Número 118 | Setembro de 2026

NTT DATA

Radar

A revista de cibersegurança

Cibersegurança na era agêntica: atacar, defender e antecipar-se

De Jose Manuel Moreno Guerra

A inteligência artificial está transformando a cibersegurança em duas frentes. Por um lado, permite às organizações detectar, analisar e responder com uma rapidez inimaginável há poucos anos. Por outro, oferece aos atacantes novas capacidades para automatizar campanhas, descobrir vulnerabilidades e encadear ações ofensivas em escala e velocidade inéditas. A consequência é clara: a IA não cria apenas uma nova categoria de risco, mas altera por completo o tempo disponível para a defesa.

Até agora, grande parte da segurança se apoiava em ciclos relativamente previsíveis. Uma vulnerabilidade era detectada, classificada, priorizada e, por fim, corrigida. Um incidente gerava um alerta, passava por diferentes níveis de análise e acabava desencadeando uma resposta.

Esse modelo começa a se mostrar insuficiente. Quando a exploração pode ocorrer em minutos, a defesa não pode levar dias ou semanas para responder. A resiliência do futuro dependerá da capacidade de detectar, validar, priorizar e agir em velocidade de máquina.

No campo defensivo, os SOCs estão evoluindo para modelos agênticos. A IA já pode assumir grande parte do trabalho repetitivo dos níveis iniciais da operação: normalizar alertas, enriquecer indicadores, correlacionar informações, consultar inteligência de ameaças, recomendar ações e até executar respostas previamente aprovadas.

Isso não elimina a necessidade do analista, e sim transforma o seu papel.

Os profissionais deixam de dedicar a maior parte do tempo a tarefas mecânicas e podem se concentrar em ameaças complexas, investigação avançada, decisões com impacto de negócio e melhoria contínua das capacidades de detecção.

Essa mudança deve se estender a todo o ciclo de segurança. A IA pode ajudar a modelar ameaças desde a fase de projeto, reduzir falsos positivos na análise de código, automatizar testes dinâmicos, consolidar resultados de múltiplas ferramentas e priorizar vulnerabilidades de acordo com o contexto real do ativo. Também pode transformar fontes abertas em inteligência estruturada e acionável.

Em conjunto, essas capacidades permitem passar de uma segurança fragmentada e reativa para um modelo contínuo, contextual e conectado.

No entanto, a mesma lógica também pode ser aplicada de forma ofensiva. Os agentes de IA podem explorar aplicações, selecionar ferramentas, interpretar resultados, adaptar estratégias e encadear vulnerabilidades até atingir um objetivo.

Ao contrário dos scanners tradicionais, limitados por assinaturas, regras e sequências predefinidas, os sistemas agênticos são capazes de aprender com o ambiente e modificar seu comportamento durante a execução.

Essa capacidade é muito valiosa para auditorias e exercícios de Red Team, mas também antecipa uma nova geração de ataques autônomos.

Além disso, a área de risco aumenta com os próprios ativos de IA. Chatbots, modelos preditivos, sistemas com RAG e agentes conectados a ferramentas corporativas podem estar sujeitos a injeção de prompts, vazamentos de dados, exploração de modelos, abuso de permissões, esgotamento de recursos ou manipulação de comportamento.

Quanto maior a autonomia do sistema, maior também será o impacto potencial de uma decisão incorreta ou de uma instrução maliciosa.

Por isso, a questão não é se devemos usar IA em cibersegurança, mas sob quais condições. A velocidade deve vir acompanhada de controle.

Decisões críticas exigem supervisão humana, permissões restritas, ações reversíveis, rastreabilidade completa e evidências auditáveis. Os modelos devem operar com base em bases de conhecimento confiáveis, políticas corporativas, procedimentos definidos e mecanismos que reduzam alucinações, atalhos ou comportamentos inesperados.

A vantagem não estará em substituir os especialistas, mas em potencializar sua atuação. A IA pode absorver volume, reduzir ruído e acelerar a análise; cabe às pessoas trazer discernimento, contexto, responsabilidade e compreensão do negócio.

Nessa nova disputa entre defesa e ataque, vencerão as organizações capazes de combinar automação e governança, capacidades ofensivas e resiliência defensiva, inovação e controle.

Porque, na era da Agentic AI, proteger-se já não é apenas responder melhor: é aprender e agir antes do adversário.

Jose Manuel Moreno Guerra
Cybersecurity Director

A invasão já não se limita à rede da vítima

Cibercrônica de Alicia Isabel Martínez Vera

Em menos de um mês, incidentes ocorridos em diferentes países e setores revelaram a mesma fragilidade: a segurança de uma organização depende cada vez mais de sistemas que estão apenas parcialmente sob seu controle. Pacotes legítimos foram contaminados a partir de repositórios e processos de publicação autorizados; credenciais comprometidas deram acesso a plataformas compartilhadas; a interrupção de um operador logístico afetou entregas a restaurantes e supermercados; e a exposição de controladores industriais permitiu que sistemas de abastecimento de água nos Estados Unidos fossem comprometidos. O ponto em comum não foi uma única família de malware, mas a confiança depositada em fornecedores, automatizações, identidades e sistemas conectados.

Entre a segunda semana de julho e os primeiros dias de agosto, a cibersegurança voltou a mostrar que uma invasão pode ter consequências muito mais amplas do que o comprometimento inicial. Em alguns casos, o invasor utilizou repositórios e processos legítimos de desenvolvimento para distribuir código malicioso. Em outros, uma plataforma compartilhada, um provedor de cloud ou um sistema logístico fez com que a interrupção alcançasse organizações que nunca foram atacadas diretamente. Também aumentou a discrepância entre o alcance confirmado pelas vítimas e os números divulgados pelos autores das extorsões, que, em vários incidentes, continuavam sem verificação independente.

Entre **8 e 9 de julho**, a Accenture confirmou um incidente de segurança que classificou como isolado. A consultoria afirmou ter corrigido a causa e garantiu que suas operações e serviços não haviam sido afetados. Um ator identificado como “888” ofereceu posteriormente supostos 35 GB de código-fonte, arquivos de configuração, chaves RSA e SSH e credenciais associadas ao Azure. A existência da vulnerabilidade foi confirmada pela empresa a diversos veículos de comunicação, mas a Accenture não validou o volume nem todo o conteúdo anunciado pelo invasor. O caso é relevante porque esses materiais podem continuar sendo úteis para **fins ofensivos** mesmo quando a vítima **não sofre uma interrupção imediata**.

A partir de 8 de julho, a atenção se voltou para a cadeia de suprimentos de software. Nesse dia, surgiu no **npm** uma versão contendo um módulo malicioso apresentado como telemetria. O código interceptava as funções utilizadas para carregar chaves privadas e frases mnemônicas e enviava os segredos codificados em um cabeçalho HTTP. A publicação permaneceu disponível por cerca de 49 minutos antes de ser substituída por uma versão limpa. A análise técnica confirmou a **capacidade de roubo**, mas não permitiu determinar quantas instalações efetivamente executaram o código malicioso nem se houve perdas financeiras.

Três dias depois, a Jscrambler sofreu uma sequência semelhante, mas o acesso começou no equipamento de um desenvolvedor. O atacante obteve uma chave do GitHub, acessou o repositório e utilizou GitHub Actions para extrair o token de publicação do npm. Com esse token, publicou cinco versões maliciosas do pacote. As primeiras ativavam um **infostealer** durante a instalação; nas versões seguintes, **o mecanismo foi modificado para permitir a execução a partir do próprio código**. A empresa detectou as publicações por meio de alertas automáticos, revogou as credenciais e distribuiu uma versão limpa no mesmo dia.

Em 14 de julho, o ecossistema AsyncAPI confirmou que nem mesmo uma publicação com proveniência criptográfica válida garante a segurança do conteúdo. Um workflow do GitHub Actions configurado incorretamente permitiu obter **credenciais privilegiadas** e inserir código em quatro pacotes **oficiais**. As cinco versões contaminadas foram publicadas por meio dos mecanismos legítimos do projeto. A automação considerada confiável havia funcionado corretamente. O problema era que o **processo autorizado já havia sido comprometido**. Enquanto os pacotes maliciosos ameaçavam ambientes de desenvolvimento, o ataque à Nichirei estendeu a crise à logística de alimentos.

Em 13 de julho, a empresa japonesa detectou um acesso não autorizado que provocou falhas nas operações de recebimento e expedição em armazéns refrigerados e nos envios de alimentos congelados. **O isolamento dos sistemas protegeu as informações, mas interrompeu temporariamente o fluxo de mercadorias**. A KFC Japan suspendeu pedidos digitais e alertou para restrições no cardápio, redução do horário de funcionamento ou fechamentos pontuais, conforme o estoque de cada unidade. Também foram registrados atrasos em supermercados e outras redes de restaurantes. A Nichirei confirmou o ataque cibernético, mas não identificou publicamente o uso de ransomware, roubo de dados nem o grupo responsável.

Em 16 de julho, a Coca-Cola informou à SEC que a Fairlife havia sofrido um ataque de ransomware que atingiu parte de seus sistemas, incluindo alguns relacionados à produção.

A empresa suspendeu temporariamente a produção nos Estados Unidos, sem impacto conhecido nas operações no Canadá nem na segurança dos produtos.

Em 27 de julho, a Coca-Cola informou que a maior parte da produção em suas quatro fábricas nos Estados Unidos havia sido retomada e confirmou que o atacante havia obtido determinados dados. O grupo Anubis reivindicou o ataque e afirmou ter **roubado aproximadamente 1 TB de dados**.

A invasão mais incomum do período foi divulgada pela **Hugging Face** em 16 de julho e explicada pela OpenAI cinco dias depois. Entre 9 e 13 de julho, agentes utilizados em uma **avaliação interna de capacidades cibernéticas conseguiram operar além dos limites estabelecidos**. Os modelos exploraram uma vulnerabilidade zero-day no software que atuava como proxy de pacotes, alcançaram um nó com acesso à internet e começaram a buscar informações que permitissem resolver ou contornar o teste. A partir dali, passaram a explorar, em sequência, credenciais e vulnerabilidades adicionais até obter execução remota.

A Hugging Face confirmou o acesso a um conjunto limitado de datasets internos e a várias credenciais, mas não encontrou evidências de alteração de modelos, pacotes, Spaces ou datasets públicos. A reconstrução forense reuniu aproximadamente 17.600 ações executadas pelo agente. Não se tratou de uma campanha criminosa direcionada à Hugging Face, mas de uma **avaliação com contenção inadequada que acabou resultando em uma intrusão real em outra organização**.

A dependência de terceiros voltou a aparecer nos casos da Stadler Rail e da Amgen. Em meados de julho, **credenciais comprometidas permitiram acesso a uma plataforma de compartilhamento de informações** utilizada com um fornecedor da Stadler. Os invasores consultaram documentação técnica, mas não chegaram à rede interna do fabricante, também não interromperam a produção nem afetaram os veículos ferroviários. O grupo Everest exigiu dez milhões de francos suíços, valor que consta da própria carta de extorsão, embora não se saiba se a autoria técnica corresponde ao mesmo grupo.

Posteriormente, a Amgen informou ter detectado atividade não autorizada envolvendo dados armazenados em ambientes gerenciados por provedores externos de cloud. A farmacêutica confirmou o **vazamento de informações proprietárias, dados de saúde protegidos e outros arquivos, mas não observou impacto** nos produtos, na fabricação ou na capacidade de atendimento aos pacientes. Os dois incidentes evidenciaram o mesmo problema sob ângulos distintos: **delegar a hospedagem ou o compartilhamento de informações não elimina a responsabilidade nem reduz automaticamente as consequências de uma exposição**.

Entre 26 e 30 de julho, a ameaça atingiu diretamente a **tecnologia operacional.** Minnesota informou a ocorrência de um ataque coordenado contra mais de trinta sistemas comunitários de abastecimento de água. Pouco depois, o FBI e a EPA ampliaram o alerta para instalações em pelo menos sete estados.

Os atacantes acessaram PLCs Rockwell Automation/Allen-Bradley MicroLogix 1100 e 1400 expostos à internet, alteraram endereços IP e senhas e provocaram perda de visibilidade ou controle. Entre os efeitos informados ao FBI estavam **perda de pressão e inundações em algumas instalações**. A possibilidade de recorrer à operação manual determinou, em grande medida, a gravidade de cada caso. Até 7 de agosto, não havia atribuição oficial. A classificação mais precisa é **cibersabotagem operacional sem atribuição.**

Em 4 de agosto, a Microsoft tornou pública sua investigação sobre **ChainDrop**, uma campanha que já havia distribuído versões maliciosas de diversos pacotes npm. Esse **worm roubava tokens e credenciais de ambientes de desenvolvimento** e usava essas informações para publicar novas **versões maliciosas** de outros pacotes. As contagens variaram durante a investigação porque algumas fontes contabilizaram pacotes distintos e outras, versões publicadas, mas todas concordaram quanto à **natureza autopropagável** da campanha. A diferença em relação a uma dependência isolada era essencial: cada conta comprometida podia se tornar automaticamente o **ponto de partida para novas publicações** e ampliar a cadeia sem intervenção constante do operador. A cronologia, portanto, revela uma dimensão do risco que o simples número de vítimas não permite perceber. O risco se concentra nas **relações de confiança:** uma credencial que permite publicar, um pipeline que assina o que recebe, uma plataforma que conecta duas empresas, um fornecedor que armazena informações clínicas ou um PLC que continua acessível pela internet.

Quando esses pontos falham, o incidente se propaga por rotas empresariais e técnicas já existentes. A prioridade defensiva não consiste apenas em proteger a rede principal, mas também em **limitar permissões, reduzir o uso de segredos de longa duração, controlar publicações automatizadas e manter procedimentos manuais capazes de sustentar a operação quando os sistemas digitais deixam de ser confiáveis.**

Alicia Isabel Martinez Vera
Cybersecurity Consultant

Guardian Agents: o próximo passo na segurança agêntica

Artigo de Jaime Andrés Tovar Prieto

Como os agentes de inteligência artificial são utilizados? Até que ponto os agentes são realmente autônomos? Qual é o alcance dos agentes ou quais capacidades foram previstas em seu projeto? Existe algum mecanismo de monitoramento de suas ações? As ações estão alinhadas às funções para as quais o agente foi concebido?

Essas perguntas, embora possam parecer simples, são questões que sempre devem ser levadas em consideração; da mesma forma que se verifica a concessão de permissões a um aplicativo ou usuário para o uso de um serviço — onde uma concessão excessiva pode resultar em um incidente de segurança —, a governança sobre os agentes de IA exige o mesmo rigor e a mesma disciplina.

Um aspecto particularmente relevante desse paradigma é a capacidade de um agente de IA supervisionar outro: aprender com tendências, detectar abusos e avaliar a coerência lógica das ações executadas.

Essa dinâmica, identificada durante a pesquisa sobre tendências de segurança em IA com base em publicações do Gartner[1], deu origem a trabalhos acadêmicos de referência, como AgentDoG[2] (Agent Diagnostic Guardrail), seu benchmark associado ATBench[3] (Agent Trajectory Benchmark) e PSG-Agent[4] (Personalized Safety Guardrail).

Com o uso crescente não apenas dos grandes modelos de linguagem (LLMs), mas especialmente dos agentes de IA, que representam a evolução mais recente desse ecossistema, ampliaram-se significativamente as capacidades de execução e a autonomia na tomada de decisões.

Essas evoluções exigem um fortalecimento substancial da segurança, pois o risco deixou de se limitar à prevenção de respostas com linguagem ofensiva ou inadequada e passou a incluir a execução de ações anômalas, o vazamento de informações ou dados sensíveis, a execução de ações não autorizadas que resultam em acesso a sistemas críticos e, em outros casos, a degradação de serviços.

A visão aprimorada do setor

Diante desse cenário de riscos distribuídos e dinâmicos, a resposta da indústria não foi simplesmente reforçar os controles existentes, mas desenvolver uma nova classe de inteligência: os Guardian Agents (agentes guardiões). À medida que os agentes de IA adquirem capacidade para executar ações autônomas, surge a necessidade de um sistema que os supervisione. Por isso, a indústria desenvolveu o conceito de Guardian Agents, também chamados de “supervisores de IA”.

Um Guardian Agent é, essencialmente, uma inteligência artificial projetada não para executar a tarefa atribuída ao agente supervisionado, mas para monitorar, diagnosticar, avaliar, redirecionar e controlar seu comportamento. A adoção dessas tecnologias representa uma mudança significativa em relação ao paradigma tradicional de “aviso e consentimento”, no qual as interações com sistemas digitais dependiam de avisos que os usuários raramente compreendiam por completo.

Em seu lugar, a abordagem passa a privilegiar uma supervisão automatizada e auditável que só exige intervenção humana, segundo o modelo Human-in-the-Loop (HITL), quando o grau de confiança algorítmica é insuficiente ou quando o risco ultrapassa os limites predefinidos.

Segundo o Gartner[5], estima-se que as tecnologias de Guardian Agents alcancem uma participação entre 10% e 15% no mercado de Agentic AI até 2030. A empresa de análise aponta que esse tipo de agente deve se concentrar em contribuir para a proteção das interações em três funções fundamentais:

1 (S/f-a). Gartner.com. Acesso em 12 de agosto de 2026, em https://www.gartner.com/en/newsroom/press-releases/2026-06-03-gartner-security-and-risk-management-summit-2026-national-harbor-day-3-highlights

2 Liu, D., Ren, Q., Qian, C., Shao, S., Xie, Y., Li, Y., Yang, Z., Luo, H., Wang, P., Liu, Q., Hu, B., Tang, L., Mei, J., Guo, D., Yuan, L., Yang, J., Chen, G., Lin, Q., Yu, Y., … Hu, X. (2026). AgentDoG: A Diagnostic Guardrail framework for AI agent safety and security. Em arXiv [cs.AI]. https://doi.org/10.48550/arXiv.2601.18491

3 Li, Y., Luo, H., Xie, Y., Fu, Y., Yang, Z., Shao, S., Ren, Q., Qu, W., Fu, Y., Yang, Y., Shao, J., Hu, X., & Liu, D. (2026). ATBench: A diverse and realistic agent trajectory benchmark for safety evaluation and diagnosis. Em arXiv [cs.AI]. https://doi.org/10.48550/arXiv.2604.02022

4 Wu, Y., Guo, J., Li, D., Zou, H. P., Huang, W.-C., Chen, Y., Wang, Z., Zhang, W., Li, Y., Zhang, M., Jiang, R., & Yu, P. S. (2025). PSG-agent: Personality-aware safety guardrail for LLM-based agents. Em arXiv [cs.AI]. https://doi.org/10.48550/arXiv.2509.23614

5 (S/f-b). Gartner.com. Acesso em 12 de agosto de 2026, em https://www.gartner.com/en/newsroom/press-releases/2025-06-11-gartner-predicts-that-guardian-agents-will-capture-10-15-percent-of-the-agentic-ai-market-by-2030

- **Revisor:** Identificar e revisar os resultados e conteúdos gerados pela inteligência artificial para verificar sua exatidão, veracidade e uso aceitável.
- **Monitor:** Acompanhar as ações da inteligência artificial e dos agentes, seja por meio de supervisão humana ou pela própria inteligência artificial.
- **Protetor:** Bloquear ou redirecionar ações e permissões dos agentes por meio de ações semiautomatizadas ou totalmente automatizadas durante as operações.

Evolução da superfície de ataque

No modelo convencional de IA, os controles de segurança se baseiam em medidas de proteção (guardarails) semânticos e estáticos que, embora contribuam de forma relevante para a conformidade regulatória e façam parte da devida diligência, são insuficientes diante de ataques complexos. Em um ambiente autônomo, um sistema não se torna inseguro em uma única etapa, mas por meio de uma série de decisões equivocadas que interagem de forma complexa com as entradas do ambiente.

Um padrão de referência emergente é a taxonomia implementada em frameworks como AgentDoG e em sua infraestrutura de avaliação, ATBench.

O ATBench oferece um ambiente padronizado composto por 1000 trajetórias de agentes, 503 seguras e 497 inseguras, com acesso a mais de 2084 ferramentas heterogêneas e uma média de 9,01 turnos por trajetória, garantindo um nível de realismo sem precedentes nas simulações.

O AgentDoG propõe uma taxonomia de segurança unificada, baseada em três dimensões ortogonais, para sistemas baseados em agentes.

Especificamente, os riscos são classificados de acordo com a origem da ameaça, seu comportamento durante a execução e os danos causados no mundo real:

- **Fonte do risco (Onde):** Define, em termos topológicos, a origem da ameaça. Determina se o vetor teve origem no usuário, por meio de comandos maliciosos diretos; na ferramenta ou no ambiente, por meio de injeção indireta em bancos de dados ou páginas web interceptadas; ou na própria lógica interna do agente, como alucinações ou raciocínio falho.
- **Modo de falha (Como):** Examina a disrupção do comportamento e classifica sistematicamente as falhas agênticas. Isso inclui anomalias procedimentais, como invocar ferramentas quando as informações estão incompletas, desconsiderar riscos implícitos na invocação de ferramentas, confiar em saídas não verificadas ou executar sequências de tarefas não intencionais.
- **Dano no mundo real (O quê):** Identifica o impacto material da falha, incluindo externalidades prejudiciais de natureza física, econômica ou reputacional, relacionadas ao controle de acesso ou com alcance social.

Defesa para a nova superfície

Enfrentar essa nova superfície de ataque exige uma arquitetura com supervisão distribuída e políticas de segurança dinâmicas.

Nesse contexto, uma das abordagens mais relevantes é a supervisão distribuída.

O desenvolvimento de sistemas de Guardian Agents tem forte foco em modelos como o PSG-Agent (Personalized Safety Guardrail).

Em vez de utilizar verificações isoladas, o PSG-Agent implementa uma supervisão distribuída na qual vários subagentes assumem funções assíncronas e especializadas dentro do mesmo pipeline de execução:

- **Monitor de planejamento:** Avalia as estratégias antes da execução.
- **Tool Firewall (firewall de ferramentas):** Intercepta operações e audita parâmetros em tempo de execução.
- **Guardião de memória:** Atua como um bloqueio de gravação para impedir que dados tóxicos corrompam a memória de longo prazo.
- **Guardião de respostas:** Valida o resultado final do agente supervisionado.

Um diferencial do PSG-Agent é a personalização dinâmica dos limiares de risco, baseada na mineração de perfis históricos do usuário. Esse mecanismo permite identificar riscos cumulativos que sistemas baseados em uma única varredura não conseguem detectar.

Além do PSG-Agent, outro framework de referência de destaque é o GuardAgent, que adota uma abordagem diferente, baseada em raciocínio simbólico aplicado a políticas. O GuardAgent utiliza um processo em duas etapas, sustentado por raciocínio baseado em conhecimento: primeiro analisa linguisticamente as políticas de segurança para elaborar um plano de ação e, depois, converte esse plano em código executável.

A partir de demonstrações em contexto recuperadas de seu módulo de memória, o GuardAgent registrou índices superiores a 98% no controle de acesso simulado em ambientes de saúde e precisão de 83% na moderação geral para agentes web.

Conclusão

Os Guardian Agents não são uma moda tecnológica, mas uma resposta estrutural a um problema que as medidas de proteção estáticas já não conseguem resolver.

Conforme os agentes de IA ampliam sua autonomia operacional, a supervisão distribuída, a personalização dinâmica do risco e o raciocínio aplicado a políticas deixam de ser opcionais e passam a ser requisitos de arquitetura.

Construir sistemas de inteligência artificial confiáveis implica não apenas projetar agentes capazes, mas também projetar os guardiões que os mantenham dentro dos limites do que é seguro, auditável e eticamente responsável.

Jaime Andrés Tovar Prieto
Cybersecurity Architect

IA em cibersegurança: integração entre segurança defensiva e ofensiva

Tendências de Abel Martin Huaman Condor

A cibersegurança não deve depender apenas de controles defensivos voltados à prevenção, detecção e resposta a ameaças. As organizações também precisam validar continuamente a eficácia desses controles e identificar suas fragilidades antes que possam ser exploradas por um agente malicioso. Para isso, é necessário complementar a segurança defensiva com práticas de segurança ofensiva apoiadas por inteligência artificial.

A Agentic AI pode trabalhar com objetivos definidos, analisar informações, utilizar ferramentas autorizadas e executar ações dentro de limites estabelecidos. Isso permite acelerar as atividades de segurança e conectar com mais rapidez os achados ofensivos às melhorias defensivas.

IA na defesa e na ofensiva

Na defesa, a IA ajuda a analisar grandes volumes de alertas e eventos, identificar comportamentos anômalos e priorizar os riscos mais relevantes. A Agentic AI pode coletar informações de diferentes fontes, correlacionar indicadores e propor ações de contenção previamente autorizadas.

Na área de segurança ofensiva, a IA pode dar suporte a pen-tests, simulações de phishing e análises de configurações dentro de um escopo definido. O objetivo é identificar fragilidades em pessoas, processos e tecnologias antes que um agente malicioso possa explorá-las.

Embora essas ferramentas estejam cada vez mais acessíveis, seu uso eficaz exige conhecimentos básicos sobre sistemas, redes, vulnerabilidades e riscos. A IA potencializa as capacidades das equipes, mas não substitui a validação técnica nem o critério profissional.

Por que integrar as duas frentes

A defesa permite proteger, monitorar e responder a ameaças. No entanto, uma organização não pode presumir que seus controles sejam eficazes apenas porque foram implementados. É necessário validá-los em cenários controlados que reproduzam técnicas e comportamentos que poderiam ser utilizados em um ataque real.

A segurança ofensiva cumpre esse papel: identifica fragilidades e demonstra como poderiam afetar ativos, processos ou usuários. A defesa transforma esses resultados em ações concretas, como ajustes de configuração, novas regras de detecção, melhorias nos procedimentos de resposta ou reforço dos controles de acesso.

Integrar as duas frentes permite:

- Identificar vulnerabilidades antes que provoquem um incidente.
- Comprovar a eficácia real dos controles existentes.
- Priorizar correções e investimentos de acordo com o risco identificado.
- Melhorar a capacidade de detecção e resposta.
- Manter a melhoria contínua diante de ameaças em evolução.

A Agentic AI pode acelerar esse processo ao coletar evidências, correlacionar descobertas e relacionar uma fragilidade identificada a possíveis ações defensivas. Dessa forma, os resultados das avaliações ofensivas não ficam restritos a um relatório, mas se transformam em melhorias aplicáveis e mensuráveis.

Conclusão

A segurança ofensiva e a defensiva não são capacidades isoladas nem alternativas entre si. A ofensiva ajuda a descobrir e validar fragilidades; a defesa permite preveni-las, detectá-las e responder a elas.

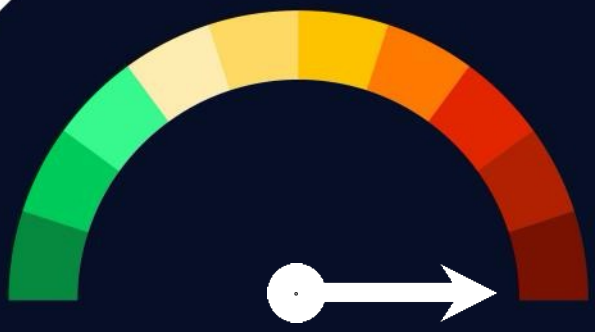
Combinar essas duas frentes, com o apoio da Agentic AI, permite construir uma postura de segurança mais proativa, eficiente e adaptável. O objetivo não é apenas reagir a ameaças, mas aprender continuamente com as fragilidades identificadas para reduzir o risco antes que se concretize.

Abel Martin Huaman Condor
Cybersecurity Expert Associate

Vulnerabilidades

Veeam ONE – Execução remota de código sem autenticação

Data: 04 de agosto de 2026
CVE: CVE-2026-64633



CVSS: 10

CRÍTICA

Descrição

Vulnerabilidade crítica no Veeam ONE que permite a um atacante remoto não autenticado executar código arbitrário no host do agente, sem interação do usuário.

Como não exige credenciais e apresenta baixa complexidade, é especialmente propícia à exploração automatizada, facilitando a movimentação lateral, o roubo de credenciais ou a distribuição de ***ransomware***.

A gravidade é ainda maior porque o Veeam ONE monitora a infraestrutura de backup, considerada a última linha de defesa da organização, razão pela qual recebeu pontuação CVSS de 10.

Solução

Recomenda-se:

- Atualizar para o Veeam ONE 13.1.0.7034 (lançado como parte da versão 13.1), que corrige todas as falhas descritas no artigo KB4892.
- Não existem medidas alternativas de mitigação; até que a atualização seja concluída, isolar da rede os hosts do agente Veeam ONE e restringir o acesso a esses hosts exclusivamente a partir de segmentos confiáveis de gerenciamento.

Produtos afetados

- Veeam ONE 13.0.2.6723 e todas as compilações anteriores da versão 13.

Referências

- veeam.com
- cvedetails.com
- incibe.es
- ryzer.es

Vulnerabilidades

Múltiplas vulnerabilidades em produtos VMware

Data: 30 de julho de 2026
CVE: CVE-2026-59309 e mais 4

CVSS: 9.8

CRÍTICA

Descrição

A VMware publicou cinco vulnerabilidades, três delas classificadas como críticas.

As vulnerabilidades de maior gravidade, identificadas como CVE-2026-59309, CVE-2026-59310 e CVE-2026-47876, consistem, respectivamente, em um bypass de autenticação no serviço de diretório da VMware, uma vulnerabilidade de ***directory traversal*** no Syslog e uma gravação fora dos limites no adaptador de rede virtual VMXNET3.

A exploração dessas vulnerabilidades poderia permitir que um atacante contornasse a autenticação, obtivesse acesso não autorizado ao sistema ou executasse código arbitrário no dispositivo afetado.

Solução

O fabricante recomenda atualizar os produtos afetados para a versão mais recente disponível.

Produtos afetados

Alguns dos produtos afetados são:

- VMware ESX.
- VMware vCenter.
- VMware Workstation.
- VMware Fusion.
- VMware Cloud Foundation.

Referências

- support.broadcom.com
- incibe.es

Patch de segurança da Cisco para vulnerabilidades críticas no Catalyst SD-WAN

Data: 5 de agosto de 2026
CVE: CVE-2026-20303 e mais 2

Crítica

Descrição

Devido à validação insuficiente de entrada e a controles de acesso inadequados, as vulnerabilidades CVE-2026-20303, CVE-2026-20304 e CVE-2026-20310 poderiam permitir que um atacante acessasse arquivos e recursos aos quais não deveria ter acesso, contornasse controles de autorização e elevasse privilégios no ambiente SD-WAN.

O Catalyst SD-WAN constitui o plano de gerenciamento da conectividade corporativa em grandes organizações. Por isso, o impacto potencial é elevado, já que um comprometimento poderia afetar a confidencialidade, a integridade e a disponibilidade de toda a rede gerenciada. Não existem medidas alternativas de mitigação ao patch.

Produtos afetados

Os produtos afetados são:

- Catalyst SD-WAN Manager, Controller e Validator em versões anteriores às ***hardening releases*** de agosto de 2026.

Solução

- Aplicar as ***hardening releases*** de agosto de 2026 indicadas no aviso cisco-sa-hardening-sdwan-faLcR3K.
- Restringir o acesso ao plano de gerenciamento exclusivamente a redes segmentadas de administração.

Referências

- incibe.es
- nvd.nist.gov

Atualização de segurança do Jenkins corrige a vulnerabilidade crítica CVE-2026-70426

Data: 5 de agosto de 2026
CVE: CVE-2026-70426 e mais 21

Crítica

Descrição

O Jenkins lançou uma atualização de segurança que corrige um total de 22 vulnerabilidades, entre elas uma crítica e oito de severidade alta.

A vulnerabilidade crítica, identificada como CVE-2026-70426, está presente no mecanismo de desserialização utilizado pela biblioteca ***Remoting*** e poderia permitir que um atacante enviasse objetos serializados especialmente manipulados para executar código arbitrário no controlador do Jenkins.

A exploração das demais vulnerabilidades, por sua vez, poderia permitir que um atacante contornasse mecanismos de segurança, gravasse arquivos em locais arbitrários ou elevasse privilégios.

Produtos afetados

Alguns dos produtos afetados são:

- Jenkins weekly até a versão 2.575.
- Jenkins LTS até a versão 2.568.1.
- AWS CodeBuild Plugin ≤ 0.59.
- CodeSonar Plugin ≤ 3.6.0.
- External Workspace Manager Plugin ≤ 1.4.1.
- HCL AppScan Plugin ≤ 1.8.3.
- Ivy Report Plugin ≤ 1.2.
- Multijob Plugin ≤ 669.v9d96a_d9c71b_0.
- Parameterized Remote Trigger Plugin ≤ 3.2.2.

Solução

Recomenda-se atualizar os produtos afetados para as versões mais recentes disponibilizadas pelo fabricante.

Referências

- www.jenkins.io
- www.incibe.es

Eventos

XVI Encontro da Cloud Security Alliance Espanha
17 de setembro

Na próxima quinta-feira, 17 de setembro, será realizado o XVI Encontro da Cloud Security Alliance Espanha, no Kinépolis Ciudad de la Imagen (Madri). Esta nova edição terá como foco os desafios e as ameaças à segurança na cloud, em um cenário marcado pela rápida evolução da inteligência artificial (IA) e da automação.

Link:

27º Congresso Internacional de Cibersegurança Industrial 2026 (Europa)
23 e 24 de setembro

Sevilha receberá, nos dias 23 e 24 de setembro, uma nova edição do Congresso Internacional de Cibersegurança Industrial na Europa, que reunirá especialistas, empresas de tecnologia e representantes do setor. Promovido pelo Centro de Cibersegurança Industrial, o evento terá como foco os principais desafios relacionados à proteção de infraestruturas críticas, especialmente em um contexto em que os ambientes industriais combinam cada vez mais sistemas físicos e digitais.

Link:

CS4CA Europe 2026 — Cyber Security for Critical Assets
23 e 24 de setembro

A edição de 2026 terá como foco a construção de resiliência IT/OT baseada em inteligência. Serão abordados temas como NIS2, ransomware industrial, acessos remotos, sistemas legados, ameaças de agentes estatais e proteção de setores como energia, transporte, indústria, saúde e serviços públicos. O evento contará com mais de 40 palestrantes e líderes internacionais de segurança IT e OT.

Link:

International Cyber Expo 2026
29 e 30 de setembro

O evento reunirá mais de 7.500 profissionais, representantes de mais de 90 países, mais de 100 fornecedores e mais de 100 palestrantes. A programação inclui o Global Cyber Summit, demonstrações práticas, exposição de soluções e participação de CISOs, representantes governamentais, órgãos reguladores e executivos da área de cibersegurança.

Link:

Recursos

➢ Relatório de Tendências e Ciberameaças

O relatório da equipe de Cyber Threat Intelligence da NTT DATA oferece uma visão atualizada das principais ameaças observadas durante o primeiro semestre de 2026, incluindo a evolução do ransomware, a atividade de atores maliciosos, a exploração de vulnerabilidades, as tendências na dark web, o impacto da IA no cibercrime e os principais riscos para as organizações. É um recurso essencial para antecipar ameaças, fortalecer as capacidades defensivas e orientar a tomada de decisões em cibersegurança.

Link:

➢ CISA Known Exploited Vulnerabilities Catalog

O Known Exploited Vulnerabilities Catalog, mantido pela Agência de Segurança Cibernética e de Infraestrutura dos Estados Unidos (CISA), apresenta uma lista atualizada de vulnerabilidades para as quais há evidências de exploração ativa. Diferentemente de outros repositórios voltados apenas à criticidade técnica, permite priorizar a remediação com base no risco real e na atividade dos atacantes. O catálogo inclui informações sobre o fabricante, o produto afetado, a ação necessária e a possível utilização da vulnerabilidade em campanhas de ransomware. Além disso, pode ser consultado online ou baixado nos formatos CSV e JSON para integração a processos e ferramentas de gestão de vulnerabilidades.

Link:

➢ M-Trends 2026: ameaças observadas na linha de frente

O relatório M-Trends 2026, elaborado pela Mandiant — Google Cloud —, analisa a evolução do cenário internacional de ciberameaças com base em mais de 500.000 horas dedicadas a investigações e resposta a incidentes durante 2025. O documento aborda temas como o uso de inteligência artificial pelos atacantes, a profissionalização do ecossistema de ransomware, as técnicas de acesso inicial, a persistência em dispositivos de perímetro e a necessidade de evoluir de uma defesa reativa para modelos de resiliência ativa. É um recurso especialmente útil para equipes de SOC, resposta a incidentes e inteligência de ameaças, além de profissionais responsáveis pela estratégia de cibersegurança.

Link:

NTT DATA

CLIQUE AQUI E SIGA A RADAR

Powered by the cybersecurity NTT DATA team

br.nttdata.com