

VERSÃO 1.0 · 07/10/2026

AvadLabs Agent Control Framework

12 controles para colocar agentes de IA em produção com
permissões, aprovação humana e auditoria.

Como usar

Os controles estão organizados em quatro perguntas. Cada controle tem um objetivo, a forma como o aplicamos e a evidência que comprova a sua existência. Para avaliar um agente, peça a evidência de cada item: uma resposta sem evidência é uma intenção, não um controle.

1. **Saber o que existe.** Inventário e dados: nenhum agente sem dono, nenhum dado sem motivo.
2. **Limitar o que o agente faz.** Permissões, aprovação, isolamento e segredos.
3. **Resistir a ataques e erros.** Injeção de prompt, custos e mudanças sem revisão.
4. **Provar e corrigir.** Auditoria, avaliação contínua e resposta a incidentes.

Saber o que existe

AC-01 Inventário de agentes

Objetivo: Saber quais agentes existem, quem responde por eles e o que fazem.

Como aplicamos:

- Registro de cada agente com dono, finalidade, modelo, ferramentas e dados acessados
- Revisão do inventário a cada mudança relevante

Evidência: Inventário versionado

AC-12 Dados e retenção

Objetivo: O agente guarda só o necessário, pelo tempo necessário.

Como aplicamos:

- Minimização e mascaramento de dados pessoais
- Prazo de retenção definido
- Base legal mapeada conforme a LGPD

Evidência: Mapa de dados do agente

Limitar o que o agente faz

AC-02 Permissão mínima

Objetivo: O agente acessa somente o que o processo exige.

Como aplicamos:

- Credencial própria por agente, nunca a de uma pessoa
- Escopos de leitura e escrita separados
- Ferramentas liberadas por lista explícita

Evidência: Política de permissões versionada

AC-03 Aprovação humana

Objetivo: Ações de valor, irreversíveis ou com impacto externo esperam uma pessoa.

Como aplicamos:

- Limites por tipo de ação: valor, destinatário e volume
- Fila de aprovação com o contexto da decisão
- Negação registrada com motivo

Evidência: Registro de aprovações e negações

AC-06 Isolamento de ferramentas

Objetivo: Uma ferramenta comprometida não alcança o restante do ambiente.

Como aplicamos:

- Execução isolada e acesso de rede restrito
- Validação de parâmetros antes de executar
- Respostas de ferramentas tratadas como dados não confiáveis

Evidência: Diagrama de arquitetura e testes de isolamento

AC-07 Segredos fora do código

Objetivo: Credenciais nunca aparecem em código, prompt ou log.

Como aplicamos:

- Cofre de segredos ou variáveis protegidas
- Rotação periódica
- Mascaramento automático em logs

Evidência: Varredura de segredos no repositório e nos logs

Resistir a ataques e erros

AC-05 Defesa contra injeção de prompt

Objetivo: Conteúdo externo nunca vira instrução para o agente.

Como aplicamos:

- Separação entre instruções e dados processados
- Ferramentas sensíveis indisponíveis ao processar conteúdo não confiável
- Testes adversariais antes de cada lançamento

Evidência: Relatório de testes adversariais

AC-09 Limites de custo e volume

Objetivo: Um erro não vira uma conta inesperada.

Como aplicamos:

- Orçamento por agente
- Limite de chamadas por período
- Alertas e corte automático

Evidência: Painel de custos e registro de alertas

AC-10 Revisão antes de publicar

Objetivo: Nenhuma mudança vai ao ar sem teste e revisão humana.

Como aplicamos:

- Ambientes separados de desenvolvimento, homologação e produção
- Revisão de código e de prompt
- Avaliação aprovada como pré-requisito

Evidência: Registro de mudanças com aprovador

Provar e corrigir

AC-04 Trilha de auditoria

Objetivo: Reconstruir o que o agente viu, decidiu e executou.

Como aplicamos:

- Log estruturado de entradas, chamadas de ferramenta, saídas e decisões
- Registro protegido contra alteração
- Retenção definida

Evidência: Consulta da trilha por sessão

AC-08 Avaliação contínua

Objetivo: A qualidade é medida antes e depois de cada mudança.

Como aplicamos:

- Suíte de casos reais com critérios de aceite
- Execução a cada mudança de modelo, prompt ou ferramenta
- Bloqueio de publicação quando a avaliação falha

Evidência: Histórico de avaliações

AC-11 Resposta a incidentes

Objetivo: Saber parar, investigar e corrigir rápido.

Como aplicamos:

- Desligamento imediato por agente
- Procedimento com severidades e prazos
- Análise pós-incidente com ações registradas

Evidência: Procedimento de incidente e relatórios

Checklist

CONTROLE	EVIDÊNCIA PEDIDA	SITUAÇÃO
AC-01 · Inventário de agentes	Inventário versionado	☐ sim ☐ parcial ☐ não
AC-02 · Permissão mínima	Política de permissões versionada	☐ sim ☐ parcial ☐ não
AC-03 · Aprovação humana	Registro de aprovações e negações	☐ sim ☐ parcial ☐ não
AC-04 · Trilha de auditoria	Consulta da trilha por sessão	☐ sim ☐ parcial ☐ não
AC-05 · Defesa contra injeção de prompt	Relatório de testes adversariais	☐ sim ☐ parcial ☐ não
AC-06 · Isolamento de ferramentas	Diagrama de arquitetura e testes de isolamento	☐ sim ☐ parcial ☐ não
AC-07 · Segredos fora do código	Varredura de segredos no repositório e nos logs	☐ sim ☐ parcial ☐ não
AC-08 · Avaliação contínua	Histórico de avaliações	☐ sim ☐ parcial ☐ não
AC-09 · Limites de custo e volume	Painel de custos e registro de alertas	☐ sim ☐ parcial ☐ não
AC-10 · Revisão antes de publicar	Registro de mudanças com aprovador	☐ sim ☐ parcial ☐ não
AC-11 · Resposta a incidentes	Procedimento de incidente e relatórios	☐ sim ☐ parcial ☐ não
AC-12 · Dados e retenção	Mapa de dados do agente	☐ sim ☐ parcial ☐ não

Quer aplicar o framework nos seus agentes? O Agent Security Review avalia agentes e copilotos contra estes 12 controles. avadlabs.com.br/contato