# Falhas de Segurança das I.A's

Decidimos trocar o tema, achamos esse tema interessante pois hoje em dia várias I.A's contem falhas que permitem pessoas as usarem pro mal, tanto para golpes quanto para fake News e etc...

## Golpes

**Phishing automatizado**: Criminosos podem usar IA para gerar e-mails ou mensagens que imitam comunicações legítimas, com o objetivo de enganar pessoas a fornecerem informações sensíveis.

**Deepfakes**: Com IA, é possível criar vídeos ou áudios falsos muito realistas, o que pode ser usado para enganar pessoas e convencê-las de que uma figura pública ou familiar disse ou fez algo que, na verdade, não aconteceu.

**Chatbots fraudulentos**: I.A's podem ser programadas para enganar pessoas em plataformas de atendimento ao cliente, imitando humanos e coletando informações pessoais ou financeiras.

## Funções que estão sendo desenvolvidas para evitar golpes

**Ferramentas de Detecção de Deepfake**

**Detecção de vídeos e áudios falsos:** Empresas e pesquisadores estão criando algoritmos que analisam características sutis de vídeos e áudios (como movimentos oculares, distorções em expressões faciais e inconsistências na entonação de voz) para identificar deepfakes.

**Exemplo:** A Microsoft lançou o **Microsoft Video Authenticator**, uma ferramenta que analisa vídeos e determina a probabilidade de serem deepfakes.

**Sistemas de Autenticação Baseados em IA**

**Autenticação biométrica avançada:** A IA está sendo usada para fortalecer autenticações biométricas, como reconhecimento facial, impressão digital, voz e até padrões comportamentais. Isso dificulta a criação de perfis falsos ou o roubo de identidade.

**Exemplo:** O uso de **IA para reconhecimento de voz** está sendo aprimorado para distinguir falsificações de voz geradas por algoritmos de deepfake.

**Ferramentas de Monitoramento de Phishing**

**Detecção automática de phishing**: Plataformas de e-mail e navegadores estão adotando soluções baseadas em IA que analisam o conteúdo de mensagens em busca de padrões comuns em golpes de phishing, alertando os usuários ou bloqueando essas mensagens.

**Exemplo**: O **Google Safe Browsing** utiliza IA para analisar páginas web e bloquear automaticamente sites maliciosos.

**Soluções de Análise Comportamental**

**Análise de comportamento:** Ferramentas de IA que monitoram o comportamento do usuário em tempo real, verificando mudanças suspeitas que podem indicar uma fraude. Isso é útil para evitar roubos de contas e fraudes financeiras.

**Exemplo:** O uso de **Behavioral Biometrics** em bancos, onde o padrão de digitação ou a maneira como o usuário interage com o sistema é monitorado para detectar anomalias.

**Blockchain para Garantia de Autenticidade**

**Blockchain** está sendo explorado para registrar de forma imutável a autenticidade de arquivos, especialmente em meios de comunicação, garantindo que vídeos, fotos ou documentos não sejam adulterados.

**Exemplo:** A **Adobe, em parceria com o New York Times**, está desenvolvendo o **Content Authenticity Initiative (CAI),** um projeto que utiliza blockchain para garantir a integridade de conteúdo digital.

**IA para Detecção de Fraudes Financeiras**

**Monitoramento de transações financeiras:** Bancos e empresas de pagamentos utilizam algoritmos de IA que monitoram grandes volumes de transações em busca de padrões anômalos, identificando possíveis fraudes em tempo real.

**Exemplo:** A **Mastercard** usa IA para monitorar e analisar milhões de transações, detectando rapidamente atividades suspeitas com base em padrões comportamentais.

**Sistemas de Autenticidade e Verificação de Notícias**

**Fact-checking automatizado:** Ferramentas de IA estão sendo desenvolvidas para verificar a veracidade de notícias e informações divulgadas nas redes sociais, especialmente no combate à disseminação de desinformação.

**Exemplo:** Plataformas como **Factmata** utilizam IA para detectar desinformação e analisar o contexto de notícias e postagens em redes sociais.

**Treinamento de IA para Detecção de Anomalias em Redes**

**Monitoramento de redes corporativas:** Sistemas de IA são treinados para monitorar tráfego de rede e identificar comportamentos anômalos que podem ser indicativos de ataques de hackers ou de golpes digitais.

**Exemplo:** O uso de **SIEM (Security Information and Event Management)** com IA ajuda a correlacionar eventos e detectar atividades maliciosas em redes.

**Verificação de Identidade com IA**

**IA para validação de identidade:** Empresas estão desenvolvendo soluções para verificação de identidade que utilizam IA para cruzar dados biométricos e comportamentais, a fim de impedir fraudes de identidade.

**Exemplo:** O serviço de verificação de identidade da **Onfido** usa IA para verificar documentos e dados biométricos em tempo real.

Essas tecnologias estão sendo cada vez mais implementadas para proteger os usuários contra as ameaças trazidas pela manipulação maliciosa de IA.

## Falhas de segurança e problemas das I.A’s

Apesar de as Inteligências Artificiais (IAs) estarem cada vez mais avançadas e sendo aplicadas para segurança, elas também apresentam vulnerabilidades que podem ser exploradas, resultando em falhas de segurança. Aqui estão algumas das principais falhas e desafios de segurança que envolvem sistemas de IA:

**Ataques Adversariais**

**Definição:** Os ataques adversariais consistem em manipular os dados de entrada de uma IA de maneira que faça o modelo cometer erros. Pequenas alterações em imagens, textos ou outros dados de entrada, imperceptíveis para os humanos, podem levar uma IA a tomar decisões incorretas.

**Exemplo:** Alterar poucos pixels de uma imagem pode enganar um sistema de reconhecimento facial, permitindo o acesso de pessoas não autorizadas.

**Exploração de Dados de Treinamento**

**Poisoning Attack:** Nesse tipo de ataque, os dados de treinamento são intencionalmente corrompidos para influenciar o comportamento da IA. Se os dados de treinamento forem manipulados, a IA pode aprender padrões errôneos ou maliciosos.

**Exemplo:** Um atacante pode injetar dados falsos em grandes conjuntos de dados públicos, levando a IA a tomar decisões prejudiciais ou enviesadas.

**Falhas de Generalização**

**Definição:** IAs treinadas com conjuntos de dados limitados ou enviesados podem apresentar problemas de generalização, ou seja, elas não conseguem tomar boas decisões em cenários que diferem dos dados de treinamento.

**Exemplo:** Um sistema de reconhecimento de voz pode não funcionar adequadamente para sotaques que não estavam representados nos dados de treinamento, o que pode ser explorado por atacantes para enganar o sistema.

**Model Stealing (Roubo de Modelo)**

**Definição:** Esse tipo de ataque envolve a extração ou cópia do comportamento de um modelo de IA treinado. Ao realizar uma série de consultas, o atacante pode reconstruir um modelo semelhante ao original sem precisar acessar diretamente o código ou os dados.

**Exemplo:** Um invasor pode replicar um modelo de IA caro, como um sistema de recomendação, e usá-lo para seus próprios fins, violando direitos de propriedade intelectual.

**Privacidade e Exposição de Dados Sensíveis**

**Definição:** IAs treinadas em grandes volumes de dados podem inadvertidamente "memorizar" informações confidenciais. Isso significa que, ao fazer consultas ao modelo, é possível recuperar dados sensíveis presentes no conjunto de treinamento.

**Exemplo:** Em modelos de linguagem como o GPT, pode ser possível extrair números de cartões de crédito ou dados pessoais se eles estiverem presentes nos dados de treinamento.

**Vieses e Discriminação**

**Definição:** Se o conjunto de dados usado para treinar a IA for enviesado, a IA herdará esses vieses, podendo tomar decisões discriminatórias ou injustas. Isso é particularmente problemático em áreas como recrutamento, concessão de crédito ou reconhecimento facial.

**Exemplo:** Algoritmos de reconhecimento facial que têm maior taxa de erro para pessoas de pele mais escura, devido à sub-representação desse grupo nos dados de treinamento.

**Insegurança em Modelos de IA Auto-Supervisionados**

**Definição:** Modelos de IA que aprendem continuamente com os dados do ambiente podem ser vulneráveis à manipulação de entradas em tempo real. Isso cria oportunidades para os atacantes influenciar o aprendizado contínuo da IA com dados falsificados ou distorcidos.

**Exemplo:** Um sistema de IA usado para detectar fraudes pode ser treinado erroneamente por dados manipulados por um atacante, reduzindo a capacidade da IA de identificar fraudes verdadeiras no futuro.

**Ataques de Replay ou Injection**

**Definição:** Ataques de injeção envolvem a introdução de comandos maliciosos ou entradas inesperadas que não foram previstas pelo sistema de IA, levando o sistema a comportar-se de maneira inesperada.

**Exemplo:** Sistemas de assistentes virtuais (como Alexa ou Siri) podem ser vulneráveis a comandos de voz adversariais que façam o sistema realizar ações indesejadas ou mesmo comprometer a privacidade do usuário.

**Exploração de Falhas em Sistemas de Decisão Automatizada**

**Definição:** Sistemas de IA que tomam decisões automáticas sem a supervisão humana podem ser explorados se suas decisões forem previsíveis ou tendenciosas. Isso pode levar a ataques em que os resultados do sistema são manipulados de forma a favorecer um atacante.

**Exemplo:** Um sistema de análise de crédito pode ser explorado para aprovar fraudulentamente pedidos de crédito, se os critérios de aprovação puderem ser manipulados.

**Dependência Excessiva de IA**

**Definição:** A confiança excessiva em sistemas de IA pode ser explorada. Quando humanos deixam de monitorar ou revisar as decisões da IA, os atacantes podem explorar falhas ou brechas, sabendo que os sistemas não estão sendo auditados adequadamente.

**Exemplo:** No setor de saúde, se diagnósticos médicos forem automatizados sem validação humana, diagnósticos errados podem ser explorados por atores maliciosos.

**Ataques de Engenharia Social Potencializados por IA**

**Definição:** IAs podem ser usadas para realizar ataques de engenharia social em grande escala, como phishing, com alto grau de personalização. Isso pode enganar vítimas de forma mais eficiente e em maior número.

**Exemplo:** A IA pode gerar e-mails falsificados com uma linguagem altamente personalizada, tornando-os mais convincentes e difíceis de detectar como fraude.

## Como Suavizar Essas Falhas:

**Auditoria contínua:** Revisar constantemente os modelos de IA e seus dados de treinamento para identificar e corrigir vieses ou vulnerabilidades.

**Atualizações de segurança:** Implementar atualizações regulares nos sistemas de IA para proteger contra novos tipos de ataques.

**Intervenção humana:** Manter sempre um nível de supervisão humana sobre as decisões críticas tomadas pela IA, especialmente em sistemas de alto impacto.

**Criptografia e anonimização:** Utilizar técnicas de privacidade, como criptografia homomórfica e aprendizado federado, para proteger os dados durante o treinamento e a execução da IA.

A segurança de sistemas de IA é um campo em constante evolução, e suavizar essas vulnerabilidades é um desafio contínuo.