Privacy Red Team: Testando a Segurança dos Dados na Era da Inteligência Artificial

O Privacy Red Teaming é uma abordagem ofensiva para testar a privacidade dos sistemas.

Who am I?

Jenniffer da Costa, tenho 25 anos

Sou formada em Analise e Desenvolvimento de Sistemas e Sistemas e Pós-Graduada em Gestão e Liderança pela pela SPTech.
Trabalho com Segurança da Informação desde 2019, trabalhei 4 anos no mercado financeiro e atualmente atualmente trabalho na área da saúde.

Meus hobbies são ir para academia, ler e fazer crochê.

Agenda

1. O que é Privacy Red Team?
2. IA e Desafios de Privacidade
3. Requisitos de Privacidade para IA
4. Principais Normas e Leis de Privacidade
5. Metodologia de Testes de Segurança
6. Recomendações Práticas

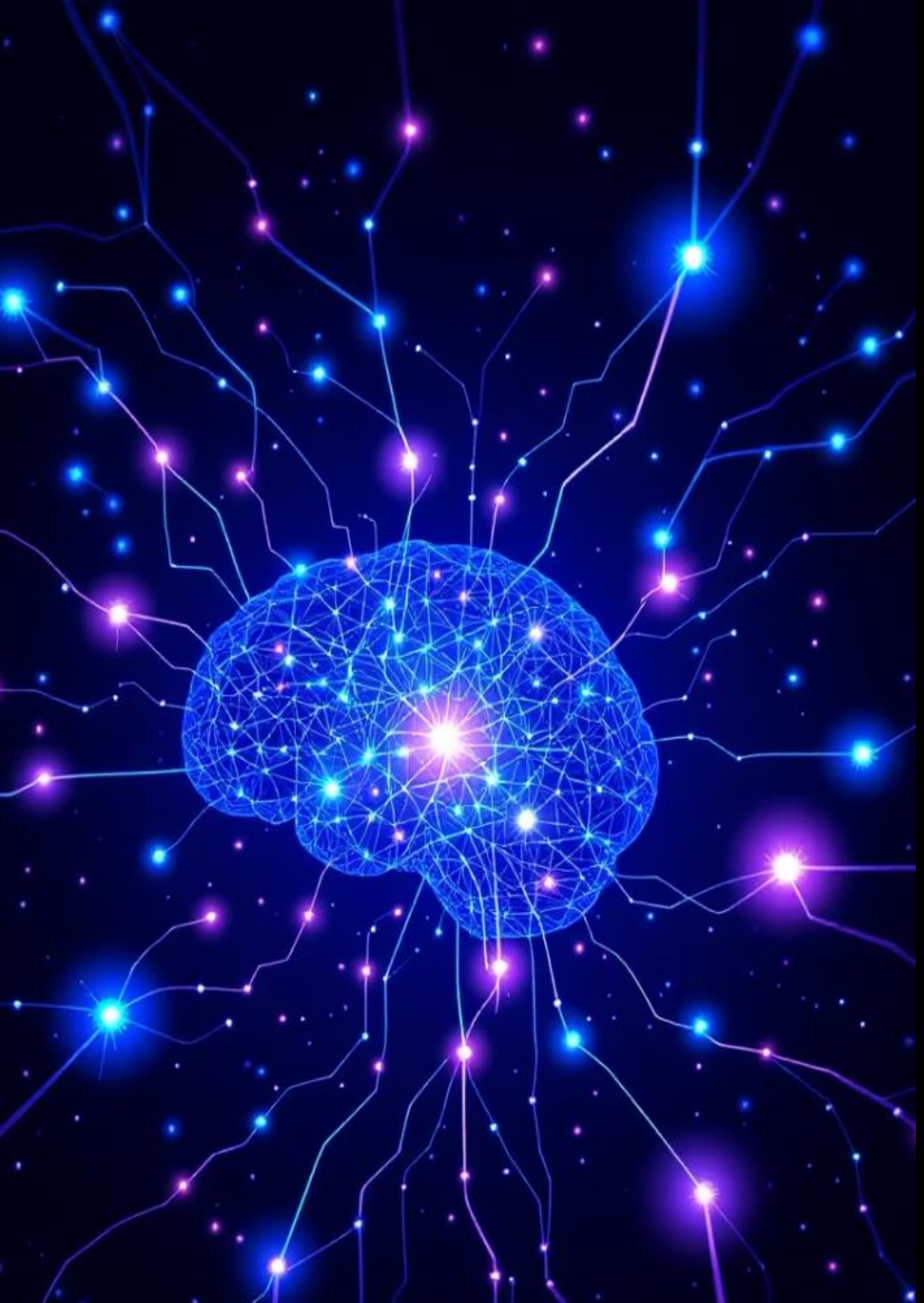
O que é o Privacy Red Team?

É uma abordagem proativa que simula ataques reais para identificar riscos à privacidade de dados em sistemas, produtos ou processos de uma organização.

Tem como objetivo encontrar falhas que possam expor dados pessoais ou sensíveis, antes que atacantes reais o façam.

Privacy Red Team vs Red Team

	RED TEAM	PRIVACY RED TEAM
OBJETIVO	Simular ataques reais para testar a resiliência da segurança	Simular abusos e usos indevidos para para testar a privacidade e proteção dos proteção dos dados
FOCO	Quebrar controles técnicos e processos de segurança	Expor riscos de privacidade e falhas na falhas na proteção de dados pessoais pessoais
EXEMPLO DE ATAQUE	Phishing, exploração de vulnerabilidades, movimentação lateral, etc.	Deanominização, inferência de atributos sensíveis, vazamento via metadados.
ABORDAGEM	Adversário externo ou interno com foco em impacto operacional	Adversário que busca reidentificar, correlacionar ou abusar de dados.
MÉTRICA DE SUCESSSO	Acesso não autorizado, comprometimento de sistemas	Violação de princípios de privacidade, como minimização, anonimato e controle do usuário.



IA e Desafios de Privacidade

Grandes Volumes de Dados

Modelos de IA dependem de enormes conjuntos de dados para aprendizado eficaz.

Proteção Inadequada Inadequada

Questões surgem quando dados sensíveis são usados sem proteções apropriadas.

Incidentes Comuns

Modelos memorizam informações pessoais. Sistemas usados sem consentimento.

Casos Reais

Brasil é o terceiro país mais atingido em vazamento de dados do ChatGPT

Por Felipe Demartini • Editado por Wallace Moté | 21/06/2023 às 17:19

Compartilhe:

Negócios | Segurança

ChatGPT confirma vazamento de dados de cartão de usuários; empresa afirma que bug foi corrigido

OpenAI, a empresa por trás do 'app-sensação' confirmou o vazamento de dados sensíveis; empresas entram em alerta e proíbem uso do site

Equipe InfoMoney
30/03/2023 13h52 • Atualizado 2 anos atrás

Requisitos de Privacidade para IA

Minimização

Capturar apenas dados necessários para a finalidade pretendida.

Não Discriminação

Evitar vieses injustos como descriminação, garantindo a equidade e respeito a todos os indivíduos.

Anonimização

Remover identificadores diretos para para impedir a reidentificação.

Ciclo de vida

Ciclo de vida dos dados, acompanhar desde a coleta até o descarte para garantir que o dado seja utilizado da melhor forma.

Principais Normas e Leis de Privacidade de Dados

GDPR (General Data Protection Regulation)

Estabelece Princípios como minimização, limitação e propósito, e direito dos titulares
Exige base legal para tratamento e comunicação transparente.

LGPD (Lei Geral de Proteção de Dados)

Define 10 bases legais para tratamento (como consentimento, legítimo interesse).

Cria a ANPD (Autoridade Nacional de Proteção de Dados) como órgão fiscalizador.

CCPA (California Consumer Privacy Act)

Garante aos consumidores o direito de saber, acessar, deletar e optar pelo não compartilhamento de dados.
de dados.
Obriga aviso prévio sobre coleta e venda de dados pessoais.

Principais Normas e Leis de Privacidade de Dados

ISSO/IEC 27001- (Gestão de Segurança da Informação)

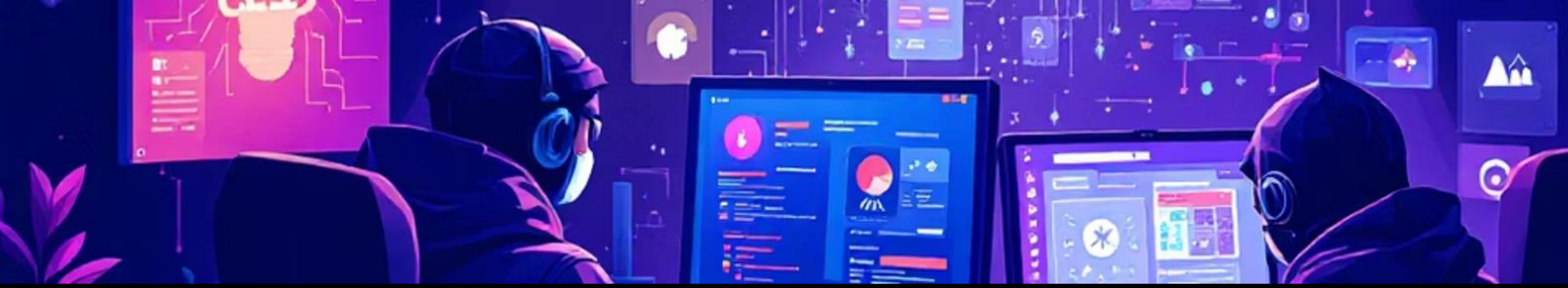
Padrão internacional para implementar um SGSI (Sistema de Gestão de Segurança da Informação).

Foco em confidencialidade, integridade e disponibilidade dos dados.

ISSO/IEC 27701 (Gestão de Privacidade da Informação)

Adiciona controles específicos para proteção de dados pessoais.

Suporta conformidade com leis como GDPR e LGPD.



Metodologia de Testes de Segurança

Mapeamento do Fluxo de Dados

Identificação de fontes e análise do ciclo de vida dos dados.

Ataques de Inversão de Modelo

Recuperação de dados de treinamento a partir da saída do modelo.

Inferência de Membros

Determinar se um dado específico foi usado para treinar o modelo.

Exfiltração por Prompts

Extrair informações privadas usando engenharia de prompts em LLMs.

Ferramentas e Estratégias

Ferramentas de Teste

- BurpSuite Pro - Extensão de IA
- Scripts para automatização de ataques

Differential Privacy

Introdução de ruído estatístico para impedir a reidentificação de dados.

Federated Learning

Modelos treinados em dispositivos locais sem compartilhamento direto de dados.

Monitoramento

Implementação de ferramentas de auditoria para identificar acessos não acessos não autorizados.

Recomendações Práticas

Obrigada!

/jenniffer-da-costa-patrocínio/