Generative AI: Challenges and Opportunities for Security and Privacy

Controls to Mitigate Security and Privacy Concerns around Generative AI

Geneva, Switzerland, 19th February 2024 11:30- 13:00

CSIR

Touching lives through innovation

CSIR

Touching lives through innovation

GENERATIVE ARTIFICIAL INTELLIGENCE TO CYBERSECURITY & PRIVACY

Dr MT Dlamini (PhD, CISSP, CCSP)

Senior Cybersecurity Researcher

Information and Cyber Security Center

Defence & Security Cluster

CSIR

Adversarial Gen AI

AI-driven cyber-attacks
- QR Code phishing – Phishing on another level

Well-crafted Spam
- Spelling mistakes in the ruleset – goes out of the window

Adversarial AI/ML
- Malware that circumvents defences at whim
- Poisoned training data – Input/prompt manipulation
- Compromised AI algorithms – goose chase/AI hallucinations
- Model inversions -- threat actors manipulate model output to infer sensitive training data - Privacy risks
- Exploratory attacks – model behaviour, vulnerabilities and sensitive data theft
- Model drift and decay –
- Model DDoS
- Compromised Gen AI 3rd party libraries/cloud services

Adversarial Gen AI

Phishing emails so real

- The difference between legit and illegitimate - blurred

Social Engineering so real

- The difference between legit and illegitimate – generative AI for voice and video

Data integrity compromises and leaking Gen AI model

- Sensitive data stealing LLMs (bogus Gen AI models)

AI, ML & Gen AI Models Under Attack

- Polymorphic and metamorphic malware on another level
- Perfectly perfect deepfakes - Lack of good integrity checkers
- Fake news on another level (misinformation & disinformation)
- Compromised AI algorithms/models – Gen AI hallucination
  - Assuring complete confidence in false information
  - False becomes relative
  - Fact or truth dead & buried
- Perfect AI algorithms/Models trained on poisoned data
  - The impact same as above

"Organizations must put real quality thinking into how to maximize the benefits of new technology and to manage the risks" **Jean Innes**, CEO of the Alan Turing Institute, UK

Cyber Defense, Detection & Prevention Controls

Fight AI-driven cyber-attacks with AI-driven defences

- Improve malware detection in encrypted traffic
- Address difficult session hijacks
- Improve continuous authentication
- Help address encrypted malware
- Improve threat detection

Automate repetitive CS tasks to improve productivity

- SOAR – Facelifted SIEM
- Address the huge skill shortage

Improve detection of APTs & other sophisticated threats

- Reduce the extended dwell time
- Identify and kick out cyber criminals lotl

Improve intrusion detection and prevention

Cyber Detection & Prevention Controls

Gen AI for Anomaly/Threat Detection

- Unusual user/process behaviour

Cybersecurity Behavioral Analytics

- Build a corpus of threat data

Secure Model Development

- Secure containers, VMs and virtual instance, secure models and data training

Real-time Monitoring & Privacy-preserving AI

- AI-powered cybersecurity solutions

Synthetic data – Privacy concerns

Cyber Incident Response & Recovery Controls

Facilitate Digital Forensic Investigations

- Sift out the needle of digital evidence in the stack of logs
- Reduce the time it takes to resolve court cases related to cybercrimes – DF readiness
- Improve the penalties associated with cybercrimes – no more slaps on the hand for serious crimes
- Improve cyber attribution for nation state threat actors

Optimize extended detection and response XDR

Optimize the SOCs

- Cyber security analysts can identify and respond to threats faster and more efficiently.

Expands the role of CS to AI

- New markets, exciting and terrifying future

CSIR
Touching lives through innovation

GENERATIVE ARTIFICIAL INTELLIGENCE TO CYBERSECURITY & PRIVACY

Thank You

TDlamini1@csir.co.za