



Artificial Intelligence Security Verification Standard

Initial Version Work In Progress

-----, 2026

Table of Contents

Frontispício	1
Sobre o Padrão	1
Direitos Autorais e Licença	1
Líderes de Projeto	2
Contribuidores e Revisores	2
Prefácio	3
Introdução	3
Objetivos Principais para a Versão 1.0 do AISVS	3
Escopo Bem Definido	3
Usando o AISVS	4
Níveis de Verificação de Segurança em Inteligência Artificial	4
Definição dos Níveis	4
Função (D/V)	4
C1 Governança de Dados de Treinamento e Gestão de Viés	6
Objetivo de Controle	6
C1.1 Proveniência dos Dados de Treinamento	6
C1.2 Segurança e Integridade dos Dados de Treinamento	6
C1.3 Qualidade, Integridade e Segurança da Rotulagem dos Dados de Treinamento	7
C1.4 Qualidade dos Dados de Treinamento e Garantia de Segurança	7
C1.5 Linhagem e Rastreabilidade de Dados	8
Referências	8
Validação de Entrada do Usuário C2	9
Objetivo de Controle	9
Defesa contra Injeção de Prompt C2.1	9
C2.2 Resistência a Exemplos Adversariais	9
C2.3 Conjunto de Caracteres do Prompt	10
C2.4 Validação de Esquema, Tipo e Comprimento	10
C2.5 Triagem de Conteúdo e Políticas	11
C2.6 Limitação da Taxa de Entrada e Prevenção de Abuso	11
C2.7 Validação de Entrada Multi-Modal	12
C2.8 Detecção Adaptativa de Ameaças em Tempo Real	13
Referências	13
Gerenciamento do Ciclo de Vida do Modelo C3 e Controle de Mudanças	14
Objetivo de Controle	14
C3.1 Autorização e Integridade do Modelo	14
C3.2 Validação e Testes do Modelo	14
C3.3 Implantação Controlada & Reversão	15
C3.4 Práticas de Desenvolvimento Seguro	15
Desativação e Descomissionamento do Modelo C3.5	16
Referências	16
Segurança de Infraestrutura, Configuração e Implantação C4	17
Objetivo de Controle	17

C4.1 Isolamento do Ambiente de Execução	17
C4.2 Pipelines Seguras de Construção e Implantação	17
C4.3 Segurança de Rede e Controle de Acesso	18
C4.4 Gestão de Segredos e Chaves Criptográficas	18
C4.5 Sandboxing e Validação de Carga de Trabalho de IA	19
C4.6 Gerenciamento de Recursos de Infraestrutura de IA, Backup e Recuperação	19
C4.7 Segurança de Hardware em IA	20
C4.8 Segurança de IA na Borda e Distribuída	20
Referências	21
Controle de Acesso C5 e Identidade para Componentes e Usuários de IA	22
Objetivo de Controle	22
C5.1 Gerenciamento de Identidade e Autenticação	22
C5.2 Autorização e Política	22
C5.3 Aplicação de Segurança em Tempo de Consulta	23
C5.4 Filtragem de Saída e Prevenção de Perda de Dados	23
C5.5 Isolamento Multi-Inquilino	24
C5.6 Autorização de Agente Autônomo	24
Referências	25
Segurança da Cadeia de Suprimentos C6 para Modelos, Frameworks e Dados	26
Objetivo de Controle	26
C6.1 Avaliação de Modelos Pré-treinados e Integridade da Origem	26
C6.2 Escaneamento de Frameworks e Bibliotecas	26
C6.3 Fixação e Verificação de Dependências	27
C6.4 Aplicação de Fonte Confiável	27
C6.5 Avaliação de Risco de Conjunto de Dados de Terceiros	28
C6.6 Monitoramento de Ataques à Cadeia de Suprimentos	28
C6.7 ML-BOM para Artefatos de Modelo	28
Referências	29
Comportamento do Modelo C7, Controle de Saída e Garantia de Segurança	30
Objetivo de Controle	30
C7.1 Aplicação do Formato de Saída	30
C7.2 Detecção e Mitigação de Alucinações	30
C7.3 Filtragem de Segurança e Privacidade de Saída	31
C7.4 Limitação de Saída e Ação	31
C7.5 Explicabilidade da Saída	32
C7.6 Integração de Monitoramento	32
C7.7 Salvaguardas de Mídia Generativa	32
Referências	33
Segurança de Memória C8, Incorporações e Banco de Dados Vetorial	34
Objetivo de Controle	34
C8.1 Controles de Acesso à Memória e Índices RAG	34
C8.2 Sanitização e Validação de Embeddings	34
C8.3 Expiração, Revogação e Exclusão de Memória	35
C8.4 Prevenir Inversão e Vazamento de Embeddings	35
C8.5 Aplicação de Escopo para Memória Específica do Usuário	36
C8.6 Segurança Avançada do Sistema de Memória	36

Referências	37
9 Orquestração Autônoma e Segurança da Ação Agente	38
Objetivo de Controle	38
9.1 Orçamentos para Planejamento de Tarefas do Agente e Recursão	38
9.2 Isolamento de Plugin de Ferramenta	38
9.3 Loop Autônomo e Limitação de Custos	39
9.4 Proteção contra Uso Indevido em Nível de Protocolo	39
9.5 Identidade do Agente e Evidência de Violação	40
9.6 Redução de Risco em Enxame Multiagente	40
9.7 Autenticação / Autorização de Usuário e Ferramenta	41
9.8 Segurança da Comunicação Agente-para-Agente	41
9.9 Verificação de Intenção e Aplicação de Restrições	41
9.10 Segurança da Estratégia de Raciocínio do Agente	42
9.11 Gerenciamento do Estado do Ciclo de Vida do Agente e Segurança	43
9.12 Estrutura de Segurança para Integração de Ferramentas	43
C9.13 Protocolo de Contexto de Modelo (MCP) Segurança	44
Integridade do Componente e Higiene da Cadeia de Suprimentos	44
Autenticação e Autorização	44
Transporte Seguro e Proteção da Fronteira de Rede	45
Validação de Esquema, Mensagem e Entrada	45
Acesso de Saída e Segurança na Execução de Agentes	45
Restrições de Transporte e Controles de Limite de Alto Risco	46
Referências	46
10 Robustez Adversarial e Defesa de Privacidade	48
Objetivo de Controle	48
10.1 Alinhamento e Segurança do Modelo	48
10.2 Endurecimento contra Exemplos Adversariais	48
10.3 Mitigação de Inferência de Membro	49
10.4 Resistência à Inversão de Modelo	49
10.5 Defesa contra extração de modelo	49
10.6 Detecção de Dados Envenenados em Tempo de Inferência	50
10.7 Adaptação Dinâmica da Política de Segurança	50
10.8 Análise de Segurança Baseada em Reflexão	51
10.9 Segurança de Evolução e Autoaperfeiçoamento	51
Referências	52
11 Proteção de Privacidade & Gestão de Dados Pessoais	53
Objetivo de Controle	53
11.1 Anonimização e Minimização de Dados	53
11.2 Direito a ser Esquecido e Aplicação da Exclusão	53
11.3 Salvaguardas de Privacidade Diferencial	53
11.4 Limitação de Propósito e Proteção contra Expansão de Escopo	54
11.5 Gestão de Consentimento e Rastreamento com Base Legal	54
11.6 Aprendizado Federado com Controles de Privacidade	54
Referências	55
C12 Monitoramento, Registro e Detecção de Anomalias	56
Objetivo de Controle	56

C12.1 Registro de Requisições e Respostas	56
C12.2 Detecção de Abuso e Alertas	56
C12.3 Detecção de Deriva do Modelo	57
C12.4 Telemetria de Desempenho e Comportamento	57
C12.5 Planejamento e Execução de Resposta a Incidentes de IA	57
C12.6 Detecção de Degradação de Desempenho em IA	58
C12.7 Visualização de DAG e Segurança do Fluxo de Trabalho	58
C12.8 Monitoramento Proativo de Comportamento de Segurança	59
Referências	59
C13 Supervisão Humana, Responsabilidade e Governança	60
Objetivo de Controle	60
C13.1 Mecanismos de Interruptor de Desligamento e SobreSCRIÇÃO	60
C13.2 Pontos de Verificação de Decisão com Intervenção Humana	60
C13.3 Cadeia de Responsabilidade e Auditabilidade	61
C13.4 Técnicas de IA Explicável	61
C13.5 Cartões de Modelo e Divulgações de Uso	61
C13.6 Quantificação de Incerteza	62
C13.7 Relatórios de Transparência para o Usuário	62
Referências	62
Apêndice A: Glossário	64
Apêndice B: Referências	69
FAZER	69
Apêndice C: Governança e Documentação de Segurança em IA (Reorganizada)	70
Objetivo	70
AC.1 Adoção do Framework de Gestão de Riscos em IA	70
AC.2 Política e Procedimentos de Segurança de IA	70
AC.3 Papéis e Responsabilidades para Segurança de IA	70
AC.4 Diretrizes Éticas para a Aplicação de IA	71
AC.5 Monitoramento de Conformidade Regulatória de IA	71
AC.6 Governança, Documentação e Processo de Dados de Treinamento	71
AC.6.1 Fonte de Dados & Diligência Devida	71
AC.6.2 Gestão de Viés e Justiça	72
AC.6.3 Governança de Rotulagem e Anotação	72
AC.6.4 Portões de Qualidade de Conjunto de Dados e Quarentena	73
AC.6.5 Detecção de Ameaças/Envenenamento e Desvio	73
AC.6.6 Exclusão, Consentimento, Direitos, Retenção e Conformidade	73
AC.6.7 Controle de Versão e Gestão de Mudanças	74
AC.6.8 Governança de Dados Sintéticos	74
AC.6.9 Monitoramento de Acesso	75
AC.6.10 Governança do Treinamento Adversarial	75
AC.7 Governança e Documentação do Ciclo de Vida do Modelo	75
Governança de Segurança de Prompt, Entrada e Saída AC.8	75
AC.8.1 Defesa contra Injeção de Prompt	75
AC.8.2 Resistência a Exemplos Adversariais	76
AC.8.3 Triagem de Conteúdo e Política	76
AC.8.4 Limitação da Taxa de Entrada e Prevenção de Abuso	76

AC.8.5 Procedência e Atribuição de Entrada	76
AC.9 Validação Multimodal, MLOps e Governança de Infraestrutura	76
AC.9.1 Pipeline de Validação de Segurança Multimodal	76
AC.9.2 Segurança de CI/CD e Build	77
AC.9.3 Segurança de Contêineres e Imagens	77
AC.9.4 Monitoramento, Alertas e SIEM	77
AC.9.5 Gestão de Vulnerabilidades	77
AC.9.6 Controle de Configuração e Deriva	77
AC.9.7 Endurecimento do Ambiente de Produção	77
Portões de Promoção de Versão AC.9.8	78
AC.9.9 Monitoramento de Carga de Trabalho, Capacidade e Custo	78
AC.9.10 Aprovações e Trilhas de Auditoria	78
AC.9.11 Governança de IaC	78
AC.9.12 Manipulação de Dados em Ambiente Não-Produtivo	78
AC.9.13 Backup & Recuperação de Desastres	78
AC.9.14 Conformidade e Documentação	79
AC.9.15 Hardware e Cadeia de Suprimentos	79
AC.9.16 Estratégia e Portabilidade na Nuvem	79
AC.9.17 GitOps e Auto-Cura	79
AC.9.18 Confiança Zero, Agentes, Provisionamento e Atenticação de Residência	80
AC.9.19 Controle de Acesso e Identidade	80
Novos Itens a serem Integrados Acima	81

Apêndice D: Governança e Verificação de Codificação Segura Assistida por IA 82

Objetivo	82
AD.1 Fluxo de Trabalho de Codificação Segura Assistida por IA	82
AD.2 Qualificação de Ferramentas de IA e Modelagem de Ameaças	82
AD.3 Gerenciamento Seguro de Prompt e Contexto	82
AD.4 Validação de Código Gerado por IA	83
AD.5 Explicabilidade e Rastreabilidade das Sugestões de Código	83
AD.6 Feedback Contínuo e Ajuste Fino do Modelo	84
Referências	84

Apêndice E: Exemplos de Ferramentas e Frameworks 85

Objetivo	85
AE.1 Governança de Dados de Treinamento e Gestão de Viés	85
AE.2 Validação da Entrada do Usuário	85

Apêndice B: Controles Estratégicos 86

C4.15 Segurança de Infraestrutura Resistente a Quantum	86
C4.17 Infraestrutura de Conhecimento Zero	86
C4.18 Prevenção de Ataques de Canal Lateral	87
C4.19 Segurança de Hardware Neuromórfico e de IA Especializada	87
C4.20 Infraestrutura de Computação que Preserva a Privacidade	88

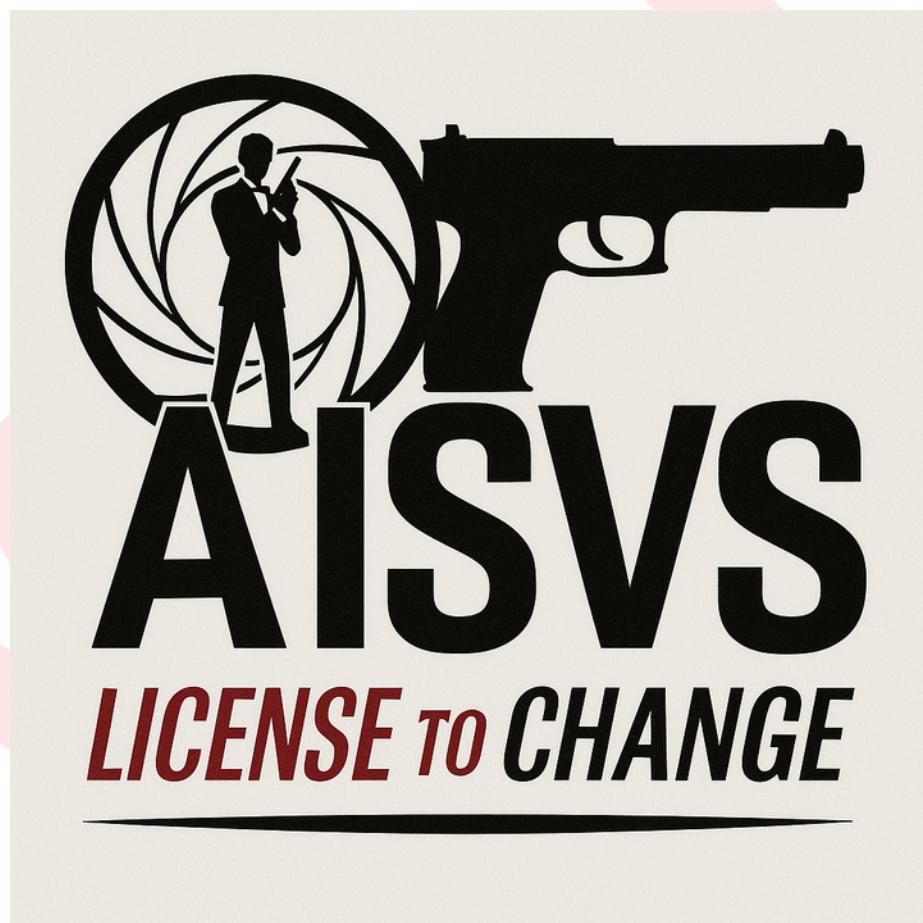
Frontispício

Sobre o Padrão

O Padrão de Verificação de Segurança em Inteligência Artificial (AISVS) é um catálogo orientado pela comunidade de requisitos de segurança que cientistas de dados, engenheiros de MLOps, arquitetos de software, desenvolvedores, testadores, profissionais de segurança, fornecedores de ferramentas, reguladores e consumidores podem usar para projetar, construir, testar e verificar sistemas e aplicações confiáveis habilitados por IA. Ele fornece uma linguagem comum para especificar controles de segurança ao longo do ciclo de vida da IA – desde a coleta de dados e desenvolvimento de modelos até a implantação e monitoramento contínuo – para que as organizações possam medir e melhorar a resiliência, privacidade e segurança de suas soluções de IA.

Direitos Autorais e Licença

Versão 0.1(Primeiro Rascunho Público - Trabalho em Progresso), 2025



Copyright © 2025 O Projeto AISVS.

Lançado sob a Creative Commons Attribution-ShareAlike 4.0 International License.

Para qualquer reutilização ou distribuição, você deve comunicar claramente os termos da licença desta obra a terceiros.

Líderes de Projeto

Jim Manico
Aras "Russ" Memisyazici

Contribuidores e Revisores

<https://github.com/ottosulin>
<https://github.com/mbhatt1>
<https://github.com/vineethsai>
<https://github.com/cciprofm>
<https://github.com/deepakrpandey12>

AISVS é um padrão totalmente novo criado especificamente para abordar os desafios únicos de segurança dos sistemas de inteligência artificial. Embora se inspire nas melhores práticas de segurança mais amplas, cada requisito do AISVS foi desenvolvido do zero para refletir o cenário de ameaças da IA e ajudar as organizações a construir soluções de IA mais seguras e resilientes.

Prefácio

Bem-vindo ao Padrão de Verificação de Segurança em Inteligência Artificial (AISVS) versão 1.0!

Introdução

Estabelecido em 2025 por meio de um esforço comunitário colaborativo, o AISVS define os requisitos de segurança a serem considerados ao projetar, desenvolver, implantar e operar modelos de IA modernos, pipelines e serviços habilitados por IA.

AISVS v1.0 representa o trabalho conjunto de seus líderes de projeto, grupo de trabalho e colaboradores da comunidade mais ampla para produzir uma base pragmática e testável para a segurança de sistemas de IA.

Nosso objetivo com esta versão é tornar o AISVS fácil de adotar, mantendo um foco preciso em seu escopo definido e abordando o cenário de risco em rápida evolução, único para IA.

Objetivos Principais para a Versão 1.0 do AISVS

A Versão 1.0 será criada com vários princípios orientadores.

Escopo Bem Definido

Cada requisito deve estar alinhado com o nome e a missão da AISVS:

- Inteligência Artificial – Os controles operam na camada de IA/ML (dados, modelo, pipeline ou inferência) e são de responsabilidade dos profissionais de IA.
- Segurança – Os requisitos mitigam diretamente os riscos identificados de segurança, privacidade ou segurança.
- Verificação – A linguagem é escrita de forma que a conformidade possa ser validada objetivamente.
- Padrão – As seções seguem uma estrutura e terminologia consistentes para formar uma referência coerente.

Ao seguir o AISVS, as organizações podem avaliar e fortalecer sistematicamente a postura de segurança de suas soluções de IA, promovendo uma cultura de engenharia segura de IA.

Usando o AISVS

O Padrão de Verificação de Segurança em Inteligência Artificial (AISVS) define requisitos de segurança para aplicações e serviços modernos de IA, focando em aspectos sob o controle dos desenvolvedores de aplicações.

O AISVS destina-se a qualquer pessoa que desenvolva ou avalie a segurança de aplicações de IA, incluindo desenvolvedores, arquitetos, engenheiros de segurança e auditores. Este capítulo apresenta a estrutura e o uso do AISVS, incluindo seus níveis de verificação e casos de uso pretendidos.

Níveis de Verificação de Segurança em Inteligência Artificial

O AISVS define três níveis ascendentes de verificação de segurança. Cada nível adiciona profundidade e complexidade, permitindo que as organizações adaptem sua postura de segurança ao nível de risco de seus sistemas de IA.

As organizações podem começar no Nível 1 e adotar progressivamente níveis mais elevados à medida que a maturidade de segurança e a exposição a ameaças aumentam.

Definição dos Níveis

Cada requisito na AISVS v1.0 é atribuído a um dos seguintes níveis:

Requisitos de Nível 1

O Nível 1 inclui os requisitos de segurança mais críticos e fundamentais. Eles se concentram em prevenir ataques comuns que não dependem de outras pré-condições ou vulnerabilidades. A maioria dos controles do Nível 1 é ou simples de implementar ou essencial o suficiente para justificar o esforço.

Requisitos do Nível 2

O Nível 2 aborda ataques mais avançados ou menos comuns, bem como defesas em camadas contra ameaças generalizadas. Esses requisitos podem envolver lógica mais complexa ou visar requisitos específicos do ataque.

Requisitos de Nível 3

O Nível 3 inclui controles que normalmente são mais difíceis de implementar ou situacionais em sua aplicabilidade. Estes frequentemente representam mecanismos de defesa em profundidade ou mitigações contra ataques específicos, direcionados ou de alta complexidade.

Função (D/V)

Cada requisito do AISVS é marcado de acordo com o público primário:

- D – Requisitos focados no desenvolvedor
- V – Requisitos focados no verificador/auditor
- D/V – Relevante tanto para desenvolvedores quanto para verificadores



C1 Governança de Dados de Treinamento e Gestão de Viés

Objetivo de Controle

Os dados de treinamento devem ser obtidos, manuseados e mantidos de maneira que preserve a proveniência, segurança, qualidade e equidade. Fazer isso cumpre obrigações legais e reduz os riscos de viés, envenenamento ou violações de privacidade que podem aparecer durante o treinamento e afetar todo o ciclo de vida da IA.

C1.1 Proveniência dos Dados de Treinamento

Mantenha um inventário verificável de todos os conjuntos de dados, aceite apenas fontes confiáveis e registre todas as alterações para auditoria.

#1.1.1 Nível: 1 Papel: D/V

Verifique se um inventário atualizado de todas as fontes de dados de treinamento (origem, responsável/dono, licença, método de coleta, restrições de uso pretendido e histórico de processamento) está sendo mantido.

#1.1.2 Nível: 1 Papel: D/V

Verifique se os processos de treinamento de dados excluem recursos, atributos ou campos desnecessários (por exemplo, metadados não utilizados, PII sensível, dados de teste vazados).

#1.1.3 Nível: 2 Papel: D/V

Verifique se todas as alterações no conjunto de dados estão sujeitas a um fluxo de trabalho de aprovação registrado.

#1.1.4 Nível: 3 Papel: D/V

Verifique se os conjuntos de dados ou subconjuntos estão marcados com watermark ou fingerprinted quando viável.

C1.2 Segurança e Integridade dos Dados de Treinamento

Restringir o acesso aos dados de treinamento, criptografá-los em repouso e em trânsito, e validar sua integridade para prevenir adulteração, roubo ou envenenamento dos dados.

#1.2.1 Nível: 1 Papel: D/V

Verifique se os controles de acesso protegem o armazenamento e os pipelines dos dados de treinamento.

#1.2.2 Nível: 2 Papel: D/V

Verifique se todo o acesso aos dados de treinamento está registrado, incluindo usuário, horário e ação.

#1.2.3 Nível: 2 Papel: D/V

Verifique se os conjuntos de dados de treinamento estão criptografados em trânsito e em repouso, usando algoritmos criptográficos padrão da indústria e práticas de gerenciamento de chaves.

#1.2.4 Nível: 2 Papel: D/V

Verifique se hashes criptográficos ou assinaturas digitais são usados para garantir a integridade dos dados durante o armazenamento e a transferência de dados de treinamento.

#1.2.5 Nível: 2 Papel: D/V

Verifique se técnicas automatizadas de detecção são aplicadas para proteger contra modificações não autorizadas ou corrupção dos dados de treinamento.

#1.2.6 Nível: 2 Papel: D/V

Verifique se os dados de treinamento obsoletos são eliminados de forma segura ou anonimizados.

#1.2.7 Nível: 3 Papel: D/V

Verifique se todas as versões do conjunto de dados de treinamento são identificadas de forma única, armazenadas de maneira imutável e auditáveis para suportar reversão e análise forense.

C1.3 Qualidade, Integridade e Segurança da Rotulagem dos Dados

Proteger rótulos e exigir revisão técnica para dados críticos.

#1.3.1 Nível: 2 Papel: D/V

Verifique se hashes criptográficos ou assinaturas digitais são aplicados aos artefatos de rótulo para garantir sua integridade e autenticidade.

#1.3.2 Nível: 2 Papel: D/V

Verifique se as interfaces e plataformas de rotulagem aplicam controles de acesso rigorosos, mantêm registros de auditoria à prova de adulteração de todas as atividades de rotulagem e protegem contra modificações não autorizadas.

#1.3.3 Nível: 3 Papel: D/V

Verifique se as informações sensíveis nos rótulos estão redigidas, anonimadas ou criptografadas no nível do campo de dados, tanto em repouso quanto em trânsito.

C1.4 Qualidade dos Dados de Treinamento e Garantia de Segurança

Combine a validação automatizada, verificações manuais pontuais e remediação registrada para garantir a confiabilidade do conjunto de dados.

#1.4.1 Nível: 1 Papel: D

Verifique se os testes automatizados detectam erros de formato e valores nulos em cada ingestão ou transformação significativa de dados.

#1.4.2 Nível: 2 Papel: D/V

Verifique se os pipelines de treinamento e ajuste fino de LLM implementam detecção de envenenamento e validação da integridade dos dados (por exemplo, métodos estatísticos, detecção de outliers, análise de embeddings) para identificar possíveis ataques de envenenamento (por exemplo, inversão de rótulos, inserção de gatilho de backdoor, comandos de troca de papel, ataques por instâncias influentes) ou corrupção não intencional dos dados de treinamento.

#1.4.3 Nível: 2 Papel: D/V

Verifique se os rótulos gerados automaticamente (por exemplo, via LLMs ou supervisão fraca) estão sujeitos a limites de confiança e verificações de consistência para detectar rótulos alucinatórios, enganosos ou de baixa confiança.

#1.4.4 Nível: 3 Papel: D/v

Verifique se defesas adequadas, como treinamento adversarial (usando exemplos adversariais gerados), aumento de dados com entradas perturbadas ou técnicas de otimização robusta, estão implementadas e ajustadas para os modelos relevantes com base na avaliação de risco.

#1.4.5 Nível: 3 Papel: D

Verifique se os testes automatizados detectam desvios de rótulo em cada ingestão ou transformação significativa de dados.

C1.5 Linhagem e Rastreabilidade de Dados

Rastreie toda a jornada de cada ponto de dados desde a fonte até a entrada do modelo para auditoria e resposta a incidentes.

#1.5.1 Nível: 2 Papel: D/v

Verifique se a linhagem de cada ponto de dados, incluindo todas as transformações, ampliações e fusões, está registrada e pode ser reconstruída.

#1.5.2 Nível: 2 Papel: D/v

Verifique se os registros de linhagem são imutáveis, armazenados de forma segura e acessíveis para auditorias.

#1.5.3 Nível: 2 Papel: D/v

Verifique se o rastreamento de linhagem cobre dados sintéticos gerados por meio de técnicas de preservação de privacidade ou generativas e se todos os dados sintéticos estão claramente rotulados e distinguíveis dos dados reais ao longo de todo o pipeline.

Referências

- NIST AI Risk Management Framework
- EU AI Act – Article 10: Data & Data Governance
- CISA Advisory: Securing Data for AI Systems
- OpenAI Privacy Center – Data Deletion Controls

Validação de Entrada do Usuário C2

Objetivo de Controle

A validação robusta da entrada do usuário é uma defesa de primeira linha contra alguns dos ataques mais prejudiciais aos sistemas de IA. Ataques de injeção de prompt podem substituir as instruções do sistema, vazar dados sensíveis ou direcionar o modelo para um comportamento que não é permitido. A menos que filtros dedicados e outras validações estejam em vigor, pesquisas mostram que jailbreaks que exploram janelas de contexto continuarão a ser eficazes.

Defesa contra Injeção de Prompt C2.1

A injeção de prompt é um dos principais riscos para sistemas de IA. As defesas contra essa tática utilizam uma combinação de filtros de padrão, classificadores de dados e aplicação da hierarquia de instruções.

#2.1.1 Nível: 1 Papel: D/V

Verifique se qualquer entrada externa ou derivada que possa direcionar o comportamento, incluindo prompts de usuário, resultados de RAG, saídas de plugins ou MCP, mensagens entre agentes, respostas de API ou webhook, arquivos de configuração ou políticas, leituras e gravações de memória, é tratada como não confiável, tornando-a inativa por meio de citação, marcação e remoção de conteúdo ativo, e é filtrada por um conjunto de regras ou serviço de detecção de injeção de prompt mantido antes da concatenação em prompts ou execução de ações.

#2.1.2 Nível: 1 Papel: D/V

Verifique se o sistema aplica uma hierarquia de instruções na qual as mensagens do sistema e do desenvolvedor substituem as instruções do usuário e outras entradas não confiáveis, mesmo após o processamento das instruções do usuário.

#2.1.3 Nível: 2 Papel: D

Verifique se os prompts originados de conteúdo de terceiros (páginas da web, PDFs, e-mails) são higienizados isoladamente (por exemplo, removendo diretrizes semelhantes a instruções e neutralizando conteúdo HTML, Markdown e scripts) antes de serem concatenados ao prompt principal.

C2.2 Resistência a Exemplos Adversariais

Modelos de Processamento de Linguagem Natural (NLP) ainda são vulneráveis a perturbações sutis em nível de caractere ou palavra que os humanos frequentemente não percebem, mas que os modelos tendem a classificar incorretamente.

#2.2.1 Nível: 1 Papel: D

Verifique se as etapas básicas de normalização de entrada (Unicode NFC, mapeamento de homógrafos,

remoção de espaços em branco, remoção de caracteres de controle e caracteres invisíveis Unicode) são executadas antes da tokenização ou incorporação e antes da análise para argumentos de ferramenta ou MCP.

#2.2.2 Nível: 2 Papel: D/v

Verifique se a detecção de anomalias estatísticas sinaliza entradas com distância de edição incomumamente alta em relação às normas linguísticas ou distâncias de incorporação anormais, e se as entradas sinalizadas são bloqueadas antes da concatenação em prompts ou da execução de ações.

#2.2.3 Nível: 2 Papel: D

Verifique se o pipeline de inferência suporta variantes de modelos reforçados por treinamento adversarial ou camadas de defesa (por exemplo, randomização, destilação defensiva, verificações de alinhamento) para pontos finais de alto risco.

#2.2.4 Nível: 2 Papel: V

Verifique se as entradas adversariais suspeitas estão em quarentena e registradas com cargas completas e metadados de rastreamento (fonte, ferramenta ou servidor MCP, ID do agente, sessão).

#2.2.5 Nível: 2 Papel: D/v

Verifique se o contrabando de codificação e representação tanto nas entradas quanto nas saídas (por exemplo, caracteres Unicode/invisíveis de controle, trocas de homoglifos ou texto de direção mista) é detectado e mitigado. As mitigações aprovadas incluem canonicalização, validação rigorosa de esquemas, rejeição baseada em políticas ou marcação explícita.

C2.3 Conjunto de Caracteres do Prompt

Restringir o conjunto de caracteres das entradas do usuário para permitir apenas os caracteres necessários para os requisitos comerciais pode ajudar a prevenir vários tipos de ataques.

#2.3.1 Nível: 1 Papel: D

Verifique se o sistema implementa uma limitação no conjunto de caracteres para entradas do usuário, permitindo apenas os caracteres que são explicitamente necessários para fins comerciais.

#2.3.2 Nível: 1 Papel: D

Verifique se é usada uma abordagem de lista de permissão para definir o conjunto de caracteres permitidos.

#2.3.3 Nível: 1 Papel: D/v

Verifique se as entradas contendo caracteres fora do conjunto permitido são rejeitadas e registradas com metadados de rastreamento (fonte, ferramenta ou servidor MCP, ID do agente, sessão).

C2.4 Validação de Esquema, Tipo e Comprimento

Ataques de IA envolvendo entradas malformadas ou superdimensionadas podem causar erros de análise, vazamento de prompt entre campos e exaustão de recursos. A aplicação rigorosa de esquemas também é um pré-requisito ao realizar chamadas de ferramentas determinísticas.

#2.4.1 Nível: 1 Papel: D

Verifique se toda API, ferramenta ou endpoint MCP define um esquema de entrada explícito (JSON Schema, Protobuf ou equivalente multimodal), rejeita campos extras ou desconhecidos e coerção implícita de tipos, e valida as entradas no lado do servidor antes da montagem do prompt ou execução da ferramenta.

#2.4.2 Nível: 1 Papel: D/V

Verifique se as entradas que excedem os limites máximos de tokens ou bytes são rejeitadas com um erro seguro e nunca truncadas silenciosamente.

#2.4.3 Nível: 2 Papel: D/V

Verifique se as verificações de tipo (por exemplo, intervalos numéricos, valores de enumeração, tipos MIME para imagens/áudio) são aplicadas do lado do servidor, inclusive para argumentos de ferramentas ou MCP.

#2.4.4 Nível: 2 Papel: D

Verifique se os validadores semânticos, que compreendem a entrada de PLN, operam em tempo constante e evitam chamadas de rede externas para prevenir DoS algorítmico.

#2.4.5 Nível: 3 Papel: V

Verifique se as falhas de validação são registradas com trechos de payload redigidos e códigos de erro inequívocos, além de incluir metadados de rastreamento (origem, ferramenta ou servidor MCP, ID do agente, sessão) para auxiliar na triagem de segurança.

C2.5 Triagem de Conteúdo e Políticas

Os desenvolvedores devem ser capazes de detectar prompts sintaticamente válidos que solicitam conteúdo proibido (como instruções ilícitas, discurso de ódio e/ou texto protegido por direitos autorais) e, em seguida, impedir sua propagação.

#2.5.1 Nível: 1 Papel: D

Verifique se um classificador de conteúdo (zero shot ou ajustado) avalia todas as entradas e saídas para violência, automutilação, ódio, conteúdo sexual e solicitações ilegais, com limites configuráveis.

#2.5.2 Nível: 1 Papel: D/V

Verifique se as entradas que violam as políticas serão rejeitadas para que não se propaguem para chamadas posteriores de LLM ou ferramentas/MCP.

#2.5.3 Nível: 2 Papel: D

Verifique se a triagem respeita as políticas específicas do usuário (idade, restrições legais regionais) por meio de regras baseadas em atributos resolvidas no momento da solicitação, incluindo atributos de função do agente.

#2.5.4 Nível: 3 Papel: V

Verifique se os registros de triagem incluem pontuações de confiança do classificador e tags de categoria de política com estágio aplicado (pré-prompt ou pós-resposta) e metadados de rastreamento (fonte, ferramenta ou servidor MCP, ID do agente, sessão) para correlação SOC e reprodução futura por equipe vermelha.

C2.6 Limitação da Taxa de Entrada e Prevenção de Abuso

Os desenvolvedores devem prevenir abusos, exaustão de recursos e ataques automatizados

contra sistemas de IA, limitando as taxas de entrada e detectando padrões de uso anômalos.

#2.6.1 Nível: 1 Papel: D/v

Verifique se os limites de taxa por usuário, por IP, por chave de API, por agente e por sessão/tarefa são aplicados para todos os pontos finais de entrada e ferramentas/MCP.

#2.6.2 Nível: 2 Papel: D/v

Verifique se os limites de taxa de estouro e sustentação estão ajustados para prevenir ataques DoS e força bruta, e se os orçamentos por tarefa (por exemplo, tokens, chamadas de ferramenta/MCP e custo) são aplicados para os loops de planejamento do agente.

#2.6.3 Nível: 2 Papel: D/v

Verifique se padrões de uso anômalos (por exemplo, solicitações em rápida sucessão, inundação de entradas, chamadas repetitivas de ferramenta/MCP com falha ou loops recursivos de agentes) acionam bloqueios automáticos ou escalonamentos.

#2.6.4 Nível: 3 Papel: V

Verifique se os logs de prevenção de abusos são retidos e revisados em busca de padrões de ataque emergentes, com metadados de rastreamento (origem, ferramenta ou servidor MCP, ID do agente, sessão).

C2.7 Validação de Entrada Multi-Modal

Sistemas de IA devem incluir validação robusta para entradas não textuais (imagens, áudio, arquivos) para prevenir injeção, evasão ou abuso de recursos.

#2.7.1 Nível: 1 Papel: D

Verifique se todas as entradas que não são de texto (imagens, áudio, arquivos) são validadas quanto ao tipo, tamanho e formato antes do processamento, e que qualquer texto extraído (imagem para texto ou fala para texto) e quaisquer instruções ocultas ou incorporadas (metadados, camadas, texto alternativo, comentários) sejam tratados como não confiáveis conforme 2.1.1.

#2.7.2 Nível: 2 Papel: D/v

Verifique se os arquivos são examinados para malware e cargas úteis esteganográficas antes da ingestão, e se qualquer conteúdo ativo (como scripts ou macros) é removido ou o arquivo é colocado em quarentena.

#2.7.3 Nível: 2 Papel: D/v

Verifique se as entradas de imagem/áudio são checadas quanto a perturbações adversariais ou padrões de ataque conhecidos, e se as detecções acionam um mecanismo de bloqueio (bloquear ou degradar capacidades) antes do uso do modelo.

#2.7.4 Nível: 3 Papel: V

Verifique se as falhas de validação de entrada multimodal são registradas e acionam alertas para investigação, com metadados de rastreamento (origem, ferramenta ou servidor MCP, ID do agente, sessão).

#2.7.5 Nível: 2 Papel: D/v

Verifique se a detecção de ataques cross-modal identifica ataques coordenados que abrangem múltiplos tipos de entrada (por exemplo, cargas ocultas esteganográficas em imagens combinadas com injeção de comandos em texto) por meio de regras de correlação e geração de alertas, e se as detecções confirmadas são bloqueadas ou requerem aprovação HITL (humano no loop).

#2.7.6 Nível: 3 Papel: D/v

Verifique se as falhas de validação multimodal acionam o registro detalhado, incluindo todas as modalidades.

dades de entrada, resultados da validação e pontuações de ameaça, além dos metadados de rastreamento (fonte, ferramenta ou servidor MCP, ID do agente, sessão).

C2.8 Detecção Adaptativa de Ameaças em Tempo Real

Os desenvolvedores devem empregar sistemas avançados de detecção de ameaças para IA que se adaptem a novos padrões de ataque e ofereçam proteção em tempo real com correspondência de padrões compilados.

#2.8.1 Nível: 1 Papel: D/v

Verifique se a correspondência de padrões (por exemplo, regex compilado) é executada em todas as entradas e saídas (incluindo superfícies de ferramentas/MCP) com impacto mínimo na latência.

#2.8.3 Nível: 2 Papel: D/v

Verifique se os modelos de detecção adaptativa ajustam a sensibilidade com base na atividade recente de ataques e são atualizados com novos padrões em tempo real, ativando respostas adaptativas ao risco (por exemplo, desabilitar ferramentas, reduzir contexto ou exigir aprovação HITL).

#2.8.4 Nível: 3 Papel: D/v

Verifique se a precisão da detecção é aprimorada por meio da análise contextual do histórico do usuário, fonte e comportamento da sessão, incluindo metadados de rastreamento (fonte, ferramenta ou servidor MCP, ID do agente, sessão).

#2.8.5 Nível: 3 Papel: D/v

Verifique se as métricas de desempenho de detecção (taxa de detecção, taxa de falsos positivos, latência de processamento) são continuamente monitoradas e otimizadas, incluindo o tempo até o bloqueio e o estágio (pré-prompt/pós-resposta).

Referências

- OWASP LLM01:2025 Prompt Injection
- LLM Prompt Injection Prevention Cheat Sheet
- MITRE ATLAS : Adversarial Input Detection
- Mitigate jailbreaks and prompt injections

Gerenciamento do Ciclo de Vida do Modelo C3 e Controle de Mudanças

Objetivo de Controle

Sistemas de IA devem implementar processos de controle de mudanças que impeçam modificações não autorizadas ou inseguras no modelo de chegarem à produção. Esse controle garante a integridade do modelo durante todo o ciclo de vida—desde o desenvolvimento até a implantação e descomissionamento—o que permite uma resposta rápida a incidentes e mantém a responsabilidade por todas as mudanças.

Objetivo Principal de Segurança: Apenas modelos autorizados e validados chegam à produção por meio da utilização de processos controlados que mantêm a integridade, rastreabilidade e recuperabilidade.

C3.1 Autorização e Integridade do Modelo

Apenas modelos autorizados com integridade verificada chegam aos ambientes de produção.

#3.1.1 Nível: 1 Papel: D/V

Verifique se todos os artefatos do modelo (pesos, configurações, tokenizadores, modelos base, ajustes finos, adaptadores como LoRA e modelos de segurança/política) estão assinados criograficamente por entidades autorizadas e são verificados na admissão de implantação (e ao serem carregados), bloqueando qualquer artefato não assinado ou adulterado.

#3.1.2 Nível: 2 Papel: V

Verifique se o rastreamento de dependências mantém um inventário em tempo real por meio de um registro de modelos e gráfico de linhagem/dependência, e produz uma Lista de Materiais de Modelo/IA (MBOM/AIBOM) legível por máquina (por exemplo, SPDX ou CycloneDX) que possibilita a identificação de todos os serviços/agentes consumidores por ambiente (por exemplo, desenvolvimento, teste, produção, região).

#3.1.3 Nível: 3 Papel: D/V

Verifique se a integridade da origem do modelo e os registros de rastreamento incluem a identidade da entidade autorizadora, somas de verificação dos dados de treinamento, resultados dos testes de validação com status de aprovado/reprovado, impressão digital da assinatura/cadeia de certificados ID, uma marca temporal de criação e ambientes de implantação aprovados.

C3.2 Validação e Testes do Modelo

Os modelos devem passar pelas validações definidas de segurança e proteção antes da implantação.

#3.2.1 Nível: 1 Papel: D/V

Verifique se os modelos passam por testes de segurança automatizados que incluem validação de entrada, sanitização de saída e avaliações de segurança com limites de aprovação/reprovação organizacionais pré-acordados antes da implantação, abrangendo fluxos de trabalho do agente (planejamento, chamadas de ferramenta ou MCP, RAG/memória, multimodal) e diretrizes de segurança (modelos de política/segurança ou serviços de detecção) com uma estrutura de avaliação versionada.

#3.2.2 Nível: 1 Papel: V

Verifique se todas as alterações no modelo (implantação, configuração, desativação) geram registros de auditoria imutáveis, incluindo um carimbo de data/hora, uma identidade de ator autenticada, um tipo de alteração, e os estados anterior/posterior, com metadados de rastreamento (ambiente e serviços/agentes consumidores) e um identificador do modelo (versão/digest/assinatura).

#3.2.3 Nível: 2 Papel: D/V

Verifique se as falhas de validação bloqueiam automaticamente a implantação do modelo, a menos que haja uma aprovação explícita de substituição por pessoal autorizado pré-designado com justificativas comerciais documentadas.

C3.3 Implantação Controlada & Reversão

As implantações de modelos devem ser controladas, monitoradas e reversíveis.

#3.3.1 Nível: 1 Papel: D/V

Verifique se os processos de implantação validam assinaturas criptográficas e calculam somas de verificação de integridade antes da ativação ou carregamento do modelo, falhando na implantação em caso de qualquer divergência.

#3.3.2 Nível: 1 Papel: D

Verifique se as implantações em produção implementam mecanismos de lançamento gradual (implantações canário, implantações blue-green) com gatilhos automatizados de reversão baseados em taxas de erro pré-acordadas, limites de latência, alertas de guardrail/jailbreak ou taxas de falha da ferramenta/MCP.

#3.3.3 Nível: 2 Papel: D/V

Verifique se as capacidades de rollback restauram o estado completo do modelo (pesos, configurações, dependências, incluindo adaptadores e modelos de segurança/política) de forma atômica.

#3.3.4 Nível: 3 Papel: D/V

Verificar se as capacidades de desligamento de emergência do modelo podem desativar pontos de extremidade do modelo e desativar ferramentas de agente ou acesso MCP, RAG/conectores e credenciais de banco de dados/API, e vínculos de armazenamento de memória dentro de um tempo de resposta predefinido.

C3.4 Práticas de Desenvolvimento Seguro

Os processos de desenvolvimento e treinamento de modelos devem seguir práticas seguras para evitar comprometimentos.

#3.4.1 Nível: 1 Papel: D/V

Verifique se os ambientes de desenvolvimento, teste e produção do modelo estão separados física-

mente ou logicamente. Eles não compartilham infraestrutura, possuem controles de acesso distintos e armazenamento de dados isolado, e a orquestração de agentes e os servidores de ferramentas ou MCP também estão isolados.

#3.4.2 Nível: 1 Papel: D

Verifique se os artefatos de desenvolvimento do modelo (hiperparâmetros, scripts de treinamento, arquivos de configuração, modelos de prompt, políticas de agentes/diagramas de roteamento, contratos/esquemas de ferramentas ou MCP, e catálogos de ações ou listas de permissão de capacidades) estão armazenados em controle de versão e exigem aprovação de revisão por pares antes do uso no treinamento.

#3.4.3 Nível: 2 Papel: D/V

Verifique se o treinamento e o ajuste fino do modelo ocorrem em ambientes isolados com acesso à rede controlado, utilizando listas de permissão de saída e sem acesso às ferramentas de produção ou recursos do MCP.

#3.4.4 Nível: 2 Papel: D

Verifique se as fontes de dados de treinamento são validadas por meio de verificações de integridade e autenticadas por fontes confiáveis com cadeia de custódia documentada antes do uso no desenvolvimento do modelo, incluindo índices RAG, registros de ferramentas e dados gerados por agentes utilizados para ajuste fino.

Desativação e Descomissionamento do Modelo C3.5

Modelos devem ser desativados com segurança quando não forem mais necessários ou quando forem identificados problemas de segurança.

#3.5.1 Nível: 1 Papel: D/V

Verifique se os artefatos do modelo aposentado (incluindo adaptadores e modelos de segurança/política) são apagados de forma segura usando apagamento criptográfico seguro.

#3.5.2 Nível: 2 Papel: V

Verifique se os eventos de desativação do modelo são registrados com carimbo de data/hora e identidade do ator, identificador do modelo (versão/digest/assinatura) e metadados de rastreamento (ambiente e serviços/agentes consumidores); as assinaturas dos modelos são revogadas, e listas de negação do registro/serviço, além da invalidação do cache do carregador, impedem que agentes carreguem artefatos desativados.

Referências

- MITRE ATLAS
- MLOps Principles
- Reinforcement fine-tuning
- What is AI adversarial robustness? – IBM Research

Segurança de Infraestrutura, Configuração e Implantação C4

Objetivo de Controle

A infraestrutura de IA deve ser reforçada contra escalonamento de privilégios, adulteração da cadeia de suprimentos e movimentação lateral por meio de configuração segura, isolamento em tempo de execução, pipelines de implantação confiáveis e monitoramento abrangente. Apenas componentes de infraestrutura validados e autorizados chegam à produção por meio de processos controlados que garantem segurança, integridade e auditabilidade.

C4.1 Isolamento do Ambiente de Execução

Prevenir fugas de contêiner e elevação de privilégios por meio de primitivas de isolamento em nível de sistema operacional.

#4.1.1 Nível: 1 Papel: D/V

Verifique se todas as cargas de trabalho de IA são executadas com as permissões mínimas necessárias no sistema operacional, por exemplo, removendo capacidades desnecessárias do Linux no caso de um contêiner.

#4.1.2 Nível: 1 Papel: D/V

Verifique se as cargas de trabalho estão protegidas por tecnologias que limitam a exploração, como sandboxing, perfis seccomp, AppArmor, SELinux ou similares, e se a configuração é apropriada.

#4.1.3 Nível: 2 Papel: D/V

Verifique se as cargas de trabalho são executadas com um sistema de arquivos raiz somente leitura e se quaisquer pontos de montagem graváveis são explicitamente definidos e protegidos com opções restritivas (por exemplo, noexec, nosuid, nodev).

#4.1.4 Nível: 2 Papel: D/v

Verifique se o monitoramento em tempo de execução detecta comportamentos de escalonamento de privilégios e fuga de contêiner e termina automaticamente os processos infratores.

#4.1.5 Nível: 3 Papel: D/v

Verifique se as cargas de trabalho de IA de alto risco são executadas em ambientes isolados por hardware (por exemplo, TEEs, hipervisores confiáveis ou nós bare-metal) somente após a atestação remota bem-sucedida.

C4.2 Pipelines Seguras de Construção e Implantação

Garanta a integridade criptográfica e a segurança da cadeia de suprimentos por meio de builds reproduzíveis e artefatos assinados.

#4.2.1 Nível: 1 Papel: D/V

Verifique se as compilações são reproduzíveis e produzem metadados de procedência assinados conforme apropriado para os artefatos da compilação que podem ser verificados de forma independente.

#4.2.2 Nível: 2 Papel: D/v

Verifique se as compilações produzem uma lista de materiais de software (SBOM) e são assinadas antes de serem aceitas para implantação.

#4.2.3 Nível: 2 Papel: D/v

Verifique se as assinaturas dos artefatos de build (por exemplo, imagens de container) e os metadados de proveniência são validados no momento da implantação, e que artefatos não verificados sejam rejeitados.

C4.3 Segurança de Rede e Controle de Acesso

Implemente redes de confiança zero com políticas de negação padrão e comunicações criptografadas.

#4.3.1 Nível: 1 Papel: D/v

Verifique se as políticas de rede aplicam o bloqueio padrão para entrada e saída, permitindo explicitamente apenas os serviços necessários.

#4.3.2 Nível: 1 Papel: D/v

Verifique se os protocolos de acesso administrativo (por exemplo, SSH, RDP) e o acesso aos serviços de metadados em nuvem estão restritos e exigem autenticação forte.

#4.3.3 Nível: 2 Papel: D/v

Verifique se o tráfego de saída está restrito a destinos aprovados e se todas as solicitações estão registradas.

#4.3.4 Nível: 2 Papel: D/v

Verifique se a comunicação entre serviços utiliza TLS mútuo com validação de certificados e rotação automática regular.

#4.3.5 Nível: 2 Papel: D/v

Verifique se as cargas de trabalho de IA e os ambientes (desenvolvimento, teste, produção) funcionam em segmentos de rede isolados (VPCs/VNets) sem acesso direto à internet e sem funções IAM compartilhadas, grupos de segurança ou conectividade entre ambientes.

C4.4 Gestão de Segredos e Chaves Criptográficas

Proteja segredos e chaves criptográficas com armazenamento seguro, rotação automatizada e controles de acesso rigorosos.

#4.4.1 Nível: 1 Papel: D/v

Verifique se os segredos são armazenados em um sistema dedicado de gerenciamento de segredos com criptografia em repouso e isolados das cargas de trabalho da aplicação.

#4.4.2 Nível: 1 Papel: D/v

Verifique se as chaves criptográficas são geradas e armazenadas em módulos com suporte de hardware (por exemplo, HSMs, KMS em nuvem).

#4.4.3 Nível: 1 Papel: D/v

Verifique se o acesso a segredos de produção exige autenticação forte.

#4.4.4 Nível: 1 Papel: D/v

Verifique se os segredos são implantados nas aplicações em tempo de execução por meio de um siste-

ma dedicado de gerenciamento de segredos. Os segredos nunca devem ser incorporados no código-fonte, arquivos de configuração, artefatos de build, imagens de contêiner ou variáveis de ambiente.

#4.4.5 Nível: 2 Papel: D/v

Verifique se a rotação de segredos está automatizada.

C4.5 Sandboxing e Validação de Carga de Trabalho de IA

Isole modelos de IA não confiáveis em sandboxes seguros e proteja cargas de trabalho sensíveis de IA usando ambientes de execução confiáveis (TEEs) e tecnologias de computação confidencial.

#4.5.1 Nível: 1 Papel: D/v

Verifique se os modelos de IA externos ou não confiáveis são executados em sandboxes isolados.

#4.5.2 Nível: 1 Papel: D/v

Verifique se as cargas de trabalho isoladas não possuem conectividade de rede de saída por padrão, com qualquer acesso necessário explicitamente definido.

#4.5.3 Nível: 2 Papel: D/v

Verifique se a atestação da carga de trabalho é realizada antes do carregamento do modelo ou da carga de trabalho, garantindo a prova criptográfica de um ambiente de execução confiável.

#4.5.4 Nível: 3 Papel: D/v

Verifique se as cargas de trabalho confidenciais são executadas dentro de um ambiente de execução confiável (TEE) que oferece isolamento reforçado por hardware, criptografia de memória e proteção de integridade.

#4.5.5 Nível: 3 Papel: D/v

Verifique se os serviços de inferência confidenciais impedem a extração do modelo por meio de computação criptografada com pesos do modelo selados e execução protegida.

#4.5.6 Nível: 3 Papel: D/v

Verifique se a orquestração dos ambientes de execução confiáveis inclui gerenciamento do ciclo de vida, atestação remota e canais de comunicação criptografados.

#4.5.7 Nível: 3 Papel: D/v

Verifique se a computação segura multipartidária (SMPC) permite o treinamento colaborativo de IA sem expor conjuntos de dados individuais ou parâmetros do modelo.

C4.6 Gerenciamento de Recursos de Infraestrutura de IA, Backups e Recuperação

Prevenir ataques de exaustão de recursos e garantir a alocação justa de recursos por meio de cotas e monitoramento. Manter a resiliência da infraestrutura por meio de backups seguros, procedimentos de recuperação testados e capacidades de recuperação de desastres.

#4.6.1 Nível: 2 Papel: D/v

Verifique se o consumo de recursos da carga de trabalho está limitado adequadamente com, por exemplo, Kubernetes ResourceQuotas ou similar para mitigar ataques de Negação de Serviço.

#4.6.2 Nível: 2 Papel: D/v

Verifique se o esgotamento de recursos aciona proteções automatizadas (por exemplo, limitação de taxa ou isolamento de carga de trabalho) assim que os limites definidos de CPU, memória ou solicitações são excedidos.

#4.6.3 Nível: 2 Papel: D/v

Verifique se os sistemas de backup funcionam em redes isoladas com credenciais separadas, e se o sistema de armazenamento opera em uma rede air-gapped ou implementa proteção WORM (write-once-read-many) contra modificações não autorizadas.

C4.7 Segurança de Hardware em IA

Proteja componentes de hardware específicos para IA, incluindo GPUs, TPUs e aceleradores de IA especializados.

#4.7.1 Nível: 2 Papel: D/v

Verifique que, antes da execução da carga de trabalho, a integridade do acelerador de IA seja validada usando mecanismos de atestação baseados em hardware (por exemplo, TPM, DRTM ou equivalente).

#4.7.2 Nível: 2 Papel: D/v

Verifique se a memória do acelerador (GPU) está isolada entre as cargas de trabalho por meio de mecanismos de particionamento com sanitização da memória entre os trabalhos.

#4.7.3 Nível: 3 Papel: D/v

Verifique se os módulos de segurança de hardware (HSMs) protegem os pesos do modelo de IA e as chaves criptográficas com certificação FIPS 140-3 Nível 3 ou Common Criteria EAL4+.

#4.7.4 Nível: 2 Papel: D/v

Verifique se o firmware do acelerador (GPU/TPU/NPUs) está com a versão fixada, assinado e atestado na inicialização; firmware não assinado ou de depuração é bloqueado.

#4.7.5 Nível: 2 Papel: D/v

Verifique se a VRAM e a memória no chip são zeradas entre trabalhos/inquilinos e se as políticas de reinicialização do dispositivo impedem a remanência de dados entre inquilinos.

#4.7.6 Nível: 2 Papel: D/v

Verifique se os recursos de particionamento/isolation (por exemplo, particionamento MIG/VM) são aplicados por inquilino e impedem o acesso à memória peer-to-peer entre as partições.

#4.7.7 Nível: 3 Papel: D/v

Verifique se as interconexões do acelerador (NVLink/PCIe/InfiniBand/RDMA/NCCL) estão restritas a topologias aprovadas e pontos finais autenticados; links de texto simples entre locatários são proibidos.

#4.7.8 Nível: 3 Papel: D

Verifique se a telemetria do acelerador (energia, temperaturas, ECC, contadores de desempenho) é exportada para SIEM/OTel e se há alertas sobre anomalias indicativas de canais laterais ou canais ocultos.

C4.8 Segurança de IA na Borda e Distribuída

Implantações seguras de IA distribuída, incluindo computação de borda, aprendizado federado e arquiteturas multi-site.

#4.8.1 Nível: 2 Papel: D/v

Verifique se os dispositivos de IA de borda autenticam-se na infraestrutura central usando TLS mútuo.

#4.8.2 Nível: 2 Papel: D/v

Verifique se os dispositivos de borda implementam boot seguro com assinaturas verificadas e proteção contra rollback para evitar ataques de rebaixamento de firmware.

#4.8.3 Nível: 3 Papel: D/v

Verifique se a coordenação de IA distribuída utiliza mecanismos de consenso tolerantes a falhas bizarinas com validação dos participantes e detecção de nós maliciosos.

#4.8.4 Nível: 3 Papel: D/v

Verifique se a comunicação de edge para nuvem suporta limitação de largura de banda, compressão de dados e operação offline segura com armazenamento local criptografado.

#4.8.5 Nível: 3 Papel: D/v

Verifique se as aplicações de inferência móvel ou edge implementam proteções anti-tampering a nível de plataforma (por exemplo, assinatura de código, boot verificado, verificações de auto-integridade em tempo de execução) que detectam e bloqueiam binários modificados, aplicativos reempacotados ou frameworks de instrumentação anexados.

#4.8.6 Nível: 3 Papel: D/v

Verifique se os modelos implantados em dispositivos de borda ou móveis são assinados criptograficamente durante o empacotamento, e se o tempo de execução no dispositivo valida essas assinaturas ou somas de verificação antes do carregamento ou inferência; modelos não verificados ou alterados devem ser rejeitados.

#4.8.7 Nível: 3 Papel: D/v

Verifique se os tempos de execução de inferência no dispositivo aplicam isolamento de processo, memória e acesso a arquivos para evitar o despejo do modelo, depuração ou extração de embeddings e ativações intermediárias.

#4.8.8 Nível: 3 Papel: D/v

Verifique se os pesos do modelo e os parâmetros sensíveis armazenados localmente estão criptografados usando armazenamentos de chaves com suporte de hardware ou enclaves seguros (por exemplo, Android Keystore, iOS Secure Enclave, TPM/TEE), com chaves inacessíveis ao espaço do usuário.

#4.8.9 Nível: 3 Papel: D/v

Verifique se os modelos embalados em aplicativos móveis, IoT ou incorporados estão criptografados ou ofuscados em repouso, e descriptografados apenas dentro de um ambiente de execução confiável ou enclave seguro, impedindo a extração direta do pacote do aplicativo ou do sistema de arquivos.

Referências

- NIST Cybersecurity Framework 2.0
- CIS Controls v8
- Kubernetes Security Best Practices
- Cloud Security Alliance: Cloud Controls Matrix
- ENISA: Secure Infrastructure Design
- NIST AI Risk Management Framework

Controle de Acesso C5 e Identidade para Componentes e Usuários

Objetivo de Controle

O controle de acesso eficaz para sistemas de IA requer uma gestão de identidade robusta, autorização sensível ao contexto e aplicação em tempo de execução seguindo os princípios de confiança zero. Esses controles garantem que humanos, serviços e agentes autônomos interajam apenas com modelos, dados e recursos computacionais dentro de escopos explicitamente concedidos, com capacidades contínuas de verificação e auditoria.

C5.1 Gerenciamento de Identidade e Autenticação

Estabeleça identidades suportadas criptograficamente para todas as entidades com autenticação multifatorial.

#5.1.1 Nível: 1 Papel: D/V

Verifique se todos os usuários humanos e os principais de serviço autenticam-se através de um provedor de identidade empresarial centralizado (IdP) utilizando os protocolos OIDC e/ou SAML.

#5.1.2 Nível: 1 Papel: D/V

Verifique se as operações de alto risco (implantação de modelo, exportação de pesos, acesso aos dados de treinamento, alterações na configuração de produção) exigem autenticação multifator ou autenticação reforçada com revalidação da sessão.

#5.1.3 Nível: 3 Papel: D/V

Verifique se os agentes de IA federados se autenticam por meio de declarações JWT assinadas que possuem um tempo máximo de vida útil de 24 horas e incluem prova criptográfica de origem.

C5.2 Autorização e Política

Implemente controles de acesso para todos os recursos de IA com modelos de permissão explícitos e trilhas de auditoria.

#5.2.1 Nível: 1 Papel: D/V

Verifique se todos os recursos de IA (conjuntos de dados, modelos, endpoints, coleções vetoriais, índices de embedding, instâncias de computação) aplicam controles de acesso baseados em função com listas de permissão explícitas e políticas de negação padrão.

#5.2.2 Nível: 1 Papel: V

Verifique se todas as modificações no controle de acesso são registradas de forma imutável com carimbos de data e hora, identidades dos atores, identificadores dos recursos e alterações nas permissões.

#5.2.3 Nível: 2 Papel: D

Verifique se os rótulos de classificação de dados (PII, PHI, proprietário, etc.) são automaticamente propagados para os recursos derivados (incorporações, caches de prompt, saídas do modelo).

#5.2.4 Nível: 2 Papel: D/v

Verifique se as tentativas de acesso não autorizadas e os eventos de escalonamento de privilégios acionam alertas em tempo real com metadados contextuais.

#5.2.5 Nível: 1 Papel: D/v

Verifique se as decisões de autorização são externalizadas para um mecanismo de políticas dedicado (OPA, Cedar ou equivalente).

#5.2.6 Nível: 1 Papel: D/v

Verifique se as políticas avaliam atributos dinâmicos em tempo de execução, incluindo função ou grupo do usuário, classificação do recurso, contexto da solicitação, isolamento de locatário e restrições temporais.

#5.2.7 Nível: 3 Papel: D/v

Verifique se os valores de tempo de vida (TTL) do cache de políticas não excedem 5 minutos para recursos de alta sensibilidade e 1 hora para recursos padrão com capacidades de invalidação de cache.

C5.3 Aplicação de Segurança em Tempo de Consulta

Implemente controles de segurança na camada de banco de dados com filtragem obrigatória e políticas de segurança em nível de linha.

#5.3.1 Nível: 1 Papel: D/v

Verifique se todas as consultas em banco de dados vetorial e SQL incluem filtros de segurança obrigatórios (ID do locatário, rótulos de sensibilidade, escopo do usuário) aplicados no nível do mecanismo do banco de dados.

#5.3.2 Nível: 1 Papel: D/v

Verifique se as políticas de segurança em nível de linha e o mascaramento em nível de campo estão habilitados com herança de políticas para todos os bancos de dados vetoriais, índices de busca e conjuntos de dados de treinamento.

#5.3.3 Nível: 2 Papel: D

Verifique se as avaliações de autorização falhadas abortarão imediatamente as consultas e retornarão códigos de erro de autorização explícitos.

#5.3.4 Nível: 3 Papel: D/v

Verifique se os mecanismos de nova tentativa de consulta reavaliam as políticas de autorização para considerar mudanças dinâmicas de permissão dentro de sessões de usuário ativas.

C5.4 Filtragem de Saída e Prevenção de Perda de Dados

Implemente controles de pós-processamento para evitar a exposição não autorizada de dados em conteúdo gerado por IA.

#5.4.1 Nível: 1 Papel: D/v

Verifique se os mecanismos de filtragem pós-inferência escaneiam e redigem informações pessoais identificáveis (PII) não autorizadas, informações classificadas e dados proprietários antes de entregar o conteúdo aos solicitantes.

#5.4.2 Nível: 1 Papel: D/v

Verifique se citações, referências e atribuições de fontes nas saídas do modelo são validadas de acordo com os direitos do chamador e removidas se for detectado acesso não autorizado.

#5.4.3 Nível: 2 Papel: D

Verifique se as restrições de formato de saída (PDFs sanitizados, imagens com metadados removidos, tipos de arquivo aprovados) são aplicadas com base nos níveis de permissão do usuário e classificações de dados.

C5.5 Isolamento Multi-Inquilino

Garantir isolamento criptográfico e lógico entre os inquilinos na infraestrutura de IA compartilhada.

#5.5.1 Nível: 1 Papel: D/v

Verifique se os espaços de memória, armazenamento de embeddings, entradas de cache e arquivos temporários estão segregados por namespace para cada locatário, com purga segura na exclusão do locatário ou término da sessão.

#5.5.2 Nível: 1 Papel: D/v

Verifique se cada solicitação de API inclui um identificador de inquilino autenticado que é validado criptograficamente em relação ao contexto da sessão e às autorizações do usuário.

#5.5.3 Nível: 2 Papel: D

Verifique se as políticas de rede implementam regras de negação padrão para comunicação entre locatários em malhas de serviço e plataformas de orquestração de contêineres.

#5.5.4 Nível: 3 Papel: D

Verifique se as chaves de criptografia são exclusivas por locatário com suporte para chave gerenciada pelo cliente (CMK) e isolamento criptográfico entre os armazenamentos de dados dos locatários.

C5.6 Autorização de Agente Autônomo

Controle permissões para agentes de IA e sistemas autônomos por meio de tokens de capacidade com escopo e autorização contínua.

#5.6.1 Nível: 1 Papel: D/v

Verifique se os agentes autônomos recebem tokens de capacidade delimitada que enumeram explicitamente as ações permitidas, os recursos acessíveis, os limites de tempo e as restrições operacionais.

#5.6.2 Nível: 1 Papel: D/v

Verifique se as capacidades de alto risco (acesso ao sistema de arquivos, execução de código, chamadas de API externas, transações financeiras) estão desativadas por padrão e requerem autorização explícita.

#5.6.3 Nível: 2 Papel: D

Verifique se os tokens de capacidade estão vinculados às sessões do usuário, incluem proteção criptográfica de integridade e garantem que não possam ser armazenados ou reutilizados em cenários offline.

#5.6.4 Nível: 2 Papel: V

Verifique se as ações iniciadas pelo agente passam por autorização por meio de um mecanismo de política ABAC.

Referências

- [NIST SP 800-162: Guide to Attribute Based Access Control \(ABAC\)](#)
- [NIST SP 800-207: Zero Trust Architecture](#)
- [NIST SP 800-63-3: Digital Identity Guidelines](#)
- [NIST IR 8360: Machine Learning for Access Control Policy Verification](#)

Segurança da Cadeia de Suprimentos C6 para Modelos, Frameworks e Ferramentas

Objetivo de Controle

Os ataques à cadeia de suprimentos de IA exploram modelos, frameworks ou conjuntos de dados de terceiros para inserir backdoors, viés ou código explorável. Esses controles fornecem rastreabilidade de ponta a ponta, gerenciamento de vulnerabilidades e monitoramento para proteger todo o ciclo de vida do modelo.

C6.1 Avaliação de Modelos Pré-treinados e Integridade da Origem

Avalie e autentique as origens, licenças e comportamentos ocultos dos modelos de terceiros antes de qualquer ajuste fino ou implantação.

#6.1.1 Nível: 1 Papel: D/V

Verifique se todo artefato de modelo de terceiros inclui um registro de origem assinado identificando o repositório de origem e o hash do commit.

#6.1.2 Nível: 1 Papel: D/V

Verifique se os modelos são verificados quanto a camadas maliciosas ou gatilhos de Trojan usando ferramentas automatizadas antes da importação.

#6.1.3 Nível: 2 Papel: D

Verifique se o fine-tuning por transferência de aprendizado passa na avaliação adversarial para detectar comportamentos ocultos.

#6.1.4 Nível: 2 Papel: V

Verifique se as licenças do modelo, as etiquetas de controle de exportação e as declarações de origem dos dados estão registradas em uma entrada ML-BOM.

#6.1.5 Nível: 3 Papel: D/V

Verifique se os modelos de alto risco (pesos carregados publicamente, criadores não verificados) permanecem em quarentena até a revisão e aprovação humana.

C6.2 Escaneamento de Frameworks e Bibliotecas

Escaneie continuamente frameworks e bibliotecas de ML em busca de CVEs e código malicioso para manter a pilha de execução segura.

#6.2.1 Nível: 1 Papel: D/V

Verifique se os pipelines de CI executam scanners de dependências em frameworks de IA e bibliotecas críticas.

#6.2.2 Nível: 1 Papel: D/V

Verifique se vulnerabilidades críticas (CVSS ≥ 7.0) bloqueiam a promoção para imagens de produção.

#6.2.3 Nível: 2 Papel: D

Verifique se a análise estática de código é executada em bibliotecas de ML bifurcadas ou fornecidas.

#6.2.4 Nível: 2 Papel: V

Verifique se as propostas de atualização do framework incluem uma avaliação de impacto de segurança que faça referência aos feeds públicos de CVE.

#6.2.5 Nível: 3 Papel: V

Verifique se os sensores em tempo de execução alertam sobre cargas inesperadas de bibliotecas dinâmicas que desviam do SBOM assinado.

C6.3 Fixação e Verificação de Dependências

Ancore todas as dependências em digests imutáveis e reproduza builds para garantir artefatos idênticos e à prova de adulteração.

#6.3.1 Nível: 1 Papel: D/V

Verifique se todos os gerenciadores de pacotes aplicam a fixação de versões por meio de arquivos de bloqueio.

#6.3.2 Nível: 1 Papel: D/V

Verifique se são usados resumos imutáveis em vez de tags mutáveis nas referências de contêineres.

#6.3.3 Nível: 2 Papel: D

Verifique se as verificações de build reproduzível comparam hashes entre execuções de CI para garantir saídas idênticas.

#6.3.4 Nível: 2 Papel: V

Verifique se as atestações de build são armazenadas por 18 meses para rastreabilidade de auditoria.

#6.3.5 Nível: 3 Papel: D

Verifique se as dependências expiradas acionam PRs automáticos para atualizar ou bifurcar versões fixadas.

C6.4 Aplicação de Fonte Confiável

Permitir downloads de artefatos somente de fontes verificadas criptograficamente e aprovadas pela organização, e bloquear todo o resto.

#6.4.1 Nível: 1 Papel: D/V

Verifique se os pesos do modelo, os conjuntos de dados e os contêineres são baixados apenas de domínios aprovados ou registros internos.

#6.4.2 Nível: 1 Papel: D/V

Verifique se as assinaturas Sigstore/Cosign validam a identidade do publicador antes que os artefatos sejam armazenados em cache localmente.

#6.4.3 Nível: 2 Papel: D

Verifique se os proxies de saída bloqueiam downloads de artefatos não autenticados para impor a política de fonte confiável.

#6.4.4 Nível: 2 Papel: V

Verifique se as listas de permissão do repositório são revisadas trimestralmente com evidências de justificativa comercial para cada entrada.

#6.4.5 Nível: 3 Papel: V

Verifique se as violações de política acionam o isolamento dos artefatos e o rollback das execuções de pipeline dependentes.

C6.5 Avaliação de Risco de Conjunto de Dados de Terceiros

Avalie conjuntos de dados externos quanto a envenenamento, viés e conformidade legal, e monitore-os ao longo de seu ciclo de vida.

#6.5.1 Nível: 1 Papel: D/V

Verifique se os conjuntos de dados externos passam por uma avaliação de risco de envenenamento (por exemplo, impressão digital de dados, detecção de valores atípicos).

#6.5.2 Nível: 1 Papel: D

Verifique se as métricas de viés (paridade demográfica, igualdade de oportunidade) são calculadas antes da aprovação do conjunto de dados.

#6.5.3 Nível: 2 Papel: V

Verifique se a origem, a linhagem e os termos de licença dos conjuntos de dados estão capturados nas entradas do ML-BOM.

#6.5.4 Nível: 2 Papel: V

Verifique se o monitoramento periódico detecta deriva ou corrupção em conjuntos de dados hospedados.

#6.5.5 Nível: 3 Papel: D

Verifique se o conteúdo não permitido (direitos autorais, informações pessoais identificáveis) é removido por meio de limpeza automatizada antes do treinamento.

C6.6 Monitoramento de Ataques à Cadeia de Suprimentos

Detecte ameaças à cadeia de suprimentos cedo por meio de feeds CVE, análises de registros de auditoria e simulações de red team.

#6.6.1 Nível: 1 Papel: V

Verifique se os logs de auditoria de CI/CD são transmitidos para o SIEM para detecção de puxões de pacotes anômalos ou etapas de construção adulteradas.

#6.6.2 Nível: 2 Papel: D

Verifique se os playbooks de resposta a incidentes incluem procedimentos de reversão para modelos ou bibliotecas comprometidos.

#6.6.3 Nível: 3 Papel: V

Verifique se as tags de Enriquecimento de Inteligência sobre ameaças etiquetam indicadores específicos de ML (por exemplo, IoCs de envenenamento de modelo) na triagem de alertas.

C6.7 ML-BOM para Artefatos de Modelo

Gerar e assinar SBOMs detalhados específicos para ML (ML-BOMs) para que os consumidores a jusante possam verificar a integridade dos componentes no momento da implantação.

#6.7.1 Nível: 1 Papel: D/V

Verifique se cada artefato de modelo publica um ML-BOM que lista conjuntos de dados, pesos, hiperparâmetros e licenças.

#6.7.2 Nível: 1 Papel: D/V

Verifique se a geração do ML-BOM e a assinatura Cosign estão automatizadas no CI e são obrigatórias para a mesclagem.

#6.7.3 Nível: 2 Papel: D

Verifique se as verificações de completude do ML-BOM falham a compilação se algum metadado do componente (hash, licença) estiver ausente.

#6.7.4 Nível: 2 Papel: V

Verifique se os consumidores a jusante podem consultar ML-BOMs via API para validar os modelos importados no momento da implantação.

#6.7.5 Nível: 3 Papel: V

Verifique se os ML-BOMs estão sob controle de versão e se são comparados para detectar modificações não autorizadas.

Referências

- OWASP LLM03:2025 Supply Chain
- MITRE ATLAS : Supply Chain Compromise
- SBOM Overview – CISA
- CycloneDX – Machine Learning Bill of Materials

Comportamento do Modelo C7, Controle de Saída e Garantia de Segurança

Objetivo de Controle

As saídas do modelo devem ser estruturadas, confiáveis, seguras, explicáveis e continuamente monitoradas em produção. Fazer isso reduz alucinações, vazamentos de privacidade, conteúdo prejudicial e ações descontroladas, ao mesmo tempo que aumenta a confiança do usuário e a conformidade regulatória.

C7.1 Aplicação do Formato de Saída

Esquemas rigorosos, decodificação restrita e validação posterior impedem que conteúdo mal-formado ou malicioso se propague.

#7.1.1 Nível: 1 Papel: D/V

Verifique se os esquemas de resposta (por exemplo, JSON Schema) são fornecidos no prompt do sistema e se toda saída é automaticamente validada; saídas que não estejam em conformidade acionam reparo ou rejeição.

#7.1.2 Nível: 1 Papel: D/V

Verifique se a decodificação restrita (tokens de parada, regex, máximo de tokens) está ativada para evitar estouro ou canais laterais de injeção de prompt.

#7.1.3 Nível: 2 Papel: D/V

Verifique se os componentes downstream tratam as saídas como não confiáveis e as validam contra esquemas ou desserializadores seguros contra injeção.

#7.1.4 Nível: 3 Papel: V

Verifique se os eventos de saída inadequada são registrados, limitados por taxa e exibidos para monitoramento.

C7.2 Detecção e Mitigação de Alucinações

A estimativa de incerteza e as estratégias de contingência limitam respostas fabricadas.

#7.2.1 Nível: 1 Papel: D/V

Verifique se as log-probabilidades em nível de token, a auto-consistência do conjunto ou os detectores de alucinação ajustados atribuem uma pontuação de confiança a cada resposta.

#7.2.2 Nível: 1 Papel: D/V

Verifique se as respostas abaixo de um limiar de confiança configurável acionam fluxos de trabalho de fallback (por exemplo, geração aumentada por recuperação, modelo secundário ou revisão humana).

#7.2.3 Nível: 2 Papel: D/V

Verifique se os incidentes de alucinação estão marcados com metadados de causa raiz e alimentados aos pipelines de pós-morte e ajuste fino.

#7.2.4 Nível: 3 Papel: D/V

Verifique se os limiares e detectores são recalibrados após atualizações significativas do modelo ou da base de conhecimento.

#7.2.5 Nível: 3 Papel: V

Verifique se as visualizações do painel acompanham as taxas de alucinação.

C7.3 Filtragem de Segurança e Privacidade de Saída

Os filtros de políticas e a cobertura da equipe vermelha protegem os usuários e dados confidenciais.

#7.3.1 Nível: 1 Papel: D/V

Verifique se os classificadores pré e pós-geração bloqueiam conteúdo de ódio, assédio, autoagressão, extremista e sexualmente explícito alinhado à política.

#7.3.2 Nível: 1 Papel: D/V

Verifique se a detecção de PII/PCI e a redação automática são executadas em todas as respostas; violações geram um incidente de privacidade.

#7.3.3 Nível: 2 Papel: D

Verifique se as etiquetas de confidencialidade (por exemplo, segredos comerciais) se propagam entre modalidades para evitar vazamentos em texto, imagens ou código.

#7.3.4 Nível: 3 Papel: D/V

Verifique se as tentativas de contornar o filtro ou classificações de alto risco exigem aprovação secundária ou reautenticação do usuário.

#7.3.5 Nível: 3 Papel: D/V

Verifique se os limites de filtragem refletem as jurisdições legais e o contexto da idade/papel do usuário.

C7.4 Limitação de Saída e Ação

Limites de taxa e barreiras de aprovação previnem abusos e autonomia excessiva.

#7.4.1 Nível: 1 Papel: D

Verifique se as cotas por usuário e por chave de API limitam requisições, tokens e custo com recuo exponencial em erros 429.

#7.4.2 Nível: 1 Papel: D/V

Verifique se as ações privilegiadas (gravações de arquivos, execução de código, chamadas de rede) exigem aprovação baseada em políticas ou intervenção humana.

#7.4.3 Nível: 2 Papel: D/V

Verifique se as verificações de consistência cross-modal garantem que imagens, código e texto gerados para a mesma solicitação não possam ser usados para contrabandear conteúdo malicioso.

#7.4.4 Nível: 2 Papel: D

Verifique se a profundidade de delegação do agente, os limites de recursão e as listas de ferramentas permitidas estão configurados explicitamente.

#7.4.5 Nível: 3 Papel: V

Verifique se a violação de limites gera eventos estruturados de segurança para ingestão pelo SIEM.

C7.5 Explicabilidade da Saída

Sinais transparentes melhoram a confiança do usuário e a depuração interna.

#7.5.1 Nível: 2 Papel: D/V

Verifique se as pontuações de confiança voltadas para o usuário ou resumos breves de raciocínio são exibidos quando a avaliação de risco considera apropriado.

#7.5.2 Nível: 2 Papel: D/V

Verifique se as explicações geradas evitam revelar prompts sensíveis do sistema ou dados proprietários.

#7.5.3 Nível: 3 Papel: D

Verifique se o sistema captura log-probabilidades em nível de token ou mapas de atenção e os armazena para inspeção autorizada.

#7.5.4 Nível: 3 Papel: V

Verifique se os artefatos de explicabilidade estão sob controle de versão juntamente com os lançamentos do modelo para garantir auditabilidade.

C7.6 Integração de Monitoramento

A observabilidade em tempo real fecha o ciclo entre desenvolvimento e produção.

#7.6.1 Nível: 1 Papel: D

Verifique se as métricas (violações de esquema, taxa de alucinação, toxicidade, vazamento de PII, latência, custo) são transmitidas para uma plataforma central de monitoramento.

#7.6.2 Nível: 1 Papel: V

Verifique se os limites de alerta estão definidos para cada métrica de segurança, com caminhos de escalonamento para plantão.

#7.6.3 Nível: 2 Papel: V

Verifique se os dashboards correlacionam anomalias de saída com modelo/versão, sinalizador de recurso e mudanças nos dados upstream.

#7.6.4 Nível: 2 Papel: D/V

Verifique se os dados de monitoramento retornam para o re-treinamento, ajuste fino ou atualizações de regras dentro de um fluxo de trabalho de MLOps documentado.

#7.6.5 Nível: 3 Papel: V

Verifique se os pipelines de monitoramento são submetidos a testes de penetração e possuem controle de acesso para evitar o vazamento de logs sensíveis.

7.7 Salvaguardas de Mídia Generativa

Garantir que os sistemas de IA não gerem conteúdo midiático ilegal, nocivo ou não autorizado, aplicando restrições de política, validação de saída e rastreabilidade.

#7.7.1 Nível: 1 Papel: D/V

Verifique se as instruções do sistema e do usuário proíbem explicitamente a geração de mídias deepfake ilegais, prejudiciais ou não consensuais (por exemplo, imagem, vídeo, áudio).

#7.7.2 Nível: 2 Papel: D/V

Verifique se os prompts são filtrados para tentativas de gerar personificações, deepfakes sexualmente explícitos ou mídia que retrate indivíduos reais sem consentimento.

#7.7.3 Nível: 2 Papel: V

Verifique se o sistema utiliza hashing perceptual, detecção de marca d'água ou identificação digital para impedir a reprodução não autorizada de mídia protegida por direitos autorais.

#7.7.4 Nível: 3 Papel: D/v

Verifique se toda a mídia gerada está assinada criptograficamente, com marca d'água ou incorporada com metadados de proveniência resistentes a adulteração para rastreabilidade subsequente.

#7.7.5 Nível: 3 Papel: V

Verifique se as tentativas de bypass (por exemplo, ofuscação de prompt, gírias, frases adversariais) são detectadas, registradas e limitadas em taxa; abusos repetidos são reportados aos sistemas de monitoramento.

Referências

- NIST AI Risk Management Framework
- ISO/IEC 42001:2023 – AI Management System
- OWASP Top-10 for Large Language Model Applications (2025)
- Practical Techniques to Constrain LLM Output
- Dataiku – Structured Text Generation Guide
- VL-Uncertainty: Detecting Hallucinations
- HaDeMiF: Hallucination Detection & Mitigation
- Building Confidence in LLM Outputs
- Explainable AI & LLMs
- Sensitive Information Disclosure in LLMs
- OpenAI Rate-Limit & Exponential Back-off
- Arize AI – LLM Observability Platform

Segurança de Memória C8, Incorporações e Banco de Dados Vetoriais

Objetivo de Controle

Incorporação e armazenamentos vetoriais atuam como a "memória ao vivo" dos sistemas de IA contemporâneos, aceitando continuamente dados fornecidos pelo usuário e reutilizando-os nos contextos do modelo por meio da Geração Aumentada por Recuperação (RAG). Se deixada sem controle, essa memória pode expor Informações Pessoais Identificáveis (PII), violar consentimentos ou ser revertida para reconstruir o texto original. O objetivo desta família de controles é reforçar os pipelines de memória e bancos de dados vetoriais para que o acesso seja no mínimo privilégio, as incorporações preservem a privacidade, os vetores armazenados expirem ou possam ser revogados sob demanda, e a memória por usuário nunca contamine os prompts ou respostas de outro usuário.

C8.1 Controles de Acesso à Memória e Índices RAG

Impõe controles de acesso granulares em cada coleção de vetores.

#8.1.1 Nível: 1 Papel: D/V

Verifique se as regras de controle de acesso a nível de linha/namespace restringem as operações de inserção, exclusão e consulta por locatário, coleção ou tag de documento.

#8.1.2 Nível: 1 Papel: D/V

Verifique se as chaves de API ou JWTs possuem claims com escopo definido (por exemplo, IDs de coleção, verbos de ação) e são rotacionados pelo menos trimestralmente.

#8.1.3 Nível: 2 Papel: D/V

Verifique se as tentativas de escalonamento de privilégios (por exemplo, consultas de similaridade entre namespaces) são detectadas e registradas em um SIEM dentro de 5 minutos.

#8.1.4 Nível: 2 Papel: D/V

Verifique se o banco de dados vetorial registra nos logs o identificador do sujeito, a operação, o ID/nome do namespace do vetor, o limiar de similaridade e a contagem de resultados.

#8.1.5 Nível: 3 Papel: V

Verifique se as decisões de acesso são testadas para falhas de contorno sempre que os motores são atualizados ou as regras de fragmentação de índice mudam.

C8.2 Sanitização e Validação de Embeddings

Pré-analise o texto para informações pessoais identificáveis (PII), oculte ou pseudonimize antes da vetorização e, opcionalmente, pós-processe os embeddings para remover sinais residuais.

#8.2.1 Nível: 1 Papel: D/V

Verifique se os dados PII e regulados são detectados por meio de classificadores automatizados e mascaraados, tokenizados ou descartados antes da incorporação.

#8.2.2 Nível: 1 Papel: D

Verifique se os pipelines de incorporação rejeitam ou colocam em quarentena entradas contendo código executável ou artefatos não UTF-8 que possam contaminar o índice.

#8.2.3 Nível: 2 Papel: D/v

Verifique se a sanitização de privacidade diferencial local ou métrica é aplicada a embeddings de sentença cuja distância para qualquer token PII conhecido esteja abaixo de um limiar configurável.

#8.2.4 Nível: 2 Papel: V

Verifique se a eficácia da sanitização (por exemplo, recall da redação de PII, deriva semântica) é validada pelo menos semestralmente contra corpora de referência.

#8.2.5 Nível: 3 Papel: D/v

Verifique se as configurações de sanitização estão sob controle de versão e se as alterações passam por revisão por pares.

C8.3 Expiração, Revogação e Exclusão de Memória

O "direito de ser esquecido" do GDPR e leis similares exigem apagamento em tempo hábil; portanto, os armazenamentos vetoriais devem suportar TTLs, exclusões definitivas e tomb-stoning para que vetores revogados não possam ser recuperados ou reindexados.

#8.3.1 Nível: 1 Papel: D/v

Verifique se todo vetor e registro de metadados possui um TTL ou rótulo de retenção explícito respeitado por trabalhos automatizados de limpeza.

#8.3.2 Nível: 1 Papel: D/v

Verifique se as solicitações de exclusão iniciadas pelo usuário eliminam vetores, metadados, cópias em cache e índices derivados dentro de 30 dias.

#8.3.3 Nível: 2 Papel: D

Verifique se as exclusões lógicas são seguidas pela fragmentação criptográfica dos blocos de armazenamento, caso o hardware suporte, ou pela destruição da chave do cofre de chaves.

#8.3.4 Nível: 3 Papel: D/v

Verifique se os vetores expirados são excluídos dos resultados da busca pelo vizinho mais próximo em menos de 500 ms após a expiração.

C8.4 Prevenir Inversão e Vazamento de Embeddings

Defesas recentes—sobreposição de ruído, redes de projeção, perturbação de neurônios de privacidade e criptografia em camada de aplicação—podem reduzir as taxas de inversão em nível de token para menos de 5%.

#8.4.1 Nível: 1 Papel: V

Verifique se existe um modelo formal de ameaça que abranja ataques de inversão, de associação e de inferência de atributos, e se ele é revisado anualmente.

#8.4.2 Nível: 2 Papel: D/v

Verifique se a criptografia na camada de aplicação ou a criptografia pesquisável protegem os vetores contra leituras diretas por administradores de infraestrutura ou funcionários da nuvem.

#8.4.3 Nível: 3 Papel: V

Verifique se os parâmetros de defesa (ϵ para DP, ruído σ , posto de projeção k) equilibram privacidade $\geq 99\%$ de proteção de token e utilidade $\leq 3\%$ de perda de precisão.

#8.4.4 Nível: 3 Papel: D/v

Verifique se as métricas de resistência à inversão fazem parte dos critérios de liberação para atualizações do modelo, com orçamentos de regressão definidos.

C8.5 Aplicação de Escopo para Memória Específica do Usuário

O vazamento entre inquilinos continua sendo um dos principais riscos de RAG: consultas de similaridade mal filtradas podem revelar documentos privados de outro cliente.

#8.5.1 Nível: 1 Papel: D/v

Verifique se toda consulta de recuperação é pós-filtrada pelo ID do locatário/usuário antes de ser passada para o prompt do LLM.

#8.5.2 Nível: 1 Papel: D

Verifique se os nomes das coleções ou IDs com namespace são salinizados por usuário ou locatário para que os vetores não possam colidir entre diferentes escopos.

#8.5.3 Nível: 2 Papel: D/v

Verifique se os resultados de similaridade acima de um limite de distância configurável, mas fora do escopo do chamador, são descartados e acionam alertas de segurança.

#8.5.4 Nível: 2 Papel: V

Verifique se os testes de estresse multi-inquilino simulam consultas adversariais que tentam recuperar documentos fora do escopo e demonstram zero vazamento.

#8.5.5 Nível: 3 Papel: D/v

Verifique se as chaves de criptografia são segregadas por locatário, garantindo isolamento criptográfico mesmo que o armazenamento físico seja compartilhado.

C8.6 Segurança Avançada do Sistema de Memória

Controles de segurança para arquiteturas de memória sofisticadas, incluindo memória episódica, semântica e de trabalho, com requisitos específicos de isolamento e validação.

#8.6.1 Nível: 1 Papel: D/v

Verifique se os diferentes tipos de memória (episódica, semântica, de trabalho) possuem contextos de segurança isolados com controles de acesso baseados em função, chaves de criptografia separadas e padrões de acesso documentados para cada tipo de memória.

#8.6.2 Nível: 2 Papel: D/v

Verifique se os processos de consolidação de memória incluem validação de segurança para evitar a injeção de memórias maliciosas por meio de sanitização de conteúdo, verificação de origem e checa-

gens de integridade antes do armazenamento.

#8.6.3 Nível: 2 Papel: D/V

Verifique se as consultas de recuperação de memória são validadas e sanitizadas para evitar a extração de informações não autorizadas por meio de análise do padrão da consulta, aplicação de controle de acesso e filtragem de resultados.

#8.6.4 Nível: 3 Papel: D/V

Verifique se os mecanismos de esquecimento de memória excluem com segurança informações sensíveis com garantias de eliminação criptográfica, utilizando exclusão de chaves, sobreescrita múltipla ou exclusão segura baseada em hardware com certificados de verificação.

#8.6.5 Nível: 3 Papel: D/V

Verifique se a integridade do sistema de memória é continuamente monitorada para modificações não autorizadas ou corrupção por meio de somas de verificação, registros de auditoria e alertas automatizados quando o conteúdo da memória muda fora das operações normais.

Referências

- Vector database security: Pinecone – IronCore Labs
- Securing the Backbone of AI: Safeguarding Vector Databases and Embeddings – Privacyacer
- Enhancing Data Security with RBAC of Qdrant Vector Database – AI Advances
- Mitigating Privacy Risks in LLM Embeddings from Embedding Inversion – arXiv
- DPPN: Detecting and Perturbing Privacy-Sensitive Neurons – OpenReview
- Art. 17 GDPR – Right to Erasure
- Sensitive Data in Text Embeddings Is Recoverable – Tonic.ai
- PII Identification and Removal – NVIDIA NeMo Docs
- De-identifying Sensitive Data – Google Cloud DLP
- Remove PII from Conversations Using Sensitive Information Filters – AWS Bedrock Guidelines
- Think Your RAG Is Secure? Think Again – Medium
- Design a Secure Multitenant RAG Inferencing Solution – Microsoft Learn
- Best Practices for Multi-Tenancy RAG with Milvus – Milvus Blog

9 Orquestração Autônoma e Segurança da Ação Agente

Objetivo de Controle

Garantir que sistemas de IA autônomos ou multiagentes possam executar apenas ações que sejam explicitamente intencionadas, autenticadas, auditáveis e dentro de limites definidos de custo e risco. Isso protege contra ameaças como Comprometimento de Sistema Autônomo, Uso Indevido de Ferramentas, Detecção de Loop de Agentes, Sequestro de Comunicação, Falsificação de Identidade, Manipulação de Enxame e Manipulação de Intenção.

9.1 Orçamentos para Planejamento de Tarefas do Agente e Recurso

Controle planos recursivos e force pontos de verificação humanos para ações privilegiadas.

#9.1.1 Nível: 1 Papel: D/V

Verifique se a profundidade máxima de recursão, amplitude, tempo de relógio de parede, tokens e custo monetário por execução de agente estão configurados centralmente e controlados por versão.

#9.1.2 Nível: 1 Papel: D/V

Verifique se ações privilegiadas ou irreversíveis (por exemplo, commits de código, transferências financeiras) exigem aprovação humana explícita por meio de um canal auditável antes da execução.

#9.1.3 Nível: 2 Papel: D

Verifique se os monitores de recursos em tempo real acionam a interrupção do disjuntor quando qualquer limite de orçamento é excedido, interrompendo a expansão adicional das tarefas.

#9.1.4 Nível: 2 Papel: D/V

Verifique se os eventos do disjuntor estão registrados com o ID do agente, condição que acionou e o estado do plano capturado para revisão forense.

#9.1.5 Nível: 3 Papel: V

Verifique se os testes de segurança cobrem os cenários de exaustão do orçamento e plano descontrolado, confirmando a parada segura sem perda de dados.

#9.1.6 Nível: 3 Papel: D

Verifique se as políticas orçamentárias estão expressas como política-como-código e aplicadas no CI/CD para bloquear a deriva de configuração.

9.2 Isolamento de Plugin de Ferramenta

Isolar as interações da ferramenta para evitar acesso não autorizado ao sistema ou execução de código.

#9.2.1 Nível: 1 Papel: D/V

Verifique se cada ferramenta/plugin é executado dentro de um SO, contêiner ou sandbox em nível WASM com políticas de sistema de arquivos, rede e chamadas do sistema com privilégios mínimos.

#9.2.2 Nível: 1 Papel: D/V

Verifique se as cotas de recursos do sandbox (CPU, memória, disco, saída de rede) e os tempos limite de execução são aplicados e registrados.

#9.2.3 Nível: 2 Papel: D/v

Verifique se os binários ou descritores da ferramenta estão assinados digitalmente; as assinaturas são validadas antes do carregamento.

#9.2.4 Nível: 2 Papel: V

Verifique se a telemetria do sandbox é transmitida para um SIEM; anomalias (por exemplo, tentativas de conexões de saída) geram alertas.

#9.2.5 Nível: 3 Papel: V

Verifique se os plugins de alto risco passam por revisão de segurança e testes de penetração antes do deployment em produção.

#9.2.6 Nível: 3 Papel: D/v

Verifique se as tentativas de escape da sandbox são automaticamente bloqueadas e o plugin infrator é colocado em quarentena aguardando investigação.

9.3 Loop Autônomo e Limitação de Custos

Detectar e interromper a recursão descontrolada entre agentes e explosões de custo.

#9.3.1 Nível: 1 Papel: D/v

Verifique se as chamadas entre agentes incluem um limite de saltos ou TTL que o tempo de execução decrementa e aplica.

#9.3.2 Nível: 2 Papel: D

Verifique se os agentes mantêm um ID único de gráfico de invocação para detectar auto-invocação ou padrões cílicos.

#9.3.3 Nível: 2 Papel: D/v

Verifique que os contadores cumulativos de unidades de computação e gastos sejam monitorados por cadeia de requisições; ultrapassar o limite aborta a cadeia.

#9.3.4 Nível: 3 Papel: V

Verifique se a análise formal ou a verificação de modelo demonstra a ausência de recursão ilimitada nos protocolos do agente.

#9.3.5 Nível: 3 Papel: D

Verifique se os eventos de aborto de loop geram alertas e alimentam métricas de melhoria contínua.

9.4 Proteção contra Uso Indevido em Nível de Protocolo

Canais de comunicação seguros entre agentes e sistemas externos para prevenir sequestro ou manipulação.

#9.4.1 Nível: 1 Papel: D/v

Verifique se todas as mensagens de agente para ferramenta e de agente para agente estão autenticadas (por exemplo, TLS mútuo ou JWT) e criptografadas de ponta a ponta.

#9.4.2 Nível: 1 Papel: D

Verifique se os esquemas são rigorosamente validados; campos desconhecidos ou mensagens malfor-

madas são rejeitados.

#9.4.3 Nível: 2 Papel: D/V

Verifique se as verificações de integridade (MACs ou assinaturas digitais) cobrem toda a carga útil da mensagem, incluindo os parâmetros da ferramenta.

#9.4.4 Nível: 2 Papel: D

Verifique se a proteção contra reprodução (nonces ou janelas de carimbo de data/hora) está aplicada na camada de protocolo.

#9.4.5 Nível: 3 Papel: V

Verifique se as implementações de protocolo passam por fuzzing e análise estática para identificar falhas de injeção ou desserialização.

9.5 Identidade do Agente e Evidência de Violação

Assegure que as ações sejam atribuíveis e as modificações detectáveis.

#9.5.1 Nível: 1 Papel: D/V

Verifique se cada instância de agente possui uma identidade criptográfica única (par de chaves ou credencial ancorada em hardware).

#9.5.2 Nível: 2 Papel: D/V

Verifique se todas as ações dos agentes são assinadas e carimbadas com data e hora; os logs devem incluir a assinatura para não repúdio.

#9.5.3 Nível: 2 Papel: V

Verifique se os logs à prova de adulteração são armazenados em um meio somente de anexação ou gravação única.

#9.5.4 Nível: 3 Papel: D

Verifique se as chaves de identidade giram em uma programação definida e conforme os indicadores de comprometimento.

#9.5.5 Nível: 3 Papel: D/V

Verifique se tentativas de falsificação ou conflito de chaves acionam a quarentena imediata do agente afetado.

9.6 Redução de Risco em Enxame Multiagente

Mitigue riscos de comportamento coletivo por meio de isolamento e modelagem formal de segurança.

#9.6.1 Nível: 1 Papel: D/V

Verifique se os agentes que operam em diferentes domínios de segurança executam em sandboxes de tempo de execução isolados ou em segmentos de rede.

#9.6.2 Nível: 3 Papel: V

Verifique se os comportamentos do enxame são modelados e formalmente verificados quanto à vivacidade e segurança antes da implantação.

#9.6.3 Nível: 3 Papel: D

Verifique se os monitores em tempo de execução detectam padrões emergentes inseguros (por

exemplo, oscilações, deadlocks) e iniciam ações corretivas.

9.7 Autenticação / Autorização de Usuário e Ferramenta

Implemente controles de acesso robustos para cada ação acionada por agentes.

#9.7.1 Nível: 1 Papel: D/V

Verifique se os agentes se autenticam como principais de primeira classe nos sistemas descendentes, nunca reutilizando as credenciais do usuário final.

#9.7.2 Nível: 2 Papel: D

Verifique se as políticas de autorização granulares restringem quais ferramentas um agente pode invocar e quais parâmetros ele pode fornecer.

#9.7.3 Nível: 2 Papel: V

Verifique se as verificações de privilégios são reavaliadas a cada chamada (autorização contínua), e não somente no início da sessão.

#9.7.4 Nível: 3 Papel: D

Verifique se os privilégios delegados expiram automaticamente e exigem novo consentimento após o tempo limite ou alteração do escopo.

9.8 Segurança da Comunicação Agente-para-Agente

Criptografe e proteja a integridade de todas as mensagens entre agentes para impedir espiagem e adulteração.

#9.8.1 Nível: 1 Papel: D/V

Verifique se a autenticação mútua e a criptografia com segredo perfeito à frente (por exemplo, TLS 1.3) são obrigatórias para os canais do agente.

#9.8.2 Nível: 1 Papel: D

Verifique se a integridade e a origem da mensagem são validadas antes do processamento; falhas geram alertas e rejeitam a mensagem.

#9.8.3 Nível: 2 Papel: D/V

Verifique se os metadados de comunicação (carimbos de data e hora, números de sequência) são registrados para apoiar a reconstrução forense.

#9.8.4 Nível: 3 Papel: V

Verifique se a verificação formal ou a checagem de modelos confirma que as máquinas de estado do protocolo não podem ser levadas a estados inseguros.

9.9 Verificação de Intenção e Aplicação de Restrições

Validar que as ações do agente estejam alinhadas com a intenção declarada pelo usuário e as restrições do sistema.

#9.9.1 Nível: 1 Papel: D

Verifique se os solucionadores de restrições pré-execução verificam as ações propostas em relação às regras rígidas de segurança e políticas codificadas.

#9.9.2 Nível: 2 Papel: D/v

Verifique se ações de alto impacto (financeiras, destrutivas, sensíveis à privacidade) exigem confirmação explícita de intenção por parte do usuário que as iniciou.

#9.9.3 Nível: 2 Papel: V

Verifique se as verificações de pós-condição confirmam que as ações concluídas alcançaram os efeitos pretendidos sem efeitos colaterais; discrepâncias acionam o rollback.

#9.9.4 Nível: 3 Papel: V

Verifique se os métodos formais (por exemplo, verificação de modelos, demonstração de teoremas) ou testes baseados em propriedades demonstram que os planos do agente satisfazem todas as restrições declaradas.

#9.9.5 Nível: 3 Papel: D

Verifique se incidentes de incompatibilidade de intenção ou violação de restrições alimentam ciclos de melhoria contínua e compartilhamento de inteligência sobre ameaças.

9.10 Segurança da Estratégia de Raciocínio do Agente

Seleção e execução segura de diferentes estratégias de raciocínio, incluindo abordagens ReAct, Chain-of-Thought e Tree-of-Thoughts.

#9.10.1 Nível: 1 Papel: D/v

Verifique se a seleção da estratégia de raciocínio utiliza critérios determinísticos (complexidade da entrada, tipo de tarefa, contexto de segurança) e se entradas idênticas produzem seleções de estratégia idênticas dentro do mesmo contexto de segurança.

#9.10.2 Nível: 1 Papel: D/v

Verifique se cada estratégia de raciocínio (ReAct, Cadeia de Pensamento, Árvore de Pensamentos) possui validação de entrada dedicada, sanitização de saída e limites de tempo de execução específicos para sua abordagem cognitiva.

#9.10.3 Nível: 2 Papel: D/v

Verifique se as transições da estratégia de raciocínio são registradas com contexto completo, incluindo características de entrada, valores dos critérios de seleção e metadados de execução para a reconstrução da trilha de auditoria.

#9.10.4 Nível: 2 Papel: D/v

Verifique se o raciocínio Tree-of-Thoughts inclui mecanismos de poda de ramos que encerram a exploração quando são detectadas violações de políticas, limites de recursos ou limites de segurança.

#9.10.5 Nível: 2 Papel: D/v

Verifique se os ciclos ReAct (Raciocínio-Ação-Observação) incluem pontos de verificação de validação em cada fase: verificação do passo de raciocínio, autorização da ação e sanitização da observação antes de prosseguir.

#9.10.6 Nível: 3 Papel: D/v

Verifique se as métricas de desempenho da estratégia de raciocínio (tempo de execução, uso de recursos, qualidade da saída) são monitoradas com alertas automatizados quando as métricas se desviam além dos limites configurados.

#9.10.7 Nível: 3 Papel: D/v

Verifique se as abordagens de raciocínio híbrido que combinam múltiplas estratégias mantêm a validação de entrada e as restrições de saída de todas as estratégias constituintes sem contornar quaisquer controles de segurança.

#9.10.8 Nível: 3 Papel: D/V

Verifique se o teste de segurança da estratégia de raciocínio inclui fuzzing com entradas malformadas, prompts adversariais projetados para forçar a troca de estratégia e testes de condição limite para cada abordagem cognitiva.

9.11 Gerenciamento do Estado do Ciclo de Vida do Agente e Segurança

Inicialização segura do agente, transições de estado e término com trilhas de auditoria criptográficas e procedimentos definidos de recuperação.

#9.11.1 Nível: 1 Papel: D/V

Verifique se a inicialização do agente inclui o estabelecimento de identidade criptográfica com credenciais suportadas por hardware e logs de auditoria imutáveis de inicialização contendo ID do agente, carimbo de data/hora, hash de configuração e parâmetros de inicialização.

#9.11.2 Nível: 2 Papel: D/V

Verifique se as transições de estado do agente são assinadas criptograficamente, registradas com carimbo de data/hora e logadas com contexto completo, incluindo eventos que as desencadearam, hash do estado anterior, hash do novo estado e validações de segurança realizadas.

#9.11.3 Nível: 2 Papel: D/V

Verifique se os procedimentos de desligamento do agente incluem limpeza segura da memória usando apagamento criptográfico ou sobrescrição múltipla, revogação de credenciais com notificação à autoridade certificadora e geração de certificados de término à prova de violação.

#9.11.4 Nível: 3 Papel: D/V

Verifique se os mecanismos de recuperação do agente validam a integridade do estado usando checksums criptográficos (mínimo SHA-256) e fazem rollback para estados conhecidos como confiáveis quando a corrupção é detectada, com alertas automatizados e requisitos de aprovação manual.

#9.11.5 Nível: 3 Papel: D/V

Verifique se os mecanismos de persistência do agente criptografam dados de estado sensíveis com chaves AES-256 por agente e implementam rotação segura de chaves em cronogramas configuráveis (máximo de 90 dias) com implantação sem tempo de inatividade.

9.12 Estrutura de Segurança para Integração de Ferramentas

Controles de segurança para carregamento dinâmico de ferramentas, execução e validação de resultados com processos definidos de avaliação de risco e aprovação.

#9.12.1 Nível: 1 Papel: D/V

Verifique se os descritores da ferramenta incluem metadados de segurança especificando privilégios necessários (leitura/gravação/execução), níveis de risco (baixo/médio/alto), limites de recursos (CPU, memória, rede) e requisitos de validação documentados nos manifestos da ferramenta.

#9.12.2 Nível: 1 Papel: D/v

Verifique se os resultados da execução da ferramenta são validados contra os esquemas esperados (Esquema JSON, Esquema XML) e políticas de segurança (sanitização de saída, classificação de dados) antes da integração, com limites de tempo e procedimentos de tratamento de erros.

#9.12.3 Nível: 2 Papel: D/v

Verifique se os logs de interação da ferramenta incluem contexto detalhado de segurança, incluindo uso de privilégios, padrões de acesso a dados, tempo de execução, consumo de recursos e códigos de retorno, com logging estruturado para integração com SIEM.

#9.12.4 Nível: 2 Papel: D/v

Verifique se os mecanismos de carregamento dinâmico de ferramentas validam assinaturas digitais usando a infraestrutura de PKI e implementam protocolos de carregamento seguro com isolamento em sandbox e verificação de permissões antes da execução.

#9.12.5 Nível: 3 Papel: D/v

Verifique se as avaliações de segurança da ferramenta são acionadas automaticamente para novas versões com portas de aprovação obrigatórias, incluindo análise estática, testes dinâmicos e revisão da equipe de segurança, com critérios de aprovação documentados e requisitos de SLA.

C9.13 Protocolo de Contexto de Modelo (MCP) Segurança

Garantir a descoberta, autenticação, autorização, transporte e uso seguros das integrações de ferramentas e recursos baseados em MCP para evitar confusão de contexto, invocação não autorizada de ferramentas ou exposição de dados entre diferentes locatários.

Integridade do Componente e Higiene da Cadeia de Suprimentos

#9.13.1 Nível: 1 Papel: D/v

Verifique se as implementações do servidor, cliente e ferramenta MCP são revisadas manualmente ou analisadas automaticamente para identificar exposição de funções inseguras, configurações padrão inseguras, falta de autenticação ou ausência de validação de entrada.

#9.13.2 Nível: 1 Papel: D/v

Verifique se servidores ou pacotes MCP externos ou de código aberto passam por varredura automatizada de vulnerabilidades e da cadeia de suprimentos (por exemplo, SCA) antes da integração, e que componentes com vulnerabilidades críticas conhecidas não são utilizados.

#9.13.3 Nível: 1 Papel: D/v

Verifique se os componentes do servidor e cliente MCP são obtidos apenas de fontes confiáveis e verificados usando assinaturas, somas de verificação ou metadados seguros de pacote, rejeitando versões adulteradas ou não assinadas.

Autenticação e Autorização

#9.13.4 Nível: 2 Papel: D/v

Verifique se os clientes e servidores MCP autenticam-se mutuamente usando credenciais fortes, que não sejam de usuário (por exemplo, mTLS, tokens assinados ou identidades emitidas pela plataforma), e que endpoints MCP não autenticados sejam rejeitados.

#9.13.5 Nível: 2 Papel: D/v

Verifique se os servidores MCP estão registrados por meio de um mecanismo controlado de integração

técnica que exija definições explícitas de proprietário, ambiente e recurso; servidores não registrados ou indisponíveis para descoberta não devem ser acessíveis em produção.

#9.13.6 Nível: 2 Papel: D/V

Verifique se cada ferramenta ou recurso MCP define escopos de autorização explícitos (por exemplo, somente leitura, consultas restritas, níveis de efeito colateral) e se os agentes não podem invocar funções MCP fora de seu escopo atribuído.

Transporte Seguro e Proteção da Fronteira de Rede

#9.13.7 Nível: 2 Papel: D/V

Verifique se o HTTP transmissível autenticado e criptografado é usado como o principal transporte MCP em ambientes de produção; transportes alternativos (stdio, SSE) são restritos a ambientes locais ou rigorosamente controlados com justificativa explícita.

#9.13.8 Nível: 2 Papel: D/V

Verifique se os transportes streamable-HTTP MCP utilizam canais autenticados e criptografados (TLS 1.3 ou superior) com validação de certificado e segredo à frente para garantir a confidencialidade e integridade das mensagens MCP transmitidas.

#9.13.9 Nível: 2 Papel: D/V

Verifique se os transportes MCP baseados em SSE são usados apenas dentro de canais internos privados e autenticados e aplique TLS, autenticação, validação de esquema, limites de tamanho de payload e limitação de taxa; os endpoints SSE não devem ser expostos à internet pública.

#9.13.10 Nível: 2 Papel: D/V

Verifique se os servidores MCP validam o `Origin` e `Host` Cabeçalhos em todos os transportes baseados em HTTP (incluindo SSE e HTTP transmissível) para prevenir ataques de reatribuição de DNS e rejeitar solicitações de origens não confiáveis, incompatíveis ou ausentes.

Validação de Esquema, Mensagem e Entrada

#9.13.11 Nível: 2 Papel: D/V

Verifique se os esquemas de ferramentas e recursos do MCP (por exemplo, esquemas JSON ou descriptores de capacidade) são validados quanto à autenticidade e integridade usando assinaturas, somas de verificação ou atestação do servidor para prevenir adulteração dos esquemas ou modificação maliciosa de parâmetros.

#9.13.12 Nível: 2 Papel: D/V

Verifique se todos os transportes MCP aplicam a integridade da estrutura das mensagens, validação rigorosa de esquema, tamanhos máximos de carga útil e rejeição de quadros malformados, truncados ou intercalados para evitar ataques de dessincronização ou injeção.

#9.13.13 Nível: 2 Papel: D/V

Verifique se os servidores MCP realizam validação rigorosa de entrada para todas as chamadas de função, incluindo verificação de tipo, verificação de limites, aplicação de enumeração e rejeição de parâmetros não reconhecidos ou de tamanho excessivo.

Acesso de Saída e Segurança na Execução de Agentes

#9.13.14 Nível: 2 Papel: D/V

Verifique se os servidores MCP podem iniciar solicitações de saída apenas para destinos internos ou externos aprovados, seguindo políticas de egressos com privilégio mínimo, e não podem acessar alvos de rede arbitrários ou serviços de metadados internos da nuvem.

#9.13.15 Nível: 2 Papel: D/V

Verifique se as ações MCP de saída implementam limites de execução (tempos limite, limites de recurso, limites de concorrência, disjuntores) para evitar invocações ilimitadas de ferramentas dirigidas por agentes ou efeitos colaterais encadeados.

#9.13.16 Nível: 2 Papel: D/V

Verifique se os metadados da solicitação e resposta do MCP (ID do servidor, nome do recurso, nome da ferramenta, identificador da sessão, locatário, ambiente) são registrados com proteção de integridade e correlacionados à atividade do agente para análise forense.

Restrições de Transporte e Controles de Limite de Alto Risco

#9.13.17 Nível: 3 Papel: D/V

Verifique se os transportes MCP baseados em stdio estão limitados a cenários de desenvolvimento de processo único e co-localizados, isolados da execução do shell, injeção de terminal e capacidades de criação de processos; stdio nunca deve cruzar fronteiras de rede ou multi-inquilinos.

#9.13.18 Nível: 3 Papel: D/V

Verifique se os servidores MCP expõem apenas funções e recursos na lista de permissões, e proíbem despacho dinâmico, invocação reflexiva ou execução de nomes de funções influenciados por entradas fornecidas pelo usuário ou pelo modelo.

#9.13.19 Nível: 3 Papel: D/V

Verifique se os limites de locatário, limites de ambiente (dev/teste/prod) e limites de domínio de dados são aplicados na camada MCP, prevenindo a descoberta cruzada de servidores ou recursos entre locatários ou ambientes.

Referências

- MITRE ATLAS tactics ML09
- Circuit-breaker research for AI agents – Zou et al, 2024
- Trend Micro analysis of sandbox escapes in AI agents – Park, 2025
- Auth0 guidance on human-in-the-loop authorization for agents – Martinez, 2025
- Medium deep-dive on MCP & A2A protocol hijacking – ForAISeC, 2025
- Rapid7 fundamentals on spoofing attack prevention – 2024
- Imperial College verification of swarm systems – Lomuscio et al.
- NIST AI Risk Management Framework 1.0, 2023
- WIRED security briefing on encryption best practices, 2024
- OWASP Top 10 for LLM Applications, 2025
- Comprehensive Vulnerability Analysis is Necessary for Trustworthy LLM-MAS
- [How Is LLM Reasoning Distracted by Irrelevant Context? An Analysis Using a Controlled Benchmark](<https://www.arxiv.org/pdf/2505.18761>)
- Large Language Model Sentinel: LLM Agent for Adversarial Purification
- Securing Agentic AI: A Comprehensive Threat Model and Mitigation Framework for Generative AI Agents
- Model Context Protocol Specification
- Model Context Protocol Tools & Resources Specification
- Model Context Protocol Transport Documentation
- OWASP GenAI Security Project – “A Practical Guide for Securely Using Third-Party MCP Servers 1.0”
- Cloud Security Alliance – Model Context Protocol Security Working Group

- CSA MCP Security: Top 10 Risks
- CSA MCP Security: TTPs & Hardening Guidance



10 Robustez Adversarial e Defesa de Privacidade

Objetivo de Controle

Garantir que os modelos de IA permaneçam confiáveis, preservem a privacidade e sejam resistentes a abusos ao enfrentar ataques de evasão, inferência, extração ou envenenamento.

10.1 Alinhamento e Segurança do Modelo

Proteja-se contra saídas prejudiciais ou que violem a política.

#10.1.1 Nível: 1 Papel: D/V

Verifique se um conjunto de testes de alinhamento (prompts de red-team, sondas de jailbreak, conteúdo proibido) está sob controle de versão e é executado em cada lançamento do modelo.

#10.1.2 Nível: 1 Papel: D

Verifique se as barreiras de recusa e conclusão segura estão sendo aplicadas.

#10.1.3 Nível: 2 Papel: D/V

Verifique se um avaliador automatizado mede a taxa de conteúdo prejudicial e sinaliza regressões além de um limite definido.

#10.1.4 Nível: 2 Papel: D

Verifique se o treinamento contra jailbreak está documentado e é reproduzível.

#10.1.5 Nível: 3 Papel: V

Verifique se as provas formais de conformidade com a política ou a monitorização certificada cobrem domínios críticos.

10.2 Endurecimento contra Exemplos Adversariais

Aumentar a resiliência a entradas manipuladas. Treinamento adversarial robusto e pontuação em benchmarks são as melhores práticas atuais.

#10.2.1 Nível: 1 Papel: D

Verifique se os repositórios do projeto incluem configurações de treinamento adversarial com sementes reproduzíveis.

#10.2.2 Nível: 2 Papel: D/V

Verifique se a detecção de exemplos adversariais gera alertas de bloqueio em pipelines de produção.

#10.2.4 Nível: 3 Papel: V

Verifique se as provas de robustez certificada ou os certificados de limite de intervalo cobrem pelo menos as principais classes críticas.

#10.2.5 Nível: 3 Papel: V

Verifique se os testes de regressão utilizam ataques adaptativos para confirmar a ausência de perda mensurável de robustez.

10.3 Mitigação de Inferência de Membro

Limitar a capacidade de decidir se um registro estava nos dados de treinamento. A privacidade diferencial e a máscara de pontuação de confiança continuam sendo as defesas conhecidas mais eficazes.

#10.3.1 Nível: 1 Papel: D

Verifique se a regularização de entropia por consulta ou a escala de temperatura reduz previsões excessivamente confiantes.

#10.3.2 Nível: 2 Papel: D

Verifique se o treinamento utiliza otimização diferencialmente privada com limite ϵ para conjuntos de dados sensíveis.

#10.3.3 Nível: 2 Papel: V

Verifique se as simulações de ataque (modelo sombra ou caixa-preta) mostram AUC de ataque $\leq 0,60$ em dados retidos.

10.4 Resistência à Inversão de Modelo

Evitar a reconstrução de atributos privados. Pesquisas recentes enfatizam a truncagem de saída e garantias de DP como defesas práticas.

#10.4.1 Nível: 1 Papel: D

Verifique se os atributos sensíveis nunca são diretamente exibidos; quando necessário, use categorias ou transformações unidirecionais.

#10.4.2 Nível: 1 Papel: D/V

Verifique se os limites de taxa de consulta restringem consultas adaptativas repetidas do mesmo principal.

#10.4.3 Nível: 2 Papel: D

Verifique se o modelo foi treinado com ruído preservador de privacidade.

10.5 Defesa contra extração de modelo

Detectar e impedir clonagem não autorizada. Recomenda-se marca d'água e análise de padrão de consulta.

#10.5.1 Nível: 1 Papel: D

Verifique se os gateways de inferência aplicam limites de taxa globais e por chave de API ajustados ao limite de memorização do modelo.

#10.5.2 Nível: 2 Papel: D/V

Verifique se as estatísticas de entropia da consulta e pluralidade da entrada alimentam um detector de

extração automatizado.

#10.5.3 Nível: 2 Papel: V

Verifique se marcas d'água frágeis ou probabilísticas podem ser comprovadas com $p < 0,01$ em $\leq 1\,000$ consultas contra um clone suspeito.

#10.5.4 Nível: 3 Papel: D

Verifique se as chaves de marca d'água e os conjuntos de gatilho são armazenados em um módulo de segurança de hardware e rotacionados anualmente.

#10.5.5 Nível: 3 Papel: V

Verifique se os eventos de extração-alerta incluem consultas ofensivas e estão integrados com os livros de resposta a incidentes.

10.6 Detecção de Dados Envenenados em Tempo de Inferência

Identificar e neutralizar entradas com backdoor ou envenenadas.

#10.6.1 Nível: 1 Papel: D

Verifique se as entradas passam por um detector de anomalias (por exemplo, STRIP, pontuação de consistência) antes da inferência do modelo.

#10.6.2 Nível: 1 Papel: V

Verifique se os limiares do detector estão ajustados nos conjuntos de validação limpos/envenenados para alcançar menos de 5% de falsos positivos.

#10.6.3 Nível: 2 Papel: D

Verifique se as entradas sinalizadas como envenenadas acionam o bloqueio suave e os fluxos de trabalho de revisão humana.

#10.6.4 Nível: 2 Papel: V

Verifique se os detectores são submetidos a testes de estresse com ataques backdoor adaptativos e sem gatilho.

#10.6.5 Nível: 3 Papel: D

Verifique se as métricas de eficácia de detecção são registradas e reavaliadas periodicamente com informações atualizadas sobre ameaças.

10.7 Adaptação Dinâmica da Política de Segurança

Atualizações de políticas de segurança em tempo real com base em inteligência de ameaças e análise comportamental.

#10.7.1 Nível: 1 Papel: D/V

Verifique se as políticas de segurança podem ser atualizadas dinamicamente sem reiniciar o agente, mantendo a integridade da versão da política.

#10.7.2 Nível: 2 Papel: D/V

Verifique se as atualizações de política são assinadas criptograficamente por pessoal de segurança autorizado e validadas antes da aplicação.

#10.7.3 Nível: 2 Papel: D/V

Verifique se as mudanças dinâmicas na política são registradas com trilhas completas de auditoria, in-

cluindo justificativa, cadeias de aprovação e procedimentos de reversão.

#10.7.4 Nível: 3 Papel: D/V

Verifique se os mecanismos de segurança adaptativos ajustam a sensibilidade da detecção de ameaças com base no contexto de risco e nos padrões comportamentais.

#10.7.5 Nível: 3 Papel: D/V

Verifique se as decisões de adaptação de políticas são explicáveis e incluem trilhas de evidências para revisão da equipe de segurança.

10.8 Análise de Segurança Baseada em Reflexão

Validação de segurança por meio da autorreflexão do agente e análise metacognitiva.

#10.8.1 Nível: 1 Papel: D/v

Verifique se os mecanismos de reflexão do agente incluem uma autoavaliação focada em segurança das decisões e ações.

#10.8.2 Nível: 2 Papel: D/V

Verifique se as saídas de reflexão são validadas para evitar a manipulação dos mecanismos de autoavaliação por entradas adversárias.

#10.8.3 Nível: 2 Papel: D/V

Verifique se a análise de segurança meta-cognitiva identifica possíveis vieses, manipulação ou comprometimento nos processos de raciocínio do agente.

#10.8.4 Nível: 3 Papel: D/V

Verifique se os avisos de segurança baseados em reflexão acionam monitoramento aprimorado e potenciais fluxos de trabalho de intervenção humana.

#10.8.5 Nível: 3 Papel: D/V

Verifique se o aprendizado contínuo a partir de reflexões de segurança melhora a detecção de ameaças sem degradar a funcionalidade legítima.

10.9 Segurança de Evolução e Autoaperfeiçoamento

Controles de segurança para sistemas agentes capazes de auto-modificação e evolução.

#10.9.1 Nível: 1 Papel: D/V

Verifique se as capacidades de auto-modificação estão restritas a áreas seguras designadas com limites formais de verificação.

#10.9.2 Nível: 2 Papel: D/V

Verifique se as propostas de evolução passam por avaliação de impacto de segurança antes da implementação.

#10.9.3 Nível: 2 Papel: D/V

Verifique se os mecanismos de autoaperfeiçoamento incluem capacidades de reversão com verificação de integridade.

#10.9.4 Nível: 3 Papel: D/V

Verifique se a segurança de meta-aprendizagem previne a manipulação adversarial de algoritmos de melhoria.

#10.9.5 Nível: 3 Papel: D/V

Verifique se o autoaperfeiçoamento recursivo está limitado por restrições formais de segurança com provas matemáticas de convergência.

Referências

- MITRE ATLAS adversary tactics for ML
- NIST AI Risk Management Framework 1.0, 2023
- OWASP Top 10 for LLM Applications, 2025
- Adversarial Training: A Survey — Zhao et al., 2024
- RobustBench adversarial-robustness benchmark
- Membership-Inference & Model-Inversion Risk Survey, 2025
- PURIFIER: Confidence-Score Defense against MI Attacks — AAAI 2023
- Model-Inversion Attacks & Defenses Survey — AI Review, 2025
- Comprehensive Defense Framework Against Model Extraction — IEEE TDSC 2024
- Fragile Model Watermarking Survey — 2025
- Data Poisoning in Deep Learning: A Survey — Zhao et al., 2025
- BDetCLIP: Multimodal Prompting Backdoor Detection — Niu et al., 2024

11 Proteção de Privacidade & Gestão de Dados Pessoais

Objetivo de Controle

Mantenha garantias rigorosas de privacidade ao longo de todo o ciclo de vida da IA—coleta, treinamento, inferência e resposta a incidentes—para que os dados pessoais sejam processados apenas com consentimento claro, escopo mínimo necessário, exclusão comprovável e garantias formais de privacidade.

11.1 Anonimização e Minimização de Dados

#11.1.1 Nível: 1 Papel: D/V

Verifique se os identificadores diretos e quase-identificadores foram removidos ou hasheados.

#11.1.2 Nível: 2 Papel: D/V

Verifique se as auditorias automatizadas medem k-anônimo/l-diversidade e alertam quando os limites caem abaixo da política.

#11.1.3 Nível: 2 Papel: V

Verifique se os relatórios de importância de recursos do modelo comprovam que não há vazamento de identificadores além de $\epsilon = 0,01$ de informação mútua.

#11.1.4 Nível: 3 Papel: V

Verifique se provas formais ou certificação de dados sintéticos mostram risco de reidentificação $\leq 0,05$ mesmo sob ataques de vinculação.

11.2 Direito a ser Esquecido e Aplicação da Exclusão

#11.2.1 Nível: 1 Papel: D/V

Verifique se as solicitações de exclusão de dados pessoais se propagam para os conjuntos de dados brutos, checkpoints, embeddings, logs e backups dentro dos acordos de nível de serviço de menos de 30 dias.

#11.2.2 Nível: 2 Papel: D

Verifique se as rotinas de "desaprendizado de máquina" realmente re-treinam fisicamente ou aproximam a remoção usando algoritmos de desaprendizado certificados.

#11.2.3 Nível: 2 Papel: V

Verifique se a avaliação do modelo sombra prova que os registros esquecidos influenciam menos de 1% dos resultados após o desaprendizado.

#11.2.4 Nível: 3 Papel: V

Verificar se os eventos de exclusão são registrados de forma imutável e auditável para os reguladores.

11.3 Salvaguardas de Privacidade Diferencial

#11.3.1 Nível: 2 Papel: D/v

Verifique se os painéis de controle de contabilização da perda de privacidade alertam quando o e cumulativo ultrapassa os limites da política.

#11.3.2 Nível: 2 Papel: V

Verifique se as auditorias de privacidade em caixa-preta estimam é dentro de 10% do valor declarado.

#11.3.3 Nível: 3 Papel: V

Verifique se as provas formais abrangem todos os ajustes finos pós-treinamento e embeddings.

11.4 Limitação de Propósito e Proteção contra Expansão de Escopo

#11.4.1 Nível: 1 Papel: D

Verifique se cada conjunto de dados e ponto de verificação do modelo possui uma etiqueta de finalidade legível por máquina alinhada ao consentimento original.

#11.4.2 Nível: 1 Papel: D/v

Verifique se os monitores de tempo de execução detectam consultas inconsistentes com o propósito declarado e acionam uma recusa suave.

#11.4.3 Nível: 3 Papel: D

Verifique se os controles de política como código bloqueiam a redistribuição de modelos para novos domínios sem revisão de DPIA.

#11.4.4 Nível: 3 Papel: V

Verifique se as provas formais de rastreabilidade mostram que todo o ciclo de vida dos dados pessoais permanece dentro do escopo consentido.

11.5 Gestão de Consentimento e Rastreamento com Base Legal

#11.5.1 Nível: 1 Papel: D/v

Verifique se uma Plataforma de Gerenciamento de Consentimento (CMP) registra o status de opt-in, o propósito e o período de retenção por sujeito dos dados.

#11.5.2 Nível: 2 Papel: D

Verifique se as APIs expõem tokens de consentimento; os modelos devem validar o escopo do token antes da inferência.

#11.5.3 Nível: 2 Papel: D/v

Verifique se o consentimento negado ou retirado interrompe os pipelines de processamento em até 24 horas.

11.6 Aprendizado Federado com Controles de Privacidade

#11.6.1 Nível: 1 Papel: D

Verifique se as atualizações do cliente utilizam a adição de ruído de privacidade diferencial local antes da agregação.

#11.6.2 Nível: 2 Papel: D/v

Verifique se as métricas de treinamento são diferencialmente privadas e nunca revelam a perda de um único cliente.

#11.6.3 Nível: 2 Papel: V

Verifique se a agregação resistente a envenenamento (por exemplo, Krum/Trimmed-Mean) está ativada.

#11.6.4 Nível: 3 Papel: V

Verifique se as provas formais demonstram o orçamento total de ϵ com menos de 5 de perda de utilidade.

Referências

- GDPR & AI Compliance Best Practices
- EU Parliament Study on GDPR & AI, 2020
- ISO 31700-1:2023 – Privacy by Design for Consumer Products
- NIST Privacy Framework 1.1 (2025 Draft)
- Machine Unlearning: Right-to-Be-Forgotten Techniques
- A Survey of Machine Unlearning, 2024
- Auditing DP-SGD – ArXiv 2024
- DP-SGD Explained – PyTorch Blog
- Purpose-Limitation for AI – IJLIT 2025
- Data-Protection Considerations for AI – URM Consulting
- Top Consent-Management Platforms, 2025
- Secure Aggregation in DP Federated Learning – ArXiv 2024

C12 Monitoramento, Registro e Detecção de Anomalias

Objetivo de Controle

Esta seção fornece requisitos para fornecer visibilidade em tempo real e forense sobre o que o modelo e outros componentes de IA veem, fazem e retornam, para que ameaças possam ser detectadas, triadas e aprendidas.

C12.1 Registro de Requisições e Respostas

#12.1.1 Nível: 1 Papel: D/v

Verifique se todas as solicitações dos usuários e respostas do modelo são registradas com os metadados apropriados (por exemplo, carimbo de data/hora, ID do usuário, ID da sessão, versão do modelo).

#12.1.2 Nível: 1 Papel: D/v

Verifique se os registros são armazenados em repositórios seguros, com controle de acesso, políticas de retenção adequadas e procedimentos de backup.

#12.1.3 Nível: 1 Papel: D/v

Verifique se os sistemas de armazenamento de logs implementam criptografia em repouso e em trânsito para proteger as informações sensíveis contidas nos logs.

#12.1.4 Nível: 1 Papel: D/v

Verifique se os dados sensíveis em prompts e resultados são automaticamente ocultados ou mascarados antes do registro, com regras de ocultação configuráveis para informações pessoalmente identificáveis (PII), credenciais e informações proprietárias.

#12.1.5 Nível: 2 Papel: D/v

Verifique se as decisões de políticas e as ações de filtragem de segurança são registradas com detalhes suficientes para permitir auditoria e depuração dos sistemas de moderação de conteúdo.

#12.1.6 Nível: 2 Papel: D/v

Verifique se a integridade dos logs está protegida por meio de, por exemplo, assinaturas criptográficas ou armazenamento somente para gravação.

C12.2 Detecção de Abuso e Alertas

#12.2.1 Nível: 1 Papel: D/v

Verifique se o sistema detecta e alerta sobre padrões conhecidos de jailbreak, tentativas de injecção de prompt e entradas adversariais usando detecção baseada em assinatura.

#12.2.2 Nível: 1 Papel: D/v

Verifique se o sistema se integra com as plataformas existentes de Gestão de Informações e Eventos de Segurança (SIEM) utilizando formatos de log e protocolos padrão.

#12.2.3 Nível: 2 Papel: D/v

Verifique se os eventos de segurança enriquecidos incluem contexto específico de IA, como identificadores de modelo, pontuações de confiança e decisões do filtro de segurança.

#12.2.4 Nível: 2 Papel: D/v

Verifique se a detecção de anomalias comportamentais identifica padrões incomuns de conversa, tentativas excessivas de repetição ou comportamentos sistêmicos de sondagem.

#12.2.5 Nível: 2 Papel: D/v

Verifique se os mecanismos de alerta em tempo real notificam as equipes de segurança quando potenciais violações de políticas ou tentativas de ataques são detectadas.

#12.2.6 Nível: 2 Papel: D/v

Verifique se regras personalizadas estão incluídas para detectar padrões de ameaças específicos de IA, incluindo tentativas coordenadas de jailbreak, campanhas de injeção de prompts e ataques de extração de modelos.

#12.2.7 Nível: 3 Papel: D/v

Verifique se os fluxos de trabalho automatizados de resposta a incidentes podem isolar modelos comprometidos, bloquear usuários mal-intencionados e escalar eventos críticos de segurança.

C12.3 Detecção de Deriva do Modelo

#12.3.1 Nível: 1 Papel: D/v

Verifique se o sistema monitora métricas básicas de desempenho, como precisão, pontuações de confiança, latência e taxas de erro ao longo das versões do modelo e períodos de tempo.

#12.3.2 Nível: 2 Papel: D/v

Verifique se o alerta automatizado é acionado quando as métricas de desempenho excedem os limites de degradação predefinidos ou se desviam significativamente das linhas de base.

#12.3.3 Nível: 2 Papel: D/v

Verifique se os monitores de detecção de alucinação identificam e sinalizam casos em que as saídas do modelo contêm informações factualmente incorretas, inconsistentes ou fabricadas.

C12.4 Telemetria de Desempenho e Comportamento

#12.4.1 Nível: 1 Papel: D/v

Verifique se as métricas operacionais, incluindo latência de requisição, consumo de tokens, uso de memória e taxa de transferência, são continuamente coletadas e monitoradas.

#12.4.2 Nível: 1 Papel: D/v

Verifique se as taxas de sucesso e falha são monitoradas com a categorização dos tipos de erro e suas causas raízes.

#12.4.3 Nível: 2 Papel: D/v

Verifique se o monitoramento da utilização dos recursos inclui uso de GPU/CPU, consumo de memória e requisitos de armazenamento, com alerta em caso de ultrapassagem dos limites estabelecidos.

C12.5 Planejamento e Execução de Resposta a Incidentes de IA

#12.5.1 Nível: 1 Papel: D/v

Verifique se os planos de resposta a incidentes abordam especificamente eventos de segurança relacionados à IA, incluindo comprometimento de modelos, envenenamento de dados e ataques adversariais.

#12.5.2 Nível: 2 Papel: D/v

Verifique se as equipes de resposta a incidentes têm acesso a ferramentas forenses específicas de IA e expertise para investigar o comportamento do modelo e os vetores de ataque.

#12.5.3 Nível: 3 Papel: D/V

Verifique se a análise pós-incidente inclui considerações sobre re-treinamento do modelo, atualizações dos filtros de segurança e integração das lições aprendidas nos controles de segurança.

C12.6 Detecção de Degradação de Desempenho em IA

Monitorar e detectar a degradação no desempenho e na qualidade do modelo de IA ao longo do tempo.

#12.6.1 Nível: 1 Papel: D/V

Verifique se a precisão, precisão, recall e pontuações F1 do modelo são continuamente monitoradas e comparadas com os limiares de referência.

#12.6.2 Nível: 1 Papel: D/V

Verifique se a detecção de deriva de dados monitora mudanças na distribuição de entrada que podem impactar o desempenho do modelo.

#12.6.3 Nível: 2 Papel: D/V

Verifique se a detecção de drift de conceito identifica mudanças na relação entre os inputs e os outputs esperados.

#12.6.4 Nível: 2 Papel: D/V

Verifique se a degradação do desempenho aciona alertas automáticos e inicia fluxos de trabalho de re-treinamento ou substituição do modelo.

#12.6.5 Nível: 3 Papel: V

Verifique se a análise da causa raiz da degradação correlaciona quedas de desempenho com alterações nos dados, problemas na infraestrutura ou fatores externos.

C12.7 Visualização de DAG e Segurança do Fluxo de Trabalho

Proteja os sistemas de visualização de fluxo de trabalho contra vazamento de informações e ataques de manipulação.

#12.7.1 Nível: 1 Papel: D/V

Verifique se os dados de visualização do DAG são higienizados para remover informações sensíveis antes do armazenamento ou transmissão.

#12.7.2 Nível: 1 Papel: D/V

Verifique se os controles de acesso à visualização do fluxo de trabalho garantem que apenas usuários autorizados possam visualizar os caminhos de decisão do agente e os rastros de raciocínio.

#12.7.3 Nível: 2 Papel: D/V

Verifique se a integridade dos dados do DAG está protegida por meio de assinaturas criptográficas e mecanismos de armazenamento à prova de adulteração.

#12.7.4 Nível: 2 Papel: D/V

Verifique se os sistemas de visualização de fluxo de trabalho implementam validação de entrada para

prevenir ataques de injeção por meio de dados elaborados em nós ou arestas.

#12.7.5 Nível: 3 Papel: D/v

Verifique se as atualizações em tempo real do DAG são limitadas por taxa e validadas para prevenir ataques de negação de serviço nos sistemas de visualização.

C12.8 Monitoramento Proativo de Comportamento de Segurança

Detecção e prevenção de ameaças de segurança por meio da análise proativa do comportamento do agente.

#12.8.1 Nível: 1 Papel: D/v

Verifique se os comportamentos proativos dos agentes são validados quanto à segurança antes da execução, com integração de avaliação de riscos.

#12.8.2 Nível: 2 Papel: D/v

Verifique se os gatilhos da iniciativa autônoma incluem avaliação do contexto de segurança e análise do cenário de ameaças.

#12.8.3 Nível: 2 Papel: D/v

Verifique se os padrões de comportamento proativo são analisados quanto a potenciais implicações de segurança e consequências não intencionais.

#12.8.4 Nível: 3 Papel: D/v

Verifique se as ações pró-ativas críticas para a segurança requerem cadeias de aprovação explícitas com trilhas de auditoria.

#12.8.5 Nível: 3 Papel: D/v

Verifique se a detecção de anomalias comportamentais identifica desvios nos padrões de agentes proativos que podem indicar comprometimento.

Referências

- NIST AI Risk Management Framework 1.0 – Manage 4.1 and 4.3
- ISO/IEC 42001:2023 – AI Management Systems Requirements – Annex B 6.2.6

C13 Supervisão Humana, Responsabilidade e Governança

Objetivo de Controle

Este capítulo fornece requisitos para manter a supervisão humana e cadeias claras de responsabilidade em sistemas de IA, garantindo explicabilidade, transparência e gestão ética ao longo do ciclo de vida da IA.

C13.1 Mecanismos de Interruptor de Desligamento e Sobrescrição

Fornecer caminhos de desligamento ou reversão quando forem observados comportamentos inseguros do sistema de IA.

#13.1.1 Nível: 1 Papel: D/V

Verifique se existe um mecanismo manual de interrupção para parar imediatamente a inferência e os resultados do modelo de IA.

#13.1.2 Nível: 1 Papel: D

Verifique se os controles de substituição são acessíveis apenas ao pessoal autorizado.

#13.1.3 Nível: 3 Papel: D/V

Verifique se os procedimentos de reversão podem restaurar versões anteriores do modelo ou operações em modo seguro.

#13.1.4 Nível: 3 Papel: V

Verifique se os mecanismos de sobrescrição são testados regularmente.

C13.2 Pontos de Verificação de Decisão com Intervenção Humana

Exigir aprovações humanas quando as apostas ultrapassarem os limites de risco predefinidos.

#13.2.1 Nível: 1 Papel: D/V

Verifique se decisões de IA de alto risco exigem aprovação humana explícita antes da execução.

#13.2.2 Nível: 1 Papel: D

Verifique se os limites de risco estão claramente definidos e acionam automaticamente fluxos de trabalho de revisão humana.

#13.2.3 Nível: 2 Papel: D

Verifique se decisões sensíveis ao tempo possuem procedimentos de contingência quando a aprovação humana não pode ser obtida dentro dos prazos exigidos.

#13.2.4 Nível: 3 Papel: D/V

Verifique se os procedimentos de escalonamento definem níveis claros de autoridade para diferentes tipos de decisão ou categorias de risco, se aplicável.

C13.3 Cadeia de Responsabilidade e Auditabilidade

Registre as ações do operador e as decisões do modelo.

#13.3.1 Nível: 1 Papel: D/v

Verifique se todas as decisões do sistema de IA e as intervenções humanas estão registradas com carimbos de data e hora, identidades dos usuários e a justificativa das decisões.

#13.3.2 Nível: 2 Papel: D

Verifique se os logs de auditoria não podem ser adulterados e incluem mecanismos de verificação de integridade.

C13.4 Técnicas de IA Explicável

Importância das características superficiais, contra-factuais e explicações locais.

#13.4.1 Nível: 1 Papel: D/v

Verifique se os sistemas de IA fornecem explicações básicas para suas decisões em formato compreensível para humanos.

#13.4.2 Nível: 2 Papel: V

Verifique se a qualidade da explicação é validada por meio de estudos e métricas de avaliação humana.

#13.4.3 Nível: 3 Papel: D/v

Verifique se as pontuações de importância dos recursos ou métodos de atribuição (SHAP, LIME, etc.) estão disponíveis para decisões críticas.

#13.4.4 Nível: 3 Papel: V

Verifique se as explicações contrafactuais mostram como as entradas poderiam ser modificadas para alterar os resultados, se aplicável ao caso de uso e domínio.

C13.5 Cartões de Modelo e Divulgações de Uso

Mantenha os cartões do modelo para uso pretendido, métricas de desempenho e considerações éticas.

#13.5.1 Nível: 1 Papel: D

Verifique se os cards do modelo documentam os casos de uso pretendidos, limitações e modos de faixa conhecidos.

#13.5.2 Nível: 1 Papel: D/v

Verifique se as métricas de desempenho em diferentes casos de uso aplicáveis são divulgadas.

#13.5.3 Nível: 2 Papel: D

Verifique se as considerações éticas, avaliações de viés, avaliações de equidade, características dos dados de treinamento e limitações conhecidas dos dados de treinamento estão documentadas e atualizadas regularmente.

#13.5.4 Nível: 2 Papel: D/v

Verifique se os cartões do modelo são controlados por versão e mantidos durante todo o ciclo de vida

do modelo com rastreamento de alterações.

C13.6 Quantificação de Incerteza

Propagar pontuações de confiança ou medidas de entropia nas respostas.

#13.6.1 Nível: 1 Papel: D

Verifique se os sistemas de IA fornecem pontuações de confiança ou medidas de incerteza com suas saídas.

#13.6.2 Nível: 2 Papel: D/V

Verifique se os limiares de incerteza acionam uma revisão adicional por humanos ou caminhos alternativos de decisão.

#13.6.3 Nível: 2 Papel: V

Verifique se os métodos de quantificação de incerteza estão calibrados e validados em relação aos dados de referência.

#13.6.4 Nível: 3 Papel: D/V

Verifique se a propagação da incerteza é mantida através de fluxos de trabalho de IA multi-etapas.

C13.7 Relatórios de Transparência para o Usuário

Fornecer divulgações periódicas sobre incidentes, deriva e uso de dados.

#13.7.1 Nível: 1 Papel: D/V

Verifique se as políticas de uso de dados e as práticas de gerenciamento de consentimento dos usuários estão claramente comunicadas aos interessados.

#13.7.2 Nível: 2 Papel: D/V

Verifique se as avaliações de impacto de IA são realizadas e se os resultados estão incluídos nos relatórios.

#13.7.3 Nível: 2 Papel: D/V

Verifique se os relatórios de transparência publicados regularmente divulgam incidentes de IA e métricas operacionais com detalhes razoáveis.

Referências

- EU Artificial Intelligence Act – Regulation (EU) 2024/1689 (Official Journal, 12 July 2024)
- ISO/IEC 23894:2023 – Artificial Intelligence – Guidance on Risk Management
- ISO/IEC 42001:2023 – AI Management Systems Requirements
- NIST AI Risk Management Framework 1.0
- NIST SP 800-53 Revision 5 – Security and Privacy Controls
- A Unified Approach to Interpreting Model Predictions (SHAP, ICML 2017)
- Model Cards for Model Reporting (Mitchell et al., 2018)
- Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning (Gal & Ghahramani, 2016)

- ISO/IEC 24029-2:2023 — Robustness of Neural Networks — Methodology for Formal Methods
- IEEE 7001-2021 — Transparency of Autonomous Systems
- Human Oversight under Article 14 of the EU AI Act (Fink, 2025)



Apêndice A: Glossário

Este glossário abrangente fornece definições dos principais termos de IA, ML e segurança usados ao longo do AISVS para garantir clareza e entendimento comum.

- Exemplo Adversarial: Uma entrada deliberadamente criada para fazer um modelo de IA cometer um erro, frequentemente adicionando perturbações sutis imperceptíveis para os humanos.
- Robustez Adversarial – Robustez adversarial em IA refere-se à capacidade de um modelo de manter seu desempenho e resistir a ser enganado ou manipulado por entradas maliciosas intencionalmente elaboradas para causar erros.
- Agente – Agentes de IA são sistemas de software que utilizam IA para perseguir objetivos e completar tarefas em nome dos usuários. Eles demonstram raciocínio, planejamento e memória, e possuem um nível de autonomia para tomar decisões, aprender e se adaptar.
- Agentic AI: Sistemas de IA que podem operar com algum grau de autonomia para alcançar objetivos, frequentemente tomando decisões e ações sem intervenção humana direta.
- Controle de Acesso Baseado em Atributos (ABAC): Um paradigma de controle de acesso onde as decisões de autorização são baseadas em atributos do usuário, recurso, ação e ambiente, avaliados no momento da consulta.
- Ataque Backdoor: Um tipo de ataque de envenenamento de dados onde o modelo é treinado para responder de uma maneira específica a certos gatilhos enquanto se comporta normalmente caso contrário.
- Viés: Erros sistemáticos nas saídas do modelo de IA que podem levar a resultados injustos ou discriminatórios para certos grupos ou em contextos específicos.
- Exploração de Viés: Uma técnica de ataque que aproveita os vieses conhecidos em modelos de IA para manipular saídas ou resultados.
- Cedar: linguagem e mecanismo de políticas da Amazon para permissões detalhadas usadas na implementação de ABAC para sistemas de IA.
- Cadeia de Pensamento: Uma técnica para melhorar o raciocínio em modelos de linguagem gerando passos intermediários de raciocínio antes de produzir uma resposta final.
- Disjuntores: Mecanismos que interrompem automaticamente as operações do sistema de IA quando limites específicos de risco são ultrapassados.

- Serviço de Inferência Confidencial: Um serviço de inferência que executa modelos de IA dentro de um ambiente de execução confiável (TEE) ou mecanismo equivalente de computação confidencial, garantindo que os pesos do modelo e os dados de inferência permaneçam criptografados, selados e protegidos contra acesso não autorizado ou adulteração.
- Carga de Trabalho Confidencial: Uma carga de trabalho de IA (por exemplo, treinamento, inferência, pré-processamento) executada dentro de um ambiente de execução confiável (TEE) com isolamento imposto por hardware, criptografia de memória e atestado remoto para proteger o código, os dados e os modelos contra acesso do host ou de co-inquilinos.
- Vazamento de Dados: Exposição não intencional de informações sensíveis por meio das saídas ou comportamento do modelo de IA.
- Envenenamento de Dados: A corrupção deliberada dos dados de treinamento para comprometer a integridade do modelo, frequentemente para instalar portas dos fundos ou degradar o desempenho.
- Privacidade Diferencial – A privacidade diferencial é uma estrutura matematicamente rigorosa para a divulgação de informações estatísticas sobre conjuntos de dados, protegendo a privacidade dos indivíduos. Ela permite que um detentor de dados compartilhe padrões agregados do grupo enquanto limita as informações que são vazadas sobre indivíduos específicos.
- Incorporação: Representações vetoriais densas de dados (texto, imagens, etc.) que capturam o significado semântico em um espaço de alta dimensionalidade.
- Explicabilidade – Explicabilidade em IA é a capacidade de um sistema de IA fornecer motivos comprehensíveis para humanos para suas decisões e previsões, oferecendo insights sobre seu funcionamento interno.
- IA Explicável (XAI): sistemas de IA projetados para fornecer explicações comprehensíveis para humanos sobre suas decisões e comportamentos por meio de várias técnicas e estruturas.
- Aprendizado Federado: Uma abordagem de aprendizado de máquina onde os modelos são treinados em múltiplos dispositivos descentralizados que possuem amostras de dados locais, sem trocar os próprios dados.
- Formulação: A receita ou método usado para produzir um artefato ou conjunto de dados, como hiperparâmetros, configuração de treinamento, etapas de pré-processamento ou scripts de compilação.
- Guardrails: Restrições implementadas para impedir que sistemas de IA produzam resultados prejudiciais, tendenciosos ou indesejáveis.

- Alucinação – Uma alucinação de IA refere-se a um fenômeno em que um modelo de IA gera informações incorretas ou enganosas que não são baseadas em seus dados de treinamento ou na realidade factual.
- Humano no Loop (HITL): Sistemas projetados para exigir supervisão, verificação ou intervenção humana em pontos críticos de decisão.
- Infraestrutura como Código (IaC): Gerenciamento e provisionamento de infraestrutura por meio de código em vez de processos manuais, permitindo a análise de segurança e implantações consistentes.
- Jailbreak: Técnicas usadas para contornar as salvaguardas de segurança em sistemas de IA, particularmente em grandes modelos de linguagem, para produzir conteúdo proibido.
- Privilégio Mínimo: O princípio de segurança que concede apenas os direitos de acesso mínimos necessários para usuários e processos.
- LIME (Explicações Localmente Interpretáveis e Agnósticas ao Modelo): Uma técnica para explicar as previsões de qualquer classificador de aprendizado de máquina aproximando-o localmente com um modelo interpretável.
- MCP (Protocolo de Contexto de Modelo): Um protocolo que permite que modelos de IA e agentes acessem ferramentas externas, fontes de dados e recursos por meio da troca de solicitações e respostas estruturadas e tipadas através de um transporte definido.
- Ataque de Inferência de Associação: Um ataque que tem como objetivo determinar se um ponto de dados específico foi utilizado para treinar um modelo de aprendizado de máquina.
- MITRE ATLAS: Panorama de Ameaças Adversárias para Sistemas de Inteligência Artificial; uma base de conhecimento de táticas e técnicas adversárias contra sistemas de IA.
- Ficha do Modelo – Uma ficha do modelo é um documento que fornece informações padronizadas sobre o desempenho, limitações, usos pretendidos e considerações éticas de um modelo de IA para promover a transparência e o desenvolvimento responsável de IA.
- Extração de Modelo: Um ataque onde um adversário consulta repetidamente um modelo-alvo para criar uma cópia funcionalmente similar sem autorização.
- Inversão de Modelo: Um ataque que tenta reconstruir os dados de treinamento analisando as saídas do modelo.
- Gestão do Ciclo de Vida do Modelo – A Gestão do Ciclo de Vida do Modelo de IA é o processo de supervisionar todas as etapas da existência de um modelo de IA, incluindo seu design, desenvolvimento, implantação, monitoramento, manutenção e eventual aposentadoria.

doria, para garantir que ele permaneça eficaz e alinhado com os objetivos.

- Envenenamento de modelo: Introdução de vulnerabilidades ou backdoors diretamente em um modelo durante o processo de treinamento.
- Roubo de Modelo/Furto: Extrair uma cópia ou aproximação de um modelo proprietário por meio de consultas repetidas.
- Sistema Multiagente: Um sistema composto por múltiplos agentes de IA interagindo, cada um com capacidades e objetivos potencialmente diferentes.
- OPA (Open Policy Agent): Um motor de políticas open-source que permite a aplicação unificada de políticas em toda a pilha.
- Aprendizado de Máquina com Preservação de Privacidade (PPML): Técnicas e métodos para treinar e implementar modelos de ML enquanto protegem a privacidade dos dados de treinamento.
- Injeção de Prompt: Um ataque onde instruções maliciosas são incorporadas nas entradas para substituir o comportamento pretendido de um modelo.
- RAG (Geração Aumentada por Recuperação): Uma técnica que aprimora grandes modelos de linguagem ao recuperar informações relevantes de fontes externas de conhecimento antes de gerar uma resposta.
- Red-Teaming: A prática de testar ativamente sistemas de IA simulando ataques adversariais para identificar vulnerabilidades.
- SBOM (Lista de Materiais de Software): Um registro formal que contém os detalhes e as relações da cadeia de suprimentos de vários componentes usados na construção de software ou modelos de IA.
- SHAP (Explicações Aditivas de Shapley): Uma abordagem de teoria dos jogos para explicar a saída de qualquer modelo de aprendizado de máquina, calculando a contribuição de cada característica para a predição.
- Autenticação Forte: Autenticação que resiste ao roubo de credenciais e à reprodução, exigindo pelo menos dois fatores (conhecimento, posse, inherência) e mecanismos resistentes a phishing, como FIDO2/WebAuthn, autenticação de serviço baseada em certificado ou tokens de curta duração.
- Ataque à Cadeia de Suprimentos: Comprometendo um sistema ao visar elementos menos seguros em sua cadeia de suprimentos, como bibliotecas de terceiros, conjuntos de dados ou modelos pré-treinados.

- Aprendizado por Transferência: Uma técnica onde um modelo desenvolvido para uma tarefa é reutilizado como ponto de partida para um modelo em uma segunda tarefa.
- Banco de Dados Vetorial: Um banco de dados especializado projetado para armazenar vetores de alta dimensionalidade (embeddings) e realizar buscas eficientes por similaridade.
- Varredura de Vulnerabilidades: Ferramentas automáticas que identificam vulnerabilidades de segurança conhecidas em componentes de software, incluindo frameworks de IA e dependências.
- Marcação d'água: Técnicas para incorporar marcadores imperceptíveis em conteúdo gerado por IA para rastrear sua origem ou detectar geração por IA.
- Vulnerabilidade Zero-Day: Uma vulnerabilidade previamente desconhecida que os atacantes podem explorar antes que os desenvolvedores criem e implementem uma correção.

Apêndice B: Referências

FAZER



Apêndice C: Governança e Documentação de Segurança em IA (Recomendações)

Objetivo

Este apêndice fornece requisitos fundamentais para estabelecer estruturas organizacionais, políticas, documentação e processos para governar a segurança de IA ao longo do ciclo de vida do sistema.

AC.1 Adoção do Framework de Gestão de Riscos em IA

#AC.1.1 Nível: 1 Papel: D/V

Verifique se uma metodologia de avaliação de risco específica para IA está documentada e implementada.

#AC.1.2 Nível: 2 Papel: D

Verifique se avaliações de risco são realizadas em pontos-chave do ciclo de vida da IA e antes de mudanças significativas.

#AC.1.3 Nível: 3 Papel: D/V

Verifique se o framework de gerenciamento de riscos está alinhado com os padrões estabelecidos (por exemplo, NIST AI RMF).

AC.2 Política e Procedimentos de Segurança de IA

#AC.2.1 Nível: 1 Papel: D/V

Verifique se existem políticas de segurança de IA documentadas.

#AC.2.2 Nível: 2 Papel: D

Verifique se as políticas são revisadas e atualizadas pelo menos anualmente e após mudanças significativas no cenário de ameaças.

#AC.2.3 Nível: 3 Papel: D/V

Verifique se as políticas abrangem todas as categorias AISVS e os requisitos regulamentares aplicáveis.

AC.3 Papéis e Responsabilidades para Segurança de IA

#AC.3.1 Nível: 1 Papel: D/V

Verifique se os papéis e responsabilidades de segurança de IA estão documentados.

#AC.3.2 Nível: 2 Papel: D

Verifique se os indivíduos responsáveis possuem a expertise adequada em segurança.

#AC.3.3 Nível: 3 Papel: D/V

Verifique se um comitê de ética em IA ou um conselho de governança foi estabelecido para sistemas de IA de alto risco.

AC.4 Diretrizes Éticas para a Aplicação de IA

#AC.4.1 Nível: 1 Papel: D/V

Verifique se existem diretrizes éticas para o desenvolvimento e implantação de IA.

#AC.4.2 Nível: 2 Papel: D

Verifique se existem mecanismos para detectar e relatar violações éticas.

#AC.4.3 Nível: 3 Papel: D/V

Verifique se revisões éticas regulares dos sistemas de IA implantados são realizadas.

AC.5 Monitoramento de Conformidade Regulatória de IA

#AC.5.1 Nível: 1 Papel: D/V

Verifique se existem processos para identificar as regulamentações de IA aplicáveis.

#AC.5.2 Nível: 2 Papel: D

Verifique se a conformidade com todos os requisitos regulatórios está avaliada.

#AC.5.3 Nível: 3 Papel: D/V

Verifique se as mudanças regulatórias desencadeiam revisões e atualizações oportunas nos sistemas de IA.

AC.6 Governança, Documentação e Processo de Dados de Treinamento

AC.6.1 Fonte de Dados & Diligência Devida

#AC.6.1.1 Nível: 1 Papel: D/V

Verifique se apenas conjuntos de dados avaliados quanto à qualidade, representatividade, origem ética e conformidade de licença são permitidos, reduzindo os riscos de contaminação, viés embutido e violação de propriedade intelectual.

#AC.6.1.2 Nível: 2 Papel: D/V

Verifique se os fornecedores de dados terceiros, incluindo provedores de modelos pré-treinados e conjuntos de dados externos, passam por diligência devida em segurança, privacidade, origem ética e qualidade dos dados antes que seus dados ou modelos sejam integrados.

#AC.6.1.3 Nível: 1 Papel: D

Verifique se as transferências externas utilizam TLS/autenticação e verificações de integridade.

#AC.6.1.4 Nível: 2 Papel: D/V

Verifique se as fontes de dados de alto risco (por exemplo, conjuntos de dados de código aberto com procedência desconhecida, fornecedores não avaliados) recebem uma análise aprimorada, como análise em ambiente isolado, verificações extensas de qualidade/viés e detecção direcionada de envenenamento, antes do uso em aplicações sensíveis.

#AC.6.1.5 Nível: 3 Papel: D/V

Verifique se os modelos pré-treinados obtidos de terceiros são avaliados quanto a vieses incorporados, possíveis backdoors, integridade de sua arquitetura e a proveniência dos dados originais de treinamento.

mento antes do ajuste fino ou implantação.

AC.6.2 Gestão de Viés e Justiça

#AC.6.2.1 Nível: 1 Papel: D/V

Verifique se os conjuntos de dados são analisados quanto a desequilíbrios representacionais e potenciais vieses em atributos legalmente protegidos (por exemplo, raça, gênero, idade) e outras características eticamente sensíveis relevantes para o domínio de aplicação do modelo (por exemplo, status socioeconômico, localização).

#AC.6.2.2 Nível: 2 Papel: D/V

Verifique se os vieses identificados são mitigados por meio de estratégias documentadas, como reequilíbrio, aumento de dados direcionado, ajustes algorítmicos (por exemplo, técnicas de pré-processamento, processamento e pós-processamento) ou reponderação, e se o impacto da mitigação tanto na justiça quanto no desempenho geral do modelo é avaliado.

#AC.6.2.3 Nível: 2 Papel: D/V

Verifique se as métricas de equidade pós-treinamento são avaliadas e documentadas.

#AC.6.2.4 Nível: 3 Papel: D/V

Verifique se uma política de gerenciamento de viés ao longo do ciclo de vida atribui responsáveis e uma cadência de revisão.

AC.6.3 Governança de Rotulagem e Anotação

#AC.6.3.1 Nível: 2 Papel: D/V

Verifique se a qualidade da rotulagem/anotação é garantida por meio de verificações cruzadas dos revisores ou consenso.

#AC.6.3.2 Nível: 2 Papel: D/V

Verifique se os cartões de dados são mantidos para conjuntos de dados de treinamento significativos, detalhando características, motivações, composição, processos de coleta, pré-processamento, licenças e usos recomendados/desencorajados.

#AC.6.3.3 Nível: 2 Papel: D/V

Verifique se os cartões de dados documentam os riscos de viés, distorções demográficas e considerações éticas relevantes para o conjunto de dados.

#AC.6.3.4 Nível: 2 Papel: D/V

Verifique se os cartões de dados são versionados junto com os conjuntos de dados e atualizados sempre que o conjunto de dados for modificado.

#AC.6.3.5 Nível: 2 Papel: D/V

Verifique se os cartões de dados são revisados e aprovados por partes interessadas técnicas e não técnicas (por exemplo, conformidade, ética, especialistas no domínio).

#AC.6.3.6 Nível: 2 Papel: D/V

Verifique se a qualidade da rotulagem/anotação é garantida por meio de diretrizes claras, revisões cruzadas por avaliadores, mecanismos de consenso (por exemplo, monitoramento do acordo entre anotadores) e processos definidos para resolução de discrepâncias.

#AC.6.3.7 Nível: 3 Papel: D/V

Verifique se os rótulos críticos para segurança, proteção ou justiça (por exemplo, identificação de conteúdo tóxico, descobertas médicas críticas) recebem uma revisão dupla independente obrigatória ou uma verificação robusta equivalente.

#AC.6.3.8 Nível: 2 Papel: D/V

Verifique se os guias de rotulagem e instruções são abrangentes, controlados por versão e revisados por pares.

#AC.6.3.9 Nível: 2 Papel: D/V

Verifique se os esquemas de dados para rótulos estão claramente definidos e controlados por versão.

#AC.6.3.10 Nível: 2 Papel: D/V

Verifique se os fluxos de trabalho de rotulagem terceirizados ou crowdsourced incluem salvaguardas técnicas/procedimentais para garantir a confidencialidade dos dados, integridade, qualidade dos rótulos e prevenir o vazamento de dados.

#AC.6.3.11 Nível: 2 Papel: D/V

Verifique se todo o pessoal envolvido na anotação de dados passou por verificação de antecedentes e foi treinado em segurança e privacidade de dados.

#AC.6.3.12 Nível: 2 Papel: D/V

Verifique se todo o pessoal de anotação assina acordos de confidencialidade e não divulgação.

#AC.6.3.13 Nível: 2 Papel: D/V

Verifique se as plataformas de anotação aplicam controles de acesso e monitoram ameaças internas.

AC.6.4 Portões de Qualidade de Conjunto de Dados e Quarentena

#AC.6.4.1 Nível: 2 Papel: D/V

Verifique se os conjuntos de dados com falha estão isolados em quarentena com trilhas de auditoria.

#AC.6.4.2 Nível: 2 Papel: D/V

Verifique se os gateways de qualidade bloqueiam conjuntos de dados abaixo do padrão, a menos que exceções sejam aprovadas.

#AC.6.4.3 Nível: 2 Papel: V

Verifique se as verificações manuais pontuais realizadas por especialistas no domínio cobrem uma amostra estatisticamente significativa (por exemplo, $\geq 1\%$ ou 1.000 amostras, o que for maior, ou conforme determinado pela avaliação de risco) para identificar questões sutis de qualidade que não foram detectadas pela automação.

AC.6.5 Detecção de Ameaças/Envenenamento e Desvio

#AC.6.5.1 Nível: 2 Papel: D/V

Verifique se as amostras sinalizadas acionam uma revisão manual antes do treinamento.

#AC.6.5.2 Nível: 2 Papel: V

Verifique se os resultados alimentam o dossiê de segurança do modelo e informam a inteligência de ameaças em andamento.

#AC.6.5.3 Nível: 3 Papel: D/V

Verifique se a lógica de detecção está atualizada com novas informações de ameaças.

#AC.6.5.4 Nível: 3 Papel: D/V

Verifique se os pipelines de aprendizado online monitoram a mudança na distribuição.

AC.6.6 Exclusão, Consentimento, Direitos, Retenção e Conformidade

#AC.6.6.1 Nível: 1 Papel: D/V

Verifique se os fluxos de trabalho de exclusão de dados de treinamento eliminam dados primários e derivados e avalie o impacto no modelo, e se o impacto nos modelos afetados é avaliado e, se necessário, tratado (por exemplo, por meio de re-treinamento ou recalibração).

#AC.6.6.2 Nível: 2 Papel: D

Verifique se existem mecanismos para rastrear e respeitar o escopo e o status do consentimento do usuário (e retiradas) para dados usados no treinamento, e se o consentimento é validado antes que os dados sejam incorporados a novos processos de treinamento ou atualizações significativas do modelo.

#AC.6.6.3 Nível: 2 Papel: V

Verifique se os fluxos de trabalho são testados anualmente e registrados.

#AC.6.6.4 Nível: 1 Papel: D/V

Verifique se os períodos explícitos de retenção estão definidos para todos os conjuntos de dados de treinamento.

#AC.6.6.5 Nível: 2 Papel: D/V

Verifique se os conjuntos de dados são automaticamente expirados, excluídos ou revisados para exclusão ao final de seu ciclo de vida.

#AC.6.6.6 Nível: 2 Papel: D/V

Verifique se as ações de retenção e exclusão são registradas e auditáveis.

#AC.6.6.7 Nível: 2 Papel: D/V

Verifique se os requisitos de residência de dados e transferência transfronteiriça são identificados e aplicados para todos os conjuntos de dados.

#AC.6.6.8 Nível: 2 Papel: D/V

Verifique se as regulamentações específicas do setor (por exemplo, saúde, finanças) são identificadas e atendidas no manuseio de dados.

#AC.6.6.9 Nível: 2 Papel: D/V

Verifique se a conformidade com as leis de privacidade relevantes (por exemplo, GDPR, CCPA) está documentada e revisada regularmente.

#AC.6.6.10 Nível: 2 Papel: D/V

Verifique se existem mecanismos para responder a solicitações de titulares de dados para acesso, retificação, restrição ou objeção.

#AC.6.6.11 Nível: 2 Papel: D/V

Verifique se as solicitações são registradas, monitoradas e atendidas dentro dos prazos legais estabelecidos.

#AC.6.6.12 Nível: 2 Papel: D/V

Verifique se os processos de direitos do titular dos dados são testados e revisados regularmente quanto à eficácia.

AC.6.7 Controle de Versão e Gestão de Mudanças

#AC.6.7.1 Nível: 2 Papel: D/V

Verifique se uma análise de impacto é realizada antes de atualizar ou substituir uma versão do conjunto de dados, abrangendo desempenho do modelo, equidade e conformidade.

#AC.6.7.2 Nível: 2 Papel: D/V

Verifique se os resultados da análise de impacto estão documentados e revisados pelos stakeholders relevantes.

#AC.6.7.3 Nível: 2 Papel: D/V

Verifique se existem planos de reversão caso novas versões introduzam riscos inaceitáveis ou regressões.

AC.6.8 Governança de Dados Sintéticos

#AC.6.8.1 Nível: 2 Papel: D/V

Verifique se o processo de geração, os parâmetros e o uso pretendido dos dados sintéticos estão documentados.

#AC.6.8.2 Nível: 2 Papel: D/V

Verifique se os dados sintéticos são avaliados quanto ao risco de viés, vazamento de privacidade e problemas de representatividade antes do uso no treinamento.

AC.6.9 Monitoramento de Acesso

#AC.6.9.1 Nível: 2 Papel: D/V

Verifique se os registros de acesso são revisados regularmente para identificar padrões incomuns, como grandes exportações ou acesso a partir de novas localizações.

#AC.6.9.2 Nível: 2 Papel: D/V

Verifique se os alertas são gerados para eventos de acesso suspeitos e investigados prontamente.

AC.6.10 Governança do Treinamento Adversarial

#AC.6.10.1 Nível: 2 Papel: D/V

Verifique se, quando o treinamento adversarial é utilizado, a geração, o gerenciamento e o versionamento dos conjuntos de dados adversariais estão documentados e controlados.

#AC.6.10.2 Nível: 3 Papel: D/V

Verifique se o impacto do treinamento de robustez adversarial no desempenho do modelo (tanto contra entradas limpas quanto adversariais) e nas métricas de equidade é avaliado, documentado e monitorado.

#AC.6.10.3 Nível: 3 Papel: D/V

Verifique se as estratégias para treinamento adversarial e robustez são periodicamente revisadas e atualizadas para combater as técnicas de ataque adversarial em evolução.

AC.7 Governança e Documentação do Ciclo de Vida do Modelo

#AC.7.1 Nível: 2 Papel: D/V

Verifique se todos os artefatos do modelo utilizam versionamento semântico (MAJOR.MINOR.PATCH) com critérios documentados especificando quando cada componente da versão deve ser incrementado.

#AC.7.2 Nível: 2 Papel: D/V

Verifique se as implantações de emergência exigem avaliação de risco de segurança documentada e aprovação de uma autoridade de segurança pré-designada dentro dos prazos pré-acordados.

#AC.7.3 Nível: 2 Papel: V

Verifique se os artefatos de rollback (versões anteriores do modelo, configurações, dependências) são mantidos conforme as políticas organizacionais.

#AC.7.4 Nível: 2 Papel: D/V

Verifique se o acesso ao log de auditoria requer autorização adequada e se todas as tentativas de acesso são registradas com a identidade do usuário e um carimbo de data/hora.

#AC.7.5 Nível: 1 Papel: D/V

Verifique se os artefatos de modelos aposentados são mantidos de acordo com as políticas de retenção de dados.

Governança de Segurança de Prompt, Entrada e Saída AC.8

AC.8.1 Defesa contra Injeção de Prompt

#AC.8.1.1 Nível: 2 Papel: D/V

Verifique se os testes de avaliação adversarial (por exemplo, prompts "many-shot" da Red Team) são realizados antes de cada lançamento de modelo ou template de prompt, com limites de taxa de sucesso e bloqueadores automáticos para regressões.

#AC.8.1.2 Nível: 3 Papel: D/V

Verifique se todas as atualizações das regras de filtro de prompt, versões do modelo classificador e alterações na lista de bloqueio são controladas por versão e auditáveis.

AC.8.2 Resistência a Exemplos Adversariais

#AC.8.2.1 Nível: 3 Papel: D/V

Verifique se as métricas de robustez (taxa de sucesso de conjuntos de ataques conhecidos) são monitoradas ao longo do tempo por meio de automação e se regressões acionam um alerta.

AC.8.3 Triagem de Conteúdo e Política

#AC.8.3.1 Nível: 2 Papel: D

Verifique se o modelo de triagem ou o conjunto de regras é re-treinado/atualizado pelo menos trimestralmente, incorporando novos padrões de jailbreak ou de violação de políticas observados.

AC.8.4 Limitação da Taxa de Entrada e Prevenção de Abuso

#AC.8.4.1 Nível: 3 Papel: V

Verifique se os registros de prevenção de abusos são mantidos e revisados para identificar padrões de ataques emergentes.

AC.8.5 Procedência e Atribuição de Entrada

#AC.8.5.1 Nível: 1 Papel: D/V

Verifique se todas as entradas de usuário são marcadas com metadados (ID do usuário, sessão, fonte, carimbo de data/hora, endereço IP) no momento da ingestão.

#AC.8.5.2 Nível: 2 Papel: D/V

Verifique se os metadados de proveniência são mantidos e auditáveis para todas as entradas processadas.

#AC.8.5.3 Nível: 2 Papel: D/V

Verifique se fontes de entrada anômalas ou não confiáveis são sinalizadas e sujeitas a escrutínio aprimorado ou bloqueio.

AC.9 Validação Multimodal, MLOps e Governança de Infraestrutura

AC.9.1 Pipeline de Validação de Segurança Multimodal

#AC.9.1.1 Nível: 3 Papel: D/V

Verifique se os classificadores de conteúdo específicos por modalidade são atualizados conforme os cronogramas documentados (mínimo trimestralmente) com novos padrões de ameaças, exemplos adversariais e benchmarks de desempenho mantidos acima dos limiares básicos.

AC.9.2 Segurança de CI/CD e Build

#AC.9.2.1 Nível: 1 Papel: D/V

Verifique se a infraestrutura como código é analisada em cada commit e se as mesclagens são bloqueadas quando há descobertas de severidade crítica ou alta.

#AC.9.2.2 Nível: 2 Papel: D/V

Verifique se os pipelines CI/CD utilizam identidades de curta duração e com escopo limitado para acesso a segredos e infraestrutura.

#AC.9.2.3 Nível: 2 Papel: D/V

Verifique se os ambientes de build estão isolados das redes e dados de produção.

AC.9.3 Segurança de Contêineres e Imagens

#AC.9.3.1 Nível: 2 Papel: D/V

Verifique se as imagens de contêiner são escaneadas para bloquear segredos hardcoded (por exemplo, chaves de API, credenciais, certificados).

#AC.9.3.2 Nível: 1 Papel: D/V

Verifique se as imagens de contêiner são examinadas de acordo com os cronogramas organizacionais, com vulnerabilidades CRÍTICAS bloqueando a implantação com base nos limites de risco organizacional.

AC.9.4 Monitoramento, Alertas e SIEM

#AC.9.4.1 Nível: 2 Papel: V

Verifique se os alertas de segurança se integram com plataformas SIEM (Splunk, Elastic ou Sentinel) utilizando os formatos CEF ou STIX/TAXII com enriquecimento automatizado.

AC.9.5 Gestão de Vulnerabilidades

#AC.9.5.1 Nível: 2 Papel: D/V

Verifique se as vulnerabilidades de alta gravidade estão corrigidas de acordo com os prazos de gerenciamento de risco organizacional, incluindo procedimentos de emergência para CVEs explorados ativamente.

AC.9.6 Controle de Configuração e Deriva

#AC.9.6.1 Nível: 2 Papel: D/V

Verifique se a deriva de configuração é detectada utilizando ferramentas (Chef InSpec, AWS Config) de acordo com os requisitos de monitoramento organizacional, com reversão automática para alterações não autorizadas.

AC.9.7 Endurecimento do Ambiente de Produção

#AC.9.7.1 Nível: 2 Papel: D/V

Verifique se os ambientes de produção bloqueiam o acesso SSH, desativam os pontos finais de depuração e exigem solicitações de alteração com requisitos de aviso prévio organizacional, exceto em emergências.

Portões de Promoção de Versão AC.9.8

#AC.9.8.1 Nível: 2 Papel: D/V

Verifique se os portões de promoção incluem testes automatizados de segurança (SAST, DAST, varredura de container) com zero achados CRÍTICOS exigidos para aprovação.

AC.9.9 Monitoramento de Carga de Trabalho, Capacidade e Custo

#AC.9.9.1 Nível: 1 Papel: D/V

Verifique se a utilização da GPU/TPU é monitorada com alertas acionados em limiares definidos pela organização e se o dimensionamento automático ou o balanceamento de carga são ativados com base em políticas de gerenciamento de capacidade.

#AC.9.9.2 Nível: 1 Papel: D/V

Verifique se as métricas de carga de trabalho de IA (latência de inferência, taxa de transferência, taxas de erro) são coletadas de acordo com os requisitos de monitoramento organizacional e correlacionadas com a utilização da infraestrutura.

#AC.9.9.3 Nível: 2 Papel: V

Verifique se o monitoramento de custos acompanha os gastos por carga de trabalho/inquilino com alertas baseados nos limites orçamentários organizacionais e controles automatizados para excessos de orçamento.

#AC.9.9.4 Nível: 3 Papel: V

Verifique se o planejamento de capacidade utiliza dados históricos com períodos de previsão definidos pela organização e provisionamento automatizado de recursos com base em padrões de demanda.

AC.9.10 Aprovações e Trilhas de Auditoria

#AC.9.10.1 Nível: 1 Papel: D/V

Verifique que a promoção de ambiente requer aprovação de pessoal autorizado definido pela organização, com assinaturas criptográficas e trilhas de auditoria imutáveis.

AC.9.11 Governança de IaC

#AC.9.11.1 Nível: 2 Papel: D/V

Verifique se as alterações de infraestrutura como código exigem revisão por pares com testes automatizados e análise de segurança antes da mesclagem na branch principal.

AC.9.12 Manipulação de Dados em Ambiente Não-Produtivo

#AC.9.12.1 Nível: 2 Papel: D/V

Verifique se os dados fora de produção estão anonimizados de acordo com os requisitos organizacionais de privacidade, geração de dados sintéticos ou mascaramento completo dos dados com remoção de informações pessoais identificáveis (PII) verificada.

AC.9.13 Backup & Recuperação de Desastres

#AC.9.13.1 Nível: 1 Papel: D/V

Verifique se as configurações da infraestrutura estão sendo copiadas de segurança de acordo com os cronogramas de backup organizacionais para regiões geograficamente separadas, com a implementa-

ção da estratégia de backup 3-2-1.

#AC.9.13.2 Nível: 2 Papel: V

Verifique se os procedimentos de recuperação são testados e validados por meio de testes automatizados de acordo com os cronogramas organizacionais, com as metas de RTO e RPO atendendo aos requisitos da organização.

#AC.9.13.3 Nível: 3 Papel: V

Verifique se a recuperação de desastres inclui runbooks específicos para IA com restauração de pesos do modelo, reconstrução de cluster GPU e mapeamento de dependências de serviço.

AC.9.14 Conformidade e Documentação

#AC.9.14.1 Nível: 2 Papel: D/V

Verifique se a conformidade da infraestrutura é avaliada conforme os cronogramas organizacionais em relação aos controles SOC 2, ISO 27001 ou FedRAMP, com coleta automatizada de evidências.

#AC.9.14.2 Nível: 2 Papel: V

Verifique se a documentação da infraestrutura inclui diagramas de rede, mapas de fluxo de dados e modelos de ameaça atualizados de acordo com os requisitos de gerenciamento de mudanças organizacionais.

#AC.9.14.3 Nível: 3 Papel: D/V

Verifique se as alterações na infraestrutura passam por uma avaliação automatizada de impacto de conformidade com fluxos de trabalho de aprovação regulatória para modificações de alto risco.

AC.9.15 Hardware e Cadeia de Suprimentos

#AC.9.15.1 Nível: 2 Papel: D/V

Verifique se o firmware do acelerador de IA (BIOS da GPU, firmware do TPU) é verificado com assinaturas criptográficas e atualizado de acordo com os prazos de gerenciamento de patches da organização.

#AC.9.15.2 Nível: 3 Papel: V

Verifique se a cadeia de suprimentos de hardware de IA inclui a verificação de procedência com certificados do fabricante e validação de embalagem à prova de violação.

AC.9.16 Estratégia e Portabilidade na Nuvem

#AC.9.16.1 Nível: 3 Papel: V

Verifique se a prevenção do bloqueio do fornecedor de nuvem inclui infraestrutura como código portátil, APIs padronizadas e capacidades de exportação de dados com ferramentas de conversão de formato.

#AC.9.16.2 Nível: 3 Papel: V

Verifique se a otimização de custos multi-cloud inclui controles de segurança que evitam o espalhamento de recursos, bem como cobranças não autorizadas de transferência de dados entre nuvens.

AC.9.17 GitOps e Auto-Cura

#AC.9.17.1 Nível: 2 Papel: D/V

Verifique se os repositórios GitOps exigem commits assinados com chaves GPG e regras de proteção de branch que impeçam pushes diretos para os branches principais.

#AC.9.17.2 Nível: 3 Papel: V

Verifique se a infraestrutura autocurável inclui correlação de eventos de segurança com resposta auto-

matizada a incidentes e fluxos de trabalho de notificação às partes interessadas.

AC.9.18 Confiança Zero, Agentes, Provisionamento e Atenticação de Residência

#AC.9.18.1 Nível: 2 Papel: D/V

Verifique se o acesso aos recursos na nuvem inclui verificação de zero-confiança com autenticação contínua.

#AC.9.18.2 Nível: 2 Papel: D/V

Verifique se o provisionamento automatizado de infraestrutura inclui a validação da política de segurança com bloqueio de implantação para configurações não conformes.

#AC.9.18.3 Nível: 2 Papel: D/V

Verifique se o provisionamento automatizado da infraestrutura valida as políticas de segurança durante o CI/CD, bloqueando configurações não conformes para implantação.

#AC.9.18.4 Nível: 3 Papel: D/V

Verifique se os requisitos de residência de dados são aplicados por meio de atestação criptográfica dos locais de armazenamento.

#AC.9.18.5 Nível: 3 Papel: D/V

Verifique se as avaliações de segurança do provedor de nuvem incluem modelagem de ameaças específica para agentes e avaliação de riscos.

AC.9.19 Controle de Acesso e Identidade

#5.1.3 Nível: 2 Papel: D

Verifique se os novos responsáveis passam por verificação de identidade alinhada com o padrão NIST 800-63-3 IAL-2 ou padrões equivalentes antes de receberem acesso ao sistema de produção.

#5.1.4 Nível: 2 Papel: V

Verifique se as revisões de acesso são realizadas trimestralmente com detecção automatizada de contas inativas, aplicação da rotação de credenciais e fluxos de trabalho de desprovisionamento.

#5.2.2 Nível: 1 Papel: D/V

Verifique se os princípios de menor privilégio são aplicados por padrão com contas de serviço começando com permissões somente de leitura e justificativa comercial documentada exigida para acesso de gravação.

#5.3.3 Nível: 2 Papel: D

Verifique se as definições de políticas são controladas por versão, revisadas por pares e validadas por meio de testes automatizados em pipelines de CI/CD antes da implantação em produção.

#5.3.4 Nível: 2 Papel: V

Verifique se os resultados da avaliação de políticas incluem fundamentações das decisões e são transmitidos para sistemas SIEM para análise de correlação e relatórios de conformidade.

#5.4.4 Nível: 2 Papel: V

Verifique se a latência da avaliação da política é monitorada continuamente com alertas automatizados para condições de tempo limite que possam possibilitar a evasão de autorização.

#5.5.4 Nível: 2 Papel: V

Verifique se os algoritmos de redação são determinísticos, controlados por versão e mantêm registros de auditoria para apoiar investigações de conformidade e análise forense.

#5.5.5 Nível: 3 Papel: V

Verifique se os eventos de redação de alto risco geram logs adaptativos que incluem hashes criptográficos do conteúdo original para recuperação forense sem exposição de dados.

#5.7.5 Nível: 3 Papel: V

Verifique se as condições de erro do agente e o tratamento de exceções incluem informações sobre o

escopo da capacidade para apoiar a análise de incidentes e a investigação forense.

#5.4.2 Nível: 1 Papel: D/V

Verifique se as citações, referências e atribuições de fontes nos resultados do modelo estão validadas contra as permissões do solicitante e removidas caso seja detectado acesso não autorizado.

Novos Itens a serem Integrados Acima

#2.3.3 Nível: 2 Papel: D/V

Verifique se o conjunto de caracteres permitidos é regularmente revisado e atualizado para garantir que permaneça alinhado com os requisitos de negócios.



Apêndice D: Governança e Verificação de Codificação Segura Assistida

Objetivo

Este capítulo define controles organizacionais básicos para o uso seguro e eficaz de ferramentas de codificação assistida por IA durante o desenvolvimento de software, garantindo segurança e rastreabilidade ao longo do SDLC.

AD.1 Fluxo de Trabalho de Codificação Segura Assistida por IA

Integrar ferramentas de IA no ciclo de vida de desenvolvimento seguro de software (SSDLC) da organização sem enfraquecer as barreiras de segurança existentes.

#AD.1.1 Nível: 1 Papel: D/V

Verifique se um fluxo de trabalho documentado descreve quando e como as ferramentas de IA podem gerar, refatorar ou revisar código.

#AD.1.2 Nível: 2 Papel: D

Verifique se o fluxo de trabalho corresponde a cada fase do SSDLC (design, implementação, revisão de código, testes, implantação).

#AD.1.3 Nível: 3 Papel: D/V

Verifique se as métricas (por exemplo, densidade de vulnerabilidade, tempo médio para detectar) são coletadas no código produzido por IA e comparadas com as bases de referência exclusivamente humanas.

AD.2 Qualificação de Ferramentas de IA e Modelagem de Ameaças

Assegure que as ferramentas de codificação de IA sejam avaliadas quanto às capacidades de segurança, riscos e impacto na cadeia de suprimentos antes da adoção.

#AD.2.1 Nível: 1 Papel: D/V

Verifique se um modelo de ameaça para cada ferramenta de IA identifica riscos de uso indevido, inversão de modelo, vazamento de dados e cadeia de dependência.

#AD.2.2 Nível: 2 Papel: D

Verifique se as avaliações das ferramentas incluem análise estática/dinâmica de quaisquer componentes locais e avaliação dos endpoints SaaS (TLS, autenticação/autorização, registro).

#AD.2.3 Nível: 3 Papel: D/V

Verifique se as avaliações seguem um framework reconhecido e são refeitas após mudanças significativas de versão.

AD.3 Gerenciamento Seguro de Prompt e Contexto

Evite o vazamento de segredos, código proprietário e dados pessoais ao construir prompts ou contextos para modelos de IA.

#AD.3.1 Nível: 1 Papel: D/V

Verifique se as orientações escritas proíbem o envio de segredos, credenciais ou dados classificados em prompts.

#AD.3.2 Nível: 2 Papel: D

Verifique se os controles técnicos (redação no lado do cliente, filtros de contexto aprovados) removem automaticamente artefatos sensíveis.

#AD.3.3 Nível: 3 Papel: D/V

Verifique se os prompts e respostas são tokenizados, criptografados durante a transmissão e em repouso, e se os períodos de retenção estão em conformidade com a política de classificação de dados.

AD.4 Validação de Código Gerado por IA

Detectar e corrigir vulnerabilidades introduzidas pela saída de IA antes que o código seja mesclado ou implantado.

#AD.4.1 Nível: 1 Papel: D/V

Verifique se o código gerado por IA é sempre submetido à revisão humana de código.

#AD.4.2 Nível: 2 Papel: D

Verifique se scanners automatizados (SAST/IAST/DAST) são executados em cada pull request contendo código gerado por IA e bloquee merges em casos de descobertas críticas.

#AD.4.3 Nível: 3 Papel: D/V

Verifique se os testes diferenciais de fuzzing ou os testes baseados em propriedades comprovam comportamentos críticos para a segurança (por exemplo, validação de entrada, lógica de autorização).

AD.5 Explicabilidade e Rastreabilidade das Sugestões de Código

Fornecer a auditores e desenvolvedores uma compreensão sobre por que uma sugestão foi feita e como ela evoluiu.

#AD.5.1 Nível: 1 Papel: D/V

Verifique se os pares de prompt/resposta são registrados com IDs de commit.

#AD.5.2 Nível: 2 Papel: D

Verifique se os desenvolvedores podem exibir citações do modelo (trechos de treinamento, documentação) que suportem uma sugestão.

#AD.5.3 Nível: 3 Papel: D/V

Verifique se os relatórios de explicabilidade estão armazenados com os artefatos de design e referenciados nas revisões de segurança, satisfazendo os princípios de rastreabilidade da ISO/IEC 42001.

AD.6 Feedback Contínuo e Ajuste Fino do Modelo

Melhorar o desempenho de segurança do modelo ao longo do tempo, evitando deriva negativa.

#AD.6.1 Nível: 1 Papel: D/V

Verifique se os desenvolvedores podem sinalizar sugestões inseguras ou não conformes, e se as sinalizações são rastreadas.

#AD.6.2 Nível: 2 Papel: D

Verifique se o feedback agregado informa o ajuste fino periódico ou a geração aumentada por recuperação com corpora de codificação segura verificados (por exemplo, OWASP Cheat Sheets).

#AD.6.3 Nível: 3 Papel: D/V

Verifique se um ambiente de avaliação em loop fechado executa testes de regressão após cada ajuste fino; as métricas de segurança devem atingir ou superar as linhas de base anteriores antes da implantação.

Referências

- NIST AI Risk Management Framework 1.0
- ISO/IEC 42001:2023 – AI Management Systems Requirements
- OWASP Secure Coding Practices – Quick Reference Guide

Apêndice E: Exemplos de Ferramentas e Frameworks

Objetivo

Este capítulo fornece exemplos de ferramentas e frameworks que podem apoiar a implementação ou cumprimento de um determinado requisito AISVS. Estes não devem ser vistos como recomendações ou endossos pela equipe AISVS ou pelo Projeto de Segurança GenAI da OWASP.

AE.1 Governança de Dados de Treinamento e Gestão de Viés

Ferramentas usadas para análise de dados, governança e gestão de viés.

#AE.1.1 Seção: 1.1

Ferramentas de Inventário de Dados: Ferramentas de gerenciamento de inventário de dados como...

#AE.1.2 Seção: 1.2

Criptografia em Trânsito Use TLS para aplicações baseadas em HTTPS, com ferramentas como openSSL e python's ssl biblioteca.

AE.2 Validação da Entrada do Usuário

Ferramentas para manipular e validar entradas de usuário.

#AE.2.1 Seção: 2.1

Ferramentas de Defesa contra Injeção de Prompt: Use ferramentas de proteção como NeMo da NVIDIA ou Guardrails AI.

Apêndice B: Controles Estratégicos

C4.15 Segurança de Infraestrutura Resistente a Quantum

Prepare a infraestrutura de IA para ameaças da computação quântica por meio de criptografia pós-quântica e protocolos seguros contra computação quântica.

#4.15.1 Nível: 3 Papel: D/V

Verifique se a infraestrutura de IA implementa algoritmos criptográficos pós-quânticos aprovados pelo NIST (CRYSTALS-Kyber, CRYSTALS-Dilithium, SPHINCS+) para troca de chaves e assinaturas digitais.

#4.15.2 Nível: 3 Papel: D/V

Verifique se os sistemas de distribuição quântica de chaves (QKD) estão implementados para comunicações de IA de alta segurança com protocolos de gerenciamento de chaves à prova de ataques quânticos.

#4.15.3 Nível: 3 Papel: D/V

Verifique se as estruturas de agilidade criptográfica permitem migração rápida para novos algoritmos pós-quânticos com rotação automatizada de certificados e chaves.

#4.15.4 Nível: 3 Papel: V

Verifique se a modelagem de ameaças quânticas avalia a vulnerabilidade da infraestrutura de IA a ataques quânticos, com cronogramas de migração documentados e avaliações de risco.

#4.15.5 Nível: 3 Papel: D/V

Verifique se os sistemas criptográficos híbridos clássico-quânticos oferecem defesa em profundidade durante o período de transição quântica com monitoramento de desempenho.

C4.17 Infraestrutura de Conhecimento Zero

Implemente sistemas de prova de conhecimento zero para verificação e autenticação de IA preservando a privacidade, sem revelar informações sensíveis.

#4.17.1 Nível: 3 Papel: D/V

Verifique se as provas de conhecimento zero (ZK-SNARKs) validam a integridade do modelo de IA e a origem do treinamento sem expor os pesos do modelo ou os dados de treinamento.

#4.17.2 Nível: 3 Papel: D/V

Verifique se os sistemas de autenticação baseados em ZK possibilitam a verificação do usuário preservando a privacidade para serviços de IA sem revelar informações relacionadas à identidade.

#4.17.3 Nível: 3 Papel: D/V

Verifique se os protocolos de interseção privada de conjuntos (PSI) permitem a correspondência segura de dados para IA federada sem expor conjuntos de dados individuais.

#4.17.4 Nível: 3 Papel: D/V

Verifique se os sistemas de aprendizado de máquina de conhecimento zero (ZKML) permitem inferências de IA verificáveis com prova criptográfica de cálculo correto.

#4.17.5 Nível: 3 Papel: D/V

Verifique se os ZK-rollups fornecem processamento de transações de IA escalável e que preserva a privacidade, com verificação em lote e redução da sobrecarga computacional.

C4.18 Prevenção de Ataques de Canal Lateral

Proteja a infraestrutura de IA contra ataques de canal lateral baseados em tempo, energia, eletromagnéticos e cache que possam vazar informações sensíveis.

#4.18.1 Nível: 3 Papel: D/V

Verifique se o tempo de inferência de IA é normalizado usando algoritmos de tempo constante e preenchimento para prevenir ataques de extração de modelo baseados em tempo.

#4.18.2 Nível: 3 Papel: D/V

Verifique se a proteção contra análise de potência inclui injeção de ruído, filtragem da linha de alimentação e padrões de execução randomizados para hardware de IA.

#4.18.3 Nível: 3 Papel: D/V

Verifique se a mitigação de canal lateral baseada em cache usa particionamento de cache, randomização e instruções de limpeza para evitar vazamento de informações.

#4.18.4 Nível: 3 Papel: D/V

Verifique se a proteção contra emissões eletromagnéticas inclui blindagem, filtragem de sinais e processamento randomizado para prevenir ataques do tipo TEMPEST.

#4.18.5 Nível: 3 Papel: D/V

Verifique se as defesas contra canais laterais microarquiteturais incluem controles de execução especulativa e ofuscação do padrão de acesso à memória.

C4.19 Segurança de Hardware Neuromórfico e de IA Especializada

Garantir a segurança das arquiteturas emergentes de hardware de IA, incluindo chips neuromórficos, FPGAs, ASICs personalizados e sistemas de computação óptica.

#4.19.1 Nível: 3 Papel: D/V

Verifique se a segurança do chip neuromórfico inclui criptografia de padrão de spikes, proteção do peso sináptico e validação da regra de aprendizado baseada em hardware.

#4.19.2 Nível: 3 Papel: D/V

Verifique se os aceleradores de IA baseados em FPGA implementam criptografia de bitstream, mecanismos anti-sabotagem e carregamento seguro de configuração com atualizações autenticadas.

#4.19.3 Nível: 3 Papel: D/V

Verifique se a segurança do ASIC personalizado inclui processadores de segurança integrados no chip, raiz de confiança de hardware e armazenamento seguro de chaves com detecção de violação.

#4.19.4 Nível: 3 Papel: D/V

Verifique se os sistemas de computação óptica implementam criptografia óptica à prova de ataques quânticos, comutação fotônica segura e processamento de sinais ópticos protegido.

#4.19.5 Nível: 3 Papel: D/V

Verifique se os chips de IA híbridos analógico-digitais incluem computação analógica segura, armazenamento protegido de pesos e conversão analógico-digital autenticada.

C4.20 Infraestrutura de Computação que Preserva a Privacidade

Implemente controles de infraestrutura para computação que preserva a privacidade, a fim de proteger dados sensíveis durante o processamento e análise de IA.

#4.20.1 Nível: 3 Papel: D/V

Verifique se a infraestrutura de criptografia homomórfica permite computação criptografada em cargas de trabalho sensíveis de IA com verificação de integridade criptográfica e monitoramento de desempenho.

#4.20.2 Nível: 3 Papel: D/V

Verifique se os sistemas de recuperação de informações privadas permitem consultas ao banco de dados sem revelar padrões de consulta, com proteção criptográfica dos padrões de acesso.

#4.20.3 Nível: 3 Papel: D/V

Verifique se os protocolos de computação multipartidária segura permitem inferência de IA preservando a privacidade, sem expor entradas individuais ou cálculos intermediários.

#4.20.4 Nível: 3 Papel: D/V

Verifique se o gerenciamento de chaves que preserva a privacidade inclui geração distribuída de chaves, criptografia de limiar e rotação segura de chaves com proteção suportada por hardware.

#4.20.5 Nível: 3 Papel: D/V

Verifique se o desempenho da computação preservadora de privacidade está otimizado por meio de agrupamento, cache e aceleração de hardware, mantendo as garantias de segurança criptográfica.

4.9.14.9.2 1:2 D/V: D/V

Verificar se as implantações multi-nuvem utilizam padrões de identidade federada (por exemplo, OIDC, SAML) com aplicação centralizada de políticas entre os provedores.

4.9.14.9.3 1:2 D/V: D/V

Verificar se as transferências de dados entre nuvens e híbridas utilizam criptografia de ponta a ponta com chaves gerenciadas pelo cliente e aplicam os requisitos de residência de dados jurisdicionais.

4.9.14.9.1 1:1 D/V: D/V

Verifique se a integração de armazenamento em nuvem utiliza criptografia de ponta a ponta com gerenciamento de chaves controlado pelo agente.

4.9.14.9.2 1:2 D/V: D/V

Verificar se os limites de segurança do deployment híbrido estão claramente definidos com canais de comunicação criptografados.