knowbe4

# A segurança da força de trabalho híbrida: protegendo pessoas e agentes de IA na nova era

WWW.KNOWBE4.COM/PT

KnowBe4 Brazil | R. Gomes de Carvalho, 911 | Sala 208 - Vila Olímpia | CEP: 04547-003 |
São Paulo-SP | Tel.: (0800) 761 2668 | E-mail: Sales@KnowBe4.com

CR-0663 0326

# Introdução

A força de trabalho mudou.

Os funcionários não trabalham mais sozinhos. Cada vez mais, eles trabalham com copilotos, assistentes e agentes autônomos de IA que ajudam a escrever código, resumir incidentes, redigir comunicações, analisar dados e esclarecer dúvidas dos clientes. O que começou como ferramentas simples de produtividade evoluiu rapidamente para uma nova classe de colegas digitais integrados aos fluxos de trabalho corporativos.

Essa mudança está avançando rapidamente. A Goldman Sachs estima que a IA agêntica poderá representar cerca de 60% do valor do mercado de software até 2030, o que indica uma mudança fundamental na forma como o trabalho é realizado. A força de trabalho moderna não é mais composta apenas por pessoas: agora, ela é formada por uma equipe híbrida de pessoas e agentes de IA que trabalham juntos para impulsionar a produtividade e a tomada de decisões.

Mas essa transformação traz um novo desafio de segurança que muitas organizações ainda não estão preparadas para enfrentar.

Durante muitas décadas, as equipes de segurança cibernética tiveram como foco a proteção de sistemas, redes e pontos de extremidade, além da redução dos riscos gerados por falhas humanas por meio de conscientização em segurança, do fortalecimento da cultura e de treinamentos. Esses esforços tinham como objetivo a defesa contra ataques de phishing, engenharia social e outras formas de manipulação cognitiva.

Hoje, essa superfície de risco aumentou. As pessoas já não são os únicos agentes que interagem com dados sigilosos, geram conteúdo ou tomam decisões dentro das empresas. Os agentes de IA agora executam muitas dessas mesmas tarefas, geralmente com rapidez e confiança, mas sem entender totalmente as políticas, o contexto ou a tolerância a riscos da organização.

Ao contrário dos funcionários, não é possível aprimorar o trabalho dos agentes de IA por meio de treinamentos tradicionais ou programas de conscientização. As organizações não conseguem inspecionar ou controlar totalmente os prompts que moldam as respostas, embora esses agentes já estejam influenciando os resultados dos negócios.

Isso cria um desafio já conhecido, mas em um novo formato. As equipes de segurança devem gerenciar a forma como as pessoas interagem com as tecnologias que usam. Na era da IA, o risco está cada vez mais concentrado na camada de interação entre pessoas e sistemas inteligentes.

Pense em um agente de IA como um estagiário extremamente capacitado, mas altamente influenciável. Ele processa grandes volumes de informação e executa instruções com rapidez e confiança. O agente de IA está sempre disposto a ajudar, mas não tem discernimento, contexto ou uma compreensão inerente dos riscos organizacionais.

Sem as proteções adequadas, até mesmo o assistente mais capacitado pode agir de maneira imprevisível.

# A superfície de ataque emergente entre pessoas e a IA

Os agentes de IA são extremamente eficientes e inerentemente vulneráveis. Os invasores já estão aprendendo a explorar a dinâmica entre as pessoas e a IA. Em muitos casos, o ataque não tem como alvo apenas a pessoa ou a IA. O alvo é a relação de confiança existente entre elas.

A inclusão da IA nos fluxos de trabalho do dia a dia criou uma nova categoria de riscos de segurança. Os agentes maliciosos estão usando a IA para ampliar a escala de ataques já conhecidos, gerando e-mails de phishing mais convincentes, produzindo conteúdos de deepfake para engenharia social e automatizando atividades de exploração. Ao mesmo tempo, eles estão desenvolvendo técnicas projetadas especificamente para manipular sistemas de IA:

- **Ataques de injeção de prompt**
  Entradas maliciosas criadas para influenciar o comportamento de um agente de IA, que podem levá-lo a contornar controles, revelar dados sigilosos ou executar ações imprevisíveis.
- **Falsificação de identidade de agentes de IA**
  Agentes maliciosos que imitam ferramentas corporativas legítimas para coletar credenciais, informações sigilosas ou dados de fluxos de trabalho de funcionários desavisados.
- **Engenharia social entre pessoas e a IA**
  Ataques que exploram a confiança em conteúdos gerados por IA, que podem transformar agentes comprometidos em ameaças internas.

Em muitos casos, não há implantação de malware nem roubo de credenciais. O funcionário simplesmente dá uma instrução, e o agente de IA executa conforme ele foi projetado. Tanto as pessoas quanto a IA se comportam como esperado, e é exatamente aí que reside a vulnerabilidade.

# O problema da confiança

As pessoas confiam instintivamente nas ferramentas que as ajudam a alcançar bons resultados. Quando sistemas de IA produzem resumos claros, recomendações úteis ou conteúdos bem elaborados, os usuários passam a confiar rapidamente em seus resultados. Com o passar do tempo, isso pode gerar o viés de automação: a tendência de confiar em decisões automatizadas mesmo quando elas podem estar incorretas ou ter sido manipuladas.

Os invasores podem explorar essa confiança. Um prompt malicioso oculto em um documento, em um conjunto de dados ou em uma página da web pode influenciar a resposta gerada por uma IA, apresentando informações manipuladas em um tom confiante e convincente. Para o funcionário, o resultado parece legítimo, mesmo tendo sido formulado por um invasor.

As arquiteturas de segurança tradicionais foram criadas para proteger sistemas, redes e pontos de extremidade. Embora esses controles continuem sendo essenciais, eles não lidam totalmente com os riscos criados pela colaboração entre as pessoas e a IA. Hoje em dia, a lacuna mais significativa está na camada de interação, quando os funcionários usam a IA para interpretar informações, gerar conteúdo e obter ajuda na tomada de decisões.

Proteger esse ambiente exige garantir a segurança de ambos os lados da força de trabalho moderna. As organizações precisam capacitar os funcionários para que eles reconheçam tentativas de manipulação e usem ferramentas de IA com segurança, além de garantir que os agentes de IA sejam resistentes a ataques de injeção de prompt, à exposição de dados e a ações não autorizadas.

# O futuro da segurança está na defesa em duas frentes

A fronteira entre as pessoas e a IA na segurança cibernética está desaparecendo. À medida que agentes de IA se integram aos fluxos de trabalho cotidianos, as organizações que adaptarem suas estratégias de segurança a essa nova realidade conseguirão preservar sua resiliência. A mensagem é clara: a segurança cibernética não se resume mais apenas a proteger os sistemas contra as pessoas ou as pessoas contra os sistemas. Trata-se de proteger a interação entre eles, pois é nessa interação que residem tanto a maior vulnerabilidade quanto a defesa mais forte.

Para enfrentar esse desafio, as organizações devem adotar uma estratégia de defesa em duas frentes que fortaleça o elemento humano e os próprios sistemas de IA.

## 1. Fortalecer a resiliência humana

Em um local de trabalho impulsionado por IA, os funcionários deixam de atuar apenas na execução de tarefas e passam a exercer funções de supervisão e tomada de decisões. Os treinamentos de conscientização em segurança precisam evoluir de acordo com essa mudança. Ensinar os usuários a identificar um e-mail de phishing já não é o suficiente. Os funcionários precisam desenvolver uma conscientização digital, com uma combinação de ceticismo saudável, de percepção contextual e da capacidade de avaliar, de maneira crítica, os resultados gerados pela IA.

O primeiro passo é o treinamento. Os funcionários precisam entender os recursos e as limitações dos agentes de IA, inclusive como eles podem ser manipulados. Eles devem ser treinados para reconhecer comportamentos anômalos ou inesperados da IA, como resultados que entrem em conflito com políticas ou com o contexto. Além disso, as organizações devem estabelecer protocolos de verificação simples, mas eficazes, para lidar com ações de alto risco, principalmente aquelas iniciadas ou auxiliadas por IA. Quando algo parecer urgente, incomum ou fora do escopo, os funcionários devem saber como parar, validar a situação e encaminhar o caso.

## 2. Aprimorar o agente de IA

Ao mesmo tempo, os sistemas de IA devem ser tratados como uma nova classe de ativos corporativos que exige governança, monitoramento e controle. Isso começa com a implementação de um processo robusto de validação de entrada para impedir que prompts maliciosos ou dados corrompidos influenciem o comportamento dos agentes. Os resultados gerados pela IA também devem ser analisados continuamente para detectar anomalias, violações de políticas ou sinais de manipulação.

Da mesma forma, é importante definir limites claros de funções para os agentes de IA. Os sistemas devem ser projetados para operar dentro de escopos específicos e recusar solicitações que estejam fora das tarefas autorizadas. Isso deve ser reforçado por políticas de uso de IA bem definidas que controlem como os agentes interagem com dados, sistemas e usuários.

# Recursos essenciais

Gerenciar riscos nesse ambiente híbrido exige visibilidade sobre como os agentes de IA operam, como interagem com os usuários e como acessam dados e sistemas. Para oferecer cobertura completa, desde a descoberta até a defesa, produtos eficazes de segurança para agentes de IA devem ser estruturados em quatro pilares principais:

### Descoberta
Identificar e catalogar automaticamente todos os agentes de IA em operação no ambiente, inclusive aqueles introduzidos por meio de plataformas autorizadas e aqueles implantados sem aprovação formal. A descoberta contínua ajuda a eliminar pontos cegos e garante que as equipes de segurança compreendam o escopo completo do uso da IA.

### Monitoramento
Registrar todas as execuções de agentes, as interações com os prompts e as chamadas de ferramentas em uma trilha de auditoria abrangente e fácil de ser pesquisada. A telemetria detalhada, que inclui metadados de assistentes digitais de IA, como o Microsoft Copilot, permite que as equipes entendam como os agentes estão sendo usados e onde possíveis riscos podem surgir.

### Detecção
Aplicar recursos avançados de detecção para identificar comportamentos de risco em tempo real. Isso inclui identificar tentativas de injeção de prompt, elevação de privilégios, exposição de dados sigilosos e atividades anômalas de agentes em diferentes idiomas e casos de uso. A análise comportamental ajuda a detectar ameaças conhecidas e emergentes que podem passar despercebidas pelos controles tradicionais.

### Proteção
Permitir que as organizações definam e apliquem a postura de segurança desejada. As equipes podem optar por monitorar atividades de maneira passiva, gerar alertas ou interceptar e bloquear ativamente as operações de alto risco em tempo real. A aplicação flexível permite que as equipes de segurança equilibrem produtividade e proteção à medida que a adoção da IA evoluir.

Outro ponto importante é que a escalabilidade é essencial para os profissionais de TI e de segurança. Gerenciar riscos em ambientes com um número crescente de usuários humanos e agentes de IA exige visibilidade centralizada e eficiência operacional. Os principais recursos devem incluir:

### Visibilidade e auditoria
Painéis centralizados oferecem uma visão clara das interações entre pessoas e agentes e entre agentes e sistemas. Logs de auditoria detalhados permitem que as equipes de segurança detectem comportamentos anômalos, investiguem incidentes e identifiquem atividades que estejam fora das políticas definidas ou dos casos de uso previstos.

### Aplicação dinâmica de políticas
As equipes de segurança podem definir políticas com base em riscos por meio de controles orientados por API, possibilitando uma governança consistente do uso de IA nos ambientes. As políticas poderão ser ajustadas à medida que novos casos de uso surgirem, garantindo que a adoção da IA permaneça alinhada aos requisitos organizacionais de segurança e conformidade.

### Nível de risco integrado
Os dados das interações enriquecem uma estrutura mais ampla de gerenciamento de riscos gerados por falhas humanas, permitindo que as organizações avaliem os riscos nos comportamentos das pessoas e dos agentes de IA. Ao incorporar interações entre pessoas e agentes aos modelos de nível de risco, os líderes de segurança obterão uma visão mais completa da exposição e poderão priorizar os esforços de mitigação com mais eficácia.

# Conclusão

As pessoas e os agentes de IA agora trabalham em conjunto. As organizações precisam de uma abordagem unificada para gerenciar esses riscos. Ao estender os princípios existentes de gerenciamento de riscos gerados por falhas humanas para incluir agentes de IA, as equipes de segurança podem promover a inovação, manter o controle e reduzir a exposição sem desacelerar o progresso.

Saiba mais

### Teste de phishing gratuito

Faça este teste de phishing gratuito e descubra qual é a Phish-prone Percentage dos seus funcionários

### Automated Security Awareness Program gratuito

Crie um programa de conscientização em segurança personalizado para sua organização

### Phish Alert Button gratuito

Agora, seus funcionários podem denunciar ataques de phishing de maneira segura com apenas um clique

### Email Exposure Check gratuito

Descubra quais e-mails de usuários estão expostos antes que os agentes mal-intencionados façam isso

### Domain Spoof Test gratuito

Descubra se os hackers conseguem falsificar um endereço de e-mail do seu domínio

## Sobre a KnowBe4

A KnowBe4 capacita a força de trabalho moderna a tomar decisões de segurança mais inteligentes todos os dias. Escolhida por mais de 70 mil organizações no mundo todo, a KnowBe4 é pioneira em segurança da força de trabalho digital, protegendo tanto agentes de IA quanto pessoas. A Plataforma KnowBe4 oferece simulação de ataques e treinamento, segurança de colaboração e proteção de agentes com base na tecnologia AIDA (Artificial Intelligence Defense Agents) e em um nível de risco proprietário. A plataforma conta com 15 anos de dados comportamentais para combater ameaças avançadas, incluindo engenharia social, injeção de prompts e IA oculta. Ao proteger pessoas e agentes, a KnowBe4 é líder do setor em confiança e defesa da força de trabalho.

Obtenha mais informações em knowbe4.com/pt. Siga a KnowBe4 no LinkedIn e no X.

Follow us on Linkedin

Follow us on X

knowbe4

KnowBe4 Brazil | R. Gomes de Carvalho, 911 | Sala 208 - Vila Olímpia | CEP: 04547-003 | São Paulo-SP
Tel.: (0800) 761 2668 | www.KnowBe4.com/pt | Sales@KnowBe4.com