


# Exame simulado

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.

# Conteúdo


| | |
|---|---|
| Introdução | 4 |
| Exame simulado | 5 |
| Gabarito de respostas | 21 |
| Avaliação | 56 |

# Introdução

Este é o exame simulado EXIN AI Security Professional based on the OWASP AI Exchange (AISP.PR). As regras e regulamentos do exame do EXIN se aplicam a este exame.

Este exame contém 40 questões de múltipla escolha. Cada questão de múltipla escolha possui um certo número de alternativas de resposta, entre as quais apenas uma resposta é a correta.

O número máximo de pontos que pode ser obtido neste exame é 40. Cada resposta correta vale 1 ponto. Você precisa de 26 pontos ou mais para passar no exame.

O tempo permitido para este exame é de 90 minutos.

Boa Sorte!

# Exame simulado

**1 / 40**
Um gerente de segurança de TI trabalha em uma empresa de logística. A empresa decidiu recentemente adotar ferramentas de IA para otimizar suas rotas de entrega. O gerente de segurança de TI lidera os esforços para organizar a segurança da IA. As duas primeiras etapas da estrutura G.U.A.R.D. já foram concluídas:

- A etapa Governar (Govern) é realizada por meio da definição de uma política de segurança da IA.
- A etapa Entender (Understand) foi realizada por meio do mapeamento de todos os ativos de IA e seus riscos.

Qual é a **próxima** etapa que o gerente de segurança de TI deve seguir?

**A)** Adaptar as práticas de segurança para incluir a modelagem de ameaças (threat modeling) específicas à IA, testes e controles da cadeia de suprimentos
**B)** Analisar os riscos de IA identificados e estabelecer uma estratégia de priorização de riscos com base em seu impacto potencial
**C)** Aplicar controles de segurança em todos os sistemas de IA para estabelecer uma segurança da IA responsável em todo o ciclo de vida da IA
**D)** Avaliar a política de segurança da IA criada durante a etapa Govern para incluir os ativos de IA mapeados e seus riscos

**2 / 40**
Uma empresa revisa sua estratégia de IA para garantir que ela beneficie os negócios, ao mesmo tempo em que seja ética e operacionalmente sólida. O responsável pela segurança da IA categoriza as duas principais dimensões da IA responsável e confiável.

Qual é a **principal** diferença entre IA responsável e IA confiável?

**A)** - A IA responsável concentra-se na ética, na sociedade e na governança.
- A IA confiável concentra-se nos aspectos técnicos e operacionais, na transparência e na explicabilidade.
**B)** - A IA responsável concentra-se na conformidade regulatória com as leis e legislações pertinentes.
- A IA confiável concentra-se na proteção de dados e na segurança cibernética para se alinhar aos princípios de privacidade.
**C)** - A IA responsável concentra-se na acurácia do sistema e nas métricas de desempenho técnico.
- A IA confiável concentra-se em diretrizes éticas e no impacto social para gerar confiança do consumidor.
**D)** - A IA responsável concentra-se na seleção de fornecedores, nos riscos e na gestão de custos.
- A IA confiável concentra-se no treinamento seguro e confiável dos usuários e na adoção de sistemas de IA.

**3 / 40**
Uma equipe de segurança cibernética já possui uma estrutura de detecção de ameaças. Agora, ela também precisa proteger um novo Modelo de Linguagem de Grande Porte (LLM) utilizado no suporte ao cliente por meio de um chatbot.

A equipe deseja utilizar as mesmas ferramentas que usa para aplicativos web:
- Regras de detecção baseadas em assinaturas
- Controles fixos de validação de entrada

Por que essa abordagem **não** é suficiente para proteger o LLM?

**A)** Porque os sistemas de IA têm superfícies de ataque diferentes e exigem controles específicos para o comportamento do LLM e o fluxo de dados
**B)** Porque a estrutura de detecção de ameaças não inclui procedimentos de resposta a incidentes, o que a torna inadequada para LLMs
**C)** Porque os LLMs geram saídas que são variáveis demais para serem avaliadas por sistemas baseados em regras e precisam ser revisadas por um ser humano
**D)** Porque os LLMs processam solicitações mais rapidamente do que as aplicações web, o que significa que as regras baseadas em assinaturas não conseguem acompanhá-los
**E)** Porque os firewalls de aplicativos web não são licenciados para uso com LLMs e exigem contratos separados com os fornecedores

**4 / 40**
Ao utilizar IA, uma empresa terá ativos específicos de IA. Cada um deles possui suas próprias ameaças principais.

Qual é a combinação **correta** entre um ativo específico de IA e sua ameaça principal?

**A)** - Ativo: registros de auditoria gerados por firewalls de rede
- Ameaça: esgotamento de recursos de IA
**B)** - Ativo: credenciais de autenticação do usuário final
- Ameaça: saída contendo injeção convencional
**C)** - Ativo: código-fonte de software armazenado em um sistema de controle de versão
- Ameaça: envenenamento direto de modelo em fase de execução (runtime)
**D)** - Ativo: dados de treinamento usados para construir um modelo
- Ameaça: divulgação de dados sensíveis na saída

**5 / 40**
Um escritório de advocacia utiliza um sistema de IA para realizar revisão e análise de documentos. A equipe de segurança aplica uma abordagem de gerenciamento de riscos. A equipe concluiu um inventário de todas as ameaças potenciais ao sistema de IA.

Qual é o **próximo** passo que a equipe de segurança deve dar?

**A)** Aplicar um processo de modelagem de ameaças (threat modeling) para transformar as ameaças em um conjunto de riscos concretos e priorizados
**B)** Designar um responsável pelo risco para cada ameaça identificada, a fim de garantir a proteção desde a concepção (by design)
**C)** Implantar controles de segurança para cada ameaça de segurança a fim de mitigar riscos adicionais ao longo do ciclo de vida do sistema de IA
**D)** Compartilhar a lista completa de ameaças com as partes interessadas relevantes para aumentar a conscientização e estabelecer transparência

**6 / 40**
Um engenheiro de segurança da IA analisa o projeto de controle de acesso para um sistema de IA multiagente. Cada agente no sistema recebe uma conta de serviço compartilhada. Essa conta possui acesso de leitura e gravação ao banco de dados de produção, ao sistema de arquivos interno e ao gateway externo da API (interface de programação de aplicações).

Os diferentes agentes de IA invocam ferramentas de forma autônoma e executam fluxos de trabalho de várias etapas nos sistemas internos. O provisionamento foi feito em uma conta compartilhada para simplificar o gerenciamento de credenciais em todo o pipeline de agentes de IA.

Por que esse projeto de conta compartilhada é **particularmente arriscado** em um contexto de IA agêntica?

**A)** Porque os agentes de IA geram volumes maiores de chamadas de API do que usuários humanos, de modo que é mais provável que controles de limitação de taxa (rate limit) sejam acionados e a conta compartilhada seja suspensa
**B)** Porque os agentes de IA não utilizam autenticação multifatorial, de modo que a credencial compartilhada representa um único ponto de falha sem outras camadas de verificação, o que um invasor pode explorar
**C)** Porque os agentes de IA operam de forma autônoma em várias etapas de raciocínio dentro de uma única sessão, de modo que um agente comprometido pode encadear ações em todos os sistemas acessíveis
**D)** Porque as estruturas regulatórias para sistemas de IA exigem que cada agente de IA receba uma credencial exclusiva, de modo que contas compartilhadas constituem uma violação direta de conformidade
**E)** Porque o compartilhamento de contas de serviço entre vários agentes de IA gera complexidade operacional excessiva durante a rotação obrigatória de credenciais e atividades de gerenciamento de acesso

**7 / 40**
Um engenheiro de aprendizado de máquina em uma instituição financeira está investigando dois incidentes distintos envolvendo o modelo de detecção de fraudes da instituição.

- **Incidente 1**: Um invasor externo, sem acesso à arquitetura do modelo, aos dados de treinamento ou aos parâmetros, envia milhares de transações manipuladas. Essas transações são refinadas iterativamente com base no fato de o alerta de fraude ter sido acionado ou não.
- **Incidente 2**: Um pesquisador usa um modelo substituto treinado localmente, com arquitetura semelhante, para gerar transações adversariais. Essas transações conseguem burlar o modelo de produção sem interação prévia.

Quais são essas duas estratégias de evasão?

**A)** - Incidente 1: evasão após envenenamento
- Incidente 2: evasão com conhecimento perfeito
**B)** - Incidente 1: evasão com conhecimento parcial
- Incidente 2: ataque de transferência
**C)** - Incidente 1: evasão com conhecimento perfeito
- Incidente 2: evasão com conhecimento zero
**D)** - Incidente 1: ataque de transferência
- Incidente 2: evasão com conhecimento parcial
**E)** - Incidente 1: evasão com conhecimento zero
- Incidente 2: ataque de transferência

**8 / 40**
Erick é usuário de uma plataforma de mídia social que modera comentários automaticamente por meio de um sistema de IA. Ele tenta fazer com que um comentário que viola as regras passe pela moderação automatizada da plataforma, pois deseja aumentar o engajamento em sua postagem. Erick não possui informações sobre a arquitetura do modelo, os parâmetros ou os dados de treinamento do modelo e vê apenas uma mensagem binária quando tenta publicar seu comentário: "bloqueado" ou "publicado". Ao fazer pequenas alterações repetidas no texto do comentário, Erick acaba encontrando uma versão que é publicada.

Que tipo de ameaça de evasão de entrada é descrita aqui?

**A)** Evasão com conhecimento zero
**B)** Evasão com conhecimento parcial
**C)** Evasão com conhecimento perfeito
**D)** Evasão após envenenamento
**E)** Ataque de transferência

**9 / 40**
Uma empresa de serviços jurídicos desenvolve um chatbot que responde às perguntas dos usuários e pode recuperar informações de recursos internos e externos, como sites e documentos da empresa. Durante uma demonstração, o comportamento do chatbot desencadeia uma investigação por parte da engenheira de segurança da IA. Ela identifica duas ameaças distintas:

- **Ameaça 1:** Uma mensagem criada para anular as instruções e regras de segurança operacional do chatbot pode fazer com que ele siga os comandos do testador em vez de sua política configurada.
- **Ameaça 2:** Uma página da web de um fornecedor recuperada pelo chatbot contém texto incorporado que funciona como instruções. Quando o chatbot acessa essa página, ele trata o conteúdo como comandos e se desvia de sua política.

Que tipos de ameaças são essas?

**A)** - Ameaça 1: evasão
- Ameaça 2: envenenamento de dados
**B)** - Ameaça 1: envenenamento de dados
- Ameaça 2: injeção direta de prompt
**C)** - Ameaça 1: injeção direta de prompt
- Ameaça 2: injeção indireta de prompt
**D)** - Ameaça 1: injeção indireta de prompt
- Ameaça 2: evasão

**10 / 40**
Uma empresa de consultoria possui um sistema de IA para auxiliar nas tarefas do dia a dia. Após uma reunião com um cliente, o assistente de IA resume a conversa. O assistente de IA começa a inserir links internos confidenciais no rascunho do e-mail a ser enviado ao cliente. No entanto, antes do envio, um detector de anomalias autônomo interrompe o envio e coloca o rascunho em quarentena.

O arquiteto de segurança da IA deseja limitar antecipadamente o acesso do assistente de IA de reunião apenas aos direitos das pessoas atendidas.

Quais camadas de proteção contra injeção de prompt estão ilustradas aqui?

**A)** Supervisão automatizada e privilégio mínimo baseado no usuário
**B)** Autorização "just-in-time" e supervisão humana
**C)** Alinhamento de modelo e privilégio mínimo baseado na intenção
**D)** Tratamento de entrada/saída da injeção de prompts e limitação de taxa (rate limit)

**11 / 40**
Um auditor está avaliando os riscos à privacidade de um Modelo de Linguagem de Grande Porte (LLM) implantado no setor de saúde e sua conformidade com as leis de proteção de dados. Para isso, o auditor testa o modelo com consultas direcionadas sobre um paciente específico e pacientes sintéticos cuidadosamente selecionados. Ao observar diferenças nos padrões de resposta e nos índices de confiança do modelo, o auditor pode determinar com elevado grau de certeza se o prontuário desse paciente foi incluído nos dados de treinamento, mesmo que nenhum detalhe médico real seja revelado.

Que tipo de ataque está sendo simulado aqui?

**A)** Divulgação de dados na saída do modelo
**B)** Injeção indireta de prompt
**C)** Inferência de associação (membership inference)
**D)** Inversão do modelo

**12 / 40**
Qual é a **principal** consequência de um ataque bem-sucedido de exfiltração de modelo?

**A)** O invasor pode extrair o conjunto de dados brutos de treinamento usado para construir o modelo original, incluindo, potencialmente, registros confidenciais.
**B)** O invasor obtém controle direto sobre o modelo implantado e pode modificar seus pesos e parâmetros sem autorização prévia.
**C)** O invasor obtém uma réplica do modelo que pode ser usada para criar ataques de evasão sem acionar as defesas do modelo original.
**D)** O modelo original fica indisponível para usuários legítimos devido ao esgotamento de recursos, causando degradação do serviço e repetidos timeouts.

**13 / 40**
Um Modelo de Linguagem de Grande Porte (LLM) hospedado na nuvem é alvo de um ataque esponja (sponge attack). Os invasores enviam entradas muito longas ou complexas que forçam cálculos adicionais, elevando os custos e causando lentidão ou interrupções no serviço. Para lidar com essa situação, o engenheiro de segurança da IA deseja implementar dois controles:

- **Controle 1** validará e sanitizará as entradas para rejeitar ou corrigir conteúdo malicioso, como entradas de tamanho excessivo.
- **Controle 2** restringirá os recursos computacionais consumidos por entrada de modelo de IA para evitar o uso excessivo.

Quais são esses dois controles?

**A)** - Controle 1: tratamento de entradas anômalas
- Controle 2: limites de confiança obscuros

**B)** - Controle 1: validação de entradas contra negação de serviço (denial-of-service, DoS)
- Controle 2: limitação de recursos

**C)** - Controle 1: controle de acesso ao modelo
- Controle 2: modelo robusto contra evasão

**D)** - Controle 1: monitoramento de uso
- Controle 2: limitação de taxa (rate limit)

**14 / 40**
Um engenheiro de segurança da IA identifica as seguintes ameaças para uma empresa:

- **Entradas adversariais** são criadas por invasores para causar classificações incorretas.
- Um **pipeline de ingestão comprometido** introduz registros com backdoor no corpus antes da construção do modelo.
- Os invasores realizam adulterações em **fluxos de trabalho de rotulagem**, alterando as anotações de referência para criar viés no comportamento do modelo.
- **Dados de treinamento** são obtidos de fontes externas ou de terceiros sem a devida validação.

De acordo com o OWASP AI Exchange, qual ameaça **não** envolve riscos de envenenamento de dados em fase de desenvolvimento (development-time)?

**A)** Entradas adversariais
**B)** Pipeline de ingestão comprometido
**C)** Fluxos de trabalho de rotulagem
**D)** Dados de treinamento

**15 / 40**
De acordo com o OWASP AI Exchange, qual é um exemplo de **envenenamento direto de modelo em fase de desenvolvimento (development-time)**?

**A)** Invasores criam entradas adversariais na fase de execução (runtime) para induzir o modelo ao erro.
**B)** Invasores manipulam pesos e um pipeline de treinamento para alterar o comportamento do modelo.
**C)** Dados sensíveis de treinamento são expostos involuntariamente por meio das saídas do modelo.
**D)** Um pico de tráfego sobrecarrega o serviço de IA, fazendo com que ele fique lento ou fique fora do ar.

**16 / 40**
De acordo com o OWASP AI Exchange, qual é um exemplo de **envenenamento de modelo na cadeia de suprimentos**?

**A)** Os invasores comprometem dados, modelos ou componentes da cadeia de suprimentos antes da integração.
**B)** Os invasores manipulam as entradas durante a fase de execução (runtime) para induzir o modelo a produzir saídas incorretas.
**C)** O modelo produz resultados incorretos devido à insuficiência de dados ou à baixa qualidade do treinamento.
**D)** O modelo vaza involuntariamente dados de treinamento sensíveis através de suas respostas.

**17 / 40**
Uma equipe de desenvolvimento de IA está realizando uma avaliação de riscos em seu ambiente de desenvolvimento. Ela identifica duas ameaças distintas:

- **Ameaça 1** envolve acesso não autorizado aos conjuntos de dados usados para treinar e testar seu modelo de IA.
- **Ameaça 2** envolve acesso não autorizado aos parâmetros e pesos do modelo de IA armazenados em um repositório comprometido.

Que tipos de ameaças são essas?

**A)** - Ameaça 1: vazamento de dados em fase de desenvolvimento (development-time)
- Ameaça 2: vazamento direto do modelo em fase de desenvolvimento
**B)** - Ameaça 1: vazamento direto de dados de aumento (augmentation data)
- Ameaça 2: vazamento do repositório
**C)** - Ameaça 1: vazamento direto de dados em fase de execução (runtime)
- Ameaça 2: vazamento direto de modelo em fase de execução
**D)** - Ameaça 1: vazamento de dados de entrada
- Ameaça 2: vazamento de código-fonte/configuração

**18 / 40**
Uma equipe de segurança analisa o panorama de ameaças em fase de execução (runtime) de um sistema de IA. O sistema de IA está integrado a um aplicativo web. A equipe identifica duas ameaças de segurança convencionais:

- **Ameaça 1:** Um invasor pode obter acesso à memória do servidor de produção e extrair os pesos e parâmetros do modelo de IA.
- **Ameaça 2:** um invasor com acesso ao registro do modelo substitui o artefato do modelo implantado por uma versão com trojan.

Quais consequências essas ameaças de segurança convencionais trazem **especificamente** para os sistemas de IA?

**A)** - A ameaça 1 faz com que o sistema se retreine automaticamente com dados maliciosos.
- A ameaça 2 força o modelo a rejeitar prompts válidos.

**B)** - A ameaça 1 altera a arquitetura do modelo em fase de execução (runtime).
- A ameaça 2 comprime silenciosamente os pesos para reduzir a precisão.

**C)** - A ameaça 1 desativa os filtros de segurança operacional, tornando o sistema mais permissivo por padrão (by design).
- A ameaça 2 torna as saídas determinísticas e previsíveis.

**D)** - A ameaça 1 possibilita o roubo do modelo e a potencial inferência de dados de treinamento.
- A ameaça 2 permite a adulteração indetectável do modelo.

**19 / 40**
Uma empresa de serviços financeiros implanta um modelo de aprendizado de máquina (ML) para detectar transações fraudulentas. O modelo de ML está hospedado em seu ambiente de produção e processa continuamente dados de transações em tempo real. Dois incidentes de segurança ocorrem:

- **Incidente A**: Um invasor obtém acesso não autorizado ao ambiente do modelo implantado e altera os parâmetros do modelo de modo que certas transações fraudulentas não sejam mais sinalizadas. A manipulação não é detectada, e o invasor se beneficia disso.
- **Incidente B**: Um invasor explora uma vulnerabilidade no sistema de produção. Isso permite que o invasor replique o modelo e analise seu comportamento para futuros ataques.

Com base no OWASP AI Exchange, quais tipos de incidentes de segurança são descritos?

**A)** - Incidente A: envenenamento de dados
- Incidente B: envenenamento direto de modelo em fase de execução (runtime)

**B)** - Incidente A: envenenamento de dados
- Incidente B: ataque de evasão

**C)** - Incidente A: envenenamento direto de modelo em fase de execução
- Incidente B: vazamento direto de modelo em fase de execução

**D)** - Incidente A: vazamento direto de modelo em fase de execução
- Incidente B: ataque de evasão

**20 / 40**
Uma organização de saúde implanta um assistente de IA para ajudar os médicos a resumir prontuários de pacientes e fornecer recomendações. O sistema se integra a bancos de dados internos e ferramentas externas, e os médicos confiam nas saídas geradas por ele durante as operações diárias. Duas ameaças são observadas:

- **Ameaça 1**: O assistente de IA gera uma resposta que inclui um script malicioso oculto no conteúdo HTML. Quando a saída é visualizada, o script é executado e envia dados da sessão ao invasor.
- **Ameaça 2**: Um invasor explora controles de acesso fracos no sistema para interceptar e recuperar dados sensíveis de pacientes.

Com base no OWASP AI Exchange, quais tipos de ameaças são descritos?

**A)** - Ameaça 1: envenenamento de dados
- Ameaça 2: vazamento de dados de entrada

**B)** - Ameaça 1: envenenamento de dados
- Ameaça 2: ataque de inversão do modelo

**C)** - Ameaça 1: saída contendo injeção convencional
- Ameaça 2: vazamento de dados de entrada

**D)** - Ameaça 1: saída contendo injeção convencional
- Ameaça 2: ataque de inversão do modelo

**21 / 40**
Uma empresa utiliza um sistema de geração aumentada por recuperação (RAG) para responder às perguntas dos funcionários sobre políticas internas. O sistema de IA recupera documentos de um banco de dados vetorial e os adiciona ao prompt antes de enviá-los ao modelo de IA.

Uma análise de segurança constata que um invasor obteve acesso não autorizado ao banco de dados vetorial. O invasor visualizou os documentos de aumento (augmentation documents) armazenados e substituiu um documento de política legítimo por outro, contendo informações falsas. Os funcionários que perguntam ao sistema de IA sobre essa política agora recebem orientações incorretas.

Quais ameaças de segurança esse cenário ilustra?

**A)** - Ameaça 1: vazamento de dados de aumento (augmentation data)
- Ameaça 2: saída contendo injeção convencional

**B)** - Ameaça 1: manipulação de dados de aumento
- Ameaça 2: vazamento de dados de aumento

**C)** - Ameaça 1: inferência de associação (membership inference)
- Ameaça 2: manipulação de dados de aumento

**D)** - Ameaça 1: saída contendo injeção convencional
- Ameaça 2: inferência de associação (membership inference)

**22 / 40**
Controles gerais de governança devem ser implementados para garantir uma supervisão eficaz da segurança dos sistemas de IA.

Qual abordagem demonstra isso **melhor**?

**A)** Definir políticas, funções e práticas de gerenciamento de riscos claras para orientar o desenvolvimento, a implantação e o monitoramento seguros de sistemas de IA
**B)** Incluir controles de segurança nos sistemas de IA para aplicar os princípios de segurança desde a concepção (by design) na fase de design do sistema de IA
**C)** Atender aos requisitos mínimos de segurança, já que os sistemas de IA são mais seguros por padrão (by default) e não exigem múltiplas camadas de segurança
**D)** Separar os controles de governança para IA dos controles de segurança da informação para garantir que esses dois conceitos diferentes sejam igualmente abrangidos

**23 / 40**
Uma agência de marketing de médio porte tem pouca experiência com IA e ainda não possui uma governança formal nessa área. Ela deseja usar uma ferramenta de IA generativa para redigir e-mails para clientes. Para tanto, a ferramenta precisa de acesso aos nomes e e-mails dos clientes. A empresa deve estabelecer uma governança de IA e definir uma supervisão de segurança antes de poder usar essa ferramenta.

Qual é o **mínimo indispensável** que a empresa deve fazer em termos de supervisão de segurança?

**A)** Realizar testes de red teaming de IA no sistema de e-mail com IA generativa para mitigar riscos de segurança antes do lançamento
**B)** Contratar um especialista em governança de IA para realizar uma avaliação de ameaças e definir políticas e controles para a implantação
**C)** Fazer um inventário do uso planejado da IA e realizar uma análise de risco para identificar ameaças, controles e responsabilidades
**D)** Solicitar assessoria jurídica de uma empresa especializada em IA e nos requisitos relacionados à Lei da IA (AI Act)
**E)** Começar com um projeto-piloto e monitorar o sistema de e-mail com IA quanto ao desempenho, usos indevidos e incidentes por meio de registros

**24 / 40**
Como a cobertura dos controles de governança na segurança da IA deve ser **melhor** compreendida?

**A)** Como um conjunto de métodos de criptografia utilizados para proteger as saídas da IA após a geração
**B)** Como um conjunto de controles abrangentes que se aplicam a todas as ameaças de IA e a todas as etapas do ciclo de vida da IA
**C)** Como controles limitados à fase de treinamento do modelo, com foco na qualidade do conjunto de dados e em técnicas de ajuste
**D)** Como medidas opcionais que as organizações podem adotar caso os sistemas de IA não processem dados sensíveis
**E)** Como salvaguardas técnicas aplicadas exclusivamente na fase de execução (runtime) para detectar entradas maliciosas

**25 / 40**
Uma empresa do setor de saúde decide utilizar um modelo de IA pronto para uso (ready-made model), fornecido por um terceiro, para um aplicativo. O aplicativo processa entradas do usuário e recupera dados externos.

O engenheiro de segurança da IA está preocupado com a segurança do modelo. A estratégia de segurança que define a divisão de responsabilidades de segurança entre o fornecedor do modelo de IA e a empresa do setor de saúde não está definida.

O que é **correto** em relação a essas responsabilidades?

**A)** A maioria dos controles de segurança pode ser delegada ao fornecedor ao utilizar modelos de IA prontos para uso, uma vez que o treinamento e a hospedagem realizados pelo fornecedor abrangem os riscos de segurança.
**B)** O responsável pela implantação é responsável por todos os controles de segurança ao integrar o modelo de IA em seus sistemas e deve considerar a segurança antes da implantação.
**C)** O fornecedor treinou o modelo de IA, portanto, ele gerencia os controles no nível do modelo, enquanto o responsável pela implantação deve gerenciar os controles no nível da aplicação.
**D)** O fornecedor é responsável por todos os controles de segurança porque treinou o modelo de IA e o hospeda, então todos os riscos e ataques devem ser tratados pelo fornecedor do modelo de IA.

**26 / 40**
Uma organização responsável pela implantação de um aplicativo de IA voltado para o cliente utilizou um Modelo de Linguagem de Grande Porte (LLM) fornecido por um fornecedor terceirizado. O modelo é acessado por meio de um serviço de API (interface de programação de aplicações) gerenciado e não foi submetido a treinamento adicional nem modificado pela organização.

A equipe de segurança descobre que entradas cuidadosamente elaboradas por usuários podem fazer com que o modelo ignore seus limites comportamentais configurados e produza saídas que violam as restrições de conteúdo da organização. A equipe avalia qual controle deve ser implementado que seja de responsabilidade da organização responsável pela implantação, e não do fornecedor terceirizado.

Qual controle a equipe de segurança deve implementar para resolver o problema?

**A)** Inserir uma camada de validação de saída que analise todas as respostas do modelo antes que elas cheguem aos usuários finais e bloquear as saídas que violarem a política de conteúdo da organização
**B)** Modificar o pipeline de inferência do fornecedor para impor restrições de geração mais rigorosas que limitem a geração de respostas e impeçam o surgimento de conteúdo não permitido
**C)** Solicitar que o fornecedor retreine e ajuste a segurança operacional do modelo fornecido usando dados adicionais de rejeição e alinhamento com políticas para melhorar a resistência contra entradas de tipo “jailbreak”
**D)** Exigir que o fornecedor revise e aplique regras de moderação de conteúdo reforçadas e sua aplicação em todo o serviço, de modo que saídas inseguras sejam filtradas globalmente antes da entrega

**27 / 40**
Uma equipe está preparando dados de treinamento para um assistente de IA generativa. A equipe constata que não precisa de todos os campos e dados que usa atualmente para o desempenho do modelo de IA. Como próximo passo, a equipe está considerando controles para limitação de dados sensíveis e melhora da confidencialidade e da integridade.

Qual é a **melhor** maneira de melhorar a confidencialidade e a integridade?

**A)** Manter todos os campos nos dados de treinamento e realizar a criptografia dos dados em repouso e nos backups
**B)** Preservar identificadores nos dados de treinamento para simplificar futuras tarefas e processos de retreinamento
**C)** Remover campos e dados desnecessários para o treinamento e o desempenho do modelo
**D)** Usar apenas dados sintéticos para todas as fases de treinamento e avaliação da IA generativa

**28 / 40**
Uma equipe de produto está treinando um modelo de IA para atendimento ao cliente. O conjunto de dados de treinamento atual inclui endereços completos dos clientes, números de telefone e registros antigos que não são necessários para atingir as metas de desempenho, mas provavelmente serão úteis no futuro. Depois que um exercício de red teaming destaca o risco de vazamento e manipulação de dados por meio de detalhes memorizados, a equipe decide reduzir a exposição. Para isso, ela remove campos desnecessários e registros inteiros, de modo que dados sensíveis não estejam presentes no conjunto de dados.

Qual controle **geral** foi implementado para limitar os dados sensíveis nessa situação?

**A)** Dados permitidos
**B)** Minimização de dados
**C)** Ocultação de dados de treinamento
**D)** Retenção curta

**29 / 40**
Uma empresa de varejo decide parar de armazenar os detalhes de pagamento dos clientes após a conclusão de cada transação.

Como essa ação **reduz** o risco de exposição de dados?

**A)** Ao eliminar a necessidade de criptografia dos dados em trânsito ou arquivados para fins de auditoria
**B)** Ao garantir a conformidade com todas as regulamentações relevantes de proteção de dados, reduzindo a complexidade
**C)** Ao garantir que nenhum usuário não autorizado possa acessar a rede e o banco de dados da empresa
**D)** Ao impedir o vazamento de dados retidos, sua reconstrução ou sua inferência a partir do sistema

**30 / 40**
Uma rede regional de hospitais utiliza um agente de IA para ler e-mails recebidos de fornecedores, redigir respostas e enviá-las automaticamente. Durante testes de segurança realizados após um recente incidente de phishing, os testadores descobriram que a injeção de prompts poderia fazer com que o agente de IA ignorasse suas regras e enviasse respostas não autorizadas.

Qual decisão de projeto funciona **melhor** para limitar os efeitos desse comportamento indesejado?

**A)** Desativar todas as integrações do agente de IA e voltar a fluxos de trabalho totalmente manuais para a maioria dos processos
**B)** Ampliar o treinamento dos usuários sobre segurança operacional de prompts e enfatizar a supervisão durante o envio de e-mails assistido pelo agente
**C)** Manter permissões amplas de envio autônomo e atualizar o sistema para um modelo generativo mais avançado
**D)** Restringir as permissões do agente de IA a escopos específicos de tarefas e exigir aprovação para ações com risco elevado

**31 / 40**
Qual é a **melhor** maneira de minimizar os efeitos de comportamentos indesejados do modelo?

**A)** Enfatizar o fortalecimento de prompts, o acesso baseado em funções e as auditorias programadas das saídas do modelo
**B)** Contar com filtros automatizados de saída, registro de exceções e verificações pontuais ocasionais quanto à equidade
**C)** Testar a presença de viés indesejado e aplicar controles como supervisão, privilégio mínimo e validação contínua
**D)** Utilizar validação robusta de entradas, permissões restritas para ferramentas e revisões periódicas com intervenção humana

**32 / 40**
Limitar comportamentos indesejados do modelo é uma estratégia crítica para a limitação de impacto e o controle do raio de impacto (blast radius).

Como essa estratégia **melhora** o desempenho e **reduz** o risco?

**A)** Ao concentrar-se em ataques que podem afetar sistemas de IA e limitar a supervisão sobre fatores de desempenho e acurácia do sistema
**B)** Ao implementar controles que eliminam a necessidade de testes e validações contínuas ao longo do tempo para reduzir a complexidade
**C)** Ao melhorar a confiabilidade operacional e a eficiência de recursos, minimizando incidentes e solicitações de computação desperdiçadas
**D)** Ao aumentar a autonomia do modelo e permitir que ele gere mais saídas para melhorar continuamente o desempenho do sistema

**33 / 40**
No âmbito mais amplo do red teaming de IA, qual é o **principal** objetivo dos testes de segurança da IA?

**A)** Avaliar se um modelo de IA é resiliente a ataques específicos, reproduzindo esses ataques em um ambiente controlado
**B)** Avaliar a acurácia e a rastreabilidade das saídas de um modelo de IA quando exposto a dados de entrada esperados
**C)** Medir se um modelo de IA demonstra a eficiência computacional esperada ao processar grandes volumes de solicitações
**D)** Verificar se um modelo de IA atende aos requisitos de negócios definidos por meio de procedimentos padrão de garantia de qualidade

**34 / 40**
Uma empresa responsável pela implantação de um assistente de IA generativa deseja definir testes de segurança da IA que abranjam mais do que apenas testes de penetração convencionais. Para isso, a equipe de testes começa a identificar as ameaças que devem ser testadas.

Quais ameaças **especificamente** associadas a sistemas de IA generativa devem ser incluídas nos testes de segurança da IA?

**A)** Ataques de evasão, tratamento inseguro de saídas e envenenamento de modelo
**B)** Ataques de exfiltração de modelo, injeção de prompt e evasão
**C)** Injeção de prompt, divulgação de dados sensíveis e tratamento inseguro de saídas
**D)** Saída de dados sensíveis do modelo, desvio do modelo e invocação insegura de ferramentas

**35 / 40**
Uma equipe de segurança da IA está testando um chatbot de IA generativa quanto à resistência à injeção de prompts. Durante o teste, o sistema bloqueou várias entradas de ataque criadas propositalmente.

Qual é a **próxima** etapa recomendada para manter a supervisão de segurança do chatbot de IA generativa?

**A)** Adicionar variações de entrada, como substituição de sinônimos, alterações de codificação ou mudanças de formatação, para tentar contornar a detecção
**B)** Concluir o ciclo de testes atual e considerar as defesas adequadas, já que a maioria das tentativas de injeção de prompt foi bloqueada
**C)** Executar o mesmo teste várias vezes para aumentar o tamanho da amostra e obter dados de desempenho das defesas mais relevantes do ponto de vista estatístico
**D)** Enviar as entradas de ataque por injeção de prompt diretamente ao modelo para treiná-lo melhor contra ameaças semelhantes em implantações futuras

**36 / 40**
O sistema de IA de uma empresa depende de grandes quantidades de dados pessoais para treinamento e tomada de decisões.

Do ponto de vista ético, qual é o risco **mais** importante que a empresa deve abordar?

**A)** Dados sendo coletados, retidos ou reutilizados para além dos fins justificados
**B)** Complexidade da interface dificultando a revisão das decisões automatizadas
**C)** Acurácia do modelo diminuindo quando os recursos são alocados de forma desigual
**D)** Atualizações de software causando problemas de compatibilidade entre plataformas

**37 / 40**
Uma livraria online utiliza um sistema de IA para analisar o histórico de navegação, as tendências de compra e as pré-vendas, a fim de prever a demanda e garantir que os livros populares estejam em estoque a tempo. Posteriormente, a empresa reutiliza esses mesmos dados de clientes para treinar um modelo de IA separado que cria perfis de marketing detalhados e envia anúncios direcionados a compradores individuais.

Qual princípio de privacidade é **mais diretamente** violado nesse caso?

**A)** Minimização de dados
**B)** Limitação da retenção de dados
**C)** Transparência e consentimento
**D)** Limitação de uso e especificação da finalidade

**38 / 40**
Uma organização deseja estabelecer uma estrutura para reger o uso responsável da IA em todas as suas operações.

Ela utiliza a norma ISO/IEC 27001 para a gestão da segurança da informação. A equipe de segurança deseja estabelecer sua governança de IA com a norma ISO/IEC que cumpre a mesma finalidade estrutural para a governança de IA que a norma ISO/IEC 27001 cumpre para a segurança da informação.

Qual é essa norma?

**A)** ISO/IEC 23894
**B)** ISO/IEC 27005
**C)** ISO/IEC 42001
**D)** ISO/IEC 5338

**39 / 40**
Uma empresa deseja utilizar um sistema de IA para selecionar e classificar automaticamente candidatos a vagas de emprego, com base em seus currículos.

De acordo com a Lei da IA (AI Act), em qual categoria o uso desse sistema de IA deve ser classificado?

**A)** Risco inaceitável
**B)** Risco elevado
**C)** Risco limitado
**D)** Risco mínimo ou nulo

**40 / 40**
Uma equipe de desenvolvimento de IA está preparando dados de treinamento para um novo modelo de aprendizado de máquina. O líder da equipe deseja garantir que a abordagem de obtenção de dados respeite a conformidade com os direitos autorais.

Qual é a **prática mais ética** de obtenção de dados?

**A)** Coletar dados apenas de domínios publicamente disponíveis
**B)** Criar todos os dados de treinamento dentro da organização
**C)** Baixar dados de fontes com permissão implícita
**D)** Usar apenas dados que pareçam ter sido amplamente compartilhados online

# Gabarito de respostas

**1 / 40**
Um gerente de segurança de TI trabalha em uma empresa de logística. A empresa decidiu recentemente adotar ferramentas de IA para otimizar suas rotas de entrega. O gerente de segurança de TI lidera os esforços para organizar a segurança da IA. As duas primeiras etapas da estrutura G.U.A.R.D. já foram concluídas:

- A etapa Governar (Govern) é realizada por meio da definição de uma política de segurança da IA.
- A etapa Entender (Understand) foi realizada por meio do mapeamento de todos os ativos de IA e seus riscos.

Qual é a **próxima** etapa que o gerente de segurança de TI deve seguir?

**A)** Adaptar as práticas de segurança para incluir a modelagem de ameaças (threat modeling) específicas à IA, testes e controles da cadeia de suprimentos
**B)** Analisar os riscos de IA identificados e estabelecer uma estratégia de priorização de riscos com base em seu impacto potencial
**C)** Aplicar controles de segurança em todos os sistemas de IA para estabelecer uma segurança da IA responsável em todo o ciclo de vida da IA
**D)** Avaliar a política de segurança da IA criada durante a etapa Govern para incluir os ativos de IA mapeados e seus riscos

**A)** Correto. Após as etapas Governar e Compreender, a próxima etapa do G.U.A.R.D. é Adaptar (Adapt), que consiste em atualizar as práticas de segurança para incorporar ameaças específicas à IA, testes de segurança e medidas relacionadas à cadeia de suprimentos. (Literatura: A, Capítulo 0)
**B)** Incorreto. Isso pertence à etapa Entender, que já foi concluída.
**C)** Incorreto. A aplicação generalizada de controles sem contexto não se alinha às etapas da estrutura G.U.A.R.D. A próxima etapa é adaptar as práticas, não aplicar controles indiscriminadamente. Os controles serão aplicados em uma etapa posterior.
**D)** Incorreto. A revisão da política faz parte da etapa Governar. A próxima etapa após Entender é Adaptar, não reformular a governança.

**2 / 40**
Uma empresa revisa sua estratégia de IA para garantir que ela beneficie os negócios, ao mesmo tempo em que seja ética e operacionalmente sólida. O responsável pela segurança da IA categoriza as duas principais dimensões da IA responsável e confiável.

Qual é a **principal** diferença entre IA responsável e IA confiável?

**A)** - A IA responsável concentra-se na ética, na sociedade e na governança.
- A IA confiável concentra-se nos aspectos técnicos e operacionais, na transparência e na explicabilidade.

**B)** - A IA responsável concentra-se na conformidade regulatória com as leis e legislações pertinentes.
- A IA confiável concentra-se na proteção de dados e na segurança cibernética para se alinhar aos princípios de privacidade.

**C)** - A IA responsável concentra-se na acurácia do sistema e nas métricas de desempenho técnico.
- A IA confiável concentra-se em diretrizes éticas e no impacto social para gerar confiança do consumidor.

**D)** - A IA responsável concentra-se na seleção de fornecedores, nos riscos e na gestão de custos.
- A IA confiável concentra-se no treinamento seguro e confiável dos usuários e na adoção de sistemas de IA.

**A)** Correto. A IA responsável enfatiza a ética, a sociedade e a governança (princípios éticos, políticas, supervisão), enquanto a IA confiável enfatiza os aspectos mais técnicos e operacionais do uso da IA em uma organização, como robustez, confiabilidade, transparência e explicabilidade. (Literatura: A, Capítulo 0)

**B)** Incorreto. Isso restringe a IA responsável à conformidade regulatória e a IA confiável à privacidade e à segurança cibernética. Esses são subconjuntos e não captam a diferença mais ampla.

**C)** Incorreto. Isso inverte o foco: exatidão/acurácia e desempenho fazem parte da confiabilidade técnica, enquanto ética e impacto social são centrais para a IA responsável.

**D)** Incorreto. A seleção de fornecedores, o custo e o treinamento de usuários são temas de gestão de mudanças, não a distinção conceitual central entre IA responsável e confiável.

**3 / 40**
Uma equipe de segurança cibernética já possui uma estrutura de detecção de ameaças. Agora, ela também precisa proteger um novo Modelo de Linguagem de Grande Porte (LLM) utilizado no suporte ao cliente por meio de um chatbot.

A equipe deseja utilizar as mesmas ferramentas que usa para aplicativos web:
- Regras de detecção baseadas em assinaturas
- Controles fixos de validação de entrada

Por que essa abordagem **não** é suficiente para proteger o LLM?

**A)** Porque os sistemas de IA têm superfícies de ataque diferentes e exigem controles específicos para o comportamento do LLM e o fluxo de dados
**B)** Porque a estrutura de detecção de ameaças não inclui procedimentos de resposta a incidentes, o que a torna inadequada para LLMs
**C)** Porque os LLMs geram saídas que são variáveis demais para serem avaliadas por sistemas baseados em regras e precisam ser revisadas por um ser humano
**D)** Porque os LLMs processam solicitações mais rapidamente do que as aplicações web, o que significa que as regras baseadas em assinaturas não conseguem acompanhá-los
**E)** Porque os firewalls de aplicativos web não são licenciados para uso com LLMs e exigem contratos separados com os fornecedores

**A)** Correto. Ferramentas convencionais de segurança cibernética, como firewalls de aplicativos web e sistemas de detecção baseados em assinaturas, são projetadas para identificar padrões de ataque conhecidos contra interfaces de software fixas. Os sistemas de IA introduzem superfícies de ataque totalmente novas: os dados de treinamento podem ser adulterados, o próprio modelo pode ser invertido para extrair informações confidenciais e as saídas podem ser manipuladas por meio de entradas adversariais que não correspondem a nenhuma assinatura conhecida. Essas ameaças exigem controles específicos para IA que atuem no comportamento do modelo e no fluxo de dados, e não apenas em sua interface de rede. (Literatura: A, Capítulos 0 e 1)
**B)** Incorreto. As estruturas convencionais de segurança cibernética incluem procedimentos de resposta a incidentes. Essa opção caracteriza erroneamente a limitação da segurança convencional para sistemas de IA, que diz respeito à cobertura da superfície de ataque, e não a lacunas procedimentais.
**C)** Incorreto. Embora as saídas dos LLMs sejam altamente variáveis, os controles convencionais baseados em regras ainda podem contribuir para defesas em camadas em ambientes de IA. A limitação é que eles não abordam adequadamente as superfícies de ataque específicas da IA, como manipulação de modelos, entradas adversariais, injeção de prompts e ataques a dados de treinamento.
**D)** Incorreto. A velocidade de processamento é uma consideração de desempenho e não é a razão pela qual a detecção baseada em assinaturas é insuficiente para sistemas de IA. A limitação aqui não é o tempo, mas a natureza da superfície de ataque.
**E)** Incorreto. Não há nenhuma restrição geral de licenciamento que impeça a aplicação de firewalls de aplicativos web em endpoints de modelos de IA. Essa opção descreve uma consideração de aquisição, e não a limitação de segurança substantiva da abordagem proposta.

**4 / 40**
Ao utilizar IA, uma empresa terá ativos específicos de IA. Cada um deles possui suas próprias ameaças principais.

Qual é a combinação **correta** entre um ativo específico de IA e sua ameaça principal?

**A)** - Ativo: registros de auditoria gerados por firewalls de rede
- Ameaça: esgotamento de recursos de IA

**B)** - Ativo: credenciais de autenticação do usuário final
- Ameaça: saída contendo injeção convencional

**C)** - Ativo: código-fonte de software armazenado em um sistema de controle de versão
- Ameaça: envenenamento direto de modelo em fase de execução (runtime)

**D)** - Ativo: dados de treinamento usados para construir um modelo
- Ameaça: divulgação de dados sensíveis na saída

**A)** Incorreto. Os registros de auditoria de firewalls de rede são um ativo geral de segurança, não um ativo específico de IA descrito no OWASP AI Exchange. O esgotamento de recursos de IA normalmente se refere ao uso excessivo de recursos computacionais necessários para tarefas de IA, o que não se relaciona diretamente com os registros de auditoria.

**B)** Incorreto. As credenciais de autenticação do usuário final são uma preocupação comum de segurança, não um ativo específico de IA descrito no OWASP AI Exchange. Ameaças de injeção convencionais, como injeção de comando ou de linguagem de consulta estruturada (SQL), são mais relevantes para a segurança de aplicativos do que para saídas de IA. Esse emparelhamento não está corretamente alinhado.

**C)** Incorreto. O código-fonte de software pode ser um ativo específico da IA se estiver relacionado a modelos de IA. No entanto, o envenenamento direto de modelo em fase de execução diz respeito à manipulação de um modelo de IA durante sua execução, e não ao seu código-fonte no controle de versão.

**D)** Correto. Os dados de treinamento são um ativo crítico específico da IA. A divulgação de dados na saída é uma ameaça legítima, especialmente se dados sensíveis ou proprietários do conjunto de treinamento aparecerem inadvertidamente nas saídas geradas pela IA generativa. Essa combinação identifica corretamente o ativo e sua ameaça associada no contexto da IA. (Literatura: A, Capítulo 0)

**5 / 40**
Um escritório de advocacia utiliza um sistema de IA para realizar revisão e análise de documentos. A equipe de segurança aplica uma abordagem de gerenciamento de riscos. A equipe concluiu um inventário de todas as ameaças potenciais ao sistema de IA.

Qual é o **próximo** passo que a equipe de segurança deve dar?

**A)** Aplicar um processo de modelagem de ameaças (threat modeling) para transformar as ameaças em um conjunto de riscos concretos e priorizados
**B)** Designar um responsável pelo risco para cada ameaça identificada, a fim de garantir a proteção desde a concepção (by design)
**C)** Implantar controles de segurança para cada ameaça de segurança a fim de mitigar riscos adicionais ao longo do ciclo de vida do sistema de IA
**D)** Compartilhar a lista completa de ameaças com as partes interessadas relevantes para aumentar a conscientização e estabelecer transparência

**A)** Correto. Após listar as ameaças, a equipe deve aplicar um processo de modelagem de ameaças, que serve como ponte entre uma lista de ameaças e um conjunto de riscos concretos e priorizados. Isso permite que a organização se concentre na aplicação de controles aos riscos mais significativos. (Literatura: A, Capítulo 0)
**B)** Incorreto. Designar responsáveis pelos riscos antes de analisar e priorizar as ameaças ignora uma etapa crítica. Sem compreender o impacto concreto e a probabilidade de cada ameaça, a designação de responsabilidades carece de contexto significativo.
**C)** Incorreto. Implantar controles para todas as ameaças sem priorização é ineficiente e pode levar à alocação incorreta de recursos. A priorização por meio da modelagem de ameaças é necessária antes de selecionar e aplicar controles.
**D)** Incorreto. Compartilhar uma lista bruta de ameaças sem priorização não ajuda as partes interessadas a tomar decisões informadas. O processo de modelagem de ameaças deve primeiro converter a lista em riscos concretos e priorizados para que a comunicação seja significativa.

**6 / 40**
Um engenheiro de segurança da IA analisa o projeto de controle de acesso para um sistema de IA multiagente. Cada agente no sistema recebe uma conta de serviço compartilhada. Essa conta possui acesso de leitura e gravação ao banco de dados de produção, ao sistema de arquivos interno e ao gateway externo da API (interface de programação de aplicações).

Os diferentes agentes de IA invocam ferramentas de forma autônoma e executam fluxos de trabalho de várias etapas nos sistemas internos. O provisionamento foi feito em uma conta compartilhada para simplificar o gerenciamento de credenciais em todo o pipeline de agentes de IA.

Por que esse projeto de conta compartilhada é **particularmente arriscado** em um contexto de IA agêntica?

**A)** Porque os agentes de IA geram volumes maiores de chamadas de API do que usuários humanos, de modo que é mais provável que controles de limitação de taxa (rate limit) sejam acionados e a conta compartilhada seja suspensa
**B)** Porque os agentes de IA não utilizam autenticação multifatorial, de modo que a credencial compartilhada representa um único ponto de falha sem outras camadas de verificação, o que um invasor pode explorar
**C)** Porque os agentes de IA operam de forma autônoma em várias etapas de raciocínio dentro de uma única sessão, de modo que um agente comprometido pode encadear ações em todos os sistemas acessíveis
**D)** Porque as estruturas regulatórias para sistemas de IA exigem que cada agente de IA receba uma credencial exclusiva, de modo que contas compartilhadas constituem uma violação direta de conformidade
**E)** Porque o compartilhamento de contas de serviço entre vários agentes de IA gera complexidade operacional excessiva durante a rotação obrigatória de credenciais e atividades de gerenciamento de acesso

**A)** Incorreto. O volume de chamadas de API é uma preocupação de desempenho e disponibilidade, não um risco de segurança específico ao design de credenciais compartilhadas. A limitação de taxa afeta todos os tipos de conta e não explica por que uma conta compartilhada representa um risco desproporcionalmente maior para agentes em comparação com usuários humanos que compartilham a mesma conta.
**B)** Incorreto. A ausência de autenticação multifatorial para contas de serviço é uma preocupação geral de segurança aplicável tanto a contas compartilhadas por usuários humanos quanto a contas compartilhadas por agentes. Isso não explica por que o risco é desproporcionalmente maior especificamente no contexto de IA agêntica.
**C)** Correto. O risco de segurança abordado na pergunta está relacionado à autonomia excessiva. Sistemas de IA agêntica podem encadear ações de forma autônoma em vários sistemas, ferramentas e etapas de raciocínio dentro de uma única sessão. Um agente comprometido que utilize uma conta compartilhada pode, portanto, realizar ações não autorizadas de amplo alcance antes que ocorra intervenção ou revisão humana. (Literatura: A, Capítulos 0 e 1)
**D)** Incorreto. Os requisitos regulatórios para o provisionamento de credenciais de agentes de IA variam de acordo com a jurisdição e o setor e não estão universalmente estabelecidos no nível de especificidade descrito. Esta resposta não aborda o risco de segurança apresentado pela pergunta.
**E)** Incorreto. Isso aborda um desafio operacional, não o risco de segurança específico criado pelo design de contas compartilhadas no contexto da IA agêntica.

**7 / 40**
Um engenheiro de aprendizado de máquina em uma instituição financeira está investigando dois incidentes distintos envolvendo o modelo de detecção de fraudes da instituição.

- **Incidente 1**: Um invasor externo, sem acesso à arquitetura do modelo, aos dados de treinamento ou aos parâmetros, envia milhares de transações manipuladas. Essas transações são refinadas iterativamente com base no fato de o alerta de fraude ter sido acionado ou não.
- **Incidente 2**: Um pesquisador usa um modelo substituto treinado localmente, com arquitetura semelhante, para gerar transações adversariais. Essas transações conseguem burlar o modelo de produção sem interação prévia.

Quais são essas duas estratégias de evasão?

**A)** - Incidente 1: evasão após envenenamento
- Incidente 2: evasão com conhecimento perfeito

**B)** - Incidente 1: evasão com conhecimento parcial
- Incidente 2: ataque de transferência

**C)** - Incidente 1: evasão com conhecimento perfeito
- Incidente 2: evasão com conhecimento zero

**D)** - Incidente 1: ataque de transferência
- Incidente 2: evasão com conhecimento parcial

**E)** - Incidente 1: evasão com conhecimento zero
- Incidente 2: ataque de transferência

**A)** Incorreto. O Incidente 1 foi identificado incorretamente como evasão após envenenamento. Não há evidências de que o invasor tenha alterado os dados de treinamento. O invasor apenas testou muitas transações e aprendeu com as respostas do modelo. O Incidente 2 foi identificado incorretamente como evasão com conhecimento perfeito. O pesquisador não tinha acesso total ao modelo real. Ele utilizou um modelo substituto.

**B)** Incorreto. O Incidente 1 foi identificado incorretamente como evasão com conhecimento parcial. O invasor não teve acesso ao projeto, aos dados ou aos parâmetros do modelo. Ele utilizou apenas as saídas do modelo. O Incidente 2 foi identificado corretamente como ataque de transferência, mas a resposta completa ainda está errada, pois o Incidente 1 não é uma evasão com conhecimento parcial.

**C)** Incorreto. O Incidente 1 foi identificado incorretamente como evasão com conhecimento perfeito. Esse ataque não concede ao invasor acesso total ao modelo. O Incidente 2 foi identificado incorretamente como evasão com conhecimento zero. O pesquisador utilizou um modelo substituto para construir o ataque, portanto, trata-se de um ataque de transferência.

**D)** Incorreto. O Incidente 1 foi identificado incorretamente como um ataque de transferência. No Incidente 1, o invasor testou o modelo real várias vezes e alterou as entradas com base nas respostas. O Incidente 2 foi identificado incorretamente como evasão com conhecimento parcial. O pesquisador utilizou um modelo separado, portanto, trata-se de um ataque de transferência.

**E)** Correto. O Incidente 1 está corretamente identificado como evasão com conhecimento zero. O invasor não tinha acesso interno e utilizou apenas as saídas do modelo. O Incidente 2 está corretamente identificado como ataque de transferência. O pesquisador construiu o ataque com base em um modelo substituto, e este também enganou o modelo real. (Literatura: A, Capítulo 2)

**8 / 40**
Erick é usuário de uma plataforma de mídia social que modera comentários automaticamente por meio de um sistema de IA. Ele tenta fazer com que um comentário que viola as regras passe pela moderação automatizada da plataforma, pois deseja aumentar o engajamento em sua postagem. Erick não possui informações sobre a arquitetura do modelo, os parâmetros ou os dados de treinamento do modelo e vê apenas uma mensagem binária quando tenta publicar seu comentário: “bloqueado” ou “publicado”. Ao fazer pequenas alterações repetidas no texto do comentário, Erick acaba encontrando uma versão que é publicada.

Que tipo de ameaça de evasão de entrada é descrita aqui?

**A)** Evasão com conhecimento zero
**B)** Evasão com conhecimento parcial
**C)** Evasão com conhecimento perfeito
**D)** Evasão após envenenamento
**E)** Ataque de transferência

**A)** Correto. Erick não tem informações sobre o modelo de IA e usa uma estratégia de ataque baseada em consultas para encontrar uma entrada de ataque bem-sucedida, conseguindo publicar seu comentário. Isso corresponde à definição de uma evasão com conhecimento zero. (Literatura: A, Capítulo 2)
**B)** Incorreto. A evasão com conhecimento parcial envolveria alguma informação adicional para orientar o ataque, informação que Erick não possui.
**C)** Incorreto. A evasão com conhecimento perfeito exigiria acesso total à arquitetura, aos parâmetros e aos dados de treinamento do modelo de IA para criar entradas com precisão.
**D)** Incorreto. A evasão após envenenamento refere-se à evasão possibilitada pelo envenenamento prévio dos dados de treinamento do modelo de IA. Não foi isso que aconteceu nesse caso.
**E)** Incorreto. Um ataque de transferência significaria criar exemplos adversariais usando um modelo substituto e, em seguida, aplicar esses exemplos adversariais ao modelo alvo.

**9 / 40**
Uma empresa de serviços jurídicos desenvolve um chatbot que responde às perguntas dos usuários e pode recuperar informações de recursos internos e externos, como sites e documentos da empresa. Durante uma demonstração, o comportamento do chatbot desencadeia uma investigação por parte da engenheira de segurança da IA. Ela identifica duas ameaças distintas:

- **Ameaça 1:** Uma mensagem criada para anular as instruções e regras de segurança operacional do chatbot pode fazer com que ele siga os comandos do testador em vez de sua política configurada.
- **Ameaça 2:** Uma página da web de um fornecedor recuperada pelo chatbot contém texto incorporado que funciona como instruções. Quando o chatbot acessa essa página, ele trata o conteúdo como comandos e se desvia de sua política.

Que tipos de ameaças são essas?

**A)** - Ameaça 1: evasão
- Ameaça 2: envenenamento de dados

**B)** - Ameaça 1: envenenamento de dados
- Ameaça 2: injeção direta de prompt

**C)** - Ameaça 1: injeção direta de prompt
- Ameaça 2: injeção indireta de prompt

**D)** - Ameaça 1: injeção indireta de prompt
- Ameaça 2: evasão

**A)** Incorreto. A ameaça 1 não é evasão, pois nenhuma entrada adversarial induziu o modelo ao erro. A ameaça 2 não é envenenamento de dados, pois os dados de treinamento do modelo não foram alterados.

**B)** Incorreto. A ameaça 1 não é envenenamento de dados, pois os dados de treinamento do modelo não foram alterados. A ameaça 2 não é injeção direta de prompt, pois as instruções vieram indiretamente do conteúdo da web.

**C)** Correto. A ameaça 1 é injeção direta de prompt, em que os prompts do usuário enganam a IA. A ameaça 2 é injeção indireta de prompt, em que o conteúdo da web contém instruções ocultas. (Literatura: A, Capítulo 2)

**D)** Incorreto. A ameaça 1 não é injeção indireta de prompt, pois as instruções vieram diretamente do usuário. A ameaça 2 não é evasão, pois o modelo não foi enganado por perturbações na entrada.

**10 / 40**
Uma empresa de consultoria possui um sistema de IA para auxiliar nas tarefas do dia a dia. Após uma reunião com um cliente, o assistente de IA resume a conversa. O assistente de IA começa a inserir links internos confidenciais no rascunho do e-mail a ser enviado ao cliente. No entanto, antes do envio, um detector de anomalias autônomo interrompe o envio e coloca o rascunho em quarentena.

O arquiteto de segurança da IA deseja limitar antecipadamente o acesso do assistente de IA de reunião apenas aos direitos das pessoas atendidas.

Quais camadas de proteção contra injeção de prompt estão ilustradas aqui?

**A)** Supervisão automatizada e privilégio mínimo baseado no usuário
**B)** Autorização “just-in-time” e supervisão humana
**C)** Alinhamento de modelo e privilégio mínimo baseado na intenção
**D)** Tratamento de entrada/saída da injeção de prompts e limitação de taxa (rate limit)

**A)** Correto. Um detector de anomalias autônomo que suspende e coloca o rascunho em quarentena é uma supervisão automatizada. Conceder ao assistente apenas as permissões dos indivíduos a quem ele atende, atribuídas antecipadamente, é o princípio do privilégio mínimo baseado no usuário. (Literatura: A, Capítulo 2)
**B)** Incorreto. Não há aprovação escalonada ou por ação que justifique a autorização “just-in-time”, e a supervisão é automatizada, não humana.
**C)** Incorreto. O cenário não menciona o alinhamento do modelo, e o princípio do privilégio mínimo descrito baseia-se em direitos pré-atribuídos à identidade do usuário, e não em privilégios dinâmicos, definidos pelo escopo da intenção.
**D)** Incorreto. Não é descrita nenhuma limitação de taxa, e o controle mostrado é um detector de anomalias autônomo (supervisão automatizada), e não medidas específicas de rotulagem de E/S.

**11 / 40**
Um auditor está avaliando os riscos à privacidade de um Modelo de Linguagem de Grande Porte (LLM) implantado no setor de saúde e sua conformidade com as leis de proteção de dados. Para isso, o auditor testa o modelo com consultas direcionadas sobre um paciente específico e pacientes sintéticos cuidadosamente selecionados. Ao observar diferenças nos padrões de resposta e nos índices de confiança do modelo, o auditor pode determinar com elevado grau de certeza se o prontuário desse paciente foi incluído nos dados de treinamento, mesmo que nenhum detalhe médico real seja revelado.

Que tipo de ataque está sendo simulado aqui?

**A)** Divulgação de dados na saída do modelo
**B)** Injeção indireta de prompt
**C)** Inferência de associação (membership inference)
**D)** Inversão do modelo

**A)** Incorreto. A divulgação de dados na saída do modelo ocorre quando o modelo revela diretamente informações confidenciais contidas nos dados de treinamento ou no contexto da fase de execução (runtime). Nesse cenário, nenhum detalhe médico real é divulgado.
**B)** Incorreto. A injeção indireta de prompt ocorre quando um terceiro incorpora instruções adversariais em um conteúdo que é recuperado e inserido em um prompt, manipulando o comportamento do modelo. O cenário não descreve tal injeção.
**C)** Correto. A inferência de associação ocorre quando um invasor apresenta ao modelo uma entrada que identifica um indivíduo específico e usa os sinais de confiança do modelo para determinar se esse indivíduo aparece no conjunto de treinamento. (Literatura: A, Capítulo 2)
**D)** Incorreto. A inversão do modelo ocorre quando um invasor reconstrói parte do conjunto de treinamento por meio de experimentação intensiva, otimizando as entradas para maximizar os sinais de confiança na saída do modelo. Esse tipo de ataque reconstrói atributos ou registros confidenciais a partir do modelo, e não apenas se um registro estava no conjunto de treinamento.

**12 / 40**
Qual é a **principal** consequência de um ataque bem-sucedido de exfiltração de modelo?

**A)** O invasor pode extrair o conjunto de dados brutos de treinamento usado para construir o modelo original, incluindo, potencialmente, registros confidenciais.
**B)** O invasor obtém controle direto sobre o modelo implantado e pode modificar seus pesos e parâmetros sem autorização prévia.
**C)** O invasor obtém uma réplica do modelo que pode ser usada para criar ataques de evasão sem acionar as defesas do modelo original.
**D)** O modelo original fica indisponível para usuários legítimos devido ao esgotamento de recursos, causando degradação do serviço e repetidos timeouts.

**A)** Incorreto. A exfiltração de modelo replica o comportamento do modelo por meio de pares de entrada-saída, mas não extrai o conjunto de dados de treinamento original. A extração de dados de treinamento é uma ameaça diferente (inversão do modelo ou inferência de associação (membership inference)).
**B)** Incorreto. A exfiltração de modelo não concede a capacidade de modificar o modelo original implantado. Ela resulta em uma cópia da funcionalidade do modelo, não no controle sobre o original.
**C)** Correto. Uma consequência importante da exfiltração de modelo é que o invasor pode usar a réplica do modelo para realizar ataques de evasão, contornando a limitação de taxa (rate limit), o controle de acesso e os mecanismos de detecção presentes no sistema original. (Literatura: A, Capítulo 2)
**D)** Incorreto. A negação de serviço (denial-of-service, DoS) é uma consequência do esgotamento de recursos, o que constitui uma ameaça distinta. A exfiltração de modelo não afeta diretamente a disponibilidade do modelo original.

**13 / 40**
Um Modelo de Linguagem de Grande Porte (LLM) hospedado na nuvem é alvo de um ataque esponja (sponge attack). Os invasores enviam entradas muito longas ou complexas que forçam cálculos adicionais, elevando os custos e causando lentidão ou interrupções no serviço. Para lidar com essa situação, o engenheiro de segurança da IA deseja implementar dois controles:

- **Controle 1** validará e sanitizará as entradas para rejeitar ou corrigir conteúdo malicioso, como entradas de tamanho excessivo.
- **Controle 2** restringirá os recursos computacionais consumidos por entrada de modelo de IA para evitar o uso excessivo.

Quais são esses dois controles?

**A)** - Controle 1: tratamento de entradas anômalas
- Controle 2: limites de confiança obscuros

**B)** - Controle 1: validação de entradas contra negação de serviço (denial-of-service, DoS)
- Controle 2: limitação de recursos

**C)** - Controle 1: controle de acesso ao modelo
- Controle 2: modelo robusto contra evasão

**D)** - Controle 1: monitoramento de uso
- Controle 2: limitação de taxa (rate limit)

**A)** Incorreto. O tratamento de entradas anômalas detecta entradas que se desviam da distribuição dos dados de treinamento e é relevante principalmente para ameaças de evasão e exfiltração de modelo. A ocultação dos limites de confiança restringe as informações de saída para reduzir a capacidade do invasor de otimizar amostras adversariais.

**B)** Correto. Esses são os dois controles apresentados na questão e projetados especificamente para lidar com o esgotamento de recursos impulsionado por entradas, complementando os controles gerais. (Literatura: A, Capítulo 2)

**C)** Incorreto. O controle de acesso ao modelo restringe quem pode interagir com o modelo e é um controle geral aplicável a múltiplas ameaças de entrada. O modelo robusto contra evasão é um controle em fase de desenvolvimento (development-time) projetado para aumentar a resiliência do modelo contra ataques de evasão.

**D)** Incorreto. O monitoramento de uso consiste em observar, correlacionar e registrar o uso do modelo, as entradas, as saídas e o comportamento do sistema para identificar eventos ou padrões que possam indicar um incidente de segurança. A limitação de taxa consiste em restringir a taxa (frequência) de acesso ao modelo.

**14 / 40**
Um engenheiro de segurança da IA identifica as seguintes ameaças para uma empresa:

- **Entradas adversariais** são criadas por invasores para causar classificações incorretas.
- Um **pipeline de ingestão comprometido** introduz registros com backdoor no corpus antes da construção do modelo.
- Os invasores realizam adulterações em **fluxos de trabalho de rotulagem**, alterando as anotações de referência para criar viés no comportamento do modelo.
- **Dados de treinamento** são obtidos de fontes externas ou de terceiros sem a devida validação.

De acordo com o OWASP AI Exchange, qual ameaça **não** envolve riscos de envenenamento de dados em fase de desenvolvimento (development-time)?

**A)** Entradas adversariais
**B)** Pipeline de ingestão comprometido
**C)** Fluxos de trabalho de rotulagem
**D)** Dados de treinamento

**A)** Correto. Isso descreve exemplos adversariais na fase da implantação, não o envenenamento de dados em fase de desenvolvimento. (Literatura: A, Capítulo 3)
**B)** Incorreto. Inserir itens manipulados no corpus coletado antes da construção do modelo cria um backdoor durante o desenvolvimento, o que representa um risco de envenenamento na fase de desenvolvimento.
**C)** Incorreto. A adulteração de rótulos durante a anotação envenena o conjunto de dados antes da construção do modelo, o que constitui um ataque na fase de desenvolvimento.
**D)** Incorreto. O uso de conjuntos de dados externos não verificados acarreta o risco de amostras envenenadas estarem incorporadas nas fontes de dados. O OWASP AI Exchange destaca os riscos da cadeia de suprimentos e a necessidade de controle de qualidade dos dados.

**15 / 40**
De acordo com o OWASP AI Exchange, qual é um exemplo de **envenenamento direto de modelo em fase de desenvolvimento (development-time)**?

**A)** Invasores criam entradas adversariais na fase de execução (runtime) para induzir o modelo ao erro.
**B)** Invasores manipulam pesos e um pipeline de treinamento para alterar o comportamento do modelo.
**C)** Dados sensíveis de treinamento são expostos involuntariamente por meio das saídas do modelo.
**D)** Um pico de tráfego sobrecarrega o serviço de IA, fazendo com que ele fique lento ou fique fora do ar.

**A)** Incorreto. Isso corresponde a um ataque adversarial na fase de execução, não a um envenenamento na fase de desenvolvimento.
**B)** Correto. O envenenamento direto de modelo em fase de desenvolvimento envolve comprometer o ambiente de desenvolvimento, o pipeline de treinamento ou os parâmetros do modelo. O invasor pode injetar lógica, modificar pesos ou alterar procedimentos de treinamento para incorporar comportamentos maliciosos. (Literatura: A, Capítulo 3)
**C)** Incorreto. Isso se refere a ameaças à confidencialidade, como inversão do modelo ou divulgação de dados, e não a envenenamento de modelo.
**D)** Incorreto. Isso corresponde a um problema de negação de serviço (denial-of-service, DoS), e não a envenenamento de modelo.

**16 / 40**
De acordo com o OWASP AI Exchange, qual é um exemplo de **envenenamento de modelo na cadeia de suprimentos**?

**A)** Os invasores comprometem dados, modelos ou componentes da cadeia de suprimentos antes da integração.
**B)** Os invasores manipulam as entradas durante a fase de execução (runtime) para induzir o modelo a produzir saídas incorretas.
**C)** O modelo produz resultados incorretos devido à insuficiência de dados ou à baixa qualidade do treinamento.
**D)** O modelo vaza involuntariamente dados de treinamento sensíveis através de suas respostas.

**A)** Correto. O ataque ocorre antes da implantação do modelo de IA, por meio de fontes de terceiros, como modelos pré-treinados, conjuntos de dados externos e componentes ou pipelines de aprendizado de máquina (ML). O OWASP AI Exchange enfatiza a necessidade de gerenciamento da cadeia de suprimentos para garantir a confiança e a integridade. (Literatura: A, Capítulo 3)
**B)** Incorreto. Isso descreve ataques de evasão ou injeção de prompt, não o envenenamento de modelo na cadeia de suprimentos. Esses ataques ocorrem durante o uso do modelo, não durante o desenvolvimento ou a aquisição.
**C)** Incorreto. Resultados incorretos devido à falta de dados ou à baixa qualidade do treinamento não são ataques. O envenenamento na cadeia de suprimentos envolve especificamente adulteração maliciosa intencional.
**D)** Incorreto. Isso se refere a ameaças à confidencialidade, como inversão do modelo ou divulgação de dados sensíveis, e não ao envenenamento de modelo.

**17 / 40**
Uma equipe de desenvolvimento de IA está realizando uma avaliação de riscos em seu ambiente de desenvolvimento. Ela identifica duas ameaças distintas:

- **Ameaça 1** envolve acesso não autorizado aos conjuntos de dados usados para treinar e testar seu modelo de IA.
- **Ameaça 2** envolve acesso não autorizado aos parâmetros e pesos do modelo de IA armazenados em um repositório comprometido.

Que tipos de ameaças são essas?

**A)** - Ameaça 1: vazamento de dados em fase de desenvolvimento (development-time)
- Ameaça 2: vazamento direto do modelo em fase de desenvolvimento

**B)** - Ameaça 1: vazamento direto de dados de aumento (augmentation data)
- Ameaça 2: vazamento do repositório

**C)** - Ameaça 1: vazamento direto de dados em fase de execução (runtime)
- Ameaça 2: vazamento direto de modelo em fase de execução

**D)** - Ameaça 1: vazamento de dados de entrada
- Ameaça 2: vazamento de código-fonte/configuração

**A)** Correto. O vazamento de dados em fase de desenvolvimento envolve acesso não autorizado a dados de treinamento e teste, violando a confidencialidade de dados sensíveis ou propriedade intelectual contidos nesses conjuntos de dados. O vazamento direto do modelo em fase de desenvolvimento envolve acesso não autorizado a atributos do modelo (parâmetros, pesos, arquitetura do modelo), violando a confidencialidade da propriedade intelectual do modelo. (Literatura: A, Capítulo 3)

**B)** Incorreto. O vazamento direto de dados de aumento não corresponde aos conjuntos de dados de treinamento e teste, e vazamento de repositório é muito vago. O problema é especificamente um vazamento de artefato do modelo, não apenas um problema genérico de repositório.

**C)** Incorreto. As ameaças descritas não são vazamentos em fase de execução. Elas ocorrem no ambiente de desenvolvimento.

**D)** Incorreto. O vazamento de dados de entrada geralmente se refere ao vazamento de entradas fornecidas pelo usuário, e não aos conjuntos de dados de treinamento e teste. Os pesos do modelo não são código-fonte/configuração, mas artefatos do modelo.

**18 / 40**
Uma equipe de segurança analisa o panorama de ameaças em fase de execução (runtime) de um sistema de IA. O sistema de IA está integrado a um aplicativo web. A equipe identifica duas ameaças de segurança convencionais:

- **Ameaça 1:** Um invasor pode obter acesso à memória do servidor de produção e extrair os pesos e parâmetros do modelo de IA.
- **Ameaça 2:** um invasor com acesso ao registro do modelo substitui o artefato do modelo implantado por uma versão com trojan.

Quais consequências essas ameaças de segurança convencionais trazem **especificamente** para os sistemas de IA?

**A)** - A ameaça 1 faz com que o sistema se retreine automaticamente com dados maliciosos.
- A ameaça 2 força o modelo a rejeitar prompts válidos.

**B)** - A ameaça 1 altera a arquitetura do modelo em fase de execução (runtime).
- A ameaça 2 comprime silenciosamente os pesos para reduzir a precisão.

**C)** - A ameaça 1 desativa os filtros de segurança operacional, tornando o sistema mais permissivo por padrão (by design).
- A ameaça 2 torna as saídas determinísticas e previsíveis.

**D)** - A ameaça 1 possibilita o roubo do modelo e a potencial inferência de dados de treinamento.
- A ameaça 2 permite a adulteração indetectável do modelo.

**A)** Incorreto. A Ameaça 1 não se refere ao envenenamento por meio do retreinamento automático, mas ao roubo do modelo e à inferência de dados. A Ameaça 2 se refere à adulteração da integridade e a backdoors, não à rejeição de prompts.

**B)** Incorreto. Mudanças na arquitetura em fase de execução e compressão silenciosa de precisão não são consequências específicas à IA decorrentes das ameaças. Elas descrevem comportamentos improváveis ou não maliciosos, em vez de roubo ou adulteração dissimulada.

**C)** Incorreto. Desativar filtros de segurança operacional corresponde à adulteração, não ao risco de exfiltração da ameaça 1. Tornar as saídas determinísticas ignora a natureza de integridade da ameaça 2.

**D)** Correto. A ameaça 1 leva ao roubo do modelo e à potencial inferência de dados de treinamento. A ameaça 2 leva a ataques ocultos à integridade (adulteração não detectada do modelo). (Literatura: A, Capítulo 4)

**19 / 40**
Uma empresa de serviços financeiros implanta um modelo de aprendizado de máquina (ML) para detectar transações fraudulentas. O modelo de ML está hospedado em seu ambiente de produção e processa continuamente dados de transações em tempo real. Dois incidentes de segurança ocorrem:

- **Incidente A**: Um invasor obtém acesso não autorizado ao ambiente do modelo implantado e altera os parâmetros do modelo de modo que certas transações fraudulentas não sejam mais sinalizadas. A manipulação não é detectada, e o invasor se beneficia disso.
- **Incidente B**: Um invasor explora uma vulnerabilidade no sistema de produção. Isso permite que o invasor replique o modelo e analise seu comportamento para futuros ataques.

Com base no OWASP AI Exchange, quais tipos de incidentes de segurança são descritos?

**A)** - Incidente A: envenenamento de dados
- Incidente B: envenenamento direto de modelo em fase de execução (runtime)

**B)** - Incidente A: envenenamento de dados
- Incidente B: ataque de evasão

**C)** - Incidente A: envenenamento direto de modelo em fase de execução
- Incidente B: vazamento direto de modelo em fase de execução

**D)** - Incidente A: vazamento direto de modelo em fase de execução
- Incidente B: ataque de evasão

**A)** Incorreto. O vazamento direto de modelo em fase de execução identifica corretamente o incidente de segurança B. O invasor extrai parâmetros e a arquitetura do modelo, comprometendo a confidencialidade do mesmo. Isso permite a replicação e exploração posterior. No entanto, o envenenamento de dados ocorre durante a fase de desenvolvimento (development time) por meio da manipulação dos dados de treinamento. O cenário descreve um comprometimento em fase de execução, não uma adulteração na fase de treinamento.

**B)** Incorreto. O envenenamento de dados ocorre durante a fase de desenvolvimento (development time) por meio da manipulação dos dados de treinamento. O cenário descreve um comprometimento em fase de execução, não uma adulteração na fase de treinamento. Um ataque de evasão envolve a manipulação de entradas, não a modificação do próprio modelo. No cenário, o modelo é alterado diretamente.

**C)** Correto. O envenenamento direto de modelo em fase de execução identifica corretamente o incidente de segurança A. O invasor modifica o modelo implantado, alterando seu comportamento. Trata-se de um ataque à integridade do modelo conhecido como envenenamento em fase de execução (reprogramação). O vazamento direto de modelo em fase de execução identifica corretamente o incidente de segurança B. O invasor extrai os parâmetros e a arquitetura do modelo, comprometendo a confidencialidade do mesmo. (Literatura: A, Capítulo 4)

**D)** Incorreto. O vazamento direto de modelo em fase de execução identifica corretamente o incidente de segurança B. O invasor extrai os parâmetros e a arquitetura do modelo, comprometendo a confidencialidade do mesmo. Isso permite a replicação e a exploração posterior. No entanto, um ataque de evasão envolve a manipulação de entradas, e não a modificação do próprio modelo. No cenário, o modelo é alterado diretamente. (Literatura: A, Capítulo 4)

**20 / 40**
Uma organização de saúde implanta um assistente de IA para ajudar os médicos a resumir prontuários de pacientes e fornecer recomendações. O sistema se integra a bancos de dados internos e ferramentas externas, e os médicos confiam nas saídas geradas por ele durante as operações diárias. Duas ameaças são observadas:

- **Ameaça 1**: O assistente de IA gera uma resposta que inclui um script malicioso oculto no conteúdo HTML. Quando a saída é visualizada, o script é executado e envia dados da sessão ao invasor.
- **Ameaça 2**: Um invasor explora controles de acesso fracos no sistema para interceptar e recuperar dados sensíveis de pacientes.

Com base no OWASP AI Exchange, quais tipos de ameaças são descritos?

**A)** - Ameaça 1: envenenamento de dados
- Ameaça 2: vazamento de dados de entrada

**B)** - Ameaça 1: envenenamento de dados
- Ameaça 2: ataque de inversão do modelo

**C)** - Ameaça 1: saída contendo injeção convencional
- Ameaça 2: vazamento de dados de entrada

**D)** - Ameaça 1: saída contendo injeção convencional
- Ameaça 2: ataque de inversão do modelo

**A)** Incorreto. A ameaça 2 está corretamente identificada como um vazamento de dados de entrada. Dados sensíveis fornecidos ao modelo são expostos ou interceptados, violando a confidencialidade. No entanto, o envenenamento de dados ocorre durante a fase de desenvolvimento (development-time), por meio da alteração dos dados de treinamento. O cenário descreve uma injeção convencional na fase de execução (runtime), e não uma manipulação do treinamento.

**B)** Incorreto. O envenenamento de dados ocorre durante a fase de desenvolvimento (development-time), por meio da alteração dos dados de treinamento. O cenário descreve uma injeção convencional em fase de execução (runtime), não uma manipulação do treinamento. A inversão do modelo envolve a inferência de dados de treinamento a partir das saídas. O cenário não descreve uma situação em que tenha ocorrido inferência do modelo.

**C)** Correto. A Ameaça 1 foi identificada corretamente como uma saída contendo injeção convencional. A saída da IA inclui conteúdo executável malicioso, nesse caso, injeção de script. A Ameaça 2 foi identificada corretamente como um vazamento de dados de entrada. Dados sensíveis fornecidos ao modelo são expostos ou interceptados, violando a confidencialidade. (Literatura: A, Capítulo 4)

**D)** Incorreto. A ameaça 1 é corretamente identificada como saída contendo injeção convencional. A saída da IA inclui conteúdo executável malicioso, nesse caso, injeção de script. No entanto, a inversão do modelo envolve a inferência de dados de treinamento a partir das saídas. O cenário não descreve uma situação em que tenha ocorrido a inferência do modelo.

**21 / 40**
Uma empresa utiliza um sistema de geração aumentada por recuperação (RAG) para responder às perguntas dos funcionários sobre políticas internas. O sistema de IA recupera documentos de um banco de dados vetorial e os adiciona ao prompt antes de enviá-los ao modelo de IA.

Uma análise de segurança constata que um invasor obteve acesso não autorizado ao banco de dados vetorial. O invasor visualizou os documentos de aumento (augmentation documents) armazenados e substituiu um documento de política legítimo por outro, contendo informações falsas. Os funcionários que perguntam ao sistema de IA sobre essa política agora recebem orientações incorretas.

Quais ameaças de segurança esse cenário ilustra?

**A)** - Ameaça 1: vazamento de dados de aumento (augmentation data)
- Ameaça 2: saída contendo injeção convencional

**B)** - Ameaça 1: manipulação de dados de aumento
- Ameaça 2: vazamento de dados de aumento

**C)** - Ameaça 1: inferência de associação (membership inference)
- Ameaça 2: manipulação de dados de aumento

**D)** - Ameaça 1: saída contendo injeção convencional
- Ameaça 2: inferência de associação (membership inference)

**A)** Incorreto. A saída contendo injeção convencional envolve conteúdo executável malicioso incorporado à saída do modelo, o que não está presente nesse cenário. Embora o vazamento de dados de aumento tenha sido identificado corretamente, a resposta como um todo está incorreta.

**B)** Correto. A manipulação de dados de aumento ocorre quando os dados de aumento armazenados em um banco de dados vetorial são modificados por um invasor, o qual insere informações falsas que alteram o comportamento do modelo. O vazamento de dados de aumento ocorre porque um invasor obteve acesso não autorizado a esses dados armazenados no banco de dados vetorial. O banco de dados vetorial representa um local adicional onde os dados de aumento podem ser acessados ou modificados, criando uma nova superfície de ataque e ampliando os riscos de exposição para além do repositório de documentos original. (Literatura: A, Capítulo 4)

**C)** Incorreto. A inferência de associação é um ataque em que um adversário determina se dados específicos aparecem no conjunto de treinamento por meio da análise de sinais de confiança. Isso não está descrito no cenário. Embora a manipulação de dados de aumento tenha sido identificada corretamente, a relação com a inferência de associação está incorreta.

**D)** Incorreto. Saída contendo injeção convencional refere-se a código malicioso incorporado à saída do modelo. A inferência de associação é um ataque em que um adversário determina se dados específicos aparecem no conjunto de treinamento por meio da análise de sinais de confiança. Nenhuma dessas ameaças de segurança é descrita no cenário.

**22 / 40**
Controles gerais de governança devem ser implementados para garantir uma supervisão eficaz da segurança dos sistemas de IA.

Qual abordagem demonstra isso **melhor**?

**A)** Definir políticas, funções e práticas de gerenciamento de riscos claras para orientar o desenvolvimento, a implantação e o monitoramento seguros de sistemas de IA
**B)** Incluir controles de segurança nos sistemas de IA para aplicar os princípios de segurança desde a concepção (by design) na fase de design do sistema de IA
**C)** Atender aos requisitos mínimos de segurança, já que os sistemas de IA são mais seguros por padrão (by default) e não exigem múltiplas camadas de segurança
**D)** Separar os controles de governança para IA dos controles de segurança da informação para garantir que esses dois conceitos diferentes sejam igualmente abrangidos

**A)** Correto. Os controles de governança tratam da definição e aplicação de políticas, funções, responsabilização e gerenciamento de riscos dentro de um Sistema de Gestão de Segurança da Informação (SGSI), garantindo supervisão de ponta a ponta ao longo do ciclo de vida da IA. (Literatura: A, Capítulo 1)
**B)** Incorreto. A aplicação dos princípios de segurança desde a concepção na fase de design dos sistemas de IA é uma parte importante dos controles de governança, mas não é suficiente. A governança envolve políticas, responsabilidade, risco, monitoramento e melhoria contínua.
**C)** Incorreto. Os sistemas de IA não são seguros por padrão. Basear-se em requisitos mínimos ignora a defesa em profundidade e a supervisão contínua necessárias para os riscos em evolução da IA.
**D)** Incorreto. Separar a governança da IA da segurança da informação cria silos e lacunas. Uma prática recomendada geral é a integração com as estruturas de risco corporativo e do Sistema de Gestão de Segurança da Informação (SGSI) para manter controles coerentes e a responsabilização e prestação de contas.

**23 / 40**
Uma agência de marketing de médio porte tem pouca experiência com IA e ainda não possui uma governança formal nessa área. Ela deseja usar uma ferramenta de IA generativa para redigir e-mails para clientes. Para tanto, a ferramenta precisa de acesso aos nomes e e-mails dos clientes. A empresa deve estabelecer uma governança de IA e definir uma supervisão de segurança antes de poder usar essa ferramenta.

Qual é o **mínimo indispensável** que a empresa deve fazer em termos de supervisão de segurança?

**A)** Realizar testes de red teaming de IA no sistema de e-mail com IA generativa para mitigar riscos de segurança antes do lançamento
**B)** Contratar um especialista em governança de IA para realizar uma avaliação de ameaças e definir políticas e controles para a implantação
**C)** Fazer um inventário do uso planejado da IA e realizar uma análise de risco para identificar ameaças, controles e responsabilidades
**D)** Solicitar assessoria jurídica de uma empresa especializada em IA e nos requisitos relacionados à Lei da IA (AI Act)
**E)** Começar com um projeto-piloto e monitorar o sistema de e-mail com IA quanto ao desempenho, usos indevidos e incidentes por meio de registros

**A)** Incorreto. Testes de red teaming de IA não são uma etapa mínima de governança. Esses testes correspondem a uma atividade de garantia de uma fase posterior que depende de um escopo definido, controles de referência e responsabilidades estabelecidas pela governança inicial.
**B)** Incorreto. A contratação de um especialista em governança de IA é uma decisão de alocação de recursos, não uma etapa mínima de governança.
**C)** Correto. Essa opção representa a medida mínima de governança de segurança da IA, identificando riscos, ameaças, controles e responsabilidades antes da implantação. (Literatura: A, Capítulo 1)
**D)** Incorreto. A análise jurídica é importante, mas faz parte da conformidade e não pode substituir a governança de segurança. Adiar a governança deixa os dados pessoais e os riscos de uso indevido sem gerenciamento.
**E)** Incorreto. A fase piloto e o monitoramento são operacionais e reativos, e a realização de um projeto-piloto em um pequeno segmento de clientes é uma estratégia de implementação. Sem governança prévia, a fase piloto e o monitoramento carecem de limites, funções e controles definidos para proteger os dados do CRM.

**24 / 40**
Como a cobertura dos controles de governança na segurança da IA deve ser **melhor** compreendida?

**A)** Como um conjunto de métodos de criptografia utilizados para proteger as saídas da IA após a geração
**B)** Como um conjunto de controles abrangentes que se aplicam a todas as ameaças de IA e a todas as etapas do ciclo de vida da IA
**C)** Como controles limitados à fase de treinamento do modelo, com foco na qualidade do conjunto de dados e em técnicas de ajuste
**D)** Como medidas opcionais que as organizações podem adotar caso os sistemas de IA não processem dados sensíveis
**E)** Como salvaguardas técnicas aplicadas exclusivamente na fase de execução (runtime) para detectar entradas maliciosas

**A)** Incorreto. A criptografia é apenas um dos possíveis controles técnicos e não representa a abrangência dos controles de governança. Os controles gerais não se limitam à criptografia ou às saídas.
**B)** Correto. A governança é transversal: políticas, funções, responsabilização e gerenciamento de riscos que se aplicam a todas as ameaças e a todas as etapas do ciclo de vida da IA. (Literatura: A, Capítulo 1)
**C)** Incorreto. A governança se aplica a todo o ciclo de vida da IA, não apenas à fase de treinamento.
**D)** Incorreto. Os controles de governança não são opcionais e não se limitam ao processamento de dados sensíveis.
**E)** Incorreto. Os controles gerais não se limitam à fase de execução ou às defesas técnicas. Eles abrangem as camadas organizacional, procedural e de governança.

**25 / 40**
Uma empresa do setor de saúde decide utilizar um modelo de IA pronto para uso (ready-made model), fornecido por um terceiro, para um aplicativo. O aplicativo processa entradas do usuário e recupera dados externos.

O engenheiro de segurança da IA está preocupado com a segurança do modelo. A estratégia de segurança que define a divisão de responsabilidades de segurança entre o fornecedor do modelo de IA e a empresa do setor de saúde não está definida.

O que é **correto** em relação a essas responsabilidades?

**A)** A maioria dos controles de segurança pode ser delegada ao fornecedor ao utilizar modelos de IA prontos para uso, uma vez que o treinamento e a hospedagem realizados pelo fornecedor abrangem os riscos de segurança.
**B)** O responsável pela implantação é responsável por todos os controles de segurança ao integrar o modelo de IA em seus sistemas e deve considerar a segurança antes da implantação.
**C)** O fornecedor treinou o modelo de IA, portanto, ele gerencia os controles no nível do modelo, enquanto o responsável pela implantação deve gerenciar os controles no nível da aplicação.
**D)** O fornecedor é responsável por todos os controles de segurança porque treinou o modelo de IA e o hospeda, então todos os riscos e ataques devem ser tratados pelo fornecedor do modelo de IA.

**A)** Incorreto. A injeção de prompts e o vazamento de dados no contexto de ferramentas externas ou da geração aumentada por recuperação (RAG) são riscos principalmente da camada de aplicação e não podem ser totalmente delegados ao fornecedor. A hospedagem e o treinamento, por si sós, não abrangem totalmente esses riscos.
**B)** Incorreto. O fornecedor treina e ajusta o modelo e define os controles de segurança operacional básicos. Portanto, ele é responsável por parte da segurança, o que significa que há uma responsabilidade compartilhada.
**C)** Correto. O fornecedor treina e ajusta o modelo, sendo responsável pelos controles no nível do modelo. O responsável pela implantação ainda deve estar ciente das ameaças ao usar modelos prontos para uso, sendo responsável pelos controles no nível da aplicação. (Literatura: A, Capítulo 0)
**D)** Incorreto. Embora o fornecedor treine e hospede o modelo, ele não é responsável por todos os controles de segurança. Ele é responsável por parte da segurança, o que significa que há uma responsabilidade compartilhada.

**26 / 40**
Uma organização responsável pela implantação de um aplicativo de IA voltado para o cliente utilizou um Modelo de Linguagem de Grande Porte (LLM) fornecido por um fornecedor terceirizado. O modelo é acessado por meio de um serviço de API (interface de programação de aplicações) gerenciado e não foi submetido a treinamento adicional nem modificado pela organização.

A equipe de segurança descobre que entradas cuidadosamente elaboradas por usuários podem fazer com que o modelo ignore seus limites comportamentais configurados e produza saídas que violam as restrições de conteúdo da organização. A equipe avalia qual controle deve ser implementado que seja de responsabilidade da organização responsável pela implantação, e não do fornecedor terceirizado.

Qual controle a equipe de segurança deve implementar para resolver o problema?

**A)** Inserir uma camada de validação de saída que analise todas as respostas do modelo antes que elas cheguem aos usuários finais e bloquear as saídas que violarem a política de conteúdo da organização
**B)** Modificar o pipeline de inferência do fornecedor para impor restrições de geração mais rigorosas que limitem a geração de respostas e impeçam o surgimento de conteúdo não permitido
**C)** Solicitar que o fornecedor retreine e ajuste a segurança operacional do modelo fornecido usando dados adicionais de rejeição e alinhamento com políticas para melhorar a resistência contra entradas de tipo “jailbreak”
**D)** Exigir que o fornecedor revise e aplique regras de moderação de conteúdo reforçadas e sua aplicação em todo o serviço, de modo que saídas inseguras sejam filtradas globalmente antes da entrega

**A)** Correto. Os responsáveis pela implantação são responsáveis pelos controles implementados em torno do aplicativo de IA e de suas saídas, enquanto os fornecedores terceirizados gerenciam o treinamento do modelo subjacente e o comportamento de inferência. Uma camada de validação de saída é um controle por parte do responsável pela implantação que pode bloquear respostas que violam políticas da empresa antes que elas cheguem aos usuários finais. (Literatura: A, Capítulo 0)
**B)** Incorreto. Modificar o pipeline de inferência do fornecedor não está sob o controle do responsável pela implantação ao usar uma API gerenciada por terceiros. Na melhor das hipóteses, ele pode ajustar parâmetros expostos, mas não alterar o pipeline.
**C)** Incorreto. O retreinamento ou o ajuste de segurança operacional do modelo são de responsabilidade do fornecedor. O responsável pela implantação pode solicitar isso, mas não pode implementar por conta própria para um modelo de API gerenciado.
**D)** Incorreto. As políticas globais de moderação e os controles de filtragem em toda a rede do fornecedor são gerenciados pelo fornecedor terceirizado, e não pela organização responsável pela implantação.

**27 / 40**
Uma equipe está preparando dados de treinamento para um assistente de IA generativa. A equipe constata que não precisa de todos os campos e dados que usa atualmente para o desempenho do modelo de IA. Como próximo passo, a equipe está considerando controles para limitação de dados sensíveis e melhora da confidencialidade e da integridade.

Qual é a **melhor** maneira de melhorar a confidencialidade e a integridade?

**A)** Manter todos os campos nos dados de treinamento e realizar a criptografia dos dados em repouso e nos backups
**B)** Preservar identificadores nos dados de treinamento para simplificar futuras tarefas e processos de retreinamento
**C)** Remover campos e dados desnecessários para o treinamento e o desempenho do modelo
**D)** Usar apenas dados sintéticos para todas as fases de treinamento e avaliação da IA generativa

**A)** Incorreto. A criptografia em repouso protege os dados armazenados, mas não reduz a exposição desnecessária de dados sensíveis nem minimiza a superfície de ataque. Ela melhora a confidencialidade apenas em repouso e contribui pouco para a integridade.
**B)** Incorreto. Manter identificadores nos dados de treinamento aumenta o risco de exposição. Isso compromete a confidencialidade e pode prejudicar a integridade.
**C)** Correto. A minimização de dados remove campos e registros desnecessários, reduzindo a superfície de ataque dos dados. Ela fortalece a integridade ao oferecer menos oportunidades para que invasores realizem adulterações nos dados e ao impedir que o modelo se baseie em padrões irrelevantes. (Literatura: A, Capítulo 1)
**D)** Incorreto. O uso de dados sintéticos nem sempre é viável ou necessário e não garante melhorias na confidencialidade e na integridade. Os dados sintéticos ainda podem preservar padrões confidenciais ou introduzir outros riscos.

**28 / 40**
Uma equipe de produto está treinando um modelo de IA para atendimento ao cliente. O conjunto de dados de treinamento atual inclui endereços completos dos clientes, números de telefone e registros antigos que não são necessários para atingir as metas de desempenho, mas provavelmente serão úteis no futuro. Depois que um exercício de red teaming destaca o risco de vazamento e manipulação de dados por meio de detalhes memorizados, a equipe decide reduzir a exposição. Para isso, ela remove campos desnecessários e registros inteiros, de modo que dados sensíveis não estejam presentes no conjunto de dados.

Qual controle **geral** foi implementado para limitar os dados sensíveis nessa situação?

**A)** Dados permitidos
**B)** Minimização de dados
**C)** Ocultação de dados de treinamento
**D)** Retenção curta

**A)** Incorreto. Dados permitidos definem quais categorias são permitidas, mas o cenário exige a exclusão de registros desnecessários, o que corresponde à minimização de dados.
**B)** Correto. A minimização de dados consiste em remover registros desnecessários do conjunto de dados de treinamento para que dados sensíveis não estejam presentes e, portanto, não possam vazar. (Literatura: A, Capítulo 1)
**C)** Incorreto. A ocultação dos dados de treinamento transforma os dados, mas ainda os retém. Portanto, os riscos podem permanecer, ao contrário do que ocorre com a remoção.
**D)** Incorreto. A retenção curta limita por quanto tempo os dados são mantidos, não quais registros são incluídos no conjunto de dados de treinamento.

**29 / 40**
Uma empresa de varejo decide parar de armazenar os detalhes de pagamento dos clientes após a conclusão de cada transação.

Como essa ação **reduz** o risco de exposição de dados?

**A)** Ao eliminar a necessidade de criptografia dos dados em trânsito ou arquivados para fins de auditoria
**B)** Ao garantir a conformidade com todas as regulamentações relevantes de proteção de dados, reduzindo a complexidade
**C)** Ao garantir que nenhum usuário não autorizado possa acessar a rede e o banco de dados da empresa
**D)** Ao impedir o vazamento de dados retidos, sua reconstrução ou sua inferência a partir do sistema

**A)** Incorreto. A minimização de dados não elimina a necessidade de criptografia. Os dados transmitidos ainda precisam ser protegidos, independentemente da quantidade de dados armazenados.
**B)** Incorreto. A minimização de dados apoia os esforços de conformidade, mas não garante automaticamente a conformidade total com todas as regulamentações de proteção de dados, as quais possuem muitos requisitos adicionais.
**C)** Incorreto. Limitar os dados armazenados não garante que usuários não autorizados não possam acessar a rede ou o banco de dados. O controle de acesso à rede é uma medida de segurança separada, não relacionada à minimização de dados.
**D)** Correto. Quando os dados não são coletados ou retidos, eles não podem ser vazados, reconstruídos ou inferidos a partir do sistema. Isso reduz diretamente o impacto do roubo de conjuntos de dados ou do acesso não autorizado. (Literatura: A, Capítulo 1)

**30 / 40**
Uma rede regional de hospitais utiliza um agente de IA para ler e-mails recebidos de fornecedores, redigir respostas e enviá-las automaticamente. Durante testes de segurança realizados após um recente incidente de phishing, os testadores descobriram que a injeção de prompts poderia fazer com que o agente de IA ignorasse suas regras e enviasse respostas não autorizadas.

Qual decisão de projeto funciona **melhor** para limitar os efeitos desse comportamento indesejado?

**A)** Desativar todas as integrações do agente de IA e voltar a fluxos de trabalho totalmente manuais para a maioria dos processos
**B)** Ampliar o treinamento dos usuários sobre segurança operacional de prompts e enfatizar a supervisão durante o envio de e-mails assistido pelo agente
**C)** Manter permissões amplas de envio autônomo e atualizar o sistema para um modelo generativo mais avançado
**D)** Restringir as permissões do agente de IA a escopos específicos de tarefas e exigir aprovação para ações com risco elevado

**A)** Incorreto. Isso remove a automação em vez de implementar controles para manter a funcionalidade com segurança.
**B)** Incorreto. A conscientização dos usuários ajuda, mas não impõe controles em fase de execução (runtime) da IA.
**C)** Incorreto. Melhorar a qualidade do modelo não reduz suficientemente o raio de impacto (blast radius) se o sistema de IA generativa ainda tiver permissões amplas.
**D)** Correto. Recomenda-se o uso de controles de privilégio mínimo e de intervenção humana para limitar o impacto caso ocorra injeção de prompt. O privilégio mínimo específico para cada tarefa limita o que o agente pode fazer, enquanto a aprovação de ações com risco elevado adiciona supervisão humana, impedindo respostas não autorizadas mesmo quando o modelo é enganado. (Literatura: A, Capítulo 1)

**31 / 40**
Qual é a **melhor** maneira de minimizar os efeitos de comportamentos indesejados do modelo?

**A)** Enfatizar o fortalecimento de prompts, o acesso baseado em funções e as auditorias programadas das saídas do modelo
**B)** Contar com filtros automatizados de saída, registro de exceções e verificações pontuais ocasionais quanto à equidade
**C)** Testar a presença de viés indesejado e aplicar controles como supervisão, privilégio mínimo e validação contínua
**D)** Utilizar validação robusta de entradas, permissões restritas para ferramentas e revisões periódicas com intervenção humana

**A)** Incorreto. Essas medidas são úteis, mas incompletas. O fortalecimento de prompts é frágil; o controle de acesso baseado em funções (RBAC) sem um programa mais amplo de privilégios mínimos é limitado; e as auditorias programadas são periódicas, e não contínuas, deixando lacunas e carecendo de testes sistemáticos de viés.
**B)** Incorreto. Filtros automatizados de saída podem ser contornados, o registro de exceções não impede danos e as verificações pontuais ocasionais quanto à equidade são insuficientes em comparação com a validação contínua e os testes estruturados de viés.
**C)** Correto. Os controles para limitar o efeito de comportamentos indesejados do modelo são: supervisão, privilégio mínimo do modelo, alinhamento do modelo, transparência da IA, explicabilidade, validação contínua e testes de viés indesejado. (Literatura: A, Capítulo 1)
**D)** Incorreto. A validação rigorosa das entradas, as permissões restritas das ferramentas e a revisão humana periódica ajudam, mas essa lista omite a validação contínua e os testes dedicados de viés. As revisões periódicas deixam lacunas e não proporcionam uma redução abrangente e contínua dos riscos.

**32 / 40**
Limitar comportamentos indesejados do modelo é uma estratégia crítica para a limitação de impacto e o controle do raio de impacto (blast radius).

Como essa estratégia **melhora** o desempenho e **reduz** o risco?

**A)** Ao concentrar-se em ataques que podem afetar sistemas de IA e limitar a supervisão sobre fatores de desempenho e acurácia do sistema
**B)** Ao implementar controles que eliminam a necessidade de testes e validações contínuas ao longo do tempo para reduzir a complexidade
**C)** Ao melhorar a confiabilidade operacional e a eficiência de recursos, minimizando incidentes e solicitações de computação desperdiçadas
**D)** Ao aumentar a autonomia do modelo e permitir que ele gere mais saídas para melhorar continuamente o desempenho do sistema

**A)** Incorreto. A transparência da IA é um dos controles para limitar comportamentos indesejados. Ela está diretamente relacionada à acurácia, pois ataques ou desvios que causam comportamentos indesejados são problemas de acurácia.
**B)** Incorreto. Testes regulares e validação de desempenho garantem que o modelo permaneça consistente com o uso pretendido ao longo do tempo e limitam o efeito de comportamentos indesejados. O teste de viés indesejado é um dos controles para limitar o efeito de comportamentos indesejados em sistemas de IA. Testes e validações contínuos ao longo do tempo são altamente recomendados.
**C)** Correto. Limitar comportamentos indesejados reduz saídas prejudiciais e incidentes operacionais, melhora a confiabilidade e aumenta a eficiência dos recursos ao reduzir novas tentativas, chamadas desnecessárias de ferramentas e desperdício de recursos computacionais. (Literatura: A, Capítulo 1)
**D)** Incorreto. Aumentar a autonomia do modelo não leva a um melhor desempenho. O nível de autonomia do modelo deve estar sob controle para que comportamentos indesejados do modelo possam ser limitados.

**33 / 40**
No âmbito mais amplo do red teaming de IA, qual é o **principal** objetivo dos testes de segurança da IA?

**A)** Avaliar se um modelo de IA é resiliente a ataques específicos, reproduzindo esses ataques em um ambiente controlado
**B)** Avaliar a acurácia e a rastreabilidade das saídas de um modelo de IA quando exposto a dados de entrada esperados
**C)** Medir se um modelo de IA demonstra a eficiência computacional esperada ao processar grandes volumes de solicitações
**D)** Verificar se um modelo de IA atende aos requisitos de negócios definidos por meio de procedimentos padrão de garantia de qualidade

**A)** Correto. Os testes de segurança de IA são parte do red teaming de IA, que verifica se o modelo de IA é capaz de resistir a determinados ataques, simulando-os. Seu objetivo principal é, portanto, verificar a resiliência do modelo por meio da reprodução controlada de ataques. (Literatura: A, Capítulo 5)
**B)** Incorreto. Avaliar a acurácia do modelo sob condições normais de entrada descreve testes funcionais ou de desempenho, e não testes de segurança da IA, que se concentram na simulação de ataques.
**C)** Incorreto. A eficiência computacional sob cargas de trabalho rotineiras é uma preocupação dos testes de desempenho ou de carga, e não do escopo dos testes de segurança da IA, que se concentram em ataques.
**D)** Incorreto. Verificar a conformidade com os requisitos de negócios por meio da garantia de qualidade está fora do escopo dos testes de segurança da IA, que visam especificamente a capacidade do modelo de IA de resistir a ataques.

**34 / 40**
Uma empresa responsável pela implantação de um assistente de IA generativa deseja definir testes de segurança da IA que abranjam mais do que apenas testes de penetração convencionais. Para isso, a equipe de testes começa a identificar as ameaças que devem ser testadas.

Quais ameaças **especificamente** associadas a sistemas de IA generativa devem ser incluídas nos testes de segurança da IA?

**A)** Ataques de evasão, tratamento inseguro de saídas e envenenamento de modelo
**B)** Ataques de exfiltração de modelo, injeção de prompt e evasão
**C)** Injeção de prompt, divulgação de dados sensíveis e tratamento inseguro de saídas
**D)** Saída de dados sensíveis do modelo, desvio do modelo e invocação insegura de ferramentas

**A)** Incorreto. Ataques de evasão são mais comumente associados a sistemas de IA preditiva, enquanto o tratamento inseguro de saídas pode afetar sistemas de IA generativa. Essa combinação não representa da melhor forma as principais ameaças específicas à IA generativa que deveriam ser incluídas nos testes de segurança da IA.
**B)** Incorreto. Embora essas sejam ameaças de segurança válidas para a IA, os ataques de evasão são mais comumente associados a sistemas de IA preditivos do que a assistentes de IA generativa.
**C)** Correto. Essas três ameaças afetam a IA generativa e, portanto, devem ser testadas. (Literatura: A, Capítulo 5)
**D)** Incorreto. O desvio do modelo é uma questão de desempenho, não uma ameaça de segurança.

**35 / 40**
Uma equipe de segurança da IA está testando um chatbot de IA generativa quanto à resistência à injeção de prompts. Durante o teste, o sistema bloqueou várias entradas de ataque criadas propositalmente.

Qual é a **próxima** etapa recomendada para manter a supervisão de segurança do chatbot de IA generativa?

**A)** Adicionar variações de entrada, como substituição de sinônimos, alterações de codificação ou mudanças de formatação, para tentar contornar a detecção
**B)** Concluir o ciclo de testes atual e considerar as defesas adequadas, já que a maioria das tentativas de injeção de prompt foi bloqueada
**C)** Executar o mesmo teste várias vezes para aumentar o tamanho da amostra e obter dados de desempenho das defesas mais relevantes do ponto de vista estatístico
**D)** Enviar as entradas de ataque por injeção de prompt diretamente ao modelo para treiná-lo melhor contra ameaças semelhantes em implantações futuras

**A)** Correto. Adicionar algoritmos de variação, como sinônimos, codificação e formatação, para tentar contornar a detecção aumenta a robustez da detecção. Ampliar a diversidade de entradas melhora a cobertura, expõe formas de contorno e mantém a supervisão de segurança. (Literatura: A, Capítulo 5)
**B)** Incorreto. O teste ainda não deve ser concluído. O sistema ainda deve ser resiliente contra tentativas deliberadas de contorno.
**C)** Incorreto. Repetir os mesmos testes sem introduzir variações provavelmente não revelará novas técnicas de contorno nem melhorará a cobertura.
**D)** Incorreto. Enviar entradas de ataque brutas diretamente para o modelo contorna os filtros de produção e os mecanismos de detecção. Além disso, alimentar o treinamento diretamente com entradas de ataque é uma atividade de aprimoramento do modelo, não a próxima etapa do teste.

**36 / 40**
O sistema de IA de uma empresa depende de grandes quantidades de dados pessoais para treinamento e tomada de decisões.

Do ponto de vista ético, qual é o risco **mais** importante que a empresa deve abordar?

**A)** Dados sendo coletados, retidos ou reutilizados para além dos fins justificados
**B)** Complexidade da interface dificultando a revisão das decisões automatizadas
**C)** Acurácia do modelo diminuindo quando os recursos são alocados de forma desigual
**D)** Atualizações de software causando problemas de compatibilidade entre plataformas

**A)** Correto. Os sistemas de IA fazem uso intensivo de dados e levantam preocupações sobre como os dados pessoais são coletados, retidos e reutilizados. Os princípios e a legislação de privacidade frequentemente exigem uma base legal ou consentimento para tal uso, juntamente com salvaguardas como limitação de uso e minimização de dados. (Literatura: A, Capítulo 6)
**B)** Incorreto. A complexidade da interface é principalmente uma questão de usabilidade. Ela não diz respeito diretamente aos riscos decorrentes de sistemas de IA que dependem de grandes quantidades de dados pessoais.
**C)** Incorreto. A alocação de recursos de hardware diz respeito ao desempenho técnico e à operação do modelo de IA. Essa não é a principal preocupação associada ao uso de dados pessoais em sistemas de IA.
**D)** Incorreto. Questões relacionadas a atualizações de software e compatibilidade dizem respeito à manutenção do sistema e à interoperabilidade. Elas não são a principal preocupação decorrente do uso de grandes quantidades de dados pessoais em sistemas de IA.

**37 / 40**
Uma livraria online utiliza um sistema de IA para analisar o histórico de navegação, as tendências de compra e as pré-vendas, a fim de prever a demanda e garantir que os livros populares estejam em estoque a tempo. Posteriormente, a empresa reutiliza esses mesmos dados de clientes para treinar um modelo de IA separado que cria perfis de marketing detalhados e envia anúncios direcionados a compradores individuais.

Qual princípio de privacidade é **mais diretamente** violado nesse caso?

**A)** Minimização de dados
**B)** Limitação da retenção de dados
**C)** Transparência e consentimento
**D)** Limitação de uso e especificação da finalidade

**A)** Incorreto. A minimização de dados refere-se à coleta apenas dos dados estritamente necessários para uma finalidade definida, e não à restrição quanto à reutilização de dados já coletados para um objetivo diferente.
**B)** Incorreto. A limitação da retenção de dados diz respeito ao tempo durante o qual os dados são armazenados antes de precisarem ser excluídos, e não à possibilidade de dados coletados para uma finalidade serem reutilizados para outra finalidade.
**C)** Incorreto. A transparência e o consentimento referem-se à informação prestada aos indivíduos sobre a coleta de dados e à obtenção de seu consentimento, e não especificamente à restrição de usar dados coletados para uma finalidade com o intuito de treinar um modelo para uma finalidade diferente.
**D)** Correto. O cenário ilustra uma violação da limitação de uso e da especificação da finalidade, pois os dados coletados para segurança operacional (safety) e segurança (security) não devem ser reutilizados como conjunto de dados de treinamento para criação de perfis ou marketing personalizado. (Literatura: A, Capítulo 6)

**38 / 40**
Uma organização deseja estabelecer uma estrutura para reger o uso responsável da IA em todas as suas operações.

Ela utiliza a norma ISO/IEC 27001 para a gestão da segurança da informação. A equipe de segurança deseja estabelecer sua governança de IA com a norma ISO/IEC que cumpre a mesma finalidade estrutural para a governança de IA que a norma ISO/IEC 27001 cumpre para a segurança da informação.

Qual é essa norma?

**A)** ISO/IEC 23894
**B)** ISO/IEC 27005
**C)** ISO/IEC 42001
**D)** ISO/IEC 5338

**A)** Incorreto. A ISO/IEC 23894 concentra-se no gerenciamento de riscos de IA, e não no fornecimento de um sistema de gestão para a governança da IA responsável.
**B)** Incorreto. A ISO/IEC 27005 trata do gerenciamento de riscos de segurança da informação e não foi projetada para servir como um sistema de gestão de governança para a IA responsável.
**C)** Correto. A ISO/IEC 42001 pode ser vista como uma norma de sistema de gestão para a governança da IA responsável em uma organização, de forma semelhante à ISO/IEC 27001, que é uma norma de sistema de gestão para a segurança da informação. (Literatura: A, Capítulo 1)
**D)** Incorreto. A ISO/IEC 5338 concentra-se na ampliação das práticas e da engenharia do ciclo de vida do software, e não em fornecer um sistema de gestão de governança para a IA responsável.

**39 / 40**
Uma empresa deseja utilizar um sistema de IA para selecionar e classificar automaticamente candidatos a vagas de emprego, com base em seus currículos.

De acordo com a Lei da IA (AI Act), em qual categoria o uso desse sistema de IA deve ser classificado?

**A)** Risco inaceitável
**B)** Risco elevado
**C)** Risco limitado
**D)** Risco mínimo ou nulo

**A)** Incorreto. A Lei da IA não impõe proibição às ferramentas de contratação baseadas em IA. Ela as classifica como de risco elevado, o que significa que estão sujeitas a requisitos e restrições específicas, mas não são totalmente proibidas.
**B)** Correto. A triagem de candidatos a emprego pode afetar o acesso ao trabalho e aos direitos fundamentais e deve ser classificada como de risco elevado. As organizações devem garantir a conformidade antes da implantação de sistemas de IA para tais fins. (Literatura: A, Capítulo 6)
**C)** Incorreto. O risco limitado abrange principalmente obrigações relacionadas apenas à transparência. A triagem de candidatos a emprego pode afetar o acesso ao trabalho e aos direitos fundamentais e deve ser classificada como de risco elevado.
**D)** Incorreto. O risco mínimo se aplica a usos sem consequências, como filtros de spam ou videogames. A triagem de candidatos a emprego pode afetar o acesso ao trabalho e aos direitos fundamentais e deve ser classificada como de risco elevado.

**40 / 40**
Uma equipe de desenvolvimento de IA está preparando dados de treinamento para um novo modelo de aprendizado de máquina. O líder da equipe deseja garantir que a abordagem de obtenção de dados respeite a conformidade com os direitos autorais.

Qual é a **prática mais ética** de obtenção de dados?

**A)** Coletar dados apenas de domínios publicamente disponíveis
**B)** Criar todos os dados de treinamento dentro da organização
**C)** Baixar dados de fontes com permissão implícita
**D)** Usar apenas dados que pareçam ter sido amplamente compartilhados online

**A)** Incorreto. A disponibilidade pública, por si só, não garante que os dados estejam licenciados para o uso pretendido pela organização. Os termos da licença devem ser analisados para confirmar se são suficientes para a aplicação planejada.
**B)** Correto. A produção interna de dados é uma das estratégias éticas reconhecidas de obtenção de dados, pois garante que a organização tenha plenos direitos sobre os dados utilizados para treinar modelos de IA. (Literatura: A, Capítulo 0)
**C)** Incorreto. A permissão implícita não é uma base válida para o uso de dados. É necessário obter permissões explícitas do proprietário dos dados antes de utilizar dados de terceiros para o treinamento de IA.
**D)** Incorreto. Conteúdos amplamente compartilhados online ainda podem estar protegidos por direitos autorais. A repetição ou ampla circulação não torna o material livre para reutilização no treinamento de IA. O foco deve estar em verificar se os dados foram criados internamente, se estão devidamente licenciados ou se foram obtidos com as permissões necessárias.

# Avaliação

A tabela a seguir mostra as respostas corretas às questões apresentadas neste exame simulado.

| Questão | Resposta | Questão | Resposta |
|---|---|---|---|
| **1** | A | **21** | B |
| **2** | A | **22** | A |
| **3** | A | **23** | C |
| **4** | D | **24** | B |
| **5** | A | **25** | C |
| **6** | C | **26** | A |
| **7** | E | **27** | C |
| **8** | A | **28** | B |
| **9** | C | **29** | D |
| **10** | A | **30** | D |
| **11** | C | **31** | C |
| **12** | C | **32** | C |
| **13** | B | **33** | A |
| **14** | A | **34** | C |
| **15** | B | **35** | A |
| **16** | A | **36** | A |
| **17** | A | **37** | D |
| **18** | D | **38** | C |
| **19** | C | **39** | B |
| **20** | C | **40** | B |

Certified for what's next

## Contato EXIN

www.exin.com