PAPER DE ORIENTAÇÃO

IA em cybersecurity: desafios para 2026

Governança, segurança de aplicações generativas e controles para autonomia operacional

Um guia executivo para equipes de segurança, tecnologia e risco

Agosto de 2026
Agion Labs & Research

Resumo executivo

Em 2026, o principal desafio da inteligência artificial aplicada à cibersegurança não é apenas a capacidade do modelo. É a capacidade da organização de controlar dados, ferramentas, decisões e evidências em todo o ciclo de vida da solução.

Sistemas generativos e agentes ampliam a superfície tradicional: instruções podem ser manipuladas, contexto pode ser contaminado, respostas podem carregar conteúdo inseguro e uma automação com permissões excessivas pode transformar um erro de modelo em impacto operacional.

A abordagem recomendada neste paper combina três perspectivas complementares: governança de risco do NIST AI RMF, modelagem de comportamento adversário do MITRE ATLAS e riscos de aplicações generativas organizados pelo OWASP GenAI Security Project.

Cinco decisões para a liderança

Decisão	Pergunta de controle
Inventário	Quais modelos, agentes, fontes de contexto e ferramentas estão autorizados?
Dados	Quais informações podem entrar em prompts, memória, RAG e telemetria?
Autonomia	Quais ações exigem confirmação humana ou aprovação em dupla?
Evidência	Como reproduzir a entrada, a decisão, a saída e a ação executada?
Resposta	Como revogar credenciais, isolar integrações e retornar a um estado seguro?

ESCOPO

Este documento é uma orientação educacional. Não substitui avaliação jurídica, regulatória, de privacidade ou teste técnico específico do ambiente.

Os desafios que definem 2026

1. Prompt injection deixa de ser um problema apenas de texto

Entradas maliciosas podem chegar por mensagens, documentos recuperados, páginas externas, anexos ou dados consumidos por ferramentas. O controle precisa separar instrução, contexto e dado, reduzir confiança implícita e validar cada transição.

2. Agentes transformam resposta em execução

Quando um sistema pode consultar identidades, abrir tickets, alterar regras ou executar código, a autorização deve ser granular e temporária. A autonomia precisa ter limites de impacto, confirmação humana e um mecanismo de interrupção independente do próprio agente.

3. Cadeia de suprimentos inclui modelos, datasets e conectores

O inventário deve alcançar provedores, versões, adaptadores, embeddings, bases vetoriais, bibliotecas e ferramentas. Mudanças precisam ser avaliadas antes de chegar ao ambiente e devem permitir rastreabilidade e reversão.

4. Saída do modelo continua sendo entrada não confiável

Texto, consultas, comandos e artefatos gerados não devem ser executados diretamente. Validação de esquema, sanitização, política de destino e revisão humana reduzem a possibilidade de a saída virar um caminho de ataque.

5. Detecção exige telemetria própria de IA

Logs tradicionais não explicam todo o encadeamento. A investigação precisa relacionar identidade, versão do modelo, fonte de contexto, ferramenta chamada, política aplicada, resultado e decisão humana.

Um modelo operacional de controle

O NIST AI RMF organiza resultados em Govern, Map, Measure e Manage. Para cybersecurity, essas funções podem ser traduzidas em responsabilidades observáveis, testáveis e ligadas ao ciclo de mudança.

Função	Aplicação em segurança	Evidência mínima
Govern	Definir responsabilidade, apetite de risco, allowlist e política de autonomia.	Dono, política, aprovação e exceções.
Map	Modelar dados, dependências, ferramentas, ameaças e impacto no negócio.	Inventário, fluxo de dados e threat model.
Measure	Testar segurança, robustez, privacidade, abuso e comportamento fora do esperado.	Casos de teste, resultados e critérios de aceite.
Manage	Priorizar riscos, aplicar controles, monitorar e responder a incidentes.	Ações, responsáveis, prazo e verificação.

Roadmap de 90 dias

Janela	Entrega	Critério de saída
0–30 dias	Inventário e classificação de dados, modelos, agentes e integrações.	Nenhum uso produtivo sem responsável e finalidade registrados.
31–60 dias	Threat model, testes de abuso e política de permissões por ferramenta.	Caminhos críticos possuem controle preventivo e evidência.
61–90 dias	Telemetria, resposta, exercícios e revisão executiva do risco residual.	Equipe consegue detectar, conter, reconstruir e aprender.

PRINCÍPIO DE OPERAÇÃO

Automatize apenas o que pode ser limitado, observado, interrompido e explicado.

Checklist de prontidão

Controle	Pronto quando…
Acesso	Contas humanas usam MFA; identidades de máquina têm escopo mínimo e rotação.
Contexto	Fontes recuperadas são classificadas, filtradas e registradas.
Ferramentas	Cada chamada possui allowlist, parâmetros validados e limite de impacto.
Saídas	Conteúdo gerado é tratado como não confiável antes de uso ou execução.
Mudança	Versões de modelo, prompt, política e conector passam por avaliação.
Monitoramento	Alertas relacionam identidade, sessão, contexto, ferramenta e resultado.
Resposta	Existe kill switch, revogação de credenciais, isolamento e rollback testado.
Governança	Risco residual e exceções têm responsável, validade e aprovação registrada.

Referências oficiais

NIST AI Risk Management Framework e AIRC

https://airc.nist.gov/airmf-resources/airmf/

NIST AI 600-1 — Generative AI Profile

https://doi.org/10.6028/NIST.AI.600-1

MITRE ATLAS — Adversarial Threat Landscape for AI Systems

https://atlas.mitre.org/

OWASP Top 10 for LLMs and Gen AI Apps — 2025

https://genai.owasp.org/llm-top-10/

Leitura e verificação das fontes: agosto de 2026.

PRÓXIMO PASSO
Escolha um caso de uso real, modele o fluxo completo e valide os controles antes de ampliar autonomia ou acesso a dados.