Artificial Intelligence Risks & AI Risk Mitigation

Policy & Oversight

Overall risks include public records compliance, ethics, hallucinations, data privacy & security and lack of accountability. Here are MITIGATIONS to those risks:

1. **Develop Clear AI Governance Policies and Guidelines.** Develop comprehensive policies, guidelines, and ethical principles to govern the development, deployment, and monitoring/review of AI systems.
2. **Auditing.** Regularly audit results to ensure outcomes match expectations, law, and local needs. The tools provide different answers, check those answers against each other.
3. **Provide AI Training & Education.** Invest in AI literacy and training for all staff and elected officials.
4. **Promote Transparency and Public Trust.** Openly communicate how and when AI is being utilized in government operations, addressing potential concerns, and fostering trust through transparency and accountability.
5. **Continuously Evaluate and Adapt.** AI is advancing quickly and will continue to create new challenges. Consider quarterly, bi-annual, or at least annual AI review, and include elected representatives.
6. **Double Signatures.** Implement secure access controls, requiring two authorized individuals to approve financial transactions through an audited double signature process.

AI Govt Cyber Risk

AI technology significantly increases cyber risks. Malicious AI cyber-attacks have automated tasks like reconnaissance, exploit development, and lateral movement, making attacks faster and more efficient. AI can be used for social engineering attacks like personalized phishing emails, deepfakes, and realistic voice calls. AI makes all these attacks harder to detect and defend against. There are tools that auto write malware, search dark web data, and pull up relevant data.

Malicious AI Tools

- **DarkBERT:** A GPT-based malware known as 'DarkBERT' has been developed, which uses the entirety of the Dark Web as its knowledge base. This adversarial tool is capable of analyzing new pieces of Dark Web content, written in its own dialects and heavily coded messages, and extracting useful information from it.
- **WormGPT**: Another AI tool that emerged on the dark web. Unlike ChatGPT, which has built-in protections against misuse, WormGPT is designed with "no ethical boundaries or limitations," making it a potent tool for hackers.
- **FraudGPT:** The developer behind the FraudGPT malicious chatbot is readying even more sophisticated adversarial tools based on generative AI and Google's Bard (Gemini) technology.

Other Illegal Activities: Nearly 3,000 dark web posts in 2023 discussed illegal activities involving ChatGPT and other large language models (LLMs). These include creating malicious versions, jailbreaking techniques, lists of harmful prompts, and discussions on stolen accounts.

Neverman
Consulting
Trustworthy Solutions for Complex Problems

august@nevermanconsulting.com 414-380-9701

Deepfake Examples

- **Binance**: The Chief Communications Officer of Binance, a blockchain ecosystem, was deepfaked. The fraudsters held 20-minute “investment” Zoom calls, trying to convince the company’s clients to turn over their Bitcoin for scammy investments. The clients were sent links to faked LinkedIn and Telegram profiles claiming to be the officer, inviting them to various meetings to talk about different listing opportunities. The criminals then used a convincing-looking holograph of the officer in Zoom calls to try and scam several representatives of legitimate cryptocurrency projects.
- **Hong Kong**: A multinational company’s Hong Kong office suffered a significant financial loss of HK$200 million (US$25.6 million) due to a sophisticated scam involving deepfake technology. The scam featured a digitally recreated version of the company’s chief financial officer, along with other employees, who appeared in a video conference call instructing an employee to transfer funds. The scammers were able to convincingly replicate the appearances and voices of targeted individuals using publicly available video and audio footage.
- **CEO Impersonation (CEO Fraud)**: A scammer used AI-powered voice technology to impersonate a German CEO. The UK CEO, believing they were interacting with their German counterpart, followed instructions that led to financial loss.
- **Other CEO Fraud:** In one case, a US-based business lost nearly $400k when the payments team received an email from the CEO asking for payments to be set up for new beneficiaries. In another case, a global commodity trading platform provider lost £920,000 ($984k) when an employee received an email from the CEO requesting a new payment. $35m was lost via deepfake audio in United Arab Emirates.
- **Other:** Deepfake scams can be used in elections, personal and corporate scams. Deepfakes are being sold as a service on the DarkWeb.

AI Cyber Security Tools

DarkTrace: AI-powered cybersecurity platform that utilizes machine learning and AI algorithms to detect and respond to cyber threats in real time across diverse digital environments. Its key features include self-learning, autonomous response, threat visualization, and machine learning insights.

CrowdStrike Facon: AI-powered endpoint protection platform that uses machine learning to detect and prevent malware, ransomware, and other threats.

Vectra AI: AI-driven threat detection and response platform that uses machine learning to identify attacker behaviors and tactics across cloud, data center, and enterprise environments. It provides automated threat hunting and triage, as well as integrations with security orchestration and automation tools.

Symantec Endpoint Protection: AI-powered endpoint security solution that uses machine learning and behavioral analysis to detect and block advanced threats, including fileless attacks and zero-day exploits.

IBM Watson for Cyber Security: AI-powered security analytics platform that uses natural language processing and machine learning to analyze structured and unstructured data from various sources.

Palo Alto Networks Cortex XDR: AI-powered extended detection and response platform that uses machine learning to detect and respond to advanced threats across endpoints, networks, and cloud environments.

Check Point Infinity: AI-powered security platform that offers proactive threat prevention, network security, cloud security, and endpoint protection.

Fortinet FortiAI: AI-powered security solution that uses machine learning and advanced analytics to detect and respond to cyber threats in real time.