Cyber Risks in Generative Artificial Intellgience Large Language Models

Understanding the Cyber Risk Landscape

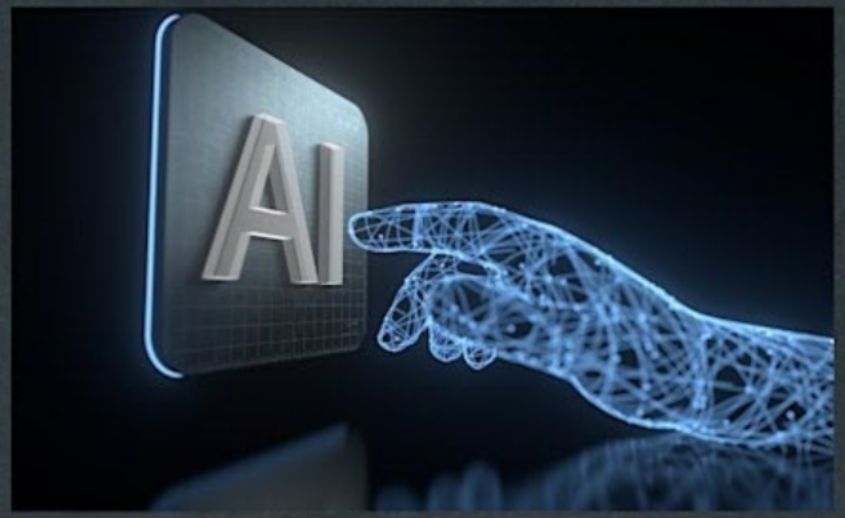
Agenda

- Understanding Generative AI
- Emerging Threat Landscape
- Examples of misuse of Generative AI
- Generic Large Language Model
- SANS AI Cyber Summit 2024 Key Takeaways
- Cyber Risk Frameworks for Generative AI
- Questions!

Emerging Threat Landscape

Cyber Risks in Gen AI Large Language Models

Understand potential risks and vulnerabilities



OT CYBERSECURITY

Prashant Prashant
Senior Cybersecurity Advisor
Enbridge Pipelines

Agenda

- Understanding Generative AI
- Emerging Threat Landscape
- Examples of misuse of Generative AI
- Generic Large Language Model
- SANS AI Cyber Summit 2024 Key Takeaways
- Cyber Risk Frameworks for Generative AI
- Questions!

Understanding Generative AI

What is Generative AI or Gen AI?

- Gen AI is a branch of Artificial Intelligence (AI) that can create *original content in response to a user's prompt or request*

- The content can include *text, images, video, audio or even software code*. Essentially, generative AI learns patterns from vast data sets and autonomously produces new content based on those patterns.

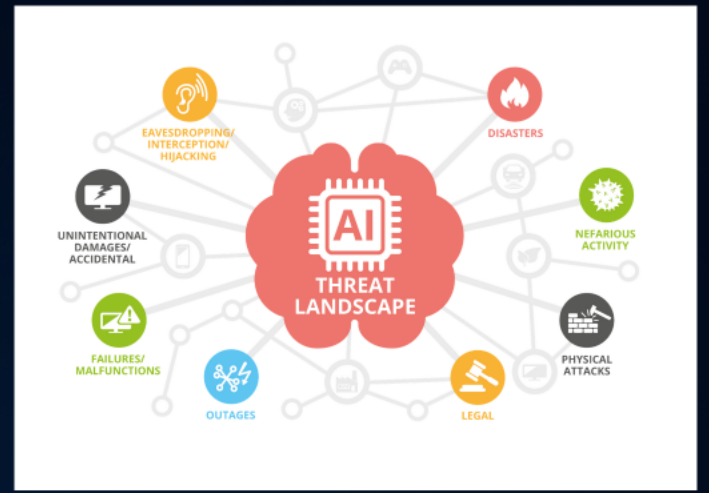
Examples and Applications:

- Language models- GPT4 (Generative Pre-trained Transformer 4 developed by Open AI), GPT-4 is a poweful LLM that can generate human-like text, images, videos based on input and prompts.

- Text Generation- Chatbots. Many chatbots use generative AI to provide helpful responses.

- Code Generation- AI can write code snippets, complete functions and even entire programs(e.g. reverse shell app)

Emerging Threat Landscape



Gen AI is attracting attention from malicious actors. Threat actors leverage AI-generated content for *disinformation, misinformation and mal-information* campaigns

- Gen AI used for intelligent reconnaissance , including scanning content of knowledge bases for leaked information.
- In the first half of 2023, Gen AI tools like *WormGPT* and *FraudGPT* began appearing in scams and cybercriminal activities, empowering adversaries to carry out more sophisticated attacks.
- Convincing fake content raises deception concerns. This can be used to shift opinion about an organization or a person.
- Copyrights and intellectual property when using someone's image or voice to create art-related content.

The Rapid Rise of AI

- Key risks when engaging with AI:
  - Data poisoning
  - Input manipulation
  - Generative AI hallucinations
  - Privacy and intellectual property concerns
  - Model stealing

- "AI systems are **highly valuable targets** for malicious cyber actors. State actors may seek **information** to advance their interests." (*Deploying AI Systems Securely*)

Examples of misuse of Gen AI

- Video of Connor McDavid bashing city of Vancouver during playoffs

- Video of Volodymyr Zelensky surrendering to Russia

- Arup(Hong Kong) $25 million deepfake video meeting

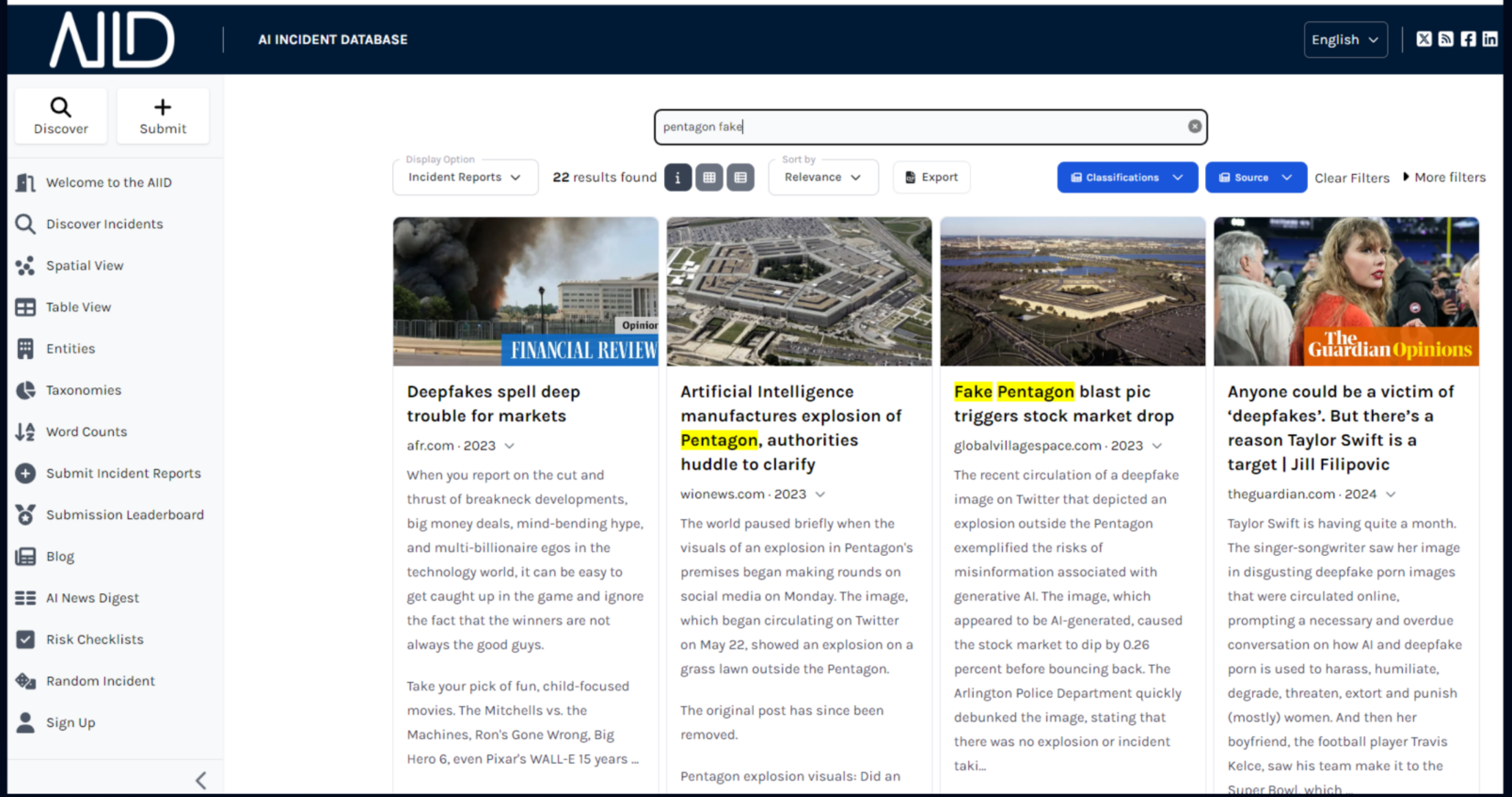
- AI-generated Drake & Weekend song

- Traylor Swift deepfake pornography lawsuit

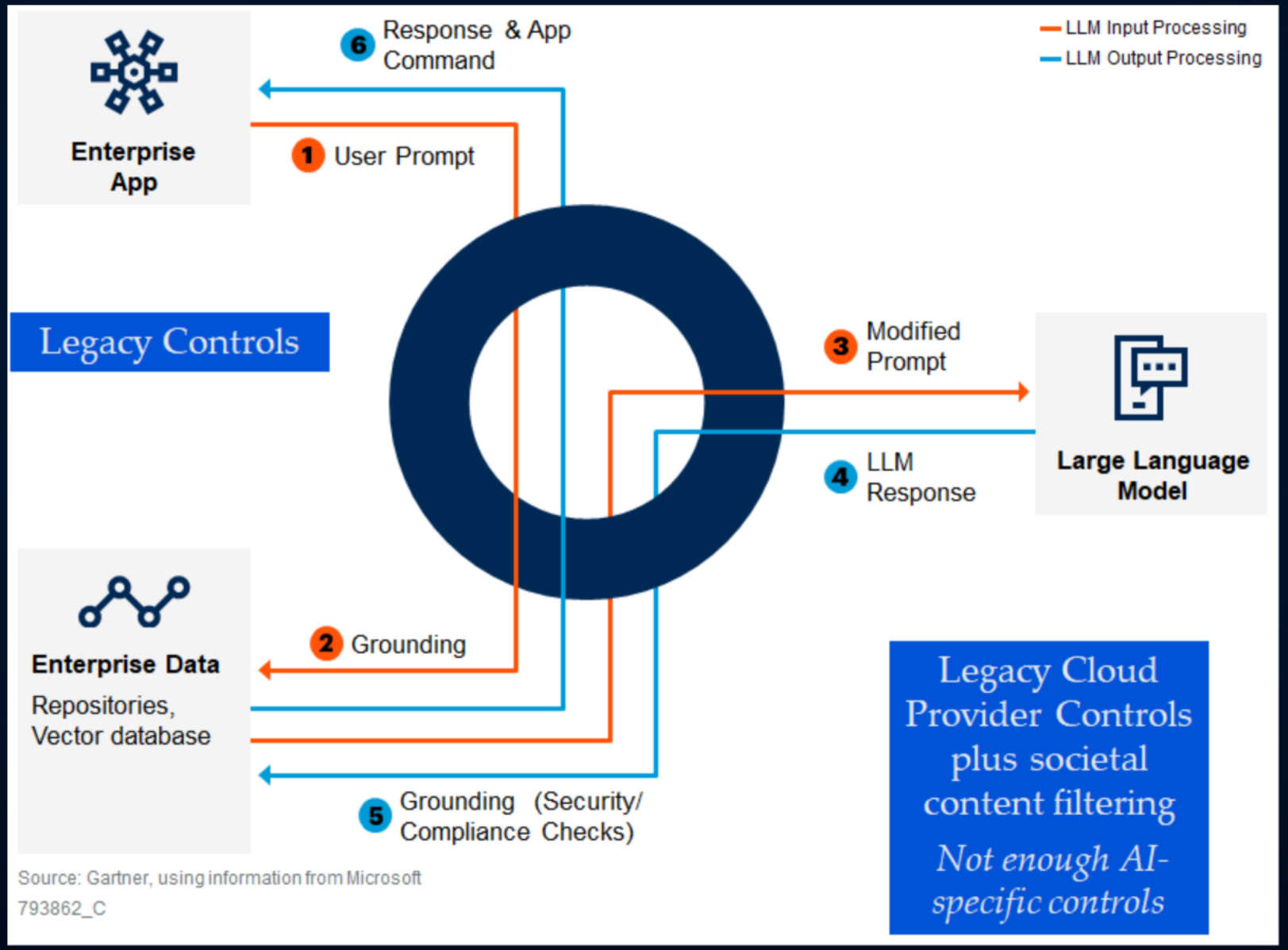
- ChatGPT 3: writing exploit code of reverse shell

- Deep Fake Pentagon images of explosion resulted in a brief Stock Market Meltdown in 2023

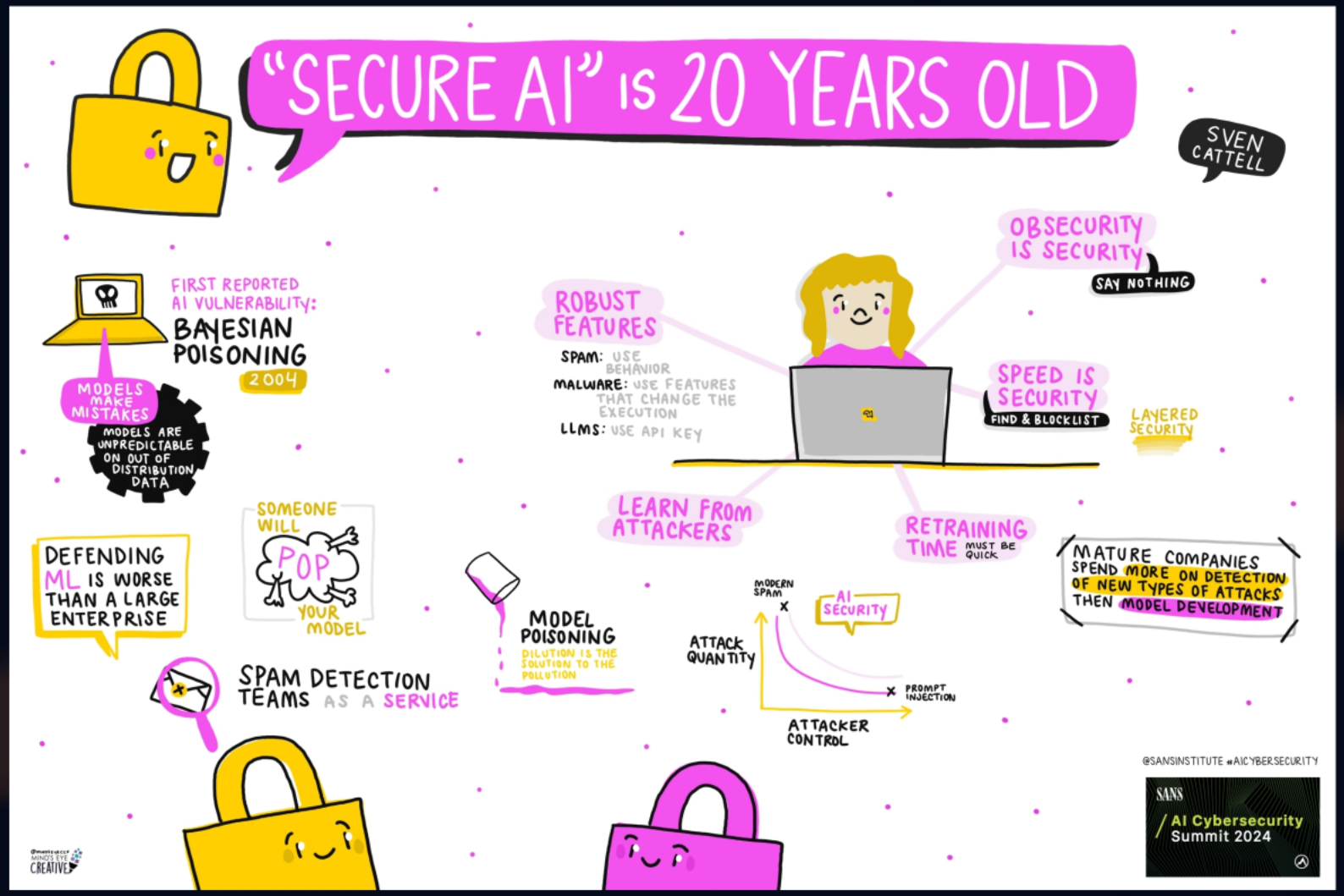
Aritificial Intelligence Incident Database
https://incidentdatabase.ai/



Generic Large Language Model-Legacy Controls are not enough



SANS AI Cyber Summit 2024 Key Takeaways*



* Reproduced by permission from SANS Institute

MAKING SECURE AI REAL

REAL THREATS, LESSONS LEARNED & FUTURE OF THE SECURE AI TECHNOLOGY STACK

KATIE BOSWELL

KRISTY HORNLAND

KPMG TRUSTED AI FRAMEWORK

TRUSTWORTHY
ETHICAL
RESPONSIBLE

INPUT

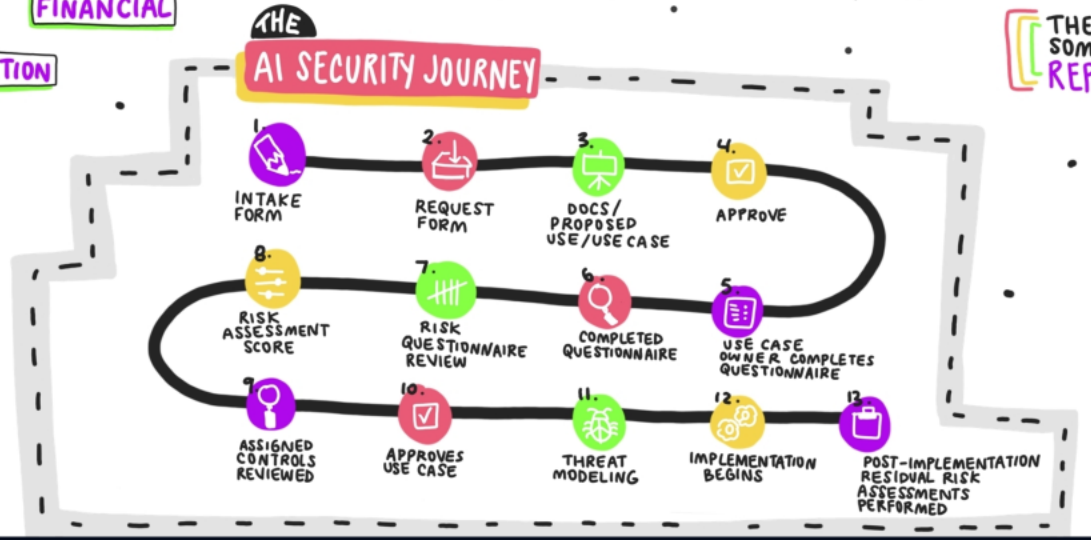
SYNTHESIZED AI SECURITY STRATEGY

ENABLEMENT

- RISK INDICATORS
- AI ECOSYSTEM
- REGULATIONS
- PRIVATE-PUBLIC SECTOR INITIATIVES

- AWARENESS
- AI SECURITY FRAMEWORK
- AI SECURITY PIPELINE MANAGEMENT
- AI SECURITY UPLIFT

THERE SHOULD BE SOME LEVEL OF REPEATABILITY



@SANSINSTITUTE #AICYBERSECURITY

SANS / AI Cybersecurity Summit 2024

@mindseyecrp MIND'S EYE CREATIVE

THE FRONTIER OF CYBERSECURITY

DEFENDING AGAINST AI-BASED THREATS

COREY WHITE

SMALL BUSINESSES

NO TOOLS OR THE RIGHT TOOLS OR HAVE THEM CONFIGURED PROPERLY

NO SECURITY PERSON

THE RISE OF AI-BASED HACKS

SECURE OF FAMILIES

AI-DRIVEN SOCIAL ENGINEERING & CYBER OPS

NATION-STATE

OUTPACING SECURITY

FRAUD GPT

- SUBSCRIPTION MODEL
- BYPASS DETECTION
- AUTOMATED SOCIAL ENGINEERING/ PLANNING
- FAKE PERSONA

DEEPFAKE ATTACKS

DECREASED COST OF AI-DRIVEN DEEPFAKE

PHISHING EMAILS

BUSINESS EMAIL COMPROMISE

ADVANCED

DATA BREACH ANALYSIS

DOCUMENT FRAUD TOOLS

AUTONOMOUS MEMORY ANALYSIS

- HARDER TO DETECT
- BYPASS TRADITIONAL SECURITY MEASURES

COUNTERMEASURES:

ROBUST ACCESS CONTROLS

AI-BASED ENDPOINT PROTECTION

BLOCK PHISHING EMAILS & FAKE WEBSITES

CONTINUOUS SCANNING & PATCHING

SECURITY AWARENESS

@SANSINSTITUTE #AICYBERSECURITY

SANS / AI Cybersecurity Summit 2024

MIND'S EYE CREATIVE

GENAI FOR DFIR IN THE REAL WORLD

PRACTICAL USE CASES

JESS GARCIA

DS4N6 PROJECT

DATA SCIENCE & ARTIFICAL INTELLIGENCE TO THE AVERAGE FORENSICATOR

WHAT CAN GEN AI DO... TO HELP ME

WORKFLOWS

THREAT HUNTING

FORENSICS

IN THE REAL WORLD

THREAT DETECTION & RESPONSE

USE CASE: SUMMARIZE LOW PRIORITY ALERTS & ENRICH THEM WITH CTI

CUSTOM GPTS + RAG

USE CASE: DATA ANALYSIS WITH CHAT GPT

CHAT GPT GENERATES CODE NEEDED TO DO ANALYSIS

ADD DETAILS TO QUERIES

ASK MORE QUESTIONS:

- WAS THERE A BRUTE FORCE ATTACK?
- WRITE REPORT WITH CONCLUSION & RECOMMENDATIONS

AUGMENTED RAG GENERATION RETRIEVAL

FORENSIC KNOWLEDGE

REPORT → CHUNKING → TOKEN-IZATION → SIMILARITY SEARCH → RETRIEVED CONTEXT

Q → RETRIEVED CONTEXT

@SANSINSTITUTE #AICYBERSECURITY

SANS AI Cybersecurity Summit 2024

@mindseyecrp MIND'S EYE CREATIVE

AI'S ACHILLES' HEEL

NAVIGATING THE OWASP TOP 10 FOR LLMS

KYRIAKOS LAMBROS

OF CEOs AGREE AIs BENEFITS OUTWEIGH THE RISKS

AND THE BUSINESS ALWAYS WINS

THE TROJAN HORSE

ATTACKERS MANIPULATE INPUTS

LEAD TO SIGNIFICANT DOWNSTREAM EFFECTS

PROMPT INJECTION

PREVENTION

IMPLEMENT STRICT INPUT VALIDATION & SANITATION

USE CONTEXT-AWARE FILTERING & OUTPUT ENCODING

THE POISONED ARROWS

ATTACKERS INJECT MALICIOUS DATA INTO TRAINING SET (PRIOR TO DEPLOYMENT)

TRAINING DATA POISONING

PREVENTION

MODEL AUDITS

VERIFY & SANITIZE TRAINING DATA SOURCES RIGOROUSLY

IMPLEMENT ROBUST MONITORING DURING TRAINING PROCESS

INSECURE OUTPUT HANDLING

MODEL DENIAL OF SERVICE

OWASP TOP 10 FOR LLMS

MODEL THEFT

SUPPLY CHAIN VULNERABILITIES

OVERRELIANCE

INSECURE PLUGIN DESIGN

SENSITIVE INFORMATION DISCLOSURE

EXCESSIVE AGENCY

UPDATING YOUR INCIDENT RESPONSE PLAN FOR WHEN AI GOES WRONG

DIFFERENTIAL PRIVACY

PREVENTION

IMPLEMENT STRICT OUTPUT FILTERING & VALIDATION MECHANISMS

AUDIT & SANITIZE TRAINING DATA FOR SENSITIVE CONTENT

THE WHISPERING WINDS

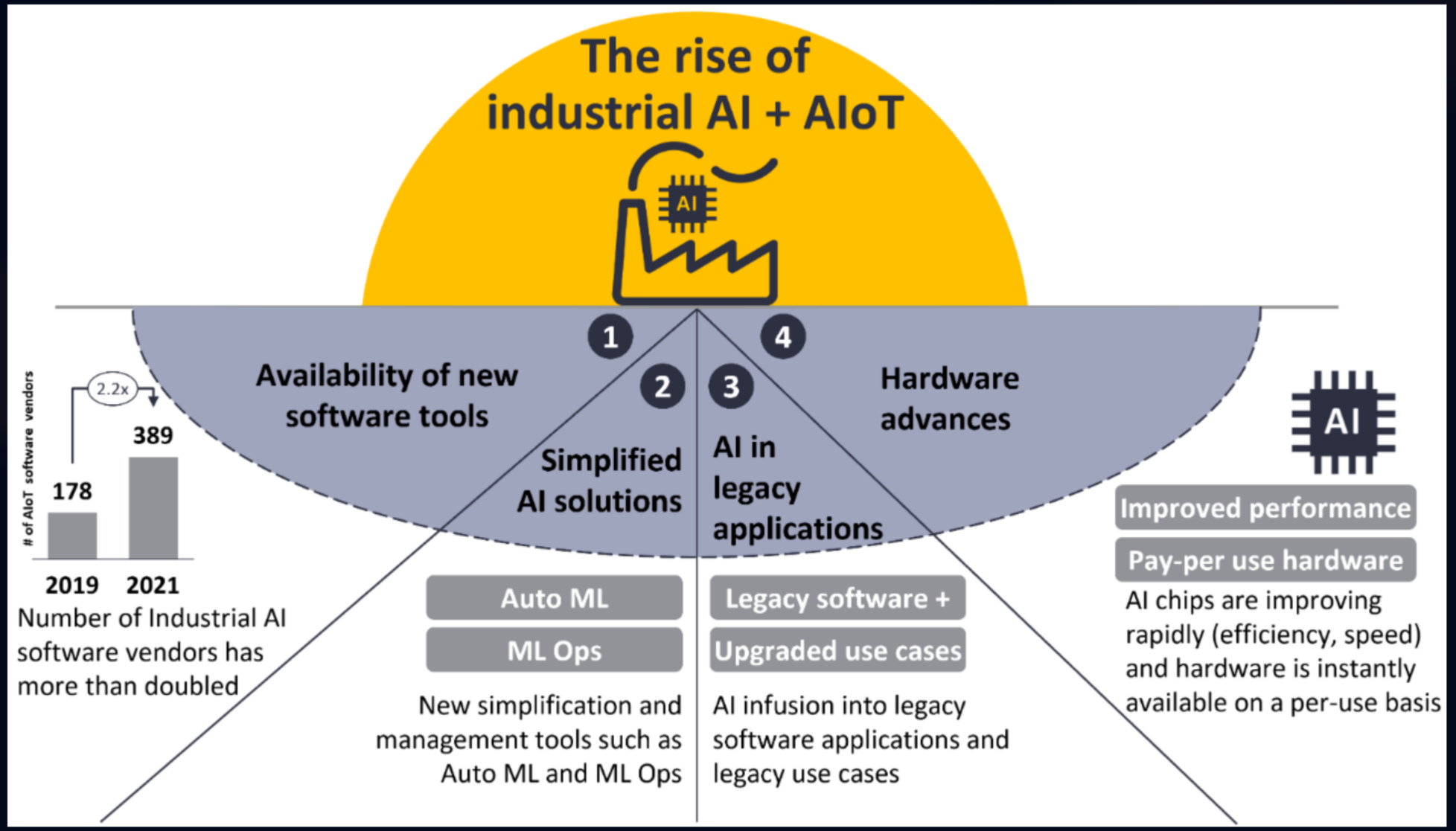
LLMs MAY INADVERTENTLY REVEAL SENSITIVE OR PROPRIETARY INFORMATION

@SANSINSTITUTE #AICYBERSECURITY

SANS / AI Cybersecurity Summit 2024

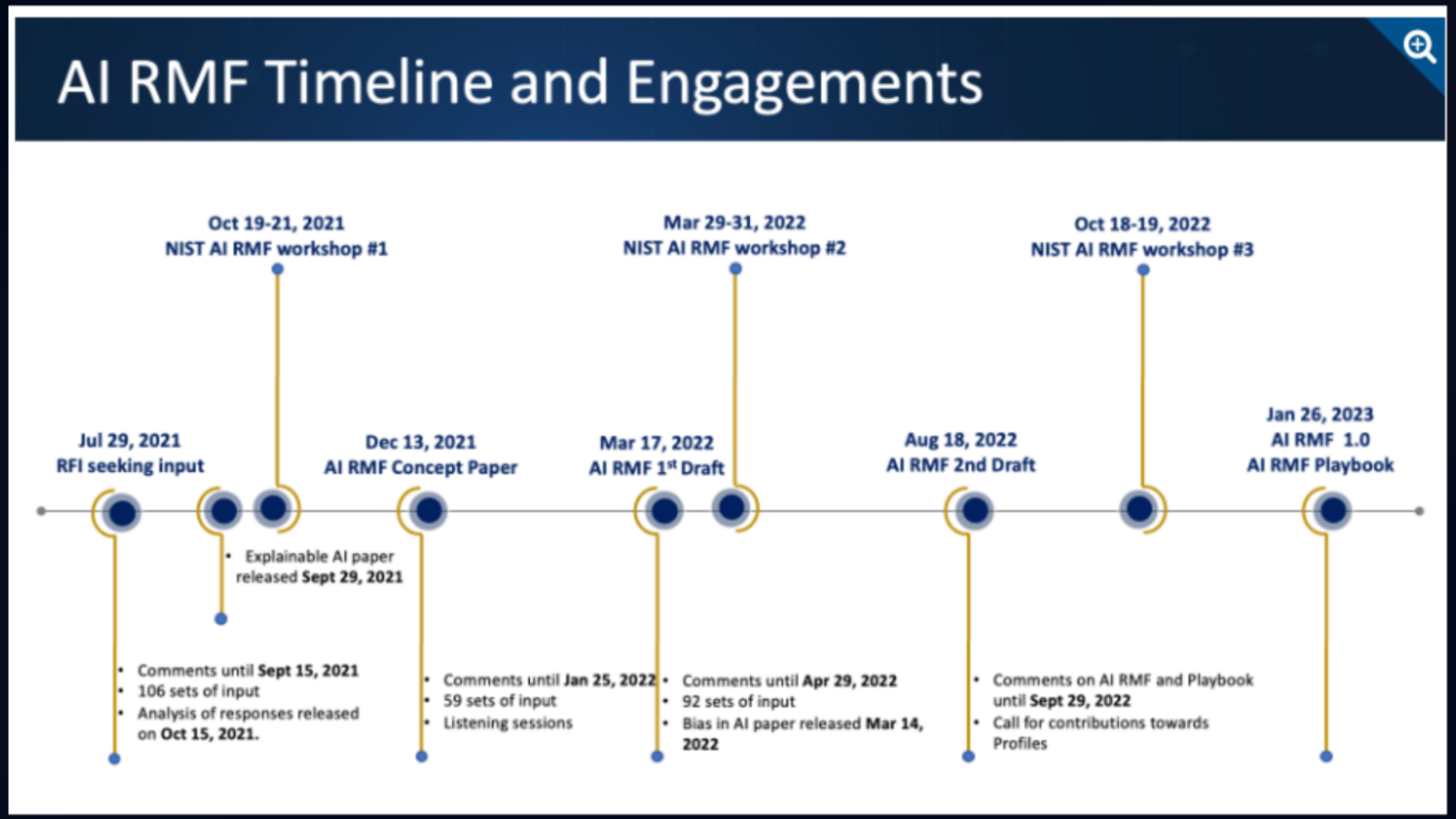
MIND'S EYE CREATIVE

Rise of the Machines

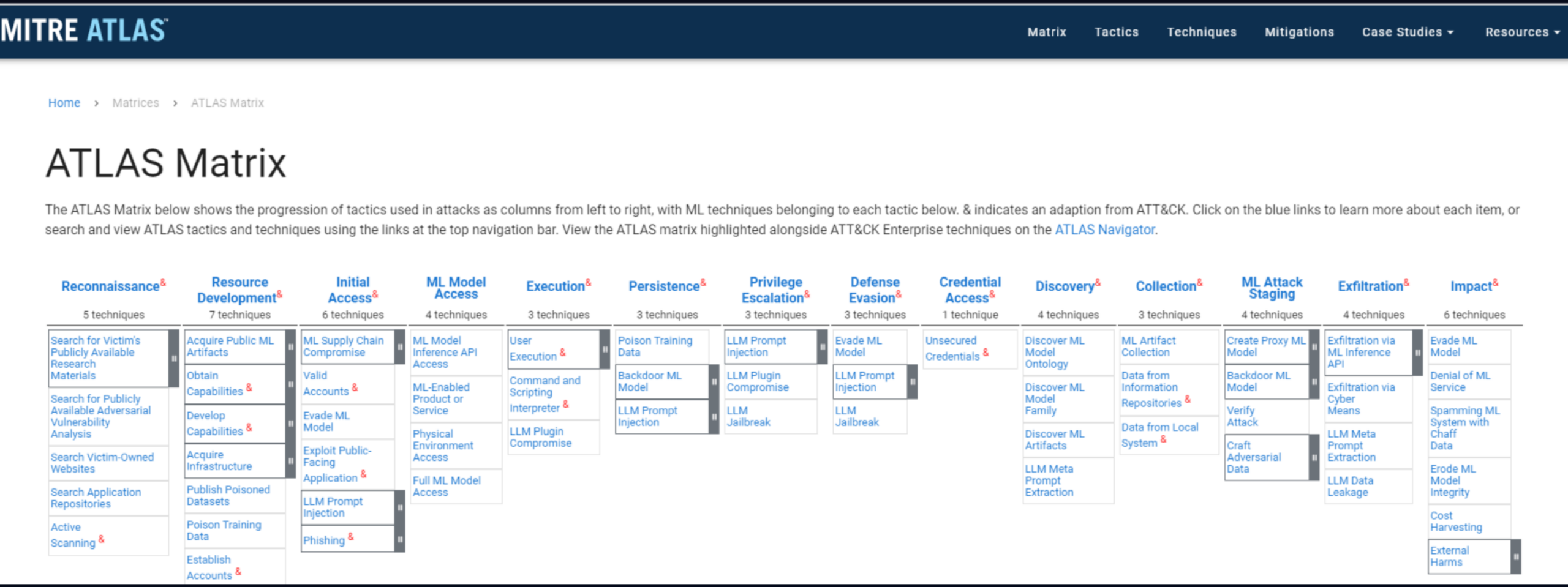


NIST AI Risk Management Framework

https://www.nist.gov/itl/ai-risk-management-framework



MITRE ATT&CK for AI (ATLAS)
https://atlas.mitre.org/



Recommendations

- Promote a *“distrust and verify” approach for content generated by Generative AI*. Users should not rely on the authenticity and accuracy of there content and should always validate the information and its sources.

- *Revise Organization’s AI Policy to be in line with good Information governance practices*. With pervasiveness of AI it is recommended to have Annual Attestation is done by Organization's users similar to IT Acceptable Use Policy and Cyber Security Policies.

- Look on the possibility of *introducing a banner on the use of Gen AI Tech in your organization.*

Strategies for Mitigation

- *Robust Authentication and Access Control:* Implement multi-factor authentication and strict access controls to prevent unauthorized access.
- *Advanced Threat Detection (using AI for DFIR):* Use AI-driven threat detection systems to identify and respond to unusual activities and potential intrusions.
- *Regular Security Audits & Penetration Testing (offensive and defensive AI):* Continuously assess and test the security of systems to identify and mitigate vulnerabilities.
- *Employee Training and Awareness for AI Threats*: Train personnel on recognizing phishing attempts and social engineering tactics to reduce the risk of human error.
- *Incident Response Planning to cater to AI Threats*: Develop and regularly update incident response plans.

References

- Generative Artificial Intelligence: Generative artificial intelligence (AI) - ITSAP.00.041 - Canadian Centre for Cyber Security
- Artificial Intelligence: Artificial Intelligence - ITSAP.00.040 - Canadian Centre for Cyber Security
- Potential Threat Implications of Generative AI for Corporate Security published 18th July 2023 by FBI
- Generative AI and Data Privacy: A Primer, Congressional Research Service, 23td May, 2023
- Microsoft is fully committed to the Responsible AI standards: Responsible AI Principles and Approach | Microsoft AI
- "Can I trust that Copilot's answers are always accurate?" Frequently asked questions about Microsoft 365 Copilot Semantic Index for Copilot: Explained by Microsoft"
- SANS AI Cyber Summit: https://www.sans.org/blog/a-visual-summary-of-sans-ai-summit-2024/
- https://atlas.mitre.org/matrices/ATLAS
- https://www.nist.gov/itl/ai-risk-management-framework

Questions?

Prashant
Senior Cyber Advisor at Fortune 500
https://www.linkedin.com/in/prashantprofile/

YegSec Slack: @guy_fawkes

https://github.com/prashant-iiitm

X (formerly Twitter): @prashant_geek

Email: prashant_iiitmg@yahoo.com

Cyber Risks in Generative Artificial Intellgience Large Language Models

Understanding the Cyber Risk Landscape

Agenda

- Understanding Generative AI
- Emerging Threat Landscape
- Examples of misuse of Generative AI
- Generic Large Language Model
- SANS AI Cyber Summit 2024 Key Takeaways
- Cyber Risk Frameworks for Generative AI
- Questions!

Emerging Threat Landscape