AI Red Team

O conceito de "AI Red Team" envolve a aplicação de práticas de red teaming, originalmente desenvolvidas em contextos militares e de segurança cibernética, aos sistemas de inteligência artificial (IA). Este campo emergente se concentra na simulação de adversários e na exploração de vulnerabilidades, visando identificar e mitigar riscos associados ao uso e implementação da IA. Com o crescimento exponencial e a integração da IA em diversos setores, as equipes de Red Team de IA tornam-se fundamentais para assegurar que os sistemas de IA sejam robustos, seguros e confiáveis, antecipando e prevenindo falhas, abusos e explorações mal-intencionadas.

Por exemplo, a equipe de Red Team de IA da Google, composta por hackers, simula uma variedade de adversários, desde estados-nação até grupos de ameaças persistentes avançadas (APT), hacktivistas, criminosos individuais e até mesmo insiders maliciosos. Essa equipe é essencial para traduzir o conceito de red teaming para inovações tecnológicas, incluindo a IA. O Red Team de IA da Google tem a responsabilidade de adaptar pesquisas relevantes para testar produtos reais e recursos que usam IA, identificando impactos em segurança, privacidade e abuso. Eles utilizam táticas, técnicas e procedimentos (TTPs) de atacantes para testar uma gama de defesas de sistemas. Exemplos desses TTPs incluem ataques de prompt, extração de dados de treinamento, backdoor no modelo, exemplos adversários, envenenamento de dados e exfiltração. As principais lições aprendidas incluem a complexidade dos ataques a sistemas de IA, a dificuldade em abordar as descobertas do red team, a importância do controle de segurança tradicional e a detecção de ataques a sistemas de IA semelhante a ataques tradicionais.

A OpenAI, por sua vez, lançou uma Rede de Red Team de IA para colaborar mais estreitamente com especialistas externos, a fim de tornar seus modelos mais seguros. A rede é uma comunidade de especialistas confiáveis e experientes que ajudam a informar a avaliação de risco e os esforços de mitigação da OpenAI. Os membros dessa rede contribuem com suas experiências e são compensados por sua participação. O objetivo é permitir uma entrada mais diversificada e contínua, tornando o red teaming um processo mais iterativo. Essa rede oferece uma oportunidade única de moldar o desenvolvimento de tecnologias e políticas de IA mais seguras, buscando especialistas com uma ampla variedade de experiências e perspectivas.

A Microsoft também possui uma equipe de Red Team de IA, dedicada a emular adversários do mundo real para identificar riscos e validar suposições, melhorando a postura de segurança geral dos sistemas. Eles compartilham as melhores práticas aprendidas para ajudar outras equipes de segurança a caçar proativamente falhas em sistemas de IA. A prática de red teaming de IA na Microsoft evoluiu para incluir não apenas a sondagem de vulnerabilidades de segurança, mas também outras falhas do sistema, como a geração de conteúdo potencialmente prejudicial. Eles colaboraram com o MITRE e parceiros da indústria e acadêmicos para desenvolver a Matriz de Ameaças de Aprendizado de Máquina Adversário e o Counterfit, uma ferramenta de automação para testar a segurança de sistemas de IA.

Já a equipe de Red Team de IA da NVIDIA é composta por profissionais de segurança ofensiva e cientistas de dados. Eles usam suas habilidades combinadas para avaliar seus sistemas de ML, identificando e ajudando a mitigar quaisquer riscos do ponto de vista da segurança da informação. A equipe usa um framework que permite abordar questões específicas em partes específicas do pipeline de ML, infraestrutura ou tecnologias. Isso inclui a expansão da evasão para incluir algoritmos ou TTPs específicos, tratamento de vulnerabilidades técnicas e cenários

de dano e abuso. Eles também consideram riscos técnicos, de reputação e de conformidade ao avaliar sistemas de ML

Em resumo, as equipes de Red Team de IA dessas organizações desempenham um papel crucial na identificação e mitigação de riscos em sistemas de IA, utilizando uma ampla gama de competências e métodos para garantir a segurança e confiabilidade desses sistemas.

Extra

Pratique suas habilidades em injeção de prompt com este desafio: https://gandalf.lakera.ai/cs50

Referências

- https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/
- https://openai.com/blog/red-teaming-network
- https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/
- https://developer.nvidia.com/blog/nvidia-ai-red-team-an-introduction/