controles existentes de segurança da planta, segurança de TO e gerenciamento de mudanças.
- **Agentes de segurança e proteção:** agentes de IA administrativos especializados que operam dentro do plano de controle de segurança (por exemplo, agentes de triagem de SOC, mecanismos de políticas dinâmicas, scanners AIASPM contínuos). Devido aos seus privilégios elevados para inspecionar telemetria e executar ações de contenção, as identidades dos agentes de proteção exigem atestação criptográfica obrigatória da carga de trabalho, separação estrita de funções e aprovação contínua do proprietário.

Substrato de identidade: o que comprova que é realmente esse agente?

- **Perfil de metadados de identidade do agente:** perfis de metadados avançados (CIMD) que definem a intenção, a origem, a linhagem e o contexto operacional do agente para orientar decisões de acesso baseadas em risco.
- **Identidade da carga de trabalho:** estruturas de produção de identidade criptográficas, independentes de plataforma e baseadas em padrões, usadas para emitir IDs seguros e verificáveis para agentes em ambientes de nuvem dinâmicos.

- **Metadados de governança do agente:** metadados estruturados vinculados ao registro do agente no diretório, incluindo o proprietário humano responsável, equipe operacional designada ou fila de plantão, status do ciclo de vida (por exemplo, ativo, suspenso, desativado), nível de risco e escopo operacional. Devem ser usados de forma contínua pelos mecanismos de política para criação de controle de acesso e avaliação em runtime.
- **Protocolos de agentes A2A:** agentes de IA que utilizam estruturas de interoperabilidade padronizadas para descobrir recursos de forma dinâmica, delegar tarefas e colaborar ponto a ponto além das fronteiras empresariais. Ao interagir entre organizações sem um diretório compartilhado ou raiz de confiança, os agentes utilizam atestação criptográfica e padrões de troca de tokens (por exemplo, WIMSE/AIMS, OAuth Token Exchange, HTTP Message Signatures) para preservar as cadeias de delegação e verificar a identidade principal entre os intermediários.

3.2 Pilar 2:

O que eles podem fazer? (acesso e direitos)

Os agentes autônomos são fundamentalmente criados para agir, o que exige que eles se conectem a dados, APIs e sistemas. Com base nas identidades verificadas estabelecidas no Pilar 1, esse pilar define os limites e permissões que governam a forma como os agentes interagem com os recursos corporativos, tanto em sessões diretas delegadas por humanos quanto em fluxos de trabalho complexos com múltiplos agentes. Isso amplia os modelos legados centrados no ser humano com a governança de IA de máquina para máquina.

3.2.1 Políticas de acesso (definição de políticas de acesso)

As políticas de acesso especificam os limites operacionais do que um agente com identidade verificada tem permissão para ler, escrever ou executar. Como a natureza dos resultados da IA generativa é imprevisível, essas políticas devem ser profundamente contextuais e adaptáveis.

- **Baixa granularidade:** controles abrangentes, baseados em funções que definem a quais ambientes de alto nível ou macrosserviços um agente pode se conectar.
- **Alta granularidade:** permissões baseadas em atributos e em nível de objeto que restringem um agente a linhas de banco de dados, arquivos ou métodos de API específicos, levando em consideração explicitamente os rótulos de classificação de dados corporativos (por exemplo, Confidencial, PII) para impor limites de dados em conformidade. As estruturas de políticas podem abranger classificação de dados, identidade do agente, escopo do risco, contexto ambiental e tipo de ação.

- **Baseado em relacionamentos:** acesso derivado dos relacionamentos de grafo de um agente: quem o gerou, em nome de quem ele age e em que ele está aninhado.
- **Baseado em intenção:** mecanismos de política avançados que analisam a intenção semântica do prompt ou do plano do agente antes de conceder acesso a recursos. É possível inspecionar um plano com as ferramentas atuais. No entanto, tratar um prompt literal do usuário como o único gatilho é uma aspiração, já que nem sempre são os prompts que levam um agente a agir. Privilégios efêmeros e com prazo determinado são criados estritamente para a duração de uma tarefa específica e revogados imediatamente após a conclusão.
- **Baseado em rede:** macrossegmentação que restringe o tráfego de agentes em VPCs corporativas, combinada com controles de microssegmentação L3/L4 que restringem a comunicação do agente com o alvo a intervalos de IP ou malhas de serviço autorizados.

- **Temporário:** as concessões com tempo determinado são limitadas a uma única tarefa ou sessão e são revogadas automaticamente após a conclusão ou o término do período de tempo limite. Como as concessões com aprovação automática e expiração automática agregam valor de controle limitado por si só, combine a definição do escopo com prazo definido com a detecção de anomalias na Seção 3.3.1.2 em vez de considerar apenas o prazo de validade como salvaguarda.
- **Cross-App Access:** estruturas de políticas e controles de limite que governam a capacidade de um agente de transitar por ecossistemas de aplicativos distintos, permitindo a tradução segura de dados, a troca de tokens e a execução de funções enquanto orquestra fluxos de trabalho em diferentes plataformas empresariais.
- **Delegação e herança:** para evitar o abuso de privilégio por delegação, o limite de permissão efetivo para um subagente é calculado como a interseção de seus próprios direitos e os direitos do agente delegador ou do humano. A política também deve limitar a amplitude e a profundidade da delegação para evitar uma disseminação descontrolada.
- **Limites operacionais:** restrições de segurança rígidas e determinísticas, bem como limites comportamentais que se sobrepõem aos circuitos de raciocínio autônomo para prevenir globalmente ações não conformes, resultados tóxicos ou etapas operacionais de alto risco, independentemente da intenção do agente ou de seus privilégios de alta granularidade.
- **Revisão contingente:** acione aprovações de fluxo de trabalho automatizadas ou com intervenção humana sempre que a ação de um agente ultrapassar limites financeiros ou operacionais definidos (por exemplo, uma compra ou reserva superior a US\$ 1.000), mesmo se a ação estiver dentro das permissões técnicas concedidas.
- **Revisões de acesso:** varredura automatizada e contínua de direitos de acesso, auxiliando os proprietários do sistema a validar permissões não conformes.
- **Solicitações de acesso:** fluxos de trabalho formais e auditados para o provisionamento de novas funcionalidades, modelos ou acesso ao sistema para um agente existente. Embora as solicitações possam ser originadas por proprietários humanos ou dinamicamente por agentes em runtime, as solicitações iniciadas por agentes devem ser avaliadas de forma rigorosa em relação aos limites de permissão estáticos. Para impedir que os agentes manipulem fluxos de trabalho automatizados para conceder acesso excessivo, elas exigem verificação obrigatória do proprietário para expansão de escopo.
- **Separação de funções (SoD):** regras algorítmicas que impedem um único agente de possuir privilégios conflitantes, como gerar e aprovar transações financeiras.
- **Entradas, movimentações e desligamentos:** aplique a mesma disciplina de integração, gerenciamento de mudanças e desprovisionamento usada para os ciclos de vida da identidade humana ao escopo, propriedade ou evolução da versão do modelo de um agente.

3.2.2 Governança

A governança fornece a base de conformidade e o ciclo de supervisão contínua, usando ferramentas automatizadas em vez de revisões manuais superficiais para manter os privilégios dos agentes alinhados com o princípio do menor privilégio na velocidade do agente.

3.3 Pilar 3:

O que eles estão fazendo? (autorização e controle em runtime)

A visibilidade e a identidade são ineficazes sem monitoramento e fiscalização inline em runtime. Esse pilar se concentra na interceptação, inspeção e autorização das ações dos agentes em tempo real à medida que executam tarefas, garantindo que não se envolvam em comportamentos maliciosos, não manipulem indevidamente dados sensíveis ou não comprometam alvos críticos da empresa.

3.3.1 Autorização e monitoramento em runtime

A autorização e monitoramento em runtime funciona como a camada de inspeção ativa para sessões de agentes em tempo real. Ela se situa diretamente no caminho de execução, analisando tanto as entradas que chegam ao agente quanto as ações ou saídas que são emitidas de volta.

3.3.1.1 Gateways (fiscalização de políticas de acesso)

Os gateways funcionam como pontos de fiscalização de políticas (PEPs) em uma cadeia de defesa em profundidade. Eles interceptam o tráfego entre a camada de agentes e o restante da pilha empresarial, analisam protocolos e bloqueiam chamadas não autorizadas antes que elas cheguem ao destino. Os API gateways downstream mantêm a autoridade de autorização final, realizando sua própria validação de cada chamada, independentemente do que os gateways de IA/agente upstream já tenham inspecionado. Gateways de IA/LLM, MCP, segurança, agente e API são frequentemente implementados em conjunto, mas cada um possui um ponto de controle distinto:

Os gateways de IA/LLM controlam o tráfego de API de modelo; os gateways de MCP controlam o tráfego de protocolo de compartilhamento de ferramentas/contexto; os gateways de segurança fornecem inspeção de ameaças independente de protocolo; os gateways de agente controlam as questões de identidade e delegação específicas do agente; e os API Gateways aplicam controles clássicos da camada de API para as ferramentas e serviços que os agentes chamam.

- **Gateway de IA/LLM:** Intercepta chamadas de API de baixo nível para modelos fundamentais, gerenciando limitação de taxa, contagem de tokens e substituições de prompts do sistema.
- **Gateway MCP (Model Context Protocol):** um proxy especializado otimizado para gerenciar protocolos de compartilhamento de contexto e integrações de padrões abertos entre agentes e aplicativos, com foco em servidores MCP.
- **Gateway de segurança:** uma camada de mitigação de ameaças inline que fornece segurança de conteúdo dedicada e verificação de protocolo para interações do agente. Ele se insere diretamente no fluxo de tráfego para realizar inspeções profundas de payloads em tempo real, com terminação TLS, interceptando e filtrando ativamente injeções complexas de prompts, jailbreaks, strings de texto adversariais ocultas e payloads de código malicioso antes que atinjam o agente ou recursos downstream. Também lida com a varredura de classificação de dados para aplicar a ocultação de dados de saída, verificações de integridade estrutural nas saídas do modelo e inspeção criptográfica de payloads criptografados do agente.

- **Gateway do agente:** funciona como o plano de controle de runtime específico do agente: vincula a identidade do agente a cada solicitação, mantém o registro de ferramentas/MCP do qual os agentes extraem informações e governa as ações do agente, não apenas o acesso, com base na identidade, no contexto e na ação específica que está sendo solicitada. Ele interrompe interações inseguras em tempo real, bloqueando ou higienizando o tráfego arriscado de modelos e negando ou exigindo aprovação para ações de ferramentas sensíveis por meio de políticas integradas e personalizadas. Além disso, injeta credenciais downstream de curta duração e sem segredos nas chamadas de API em tempo de execução para que os agentes nunca manipulem, armazenem ou exponham chaves de API corporativas permanentes ou segredos de contas de serviço. Pode fornecer telemetria governada em interações multiagentes compatíveis e detectar ou mascarar dados confidenciais, incluindo PII, de acordo com políticas e recursos configurados, ao mesmo tempo em que oferece proteção contra comportamentos inseguros e ataques maliciosos baseados em prompts.
- **API gateway:** impõe controles clássicos da camada de API, limitação de taxa, cotas, validação de esquema, mTLS, WAF, modelagem de tráfego e revogação no gateway para as ferramentas e serviços que os agentes e servidores MCP utilizam, realizando sua própria verificação final de autorização em cada chamada em vez de presumir que os gateways de IA/agente upstream já detectaram todas as anomalias.

3.3.1.2 Monitoramento (monitoramento do runtime)

O monitoramento proporciona observabilidade profunda, detecção de ameaças e fiscalização ativa, analisando a telemetria em runtime em busca de comportamentos anômalos, vazamento de dados e tentativas de exploração externa, além de aplicar controles como sandbox e isolamento de endpoints em tempo real.

- **Sandboxing:** isolamento da execução de agentes (especialmente agentes não confiáveis, de terceiros ou recém-atualizados) em runtimes seguros e restritos para evitar movimentação lateral e conter riscos de execução por meio de limites estruturais.
- **Isolamento e listas de permissões de endpoints:** controles de runtime locais para agentes hospedados em endpoints, incluindo sandboxing em nível de sistema operacional, listas de permissões de aplicativos de endpoint e verificações de postura de dispositivo local para restringir a execução de comandos do agente local e o acesso a arquivos.
- **Acesso a ferramentas/dados:** inspeção estrutural e auditoria em tempo real de argumentos, payloads, esquemas e consultas passados para ferramentas e repositórios downstream. A realização de inspeções dentro ou diretamente adjacentes ao limite de execução ajuda a garantir que os planos de execução dinâmicos não se desviem dos parâmetros autorizados nem introduzam mutações não intencionais durante a sessão.
- **Detecção de ameaças/anomalias:** mecanismos de aprendizado de máquina rastreiam os parâmetros comportamentais dos agentes para sinalizar picos incomuns no acesso a dados, loops rápidos de API ou padrões de execução bizarros.

- **Injeção de prompts:** scanners inline desenvolvidos para detectar sequências de texto maliciosas, tanto diretas quanto indiretas, destinadas a sequestrar a lógica principal do agente ou a substituir as instruções do sistema.
- **DLP/filtragem de saída:** ferramentas de prevenção de perda de dados (DLP) que analisam saídas de agentes para mitigar o risco de vazamento de código-fonte proprietário, PII ou segredos comerciais para fornecedores externos ou usuários não autorizados.
- **Rastreamento:** mapeamento abrangente de transações que descreve o caminho exato de execução, as pilhas de chamadas e as invocações de subagentes acionadas por um prompt de usuário principal.
- **Custo e desempenho:** Rastreamento do uso de tokens, custos operacionais, latência e consumo de recursos para evitar loops descontrolados de agentes ou ataques de esgotamento de recursos.
- **Fluxo de trabalho integrado:** gatilhos programáticos que pausam a execução do agente para ações de alto risco, encaminhando-as para verificações cruzadas de políticas automatizadas ou aprovação humana explícita antes de prosseguirem.
- **Monitoramento contínuo:** monitoramento contínuo e possibilidade de exportação da atividade do agente, incluindo rastreamento granular, registro de trilha de auditoria e não repúdio das ações do agente.

3.3.1.3 Fiscalização contínua de intenção e alinhamento

A fiscalização contínua de intenção e alinhamento determina se um agente permanece dentro dos limites da tarefa delegada durante toda a execução ou se desviou para uma expansão de propósito não autorizada. Ela complementa as políticas de acesso inicial: a política baseada em intenção determina o que um agente pode acessar inicialmente, enquanto a fiscalização contínua do alinhamento avalia se os comportamentos evolutivos e de várias etapas permanecem apropriados ao longo do tempo.

As principais funcionalidades incluem:

- **Vinculação de finalidade:** garantir uma ligação estrita entre o usuário que delega, a identidade do agente, o fluxo de trabalho aprovado e a tarefa solicitada.
- **Avaliação de plano versus ação:** comparação em tempo de execução das chamadas de ferramentas planejadas com as ações observadas para detectar desvios de tarefas, desvios causados por injeção e abuso de privilégio por delegação.
- **Fiscalização contextualizada:** execução de ações em tempo real, incluindo permitir, negar, ocultar, aprovação com aumento de nível, redução de escopo, suspensão ou contenção.
- **Comprovação rastreável das decisões:** manter registros de auditoria irrefutáveis que expliquem por que cada ação relevante foi permitida, modificada ou bloqueada.

3.3.2 Acesso a recursos

O acesso a recursos define o raio de impacto real de um agente. Ele categoriza os diversos alvos downstream com os quais um agente pode interagir, ajudando a garantir que as camadas de segurança apropriadas cerquem cada interface.

3.3.2.1 Recursos-alvo (componentes que estão sendo acessados)

Os recursos-alvo representam os destinos das ações de um agente. Seja para obter contexto ou para implementar alterações, esses endpoints devem ser catalogados e protegidos de forma contínua.

- **Servidores MCP:** repositórios de dados e servidores utilitários explicitamente configurados para compartilhar informações contextuais (dados estruturados, documentos, embeddings, ferramentas e estado) com agentes por meio do Model Context Protocol.
- **Apps SaaS:** aplicativos de produtividade empresarial e negócios essenciais que contêm vastos repositórios de dados corporativos sensíveis.
- **Repositórios de dados e de conhecimentos:** repositórios estruturados e não estruturados, incluindo bancos de dados vetoriais, data lakes, data warehouses e sistemas de armazenamento transacional, utilizados por agentes para extrair contexto de negócios ou confirmar resultados operacionais. Esses repositórios são regidos por práticas de gerenciamento da postura de segurança de dados (DSPM), rastreando a linhagem e a classificação à medida que os agentes acessam cada repositório.

- **Ferramentas CLI:** interfaces de linha de comando que os agentes usam para executar scripts, configurações ou comandos operacionais em nível de sistema.
- **Outros agentes:** subagentes downstream ou agentes especializados encarregados de executar componentes fracionários de uma carga de trabalho mais ampla.
- **Recursos privilegiados:** ambientes de destino altamente sensíveis, incluindo diretórios raiz, repositórios de identidade e consoles de configuração mestre.
- **Endpoints padrão e legados:** APIs REST/GraphQL, bancos de dados e aplicativos legados empresariais que não oferecem suporte nativo a protocolos de contexto de agentes, como o MCP.
- **Sistemas de pagamento:** redes financeiras, gateways de transação e livros contábeis exigem o mais alto grau de validação criptográfica e supervisão humana.
- **Navegadores:** sessões web automatizadas utilizadas por agentes para coletar informações públicas ou interagir com consoles externos baseados na web.
- **Runtimes de habilidades e executáveis de código:** interpretadores de código, sandboxes para execução de scripts e módulos procedurais invocados dinamicamente por agentes para executar código gerado, exigindo limites de computação isolados para mitigar riscos de fuga em runtime.
- **Modelos de base:** os provedores de inteligência bruta onde os pesos são hospedados, exigindo chaves de API de modelo seguras e validadas.
- **Canais de comunicação e colaboração:** e-mail, chat e sistemas de tíquetes pelos quais os agentes interagem, representando caminhos de entrada de prompts não confiáveis sujeitos a inspeção e varredura profunda de payload.

3.3.2.2 Segurança de dados baseada em finalidade

A segurança de dados baseada em finalidade avalia se um determinado agente deve acessar, processar, transformar, resumir, divulgar ou transmitir dados específicos com base na finalidade delegada, na autoridade do usuário, na identidade do agente, no estado do fluxo de trabalho, na classificação dos dados, no destinatário, no destino e na ação downstream. Ela aplica controles de dados no momento do uso, não apenas no momento do acesso. Isso é importante porque os agentes combinam dados de diferentes sistemas, inferem informações sensíveis a partir de entradas não sensíveis e divulgam os resultados por meio de saídas, APIs, mensagens, arquivos e chamadas de subagentes. A classificação e a prevenção contra perda de dados determinam o que são os dados e se eles são sensíveis. Por sua vez, a segurança de dados com reconhecimento de finalidade determina se o uso desses dados é apropriado ao contexto. As principais funcionalidades incluem a vinculação de finalidade entre a solicitação do usuário, a tarefa do agente, a fonte de dados e o uso permitido; a avaliação em runtime para determinar se o acesso e a divulgação são adequados à etapa atual do fluxo de trabalho; controles para sumarização, transformação, exportação, movimentação entre apps e divulgação a agentes de terceiros; detecção de acesso indevido a dados, recuperação excessiva, inferência não autorizada e compartilhamento inadequado; e fiscalização que abrange ocultação, mascaramento, resposta parcial, fluxo de trabalho de aprovação, registro, bloqueio ou contenção.

3.4 Pilar 4:

Como responder? (contenção ativa)

O pilar final do blueprint transforma a telemetria em contenção imediata. Quando os sistemas de monitoramento sinalizam comportamentos anômalos, credenciais comprometidas ou um ataque de injeção de prompts, a empresa deve dispor de mecanismos de mitigação altamente automatizados e precisos que priorizem respostas proporcionais e menos disruptivas (limitação de taxa, quarentena) em vez de ações destrutivas que possam causar uma interrupção autoinfligida.

3.4.1 Resposta/fiscalização (throttling e contenção)

O circuito de controle operacional automatizado foi projetado para responder a sinais de risco em tempo real. Ele funciona como uma camada de controle coordenada, e não como um único ponto de controle centralizado, executando respostas de segurança direcionadas e proporcionais em toda a pilha tecnológica.

A fiscalização é mais eficaz quando combinada com um protocolo de notificação padronizado: o sistema que detecta o risco publica um evento de mecanismo de desativação, os sistemas inscritos agem e o resultado é publicado novamente como um sinal de confirmação.

3.4.1.1 Ações de fiscalização (mecanismos de mitigação)

As ações de fiscalização representam as alavancas concretas e de várias camadas que equipes de segurança e orquestradores automatizados podem acionar para isolar, conter ou neutralizar completamente um agente problemático ou comprometido. Essas ações formam uma biblioteca de opções de resposta pertencente ao SOC, e não interruptores isolados de acionamento único. Os respondentes selecionam a ação menos disruptiva proporcional ao sinal de risco agregado.

As ações de contenção acionadas por agentes guardiões (suspender, isolar, encerrar) exigem que o agente de destino opere dentro de um limite de execução que suporte controle granular em nível de sessão, seja em execução em um serviço na nuvem gerenciado, aplicativo de SaaS ou ambiente de execução em contêiner. Sem uma plataforma subjacente ou suporte de tempo de execução para pausar um agente sem impacto colateral nos Workflows adjacentes, o confinamento baseado em API proporciona uma falsa sensação de segurança. Ao conter agentes ligados a equipamentos físicos ou tecnologia operacional, os gatilhos automatizados devem ser calibrados para garantir que o isolamento abrupto não crie riscos à segurança física ou operacional no chão de fábrica.

- **Revogação de token:** invalida tokens de acesso OAuth ou de API ativos instantaneamente, congelando de modo imediato a capacidade do agente de chamar aplicativos conectados.
- **Revogação de token de sessão:** invalidação de tokens OAuth ativos e revogação de contextos de sessão downstream em serviços empresariais conectados e/ou encerramento da conexão de runtime ativa de uma única string de interação específica, sem necessariamente excluir a identidade geral do agente.
- **Autorização contínua:** reavaliação dinâmica e em tempo real dos contextos de segurança que reduz instantaneamente as permissões no exato momento em que o nível de risco se altera.

- **Encerramento do processo:** fechamento forçado dos runtimes de computação como último recurso quando a quarentena falha. Como o encerramento sacrifica a telemetria forense em memória, dê preferência à quarentena de rede, suspensão da sessão ou suspensão do runtime sempre que possível.
- **Quarentena de rede:** utilização de regras de microssegmentação para isolar o ambiente de hospedagem do agente, interrompendo sua comunicação tanto com a rede interna quanto com a internet aberta.
- **Encerramento do serviço na nuvem:** desligamento instantâneo das instâncias de nuvem subjacentes, funções sem servidor ou clusters de máquinas virtuais que hospedam o agente não autorizado.
- **Encerramento da plataforma do agente:** um comando administrativo de alto nível que instrui a plataforma centralizada de orquestração de agentes a suspender a conta operacional do agente e interromper todas as tarefas agendadas.
- **Throttling e limitação de taxa:** restrição progressiva da taxa de requisições, do orçamento de tokens ou da concorrência de um agente como primeira resposta proporcional, diminuindo a capacidade sem encerrar completamente a sessão.

3.4.2 Recuperação de acesso

Um mecanismo de desativação representa apenas metade do ciclo de controle: cada ação de contenção precisa de um caminho correspondente de volta a um último estado válido e verificado. A recuperação de acesso define como um agente contido, em quarentena ou encerrado é restaurado ao serviço, confirmando que a reintegração é deliberada, documentada e rastreável, em vez de uma reversão automática da ação de fiscalização. Fluxos de trabalho de recuperação totalmente automatizados e padronizados são um objetivo a ser alcançado. Hoje, a maioria das organizações ainda exige aprovação manual em uma ou mais das etapas abaixo.

- **Reatestação:** nova verificação da identidade, do código e da configuração do agente em relação a um parâmetro confiável antes que qualquer acesso seja restaurado em uma nova instância de execução isolada.
- **Reinscrição em etapas:** restauração gradual do acesso (por exemplo, somente leitura antes de escrita, em sandbox antes do ambiente de produção) em vez de restabelecer todos os privilégios de uma só vez.
- **Aprovação da causa raiz:** exigir que o responsável humano ou a equipe de segurança documente a causa raiz e a remediação antes de reativar o agente.
- **Encerramento da trilha de auditoria:** vinculação do evento original do mecanismo de desativação, das ações de remediação e da decisão de reinscrição em um único registro de incidente fechado e auditável.

3,5

Componentes fundamentais transversais

Essas camadas arquitetônicas não existem isoladamente: elas se situam abaixo e ao longo dos quatro pilares, agindo como o tecido que conecta visibilidade, identidade, monitoramento e resposta em uma estrutura de segurança unificada.

3.5.1 Contexto de execução e sinais de risco

O contexto de execução e os sinais de risco funcionam como o tecido analítico conectivo do blueprint. Alimentado continuamente por telemetria de execução em tempo real, avaliações de postura de segurança e sinais de identidade do usuário delegados (como indicadores de comprometimento de conta ou risco elevado do usuário principal), esse mecanismo de orquestração avalia as ameaças de forma dinâmica, correlacionando condições de baixo risco individuais em combinações tóxicas. Também publica um contexto de execução positivo, uma confirmação estrutural do que já se sabe ser válido, e não apenas alertas sobre o que deu errado. Quando um limite é ultrapassado (por exemplo, uma injeção de prompts repentina combinada com uma tentativa de exfiltrar dados financeiros), os sinais de risco acionam os mecanismos de resposta/fiscalização correspondentes de forma direta e programaticamente. Para ajudar a garantir a interoperabilidade em tempo real e respostas coordenadas em ambientes heterogêneos de fornecedores, esse sistema pode utilizar padrões abertos de engenharia de segurança, como a Shared Signals Framework (SSF) e o Continuous Access Evaluation Profile (CAEP), para publicar e consumir modificações de estado de Zero Trust de forma assíncrona.

3.5.2 Telemetria, registro e observabilidade de agentes de IA

Esta é a camada de base fundamental que sustenta todos os pilares da arquitetura de referência. Funciona como o plano de dados imutável, capturando metadados de execução em toda a pilha empresarial, com o respaldo de ocultação obrigatória de payload inline para impedir que PII, segredos e credenciais entrem no log lake. Sem esse log lake estruturado e de alta fidelidade, a detecção automatizada de ameaças torna-se impossível, as avaliações de postura ficam sem dados e os responsáveis pela resposta a incidentes forenses ficam completamente no escuro. Para ajudar a garantir a consistência e reduzir os silos de fornecedores em interações complexas com múltiplos agentes, esta camada pode utilizar padrões abertos, como o **Open Cybersecurity Schema Framework (OCSF)**, para normalizar diversos fluxos de registros em uma taxonomia de segurança unificada.

- Origem da telemetria resistente à adulteração: capturar a telemetria no limite da execução, em vez de depender de registros de agentes autodeclarados, ajuda a garantir o não repúdio, impedindo que um agente comprometido suprima, modifique ou falsifique seu registro de execução.

4

Operacionalização com Global System Integrators (GSIs)

Um blueprint arquitetônico requer um modelo de implantação operacional para ter valor empresarial duradouro. As seções abaixo definem como as equipes de segurança e os parceiros de consultoria operacionalizam os quatro pilares nos fluxos de trabalho empresariais existentes. As organizações podem operacionalizar a arquitetura por meio de equipes empresariais internas, GSIs, Managed Service Providers (MSPs) ou outros parceiros qualificados, dependendo das capacidades e requisitos organizacionais. Uma arquitetura de referência publicada sem um modelo operacional se torna obsoleta em menos de um ano. As seções abaixo definem como essa arquitetura é de fato implementada, estruturada e mantida dentro de uma empresa.

4.1 Consultoria de implantação e capacitação em arquitetura

Os GSIs convertem os quatro pilares em um rollout sequencial e específico para cada organização, em vez de uma implantação simultânea e de uma só vez. Isso inclui:

- **Avaliação do estado atual:** mapeamento da pilha de IAM, rede e monitoramento existente de uma empresa em comparação com a arquitetura de referência em formato blueprint para identificar lacunas reais e os recursos que já existem sob um nome diferente.
- **Design do modelo operacional de destino:** definição de responsabilidades operacionais claras em todos os pilares para evitar lacunas de responsabilidade entre as equipes. Normalmente, esse modelo atribui gateways e monitoramento à engenharia de segurança; Agent Directory, políticas de acesso e governança ao IAM; resposta/fiscalização ao SOC e limites de execução (sandboxing, runtimes de isolamento, gerenciamento do ciclo de vida) à engenharia de computação e plataforma.
- **Planejamento faseado do rollout:** sequenciamento da implantação adaptado aos investimentos em ferramentas já existentes do cliente e ao cronograma regulatório.
- **Integração de interoperabilidade de fornecedores:** configurar os gateways de IA/LLM, MCP, segurança, agentes e API específicos que uma empresa já licenciou para interoperar por meio da camada compartilhada de telemetria e sinais de risco, em vez de exigir uma substituição completa da infraestrutura existente.

4.2 Matrizes padronizadas de pontuação de risco

Nem todos os agentes exigem o mesmo nível de controle. Os GSIs fornecem estruturas de pontuação calibradas que impedem que as organizações exagerem na engenharia de implantações de baixo risco ou ofereçam proteção inadequada para as de alto risco. Isso inclui:

- **Modelos de classificação de risco** que avaliam os agentes de acordo com o escopo operacional (somente leitura ou transacional), classificações de dados, nível de autonomia (OBO ou totalmente autônomo) e raio de impacto dos recursos-alvo.
- **Limites calibrados pelo setor:** um agente de serviços financeiros que executa fluxos de trabalho de pagamento e um agente de manufatura que lê dados de sensores de tecnologia operacional justificam parâmetros de risco materialmente diferentes. Os GSIs trazem a biblioteca de padrões intersetoriais para calibrar esses limiares, em vez de começar do zero.
- **Avaliação do impacto nos negócios para governança semântica:** definição dos limites financeiros, regulatórios ou de reputação que acionam a aprovação obrigatória com intervenção humana, ajustada ao apetite real ao risco do cliente, em vez de um padrão genérico.
- **Profundidade de isolamento dimensionada para o nível de risco:** a categorização por nível de risco deve adequar a robustez do limite de execução estrutural (por exemplo, sandboxing em nível de kernel, ciclos de vida de runtime efêmeros) em conjunto com mecanismos de revisão, garantindo que um agente de alto risco não compartilhe o mesmo raio de impacto físico que uma ferramenta de baixo risco caso seja comprometido.

4.3 Gerenciamento contínuo do ciclo de vida

A governança de agentes não se resume a uma certificação única. Os GSIs operam os processos contínuos de ciclo de vida que mantêm o Agent Directory e as políticas de acesso precisos à medida que o ambiente muda:

- **Ciclos automatizados de certificação de acesso:** o serviço gerenciado consulta o Agent Directory e o estado das políticas de acesso, sinalizando os agentes com responsáveis humanos que deixaram a organização ou com direitos de acesso que se desviaram do princípio do menor privilégio.
- **Paridade entre entradas, movimentações e desligamentos para agentes:** aplicação da mesma disciplina de desprovisionamento usada para o desligamento de funcionários humanos às mudanças de propriedade do agente, upgrades de versão do modelo e aposentadoria de subagentes para que credenciais órfãs não se acumulem como a principal fonte da lacuna de identidade.
- **Roteamento de exceções de Separação de Funções (SoD):** operar os fluxos de trabalho que encaminham exceções de privilégios conflitantes (por exemplo, um agente que pode gerar e aprovar uma transação financeira) para o proprietário correto do sistema em busca de resolução, na mesma cadência da governança de identidade humana.
- **Gerenciamento de desvios no ambiente de execução:** aplicação programada de patches, atualização de dependências e reatestação de runtimes de agentes e configurações de plataforma subjacentes com o objetivo de eliminar a deriva de vulnerabilidades em implantações hospedadas, em nuvem e em contêineres.

4.4 Alinhamento regulatório e de conformidade

Os GSIs fazem a ponte entre a arquitetura e os regimes regulatórios específicos dentro dos quais uma empresa opera, convertendo os controles técnicos em evidências criadas para atender aos requisitos regulatórios, de conformidade e de auditoria:

- **Mapeamento de controles para estruturas regulatórias:** alinhamento dos resultados de telemetria, trilha de auditoria e revisão de acesso aos requisitos específicos do setor (por exemplo, gerenciamento de risco de modelos de serviços financeiros, tratamento de dados de saúde, proteção de infraestrutura crítica) à medida que evoluem juntamente com as diretrizes específicas de IA agêntica.
- **Pacotes de evidências prontos para auditoria:** estruturar a saída da camada de telemetria e criação de registros para que as revisões de acesso, eventos de contenção e throttling e registros de recuperação de acesso (reavaliação, aprovação da causa raiz, fechamento da trilha de auditoria) cumpram os padrões de documentação de examinadores e auditores, e não apenas às necessidades internas do SOC.
- **Orientações sobre operações transfronteiriças e residência de dados:** aconselhamento sobre como os canais do Agent Import (agentes SaaS, agentes próprios que cruzam fronteiras organizacionais) se interligam com os requisitos de residência de dados e transferência transfronteiriça específicos de cada jurisdição em que a empresa opera.
- **Proveniência de auditoria inviolável:** estruturar a coleta de telemetria para que reguladores e examinadores recebam registros de auditoria não repudiáveis gerados diretamente no limite de execução, reduzindo a dependência da reconstrução secundária de registros a partir da infraestrutura de rede circundante.

4.5 Guias práticos específicos por setor

Os requisitos de controle variam conforme o setor, exigindo que os GSIs mantenham guias práticos adaptados a diferentes ambientes operacionais:

- **Serviços financeiros:** prioriza o isolamento rigoroso da execução, limites de transação de alta segurança e controles de dados regulatórios.
- **Manufatura e indústria:** enfatiza a descoberta de OT, IoT e agentes de borda em hardware legado.
- **Varejo e comércio:** focado na defesa contra injeção de prompts B2C, proteção da marca e vazamento de PII de clientes.

Os membros fundadores atualizam esses guias práticos de forma contínua à medida que os resultados conjuntos de interoperabilidade são publicados, verificando se o modelo operacional se adapta às capacidades agênticas emergentes.

Conclusão: definindo o futuro da segurança em IA

A Blueprint Alliance regula o comportamento dos agentes em tempo de execução de produção. Integrada diretamente à infraestrutura corporativa, em vez de ser adaptada como uma camada externa, essa arquitetura fornece as salvaguardas estruturais necessárias para implantar a automação baseada em agentes de forma segura e em grande escala.

Para garantir a verdadeira interoperabilidade entre múltiplos fornecedores, os membros fundadores validam ativamente suas plataformas em relação a padrões abertos essenciais, incluindo o Model Context Protocol (MCP) para interação com ferramentas, o Open Cybersecurity Schema Framework (OCSF) para registro unificado, o Shared Signals and Events (SSF/CAEP) para troca de informações de risco em tempo real e o HTTP Message Signatures (RFC 9421) para autenticação de solicitações. Esse parâmetro de telemetria compartilhado ajuda a garantir que uma violação de limite ou um sinal de ameaça detectado em um runtime possa ser incorporado e processado nativamente em todo o plano de controle. As funcionalidades variam de acordo com o fornecedor e a implementação; integrações de referência e resultados de interoperabilidade validados serão publicados assim que os testes forem concluídos.

Com a publicação regular de resultados conjuntos de interoperabilidade e integrações de referência, a Alliance proporciona um ecossistema de múltiplos fornecedores em constante evolução, que acompanha a velocidade da inovação agêntica.

O blueprint em ação

Os cinco cenários abaixo mostram como os pilares da Seção 3 funcionam juntos na prática. Cada um é um composto e construído a partir de padrões observados nos primeiros trabalhos da Alliance, em vez de um cliente específico nomeado. Esses exemplos ilustram o comportamento pretendido da arquitetura; eles não são evidências de resultados implantados na escala dos membros da Alliance atualmente. Leia-os como uma arquitetura em movimento, não como um estudo de caso de uma única implantação.

De agente paralelo a identidade registrada. Um analista financeiro conecta um assistente de IA baseado em navegador à sua sessão corporativa de SSO para resumir contratos com fornecedores. O Shadow AI Discovery identifica o tráfego de saída não reconhecido logo em sua primeira conexão. O AI Agent Security Posture Management verifica se há vulnerabilidades conhecidas na ferramenta e a encaminha para o Agent Directory. Antes mesmo que dados sensíveis sejam acessados, o assistente é registrado como uma identidade governada com um perfil de metadados verificado e acesso com escopo definido ou bloqueado na camada de rede. Aquilo que antes permanecia invisível por meses agora é visível, verificado e decidido antes mesmo de entrar em contato com dados sensíveis.

Preservação da delegação em um fluxo de trabalho com múltiplos agentes. Um lead de operações de vendas solicita a um agente de orquestração que atualize um conjunto de registros de CRM e notifique uma lista de distribuição. O orquestrador cria um subagente para lidar com a escrita e outro para redigir a notificação. Cada salto carrega uma declaração de ator aninhada de volta para a sessão humana original para que o Agent Directory e o mecanismo de políticas de acesso possam impor restrições OBO em cada etapa. Se um subagente tentar acessar um sistema para o qual o humano original não tinha permissão, os limites principais de IAM e políticas, incluindo o Cross-App Access, o bloqueiam. Como cada subagente é executado em seu próprio contexto de execução isolado, um ataque bloqueado não pode afetar a sessão do orquestrador nem as tarefas irmãs, enquanto todo o fluxo de trabalho permanece unificado em um único rastreamento de execução.

Dimensionamento do acesso de acordo com aquilo que o agente realmente está tentando fazer. Um agente de suporte recebe acesso de leitura amplo a uma plataforma de dados do cliente. Quando ele tenta emitir um reembolso, o mecanismo de política baseado em intenção detecta a transição de consulta para transação e exige aprovação humana antes de conceder um privilégio efêmero e específico para aquele pedido. Executar o reembolso em um runtime efêmero, específico para aquela tarefa aprovada, ajuda a garantir que os privilégios elevados e o ambiente de runtime sejam desativados após a conclusão da tarefa, reduzindo permissões permanentes e superfícies de ataque persistentes. O privilégio elevado expira instantaneamente após a conclusão da tarefa.

Da anomalia à contenção automatizada. O monitoramento em runtime (geralmente um modelo tradicional de detecção de anomalias por ML em vez de um LLM) detecta um agente fazendo um volume excepcionalmente alto de chamadas para uma API interna que ele raramente utiliza; um Guardian AI Agent inline sinaliza separadamente uma assinatura de injeção de prompts em sua entrada mais recente. O Risk Signals correlaciona ambos os eventos por meio do Shared Signals Framework, ultrapassa um limite de risco definido e aciona automaticamente a revogação do token e a quarentena da rede, além da suspensão do ambiente de execução ativo do agente, preservando seu estado na memória para análise forense em vez de encerrá-lo completamente, tudo sem esperar que um humano perceba o alerta. A equipe de segurança recebe o rastreamento completo como um registro de confirmação de circuito fechado, reduzindo a contenção de uma investigação manual de várias horas para uma resposta automatizada e quase instantânea.

Transformação das revisões de acesso em um processo contínuo. Durante cada ciclo contínuo de certificação de acesso, o serviço gerenciado de um Global System Integrator extrai o estado completo do Agent Directory e das políticas de acesso por meio da camada de telemetria aberta do modelo, sinaliza os agentes com proprietários que deixaram a empresa e encaminha as exceções de separação de funções para o proprietário correto do sistema. Uma revisão manual, baseada em planilhas, se transforma em um fluxo de trabalho contínuo e pronto para auditoria.

Boas práticas para começar

Você não precisa de todas as funcionalidades deste blueprint logo no primeiro dia. Confira abaixo a ordem que as implantações dos membros da Alliance seguiram na prática.

Implemente primeiro um sistema unificado de registro e telemetria.

Não é possível depurar a fiscalização de política ou o comportamento de referência sem um único registro normalizado. Considere isso como pré-requisito em todas as etapas a seguir.

Comece pela descoberta, não pela política. Você não pode governar o que não pode ver. Implemente o Shadow AI Discovery e um inventário inicial de agentes antes de escrever uma única política de acesso. A maioria das equipes se surpreende com a quantidade de agentes que já estão em execução.

Registre antes de restringir. Reúna todos os agentes descobertos em um único Agent Directory com identidades e proprietários verificados antes de adicionar políticas de alta granularidade ou baseadas em intenção. Uma política sobre um agente não registrado é uma política sobre nada.

Faça autorização em camadas. Controles de alta granularidade (FGA) e baseados em relacionamento (ReBAC) precisam ser implantados antes que a autorização baseada em intenção possa restringir corretamente as permissões do agente.

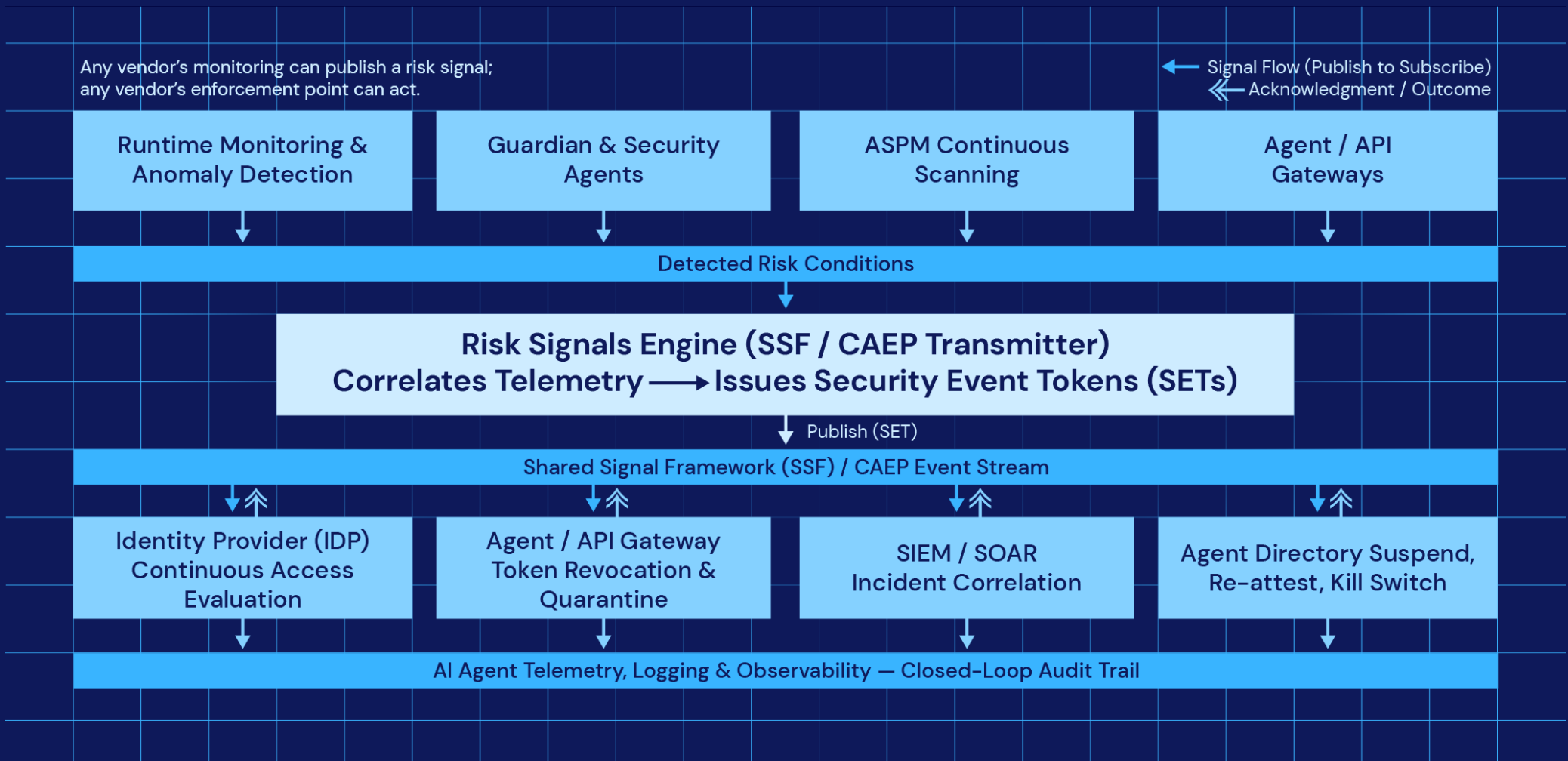
Instrumente o runtime antes de automatizar a resposta. Implante monitoramento, rastreamento e filtragem de saída/DLP de forma ampla o suficiente para estabelecer um parâmetro comportamental real antes de integrar a fiscalização automatizada ao Risk Signals. A automação da contenção contra telemetria ruidosa ou incompleta pode resultar em falsos positivos, corroendo a confiança em todo o sistema.

Faça um teste de pressão no mecanismo de desativação antes de precisar usá-lo. Valide a revogação de token, o encerramento da sessão e o isolamento de rede em um agente de não produção em intervalos regulares. Um mecanismo de desativação que você nunca testou é uma hipótese, não um controle.

Trate a governança como um processo contínuo, e não como um projeto. As revisões de acesso, solicitações de acesso e verificações de separação de funções devem ser executadas com a mesma frequência que a governança de identidades humanas, e não como uma iniciativa pontual de IA agêntica. Um Global System Integrator pode ajudar a fazer essa transição do modelo de projeto para o modelo operacional.

Este material tem caráter meramente informativo e não constitui aconselhamento jurídico, comercial, de privacidade, de segurança ou de conformidade. O conteúdo pode não refletir os desenvolvimentos mais recentes nas áreas de segurança, legislação ou regulamentação. É de sua inteira responsabilidade obter aconselhamento de seus próprios consultores jurídicos e/ou profissionais. Você não deve se basear nestes materiais como uma substituição para a consultoria profissional. Os autores não oferecem nenhuma garantia ou declaração em relação a este conteúdo nem se responsabilizam por perdas ou danos resultantes da implementação destas recomendações por você.

Fluxo de dados em runtime: como os sinais de risco se propagam (SSF / CAEDP)



Visão unificada: fluxo de dados + fluxo de sinal no plano de controle

