

Glossário

Edição 202610

Copyright © EXIN Holding B.V. 2026. All rights reserved.
EXIN® is a registered trademark of EXIN.

No part of this publication may be reproduced, stored, utilized or transmitted in any form or by any means, electronic, mechanical, or otherwise, without the prior written permission from EXIN.

Introdução

Este glossário reúne os principais termos e conceitos relacionados à certificação EXIN AI Security Professional based on the OWASP AI Exchange.

O glossário apresenta as definições dos conceitos básicos incluídos no guia de preparação e abordados no exame, auxiliando os candidatos a se prepararem adequadamente.

As definições apresentadas estão alinhadas ao conteúdo do OWASP AI Exchange.

Glossário

Os termos e suas respectivas definições em português são apresentados segundo a ordem alfabética das denominações originais em inglês.

Termo	Definição
responsabilização e prestação de contas *accountability*	Princípio segundo o qual indivíduos ou organizações respondem por suas ações e decisões, podendo ser chamados a explicá-las, justificá-las e assumir a responsabilidade pelos resultados.
ataque adversarial *adversarial attack*	Termo com múltiplos significados, às vezes usado para se referir a qualquer ataque e, outras vezes, a uma tentativa de criar ou manipular deliberadamente entradas, dados ou um modelo, a fim de fazer com que um sistema de aprendizado de máquina produza saídas incorretas, prejudiciais ou escolhidas pelo atacante.
testes adversariais *adversarial testing*	Testes realizados por meio de ataques adversariais, com o objetivo de identificar vulnerabilidades exploráveis antes que invasores reais o façam.
IA agêntica *agentic AI*	Sistemas construídos em torno de agentes de IA capazes de planejar, realizar ações e interagir com ferramentas, dados ou outros agentes para atingir objetivos com intervenção humana limitada.
inteligência artificial (IA) *AI (artificial intelligence)*	A capacidade de um sistema de computador de realizar tarefas que normalmente exigem inteligência semelhante à humana, como raciocínio, aprendizagem, reconhecimento de padrões ou criação de conteúdo, seja por meio de métodos baseados em regras ou de aprendizado de máquina.
agente de IA *AI agent*	Um componente de software capaz de decidir e executar ações de forma autônoma — como chamar funções ou ferramentas — para atingir um objetivo, em vez de apenas produzir uma resposta.
ativos de IA *AI assets*	Os elementos específicos de IA pertencentes a um determinado sistema ou organização que precisam ser protegidos, como dados de treinamento e parâmetros de modelos.
jailbreak de IA *AI jailbreak*	Um ataque de injeção de prompt que contorna as restrições de segurança operacional ou de conteúdo integradas a um modelo para fazê-lo gerar saídas que ele foi projetado para recusar.
sistema de gestão de IA (SGIA) *AI management system (AIMS)*	Uma estrutura organizacional de políticas e processos, conforme definido, por exemplo, pela norma ISO/IEC 42001, destinada a promover a governança responsável e segura da IA.
modelo de IA *AI model*	Um artefato matemático ou computacional que mapeia entradas para saídas, aprendido a partir de dados ou definido por regras.
fornecedor do modelo de IA *AI model provider*	Uma organização ou entidade que desenvolve, treina ou fornece um modelo de IA para utilização direta ou como base para outros sistemas.

Termo	Definição
princípios de privacidade da IA *AI privacy principles*	Um conjunto de princípios sobre como os dados pessoais devem ser tratados, abrangendo aspectos como exatidão, consentimento, minimização, explicabilidade, equidade, transparência e limitação.
• **exatidão (dos dados) / acurácia (do modelo)** • *accuracy*	No contexto dos dados pessoais, exatidão é o princípio segundo o qual os dados devem estar corretos e, quando necessário, atualizados. No contexto dos modelos, acurácia é o grau em que as saídas produzidas pelo sistema de IA correspondem aos resultados esperados ou corretos.
• **consentimento** • *consent*	Manifestação livre, específica, informada e inequívoca pela qual o titular concorda com o tratamento de seus dados pessoais para uma finalidade determinada.
• **minimização de dados e limitação de armazenamento** • *data minimization and storage limitation*	O princípio de coletar apenas os dados necessários para uma finalidade específica e de mantê-los apenas pelo tempo necessário para essa finalidade.
• **direitos do titular dos dados** • *data subject rights*	Direitos individuais sobre os dados pessoais, tais como acesso, retificação e exclusão.
• **explicabilidade** • *explainability*	O princípio de que os indivíduos devem ser informados sobre como seus dados são utilizados e, quando decisões automatizadas os afetam, devem ser capazes de compreender através de que etapas essas decisões foram tomadas. Em outras palavras, é a capacidade de apresentar explicações compreensíveis sobre os fatores, fundamentos ou processos que levaram um sistema de IA a produzir determinado resultado, considerando o público destinatário.
• **equidade e licitude** • *fairness and lawfulness*	O princípio de que os dados pessoais devem ser tratados de forma legalmente justificada e que não prejudique injustamente indivíduos ou grupos.
• **transparência** • *transparency*	O princípio de que as pessoas devem ser informadas sobre como seus dados são coletados, utilizados, compartilhados e retidos e, quando decisões automatizadas as afetam, devem poder compreender a abordagem utilizada para chegar a essas decisões.
• **limitação de uso e especificação da finalidade** • *use limitation and purpose specification*	O princípio de privacidade segundo o qual os dados pessoais coletados para uma finalidade não devem ser simplesmente reutilizados para treinar um modelo de IA com uma finalidade diferente e incompatível, sem uma base legal (hipótese de tratamento, no caso da LGPD) adequada.

Termo	Definição
red teaming de IA *AI red teaming*	Um exercício estruturado de testes adversariais, no qual os testadores simulam a atuação de invasores reais para identificar fragilidades relacionadas à segurança (security), à segurança operacional (safety) ou à ética em um sistema de IA.
segurança da IA *AI security*	A disciplina de proteger sistemas de IA, seus dados e sua infraestrutura de suporte contra ameaças que exploram características específicas da IA, além das ameaças convencionais de cibersegurança.
governança de segurança da IA *AI security governance*	As estruturas organizacionais, políticas e processos de supervisão destinados a garantir que os riscos de segurança relacionados à IA sejam identificados e gerenciados com responsabilidades definidas e mecanismos de prestação de contas.
abordagem de testes de segurança de IA *AI security testing approach*	Uma metodologia estruturada para planejar e executar testes de segurança em um sistema de IA, abrangendo escopo, técnicas e critérios de sucesso.
sistema de IA *AI system*	Um sistema completo construído em torno de um ou mais modelos de IA, incluindo seus fluxos de dados, interfaces e infraestrutura de suporte.
fornecedor do sistema de IA *AI system provider*	A organização que desenvolve ou monta um sistema de IA e o disponibiliza para implantação ou uso.
anonimização *anonymization*	Tratamento de dados pessoais que torna o titular não identificável, direta ou indiretamente, sem possibilidade de reidentificação pelos meios razoavelmente disponíveis no contexto do tratamento.
API (interface de programação de aplicações) *API (application programming interface)*	Conjunto definido de protocolos e ferramentas que permite que diferentes programas ou componentes de software se comuniquem e troquem funcionalidades.
auditoria *audit*	Uma análise independente de um sistema, processo ou organização para verificar a conformidade com normas, políticas ou regulamentações.
dados de aumento (augmentation data) *augmentation data*	Dados ou informações externas adicionadas ao contexto de um modelo em algum momento de seu ciclo de vida para melhorar ou ampliar sua saída, como documentos recuperados para a geração aumentada por recuperação (RAG).
supervisão automatizada *automated oversight*	Monitoramento ou controle de um sistema realizado por ferramentas automatizadas, em vez de por pessoas.
automação *automation*	O uso de tecnologia para executar tarefas ou processos com intervenção humana reduzida.

Termo	Definição
viés *bias*	Um desvio sistemático em dados, decisões ou saídas frequentemente introduzido de forma não intencional por meio de escolhas relacionadas a dados ou ao projeto. Dependendo do caso, o viés pode levar a um tratamento injusto ou impreciso de determinados grupos ou casos.
raio de impacto (blast radius) *blast radius*	A extensão dos danos, do comprometimento ou da exposição que resultariam caso um determinado ativo, componente ou credencial fosse atacado com sucesso ou apresentasse falha.
chatbot *chatbot*	Um aplicativo de software que interage com usuários por meio de conversas por texto ou voz, realizando a automação de respostas que, de outra forma, exigiriam a intervenção humana.
Chief AI Officer (CAIO) *Chief AI Officer (CAIO)*	Função organizacional responsável pela estratégia geral de IA, governança e supervisão de riscos de uma organização.
Chief Information Security Officer (CISO) *Chief Information Security Officer (CISO)*	O executivo sênior responsável pela estratégia e pelo programa de segurança da informação e cibersegurança de uma organização.
código *code*	Conjunto de instruções, escritas em uma linguagem de programação, que implementa a lógica de um programa de software.
conformidade *compliance*	Aderência às leis, regulamentos, normas e políticas internas aplicáveis que regem as atividades de uma organização.
confidencialidade *confidentiality*	A propriedade pela qual as informações são acessíveis apenas às pessoas autorizadas a acessá-las. Um dos três objetivos principais da segurança da informação, juntamente com a integridade e a disponibilidade.
vazamento de configuração *configuration leak*	A exposição não intencional de detalhes de configuração de um sistema, como ajustes ou credenciais, a terceiros não autorizados.
direitos autorais *copyright*	Forma de proteção legal concedida aos criadores de obras originais, conferindo-lhes direitos exclusivos sobre a reprodução, distribuição e uso dessas obras.
vazamento de dados *data leak*	A exposição não autorizada ou não intencional de dados a pessoas que não deveriam ter acesso a eles.
proteção de dados *data protection*	No contexto dos dados pessoais, medidas e práticas destinadas a proteger os dados pessoais contra acesso, uso ou divulgação não autorizados.
avaliação de impacto sobre a proteção de dados (DPIA) *data protection impact assessment (DPIA)*	Um processo estruturado, exigido em determinadas circunstâncias pela legislação aplicável, para identificar e mitigar riscos à privacidade antes de realizar um tipo de tratamento de dados pessoais suscetível de resultar em risco elevado para as pessoas físicas.

Termo	Definição
segurança de dados *data security*	Proteção de dados contra acesso, alteração ou perda não autorizados ao longo de todo o seu ciclo de vida.
titular dos dados *data subject*	Pessoa física identificada ou identificável a quem se referem os dados pessoais.
negação de serviço (denial-of-service, DoS) *denial-of-service (DoS)*	Um ataque que visa tornar um sistema ou serviço indisponível para usuários legítimos, normalmente sobrecarregando-o com tráfego ou solicitações.
negação de carteira (denial-of-wallet, DoW) *denial-of-wallet (DoW)*	Um ataque que aumenta deliberadamente os custos operacionais de um alvo, por exemplo, provocando solicitações excessivas mediante o consumo excessivo de serviços cobrados conforme o uso, tendo como impacto principal o prejuízo financeiro (e não o prejuízo quanto à disponibilidade).
responsável pela implantação *deployer*	Pessoa ou organização que coloca um sistema em operação, distinguindo-se da parte que o desenvolveu.
fase de desenvolvimento (development-time) *development-time*	A fase do ciclo de vida de um sistema em que ele é projetado, construído e treinado ou testado, em oposição à fase em que está implantado e em execução.
ameaças de segurança em fase de desenvolvimento (development-time) *development-time security threats*	Ameaças de segurança que podem se materializar enquanto um sistema está sendo projetado, construído ou treinado, como comprometimento de código, inserção de backdoors em componentes ou corrupção de dados.
• **envenenamento de dados em fase de desenvolvimento** • *data poisoning development-time*	Corromper ou manipular deliberadamente os dados usados para treinar um modelo de aprendizado de máquina, de modo que o modelo resultante aprenda um comportamento falho ou controlado por invasores.
• **envenenamento de modelo em fase de desenvolvimento** • *model poisoning development-time*	Adulteração direta dos parâmetros, da arquitetura do modelo ou do processo de treinamento de um modelo para incorporar comportamentos ocultos ou maliciosos.
• **vazamento de dados sensíveis em fase de desenvolvimento** • *sensitive data leak development-time*	A exposição não intencional de dados sensíveis, como dados de origem, credenciais ou segredos de configuração, durante o processo de desenvolvimento de um sistema.

Termo	Definição
embedding (representação vetorial) *embedding*	Uma representação vetorial numérica de dados, como texto ou imagens, que captura seu significado semântico para que os modelos de aprendizado de máquina possam processá-los, compará-los ou pesquisá-los.
criptografia *encryption*	Processo de codificação de dados para que possam ser lidos apenas por partes que possuam a chave de descriptografia correta, utilizado para proteger a confidencialidade dos dados.
princípios éticos *ethical principles*	Um conjunto de valores, como justiça, honestidade, responsabilização e prestação de contas, que orientam o que é considerado conduta correta ou aceitável em um determinado campo ou atividade.
União Europeia (UE) *European Union (EU)*	União política e econômica de Estados-membros europeus que adota leis e regulamentos, como o GDPR e a Lei da IA (AI Act), vinculativos para seus membros.
G.U.A.R.D (Governar, Entender, Adaptar, Reduzir, Demonstrar) *G.U.A.R.D (Govern, Understand, Adapt, Reduce, Demonstrate)*	Estrutura de cinco etapas do OWASP AI Exchange destinada a orientar a abordagem de uma organização quanto à gestão de riscos de segurança relacionados à IA: Governar, Entender, Adaptar, Reduzir, Demonstrar.
Regulamento Geral de Proteção de Dados (GDPR) *General Data Protection Regulation (GDPR)*	O regulamento da União Europeia que estabelece regras para o tratamento de dados pessoais e protege os direitos das pessoas a quem esses dados se referem.
IA generativa *generative AI*	Sistemas de IA que criam novos conteúdos, como texto, imagens ou código, normalmente por meio do aprendizado de padrões a partir de grandes conjuntos de dados de treinamento.
sequestro *hijacking*	Um ataque no qual um adversário assume o controle não autorizado de um sistema, sessão, recurso ou canal de comunicação e o redireciona para seus próprios fins.
supervisão humana *human oversight*	O envolvimento de pessoas no monitoramento, na revisão ou na intervenção nas decisões e ações de um sistema automatizado ou autônomo.
fase de inferência (inference-time) *inference-time*	A etapa do ciclo de vida de um modelo de aprendizado de máquina na qual ele é utilizado ativamente para produzir saídas a partir de novas entradas, em oposição ao treinamento.
sistema de gestão de segurança da informação (SGSI) *information security management system (ISMS)*	Uma estrutura sistemática de políticas, processos e controles, nos termos da norma ISO/IEC 27001, para gerenciar os riscos de segurança da informação de uma organização.

Termo	Definição
dados de entrada *input data*	Dados fornecidos a um sistema ou processo com o objetivo de produzir uma saída. No contexto da IA, os dados utilizados como entrada para o modelo de IA.
ameaças de segurança por meio de entrada *input security threats*	Ameaças de segurança decorrentes de dados ou conteúdos fornecidos como entrada a um sistema, incluindo conteúdos maliciosos ou manipulados com o objetivo de alterar seu comportamento.
• **evasão** • *evasion*	Ataque que cria entradas maliciosas na forma de dados, e não de instruções, projetadas especificamente para burlar um sistema de detecção ou classificação, fazendo com que ele produza um resultado incorreto. Por exemplo, um sinal de trânsito sutilmente alterado que um modelo de visão interpreta incorretamente.
• **exfiltração de modelo** • *model exfiltration*	Extração de uma cópia ou equivalente funcional de um modelo proprietário, frequentemente por meio de consultas repetidas, sem autorização.
• **injeção de prompt** • *prompt injection*	Um ataque que manipula as instruções fornecidas a um modelo de IA generativa, por meio do prompt ou do conteúdo injetado, para substituir o comportamento pretendido.
• **esgotamento de recursos** • *resource exhaustion*	Um ataque ou condição que consome recursos computacionais, de memória ou de armazenamento, a ponto de degradar ou negar o serviço a usuários legítimos ou causar grande impacto financeiro.
• **divulgação de dados sensíveis por meio do uso** • *sensitive data disclosure through use*	Divulgação de dados sensíveis de treinamento ou outros dados por meio do uso normal ou abusivo do sistema.
verificação de integridade *integrity verification*	Verificação de que dados, códigos ou outros artefatos digitais não foram alterados ou submetidos a adulteração, normalmente por meio de técnicas como somas de verificação, hashes ou assinaturas digitais.
propriedade intelectual (PI) *intellectual property (IP)*	Criações intelectuais protegidas legalmente, como segredos comerciais, patentes, obras protegidas por direitos autorais e marcas registradas, sobre as quais seu criador ou proprietário detém direitos exclusivos.
ISO/IEC 42001 *ISO/IEC 42001*	Norma internacional que especifica os requisitos para um sistema de gestão de IA, a fim de governar a IA de forma responsável e segura.
modelo de linguagem de grande porte (LLM) *large language model (LLM)*	Um modelo de aprendizado de máquina treinado com vastas quantidades de texto, capaz de interpretar e gerar linguagem natural.

Termo	Definição
vazamento *leakage*	A divulgação não autorizada ou não intencional de informações que deveriam ter permanecido confidenciais ou protegidas.
aprendizado de máquina *machine learning*	Um ramo da IA no qual os sistemas aprendem padrões a partir de dados, em vez de serem explicitamente programados com regras.
inferência de associação (membership inference) *membership inference*	Um ataque que tenta inferir se um registro específico fazia parte dos dados de treinamento de um modelo, revelando potencialmente informações confidenciais sobre indivíduos.
arquitetura do modelo *model architecture*	O projeto estrutural de um modelo, incluindo seus componentes e como eles estão conectados.
inversão do modelo *model inversion*	Um ataque que reconstrói ou aproxima dados de treinamento sensíveis por meio da análise das saídas de um modelo.
vazamento do modelo *model leak*	A exposição não autorizada ou o roubo dos parâmetros, pesos ou arquitetura do modelo.
código aberto *open source*	Software ou outros artefatos técnicos cujo código-fonte ou componentes subjacentes são disponibilizados publicamente para que qualquer pessoa possa usá-los, inspecioná-los, modificá-los e redistribuí-los.
dados de saída *output data*	O resultado produzido por um sistema ou processo em resposta a uma determinada entrada. No contexto da IA, normalmente a saída de um modelo de IA.
sanitização de saída *output sanitization*	Filtragem ou limpeza da saída de um sistema antes que ela seja transmitida ou exibida, a fim de remover conteúdo prejudicial, sensível ou malicioso.
validação de saída *output validation*	Verificação da saída de um sistema em relação a critérios definidos para garantir que ela seja correta, segura e adequada ao uso pretendido antes de ser utilizada ou encaminhada para etapas posteriores.
dados pessoais *personal data*	Qualquer informação relacionada a uma pessoa física identificada ou identificável.
IA preditiva *predictive AI*	Sistemas de IA projetados para prever resultados ou classificar dados com base em padrões aprendidos a partir de dados históricos, em oposição à geração de novos conteúdos.
privacidade *privacy*	O direito de um indivíduo de controlar como suas informações pessoais são coletadas, utilizadas e compartilhadas.
privacidade por padrão (by default) *privacy by default*	O princípio de que as configurações padrão de um sistema devem ser a opção disponível que mais protege a privacidade, sem exigir nenhuma ação adicional por parte do indivíduo, um conceito complementar à privacidade desde a concepção (by design).
privacidade desde a concepção (by design) *privacy by design*	Incorporar proteções de privacidade à arquitetura e aos processos de um sistema desde as fases iniciais de desenvolvimento, em vez de adicioná-las posteriormente.

Termo	Definição
avaliações de impacto sobre a privacidade (PIA) *privacy impact assessments (PIA)*	Uma avaliação utilizada para identificar e abordar os riscos à privacidade que um projeto, sistema ou processo possa representar para os indivíduos.
incidente de privacidade *privacy incident*	Um evento no qual dados pessoais são tratados, acessados ou divulgados de forma que viole as proteções à privacidade.
tratamento (de dados pessoais) *processing (of personal data)*	Qualquer operação realizada sobre dados pessoais, como coleta, armazenamento, uso ou exclusão.
prompt *prompt*	O texto de entrada ou as instruções fornecidas a um modelo de IA generativa para obter uma resposta.
pseudonimização *pseudonymization*	Substituição de informações identificadoras nos dados por identificadores artificiais, de modo que sejam necessárias informações adicionais, mantidas separadamente, para reidentificar o indivíduo.
acessível ao público *publicly accessible*	Dados, sistemas ou recursos que podem ser acessados por qualquer pessoa sem autenticação ou permissão especial.
modelo pronto para uso (ready-made model) *ready-made model*	Um modelo pré-treinado obtido de terceiros e utilizado tal como está ou com adaptações, em vez de ser treinado do zero.
risco residual *residual risk*	O risco que permanece após a aplicação de controles de segurança e medidas de mitigação.
IA responsável *responsible AI*	Uma abordagem para o desenvolvimento e a operação de sistemas de IA que considera a ética, a equidade, a segurança operacional (safety) e a segurança (security), além do desempenho.
geração aumentada por recuperação (RAG) *retrieval-augmented generation (RAG)*	Técnica a partir da qual um modelo de IA generativa recupera documentos ou dados externos relevantes durante a inferência para fundamentar e enriquecer suas respostas.
risco *risk*	A possibilidade de uma ameaça explorar uma vulnerabilidade e causar danos, normalmente avaliada em termos de probabilidade e impacto.
análise de risco *risk analysis*	O processo de identificar e avaliar riscos para determinar sua probabilidade e impacto potencial.

Termo	Definição
níveis de risco para sistemas de IA com base na Lei da IA (AI Act) *risk levels for AI systems based on AI Act*	A classificação da Lei da IA da UE dos sistemas de IA em níveis de risco regulatório, que determina quais obrigações legais se aplicam a um determinado sistema.
• **risco mínimo ou nulo** • *minimal or no risk*	O nível previsto na Lei da IA para sistemas de IA considerados de risco insignificante à saúde, à segurança ou aos direitos fundamentais, sujeitos a pouca ou nenhuma regulamentação.
• **risco limitado** • *limited risk*	O nível previsto na Lei da IA para sistemas de IA que representam algum risco, normalmente exigindo obrigações específicas de transparência, como a divulgação de que o conteúdo é gerado por IA.
• **risco elevado** • *high risk*	Nível previsto na Lei da IA para sistemas de IA que podem afetar significativamente a saúde, a segurança ou os direitos fundamentais, sujeitos a obrigações rigorosas antes e após a implantação.
• **risco inaceitável** • *unacceptable risk*	Nível previsto na Lei da IA para práticas de IA consideradas de risco tão grave que são totalmente proibidas.
gerenciamento de riscos *risk management*	Processo contínuo de identificação, avaliação, tratamento e monitoramento de riscos para mantê-los dentro de níveis aceitáveis.
mitigação de riscos *risk mitigation*	A seleção e implementação de medidas para reduzir a probabilidade ou o impacto de um risco identificado.
classificação de riscos *risk rating*	Uma pontuação ou classificação atribuída a um risco com base em sua probabilidade e impacto avaliados.
fase de execução (runtime) *runtime*	A fase do ciclo de vida de um sistema em que ele está implantado e em operação, processando entradas reais em um ambiente real.
ameaças de segurança em fase de execução (runtime) *runtime security threats*	Ameaças de segurança que podem se materializar enquanto um sistema está operando em um ambiente de produção, em oposição às ameaças presentes durante o desenvolvimento.
• **manipulação de dados de aumento (augmentation data)** • *augmentation data manipulation*	Adulteração de dados externos de contexto recuperados ou inseridos por um sistema durante a fase de execução (runtime), como documentos utilizados por sistemas de geração aumentada por recuperação (RAG), com o objetivo de influenciar os dados de saída.

Termo	Definição
• **vazamento direto de dados de aumento (augmentation data)** • *direct augmentation data leak*	Exposição de dados aumentados por meio de acesso direto não autorizado durante sua transmissão ou seu armazenamento.
• **vazamento direto de modelo em fase de execução (runtime)** • *direct runtime model leak*	A exposição não autorizada dos parâmetros ou da arquitetura do modelo por meio de acesso direto não autorizado, seja em trânsito ou em repouso.
• **envenenamento direto de modelo em fase de execução (runtime)** • *direct runtime model poisoning*	Manipulação, por parte de um invasor, do comportamento de um modelo implantado por meio de acesso direto não autorizado durante a fase de execução, em trânsito ou em repouso.
• **vazamento de dados de entrada** • *input data leak*	A exposição não autorizada de dados fornecidos como dados de entrada para um sistema.
• **saída contendo injeção convencional** • *output containing conventional injection*	Saída de um sistema que, por si só, contém conteúdo malicioso, como fragmentos de script ou de consulta, o que desencadeia um ataque de injeção convencional, como injeção de SQL ou cross-site scripting, quando transmitida a um sistema a jusante.
segurança operacional de um sistema de IA *safety of an AI system*	A propriedade de um sistema de IA de não causar danos inaceitáveis a pessoas, bens ou ao meio ambiente por meio de seu funcionamento.
ataque (de segurança) *(security) attack*	Ação deliberada de um adversário com o objetivo de comprometer a confidencialidade, a integridade ou a disponibilidade de um sistema ou de seus ativos.
violação de segurança *security breach*	Um incidente no qual os controles de segurança são contornados ou falham, resultando em acesso não autorizado, divulgação ou danos.

Termo	Definição
controles de segurança *security controls*	As medidas técnicas, procedimentais ou organizacionais implementadas para reduzir os riscos de segurança.
• **controles para limitação de dados sensíveis** • *controls for sensitive data limitation*	Controles de segurança voltados especificamente para restringir a coleta, a exposição ou a retenção de dados sensíveis.
• **controles para limitar o efeito de comportamento indesejado** • *controls to limit the effect of unwanted behavior*	Controles de segurança projetados para conter ou reduzir o impacto de um comportamento indesejado ou prejudicial de um sistema, uma vez que ele ocorra.
• **controles de governança** • *governance controls*	Controles em nível organizacional, como políticas, estruturas de supervisão e mecanismos de responsabilização e prestação de contas, que gerenciam riscos em nível estratégico.
incidente de segurança *security incident*	Um evento que indica uma possível violação da política de segurança, falha de controle ou comprometimento de um ativo.
riscos de segurança *security risks*	A possibilidade de ameaças de segurança explorarem vulnerabilidades e causarem danos.
testes de segurança *security testing*	O processo de avaliar um sistema para identificar pontos fracos de segurança antes que possam ser explorados.
ameaças de segurança *security threats*	Circunstâncias ou eventos potenciais capazes de comprometer a confidencialidade, a integridade ou a disponibilidade de um sistema ou de seus ativos. No contexto dos sistemas de IA, isso inclui ameaças de entrada, ameaças na fase de desenvolvimento (development-time) e ameaças convencionais na fase de execução (runtime).
dados sensíveis *sensitive data*	Dados que exigem proteção especial devido ao seu potencial de causar danos caso sejam expostos, como informações pessoais, confidenciais ou classificadas.
prestador de serviços *service provider*	Um terceiro que fornece um serviço, como hospedagem, infraestrutura ou software, do qual uma organização ou sistema depende.
código-fonte *source code*	As instruções legíveis por humanos, escritas em uma linguagem de programação, que compõem um programa de software.
vazamento de código-fonte *source code leak*	A exposição não autorizada do código-fonte de um sistema a terceiros não autorizados.

Termo	Definição
ataque esponja (sponge attack) *sponge attack*	Um ataque que alimenta um modelo com entradas projetadas para maximizar seu tempo de computação ou uso de recursos, prejudicando seu desempenho ou disponibilidade.
parte interessada *stakeholder*	Pessoa ou organização que tenha interesse em um projeto, sistema ou decisão, ou que seja afetada por um deles.
startup *startup*	Empresa em estágio inicial, normalmente fundada para desenvolver e expandir um novo produto, tecnologia ou modelo de negócios.
linguagem de consulta estruturada (SQL) *structured query language (SQL)*	Linguagem padrão usada para criar, consultar e gerenciar dados armazenados em bancos de dados relacionais.
adulteração *tampering*	Modificação não autorizada de dados, código ou um sistema com o objetivo de alterar seu comportamento ou integridade.
terceiro *third party*	Qualquer organização ou indivíduo externo, que não seja o fornecedor principal nem o usuário, mas que, mesmo assim, esteja envolvido em um sistema ou o afete, por exemplo, por meio de sua cadeia de suprimentos.
modelagem de ameaças (threat modeling) *threat modeling*	Um processo estruturado para identificar ameaças potenciais, vulnerabilidades e caminhos de ataque contra um sistema, utilizado para priorizar medidas de segurança.
token *token*	A unidade básica de texto, como uma palavra, parte de uma palavra ou caractere, na qual a linguagem é dividida para processamento por um modelo de linguagem. No contexto de serviços de IA na nuvem, os tokens também são a unidade de cobrança.
tokenização *tokenization*	Processo de dividir o texto em unidades menores para processamento por um modelo de linguagem, ou de substituir elementos de dados sensíveis por valores substitutos não sensíveis para proteger os dados originais.
ferramenta *tool*	No contexto da IA agêntica, uma função externa, API (interface de programação de aplicações) ou recurso que um agente de IA pode invocar para realizar ações além da geração de texto.
dados de treinamento *training data*	Os dados utilizados para treinar um modelo de aprendizado de máquina a reconhecer padrões ou realizar sua tarefa.
IA confiável *trustworthy AI*	IA projetada e operada para ser lícita, ética e tecnicamente robusta, incluindo requisitos de segurança.
comportamento indesejado *unwanted behavior*	Qualquer saída, decisão ou ação de um sistema que se desvie de seu comportamento pretendido, seguro ou esperado.
vulnerabilidade *vulnerability*	Uma fragilidade em um sistema, processo ou controle que pode ser explorada por uma ameaça para causar danos.

Certified for what's next

Contato EXIN

www.exin.com