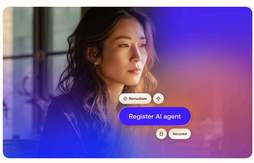
O plano para uma empresa agêntica segura

Três perguntas que todos os líderes de segurança e TI devem responder antes que os agentes saiam do controle



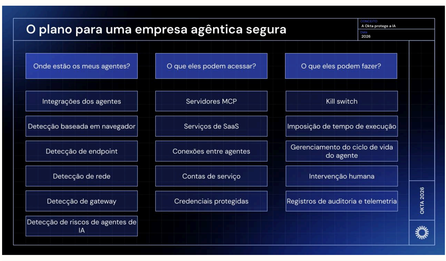
A brecha de identidade no centro da segurança de IA

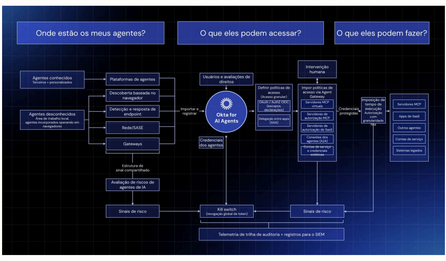
Os softwares apenas faziam as tarefas para as quais eram programados. Agora eles decidem, agem e se conectam por conta própria. Os agentes de IA já estão ajudando funcionários, atendendo clientes e operando em toda a cadeia de suprimentos. E adoção crescente desses agentes faz com que eles acabem se expandindo mais rápido do que a segurança ao redor. Na última década, graças ao uso de mecanismos como o armazenamento seguro, o princípio do menor privilégio e a autenticação contínua, as organizações reforçaram a segurança da identidade para seres humanos. Entretanto, a ascensão meteórica dos agentes de IA está criando uma nova lacuna de identidade. Qualquer pessoa pode criar um agente, sem falar que, inclusive, os agentes podem gerar mais agentes, que, por sua vez, conectam-se separadamente a apps, APIs, ferramentas de SaaS e sistemas de dados. O resultado são milhares de novas entidades com acesso privilegiado operando na velocidade das máquinas e, frequentemente, fora dos controles de segurança existentes.

É por isso que os agentes devem ser tratados como identidades de primeira classe. Uma empresa agêntica segura começa com o estabelecimento de responsabilidades bem definidas e visibilidade clara, antes que os agentes saiam do controle. As organizações devem ser capazes de responder a três perguntas:

1. Onde estão os meus agentes?
2. O que eles podem acessar?
3. O que eles podem fazer?

Essas perguntas definem o plano operacional para garantir a segurança dos agentes de IA. Para respondê-las, não basta fazer um inventário, é preciso ter os sistemas certos, controles de identidade e um modelo de governança adequado para operar com segurança uma empresa agêntica.





Pergunta 1: onde estão os meus agentes?

A visibilidade é a maior preocupação que ouvimos dos nossos clientes.

Pergunte à sua equipe quantos agentes estão no seu ambiente. A maioria não consegue fornecer um número preciso. Em geral, os agentes que os funcionários iniciam em um navegador ou que são executados silenciosamente em computadores são desconhecidos e não podem ser controlados.

Por isso, é necessário ter a capacidade de descobrir agentes em qualquer lugar em que tenham sido criados ou implantados.

- **Integrações da plataforma agêntica:** registre agentes das principais plataformas de terceiros e seus próprios agentes personalizados no seu provedor de identidade. Se não for possível vê-los na criação, não será possível governá-los.
- **Detecção baseada em navegador:** descubra agentes ocultos que operam por meio de navegadores e extensões fora do seu provedor de identidade. Esses são os agentes que os funcionários criam sem pedir permissão.
- **Detecção de endpoint:** identifica agentes em execução em dispositivos gerenciados. Isso deve se integrar a qualquer sistema que você utilize para gerenciamento de dispositivos móveis ou segurança de endpoint.
- **Detecção de rede:** identifique tráfego não autorizado entre agentes e recursos na camada de rede. Os agentes se comunicam entre eles e com os serviços, dessa forma, é necessário visualizar todas as conexões deles.
- **Detecção de gateway:** identifique e gerencie agentes de IA não registrados e clientes OAuth que interagem com sua API, MCP e gateways de agentes. Se agentes estiverem chamando suas APIs, eles devem ser autenticados e registrados.
- **Avaliação de riscos de agentes de IA:** monitore e identifique continuamente as configurações incorretas, pois elas deixam os agentes de IA vulneráveis à exploração. Ao analisar a postura de segurança da identidade de cada agente, identifique os riscos de modo proativo.

Não importa qual pilha de tecnologias está sendo usada, é necessário ingerir sinais de todas essas fontes. A camada de descoberta deve funcionar em diferentes plataformas, ferramentas e equipes. Visibilidade fragmentada é o mesmo que não ter visibilidade.

Pergunta 2: o que eles podem acessar?

Depois de visualizar um agente, é necessário mapear todos os recursos que ele pode acessar e aplicar políticas de acesso. Sem controle centralizado sobre os caminhos de conexão, um único agente comprometido pode encadear acessos em todo o ambiente na velocidade da máquina.

O raio de explosão de um agente comprometido é definido pelas conexões dele.

- **Servidores e recursos MCP:** os servidores do Model Context Protocol fornecem aos agentes acesso a ferramentas e fontes de dados, tanto internas (apps, APIs, bancos de dados, propriedade intelectual) quanto externas (MCPs de terceiros, como Slack, GitHub e Notion). O perímetro de segurança, agora, abrange todos os recursos que podem ser acessados pelo agente.
- **Aplicativos de SaaS:** os agentes se conectam às mesmas ferramentas de SaaS que seus funcionários usam diariamente. A diferença é que os agentes trabalham muito mais rápido e acessam mais dados. Um agente comprometido pode exfiltrar dados ou realizar alterações em todos os apps de SaaS conectados mais rapidamente do que um ser humano conseguiria.
- **Conexões entre agentes:** garanta a segurança do handshake e da autorização entre entidades autônomas. O movimento lateral começa quando um agente chama outro agente. É necessário verificar a identidade de ambos antes da troca de dados ou delegação de tarefas.
- **Contas de serviço:** elimine a proliferação de credenciais estáticas de longa duração. Elas funcionam como "chaves mestras", dando amplo acesso aos agentes, que as herdam dos padrões legados de máquina para máquina. Cada credencial estática é uma porta dos fundos persistente pronta para ser explorada.
- **Credenciais protegidas:** proteja e rotacione senhas automaticamente. Um token não rotacionado é uma porta aberta para invasores. As credenciais devem ser armazenadas em um cofre, emitidas de forma dinâmica e rotacionadas com frequência. Nenhum agente deve operar com uma credencial que exceda a duração de sua tarefa.

Todas essas conexões precisam ser registradas no SIEM. Você não pode proteger o que não pode ver. Cada conexão de agente (incluindo o que foi acessado, quando e com quais credenciais) deve ter seu fluxo direcionado ao centro de operações de segurança, para monitoramento e investigação.

Pergunta 3: o que eles podem fazer?

Saber onde os agentes estão e o que eles podem acessar não será suficiente se não for possível controlar e interromper o que eles realmente fazem. Quando um agente começa a exfiltrar dados ou a gerar processos não autorizados, é necessário responder rapidamente.

- **Botão de desativação:** se um agente se desviar de sua missão pretendida, acessar dados confidenciais inesperadamente ou quando uma ameaça for detectada, será necessário ter capacidade de revogar o acesso de modo instantâneo, em todos os sistemas, para conter o risco.
- **Imposição em tempo de execução:** autorize os agentes com base no que eles estão tentando realizar em tempo real. Avalie o contexto, a sequência e o volume. Uma consulta para 10 registros de clientes é diferente de uma consulta para 10 mil. Detecte ataques de injeção de prompts e implemente políticas no nível da ferramenta antes da execução das ações.
- **Gerenciamento do ciclo de vida do agente:** as permissões de agente que fazem sentido no primeiro dia raramente fazem sentido no nonagésimo dia. Revise continuamente o acesso para aplicar o menor privilégio, automatize as certificações e revogue o acesso imediatamente quando os agentes forem desativados ou os funcionários deixarem a empresa..
- **Aprovações com intervenção humana:** exija aprovação humana para ações de agentes sensíveis ou potencialmente arriscadas. Impeça operações destrutivas, acesso a dados em massa ou elevação de privilégios de agentes.
- **Registros de auditoria e telemetria:** todas as ações do agente precisam ser registradas e enviadas ao seu SIEM. Cada chamada de ferramenta, cada decisão sobre autorização, cada tentativa de acesso. A aplicação de medidas em tempo de execução e os mecanismos de desativação só serão eficientes se houver visibilidade completa.

O plano é a referência

As três perguntas — onde estão os meus agentes, a que eles podem se conectar, o que eles podem fazer — não são aspiracionais. São o padrão mínimo de que você precisa para operar agentes de IA em produção.

As organizações que não conseguem responder a essas perguntas estão operando sem visão. E, quando esse questionamento partirem do conselho ou dos auditores, ou do órgão regulador quando ocorrer uma violação, "não sabemos" não é uma resposta aceitável.

Os indivíduos mais ágeis já perceberam isso. As empresas líderes não esperam o primeiro incidente para se colocarem na obrigação de agir. Elas estão tratando agentes como identidades de primeira classe agora. Estão integrando descoberta, aplicação de políticas e governança em suas implantações de agentes desde o primeiro dia. Estão respondendo às três perguntas antes que os agentes saiam do controle.

O plano não é uma auditoria única. À medida que a base de agentes cresce de dezenas para milhares, responder a essas três perguntas é uma disciplina operacional contínua. Você passou uma década construindo segurança de identidade para humanos. Não deixe que os agentes desfaçam isso.

A Okta construiu esse modelo com base no trabalho contínuo com empresas líderes em segurança de agentes de IA em grande escala. Saiba como a Okta Platform o implementa em okta.com/ai-agents.