TRIBUNAL DE CONTAS DA UNIÃO
Secretaria-Geral da Presidência - SEGEPRES
Secretaria de Relações Internacionais - Serint

RELATÓRIO DE VIAGEM

DADOS DO EVENTO

DATA DE INÍCIO	DATA DE TÉRMINO	NOME DO EVENTO	CIDADE/PAÍS
3 de agosto de 2024	8 de agosto de 2024	Black Hat USA 2024	Las Vegas / Estados Unidos

RESUMO DO EVENTO

ENTIDADE ORGANIZADORA	PROCESSO	PARTICIPANTES
Informa Tech	016.507/2024-0	Leandro Resende Gomes

JUSTIFICATIVA (RESUMO)

[Neste campo demonstre a relevância do trabalho desempenhado na cidade de realização do evento, para a obtenção da autorização de viagem. Procure ser objetivo (a), apresentando apenas o que for importante para análise do pleito. Lembre-se que este relatório será publicado.]

Com a massiva expansão das tecnologias de IA generativa, o evento Black Hat 2024 dedicou uma trilha de conhecimento específica para os trabalhos área de segurança de IA. Além de outras trilhas técnicas como: segurança de dispositivos móveis; segurança em endpoint; segurança de aplicações; segurança em nuvem; criptografia; forense digital e outros. Contudo, meu objetivo era concentrar nos conhecimentos técnicos na área de IA, participando de 12 palestras sobre o assunto e também do treinamento "IA Red Teaming in Practice", que foi um dos dois treinamentos focados em IA disponíveis no evento.

RELATO

[Descreva o evento de forma sucinta, destacando aquilo que posa ser útil a outros colegas, como transferência de conhecimento. Evite elogios e juízo de valor, tais como: "O evento foi muito proveitoso", "Os anfitriões são muito acolhedores", "O evento não foi bem organizado" etc.

Caso o evento deixe de trazer algo de novo, causando frustação em termos de expectativas, relate os pontos considerados altos em relação ao conhecimento transferido, ou faça um paralelo daquilo que foi apresentado e a situação do TCU]

O evento Black Hat USA é o maior evento de segurança da atualidade. As maiores empresas globais da área de segurança da informação marcam presença. O evento conta com uma estrutura que atende muito bem o lado do networking e também o lado técnico. Na parte de networking, é uma oportunidade de conhecer novas empresas e produtos e de fortalecer laços com empresas que já prestam serviços para a administração pública. Para citar algumas: estava o CEO da Qualys, os representantes comerciais da Microsoft Brasil, empresários do ramo de pentest da HackerOne dentre dezenas de outras empresas. A agenda do evento é bem cheia e tudo ocorre em paralelo. A área dos estandes (booths) é muito grande e para aproveitar bem o networking o participante deve dedicar bastante tempo nessas visitas e também nos eventos de confraternização promovidos pelas empresas. O meu foco foi todo voltado para a área técnica, então me concentrei no treinamento e nas palestras.

O treinamento “AI Red Teaming in Practice” foi promovido por dois funcionários da Microsoft que atuam exclusivamente em um time de ataque à IA (red team). A maior parte do treinamento são exercícios hands-on, utilizando um ambiente Google Colab já preparado pelos instrutores e também uma plataforma de exercícios em que o aluno tenta quebrar a IA e é pontuado conforme vai cumprindo com os desafios. Não há downtime para preparação de ambientes, o aluno chega com seu notebook, é dado uma credencial de acesso e em poucos minutos está apto a enfrentar os exercícios.

Os instrutores usam slides para passar os conceitos, mas logo em seguida o curso volta para aplicar os conceitos na prática. É um ramo de conhecimento muito especializado, em tentar burlar os controles de segurança da IA por meio de técnicas de prompt injection e jailbreaking. Alguns dos desafios exigiam o uso de uma ferramenta avançada de testes de IA, chamada PyRit. O projeto python dessa ferramenta é de código aberto e é mantido principalmente por este time da Microsoft.

Alguns trabalhos acadêmicos são mostrados e referenciados durante o curso. Questionei a respeito de livros sobre o tema e indicaram uma literatura base sobre segurança de machine learning, mas não algo focado em IA generativa ainda. Os próprios instrutores têm planos de escrever um livro sobre o assunto.

Os outros dois dias de palestras eu continuei o foco na trilha de conhecimento IA. Participei de 12 palestras ao todo, quais sejam:

- (Abertura) Keynote: Democracy's Biggest Year: The Fight for Secure Elections Around the World
- Practical LLM Security: Takeaways From a Year in the Trenches
- 15 Ways to Break Your Copilot
- Predict, Prioritize, Patch: How Microsoft Harnesses LLMs for Security Response
- Deep Backdoors in Deep Reinforcement Learning Agents
- Al Safety and You: Perspectives on Evolving Risks and Impacts
- Isolation or Hallucination? Hacking Al Infrastructure Providers for Fun and Weights
- From MLOps to MLOops - Exposing the Attack Surface of Machine Learning Platforms
- Uncovering Supply Chain Attack with Code Genome Framework
- Bypassing ARM's Memory Tagging Extension with a Side-Channel Attack
- Threat Hunting with LLM: From Discovering APT SAAIWC to Tracking APTs with AI
- Ignore Your Generative Al Safety Instructions. Violate the CFAA

Destaque para a palestra “Isolation or Hallucination? Hacking Al Infrastructure Providers for Fun and Weights”, promovida por dois técnicos de Israel. Eles mostraram como conseguiram invadir três grandes provedores de plataformas de IA incluindo Hugging Faces. Ao fazer com que códigos python fossem executas pelos servidores, conseguiram fazer movimentação lateral e acessar modelos privados e executar códigos remotos, tomando o domínio do ambiente. Os palestrantes entraram em contato com as empresas e ajudaram a corrigir as falhas, que não eram mais passíveis de exploração no momento da apresentação.

ENCAMINHAMENTOS POSSÍVEIS, NO ÂMBITO DO TCU, DECORRENTES DESTA AÇÃO

[Baseado em sua experiência e as novas informações/os novos conteúdos assimilados, proponha pontos de melhoria para o Tribunal atingir a sua missão precípua ou para sua Unidade, caso a ação seja específica para o seu trabalho.]

Tive a oportunidade de conhecer o funcionamento de um red team (time de ataque) focado em AI generativa. Na Microsoft e em outras grandes empresas existe um time dedicado a essa tarefa, testando e melhorando os modelos. O TCU está posicionado como um Órgão referência na adoção e uso de IA generativa, segundo apontamento da OCDE feito em abril de 2024. Contudo, institucionalizar um time próprio para segurança de IA talvez não seja viável frente aos vários desafios que as equipes técnicas possuem: migração de workloads para nuvem, evolução da IA, segurança do ambiente computacional.

Por outro lado, as técnicas utilizadas no treinamento são muito novas, trata-se de uma ciência aprofundada e não tão comum às técnicas de ataques convencionais, pois o resultado gerado pela IA não é tão determinístico como um

ataque focado em explorar um sistema com entradas e saídas esperadas. Ou seja, existe um esforço significativo para testar a segurança da IA, seja na construção do ambiente de testes (no curso foi mostrado o PyRit como framework de ataque), seja no aprendizado e refinamento das novas técnicas de ataque.

O que eu considero viável a curto prazo para o TCU é realizar uma disseminação do tema através de palestras técnicas. Amadurecer o tema e o conhecimento junto ao núcleo de IA e vislumbrar possíveis trabalhos nessa área de segurança de IA. Para tanto, irei estudar o tema de forma mais aprofundada e preparar uma apresentação. Estou cursando a pós-graduação do ISC e com essa pesquisa acadêmica que irei desempenhar terei condições de avaliar a pertinência de propor esse tema como tese de conclusão do curso.

O material das palestras e dos treinamentos serão baixados assim que disponibilizados pela organização do evento. Pretendo armazenar e compartilhar todo material em um ambiente interno.

A respeito das palestras, elas reforçaram que o tema de segurança de IA é algo muito relevante. Além de apresentações técnicas bem aprofundadas, como a já mencionada apresentação de Israel, o setor já discute implicações no âmbito legal: o que pode ou não ser considerado como uma infração legal quando se burla uma IA generativa.

A respeito do networking, como minha missão envolvia principalmente obter conhecimento técnico e tudo acontecia de forma simultânea, fiz alguns contatos com o pessoal da Microsoft, HackerOne e Clavis. Mas a oportunidade poderia ter sido melhor aproveitada caso a necessidade maior fosse ampliar a rede de contatos e parcerias.