

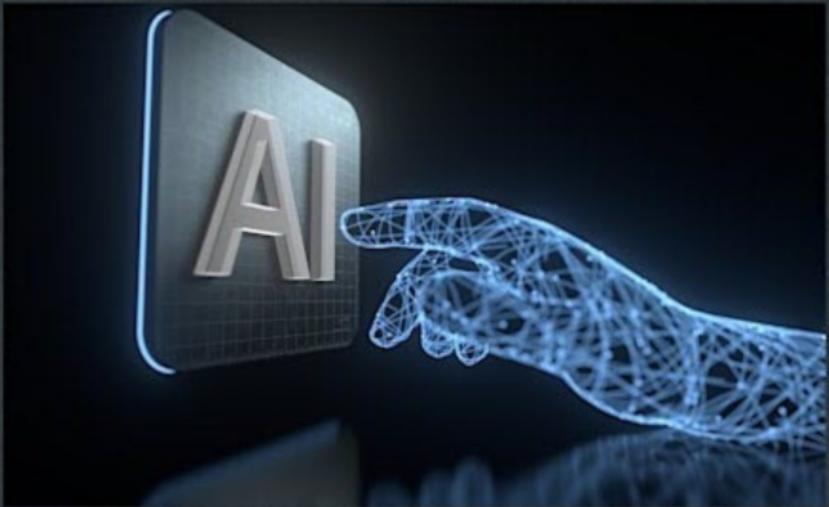
# Cyber Risks in Generative Artificial Intelligence Large Language Models

Understanding the Cyber Risk 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 powerful 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*)

The image shows the cover of a report titled "Joint Cybersecurity Information" at the top, with "TLP: CLEAR" below it. The cover features logos for several organizations: Communications Security Establishment Canada, Canadian Centre for Cyber Security, Australian Signals Directorate (ASD), and National Cyber Security Centre (UK). The title "Deploying AI Systems Securely" is prominently displayed, followed by "Best Practices for Deploying Secure and Resilient AI Systems". Below this is the "Executive summary" section. The main body of the report discusses the deployment of AI systems, mentioning the "Guidelines for secure AI system development" and incorporating mitigation considerations from "Engaging with Artificial Intelligence (AI)". It notes that the report is for organizations deploying and operating AI systems designed and developed by another entity, and that mitigations should be adapted to specific use cases and threat profiles. The report was authored by the U.S. National Security Agency's Artificial Intelligence Security Center (AISC), the Cybersecurity and Infrastructure Security Agency (CISA), the Federal Bureau of Investigation (FBI), the Australian Signals Directorate's Australian Cyber Security Centre (ACSC), the Canadian Centre for Cyber Security (CCCS), the New Zealand National Cyber Security Centre (NCSC-NZ), and the United Kingdom's National Cyber Security Centre (NCSC-UK). The goals of the AISC and the report are to:

1. Improve the confidentiality, integrity, and availability of AI systems;
2. Assure that known cybersecurity vulnerabilities in AI systems are appropriately mitigated; and
3. Provide methodologies and controls to protect, detect, and respond to malicious activity against AI systems and related data and services.

The cover also features a blue abstract graphic of interconnected dots forming a wave pattern.

# 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

Nick Waters  
@N\_Waters89 · Follow  
Confident that this picture claiming to show an "explosion near the pentagon" is AI generated.  
Check out the frontage of the building, and the way the fence melds into the crowd barriers. There's also no other images, videos or people posting as first hand witnesses.

An aerial view of the Pentagon on May 21, 2023, showing a massive plume of smoke rising from the building. A red X is drawn over the image to indicate it is a fake. Below the main image are two smaller images: one of a building with a red X over it, and another of a fence and crowd barriers.

# Artificial Intelligence Incident Database

<https://incidentdatabase.ai/>

AI INCIDENT DATABASE

English | [X](#) [RSS](#) [F](#) [In](#)

Discover + Submit

Welcome to the AIID

Discover Incidents

Spatial View

Table View

Entities

Taxonomies

Word Counts

Submit Incident Reports

Submission Leaderboard

Blog

AI News Digest

Risk Checklists

Random Incident

Sign Up

pentagon fake

Display Option: Incident Reports | 22 results found | Sort by: Relevance | Export | Classifications | Source | Clear Filters | More filters

**Deepfakes spell deep trouble for markets**  
afr.com · 2023 ▾  
When you report on the cut and thrust of breakneck developments, big money deals, mind-bending hype, and multi-billionaire egos in the technology world, it can be easy to get caught up in the game and ignore the fact that the winners are not always the good guys.

Take your pick of fun, child-focused movies. The Mitchells vs. the Machines, Ron's Gone Wrong, Big Hero 6, even Pixar's WALL-E 15 years ...

**Artificial Intelligence manufactures explosion of Pentagon, authorities huddle to clarify**  
wionews.com · 2023 ▾  
The world paused briefly when the visuals of an explosion in Pentagon's premises began making rounds on social media on Monday. The image, which began circulating on Twitter on May 22, showed an explosion on a grass lawn outside the Pentagon.

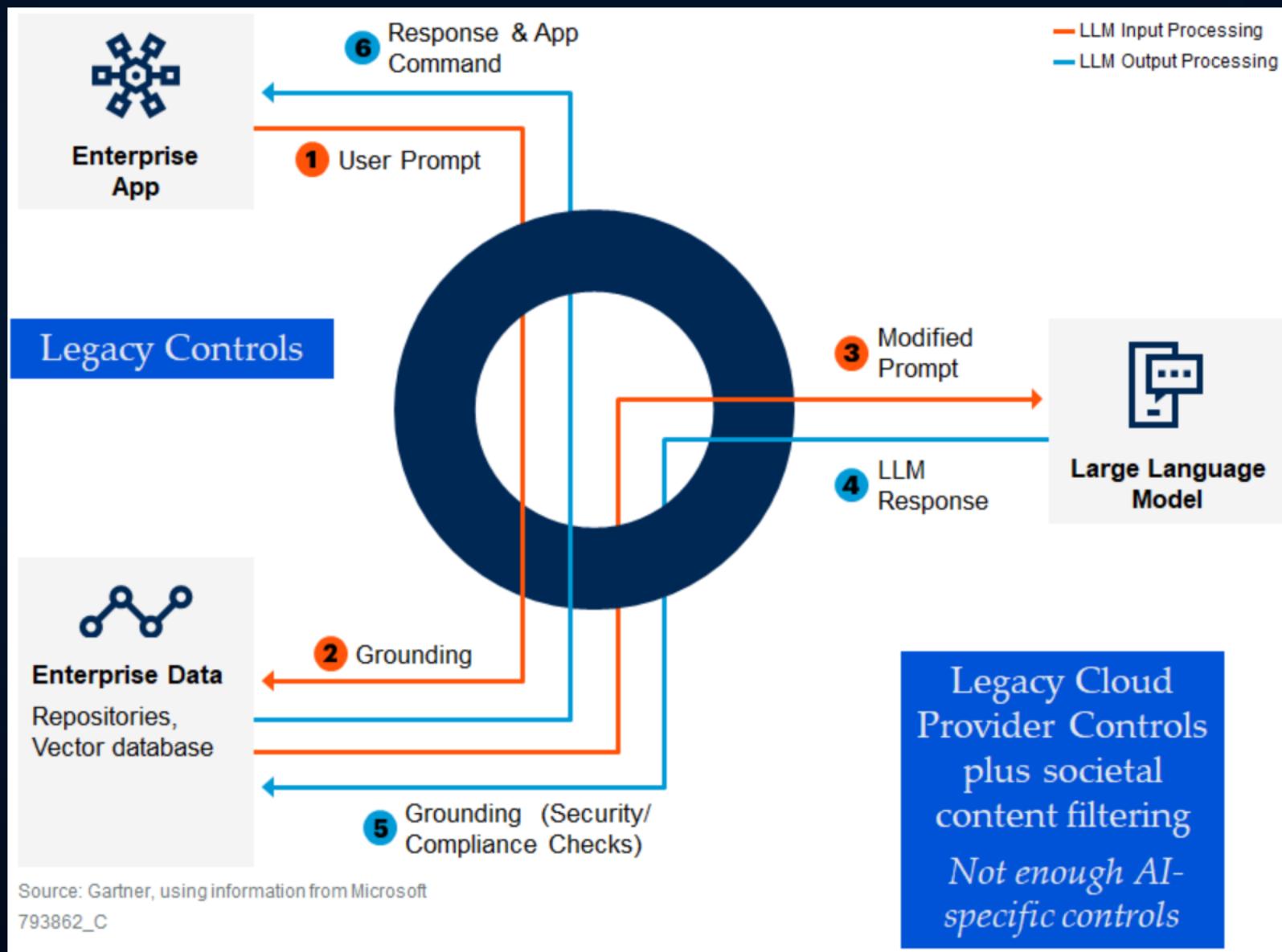
The original post has since been removed.

Pentagon explosion visuals: Did an

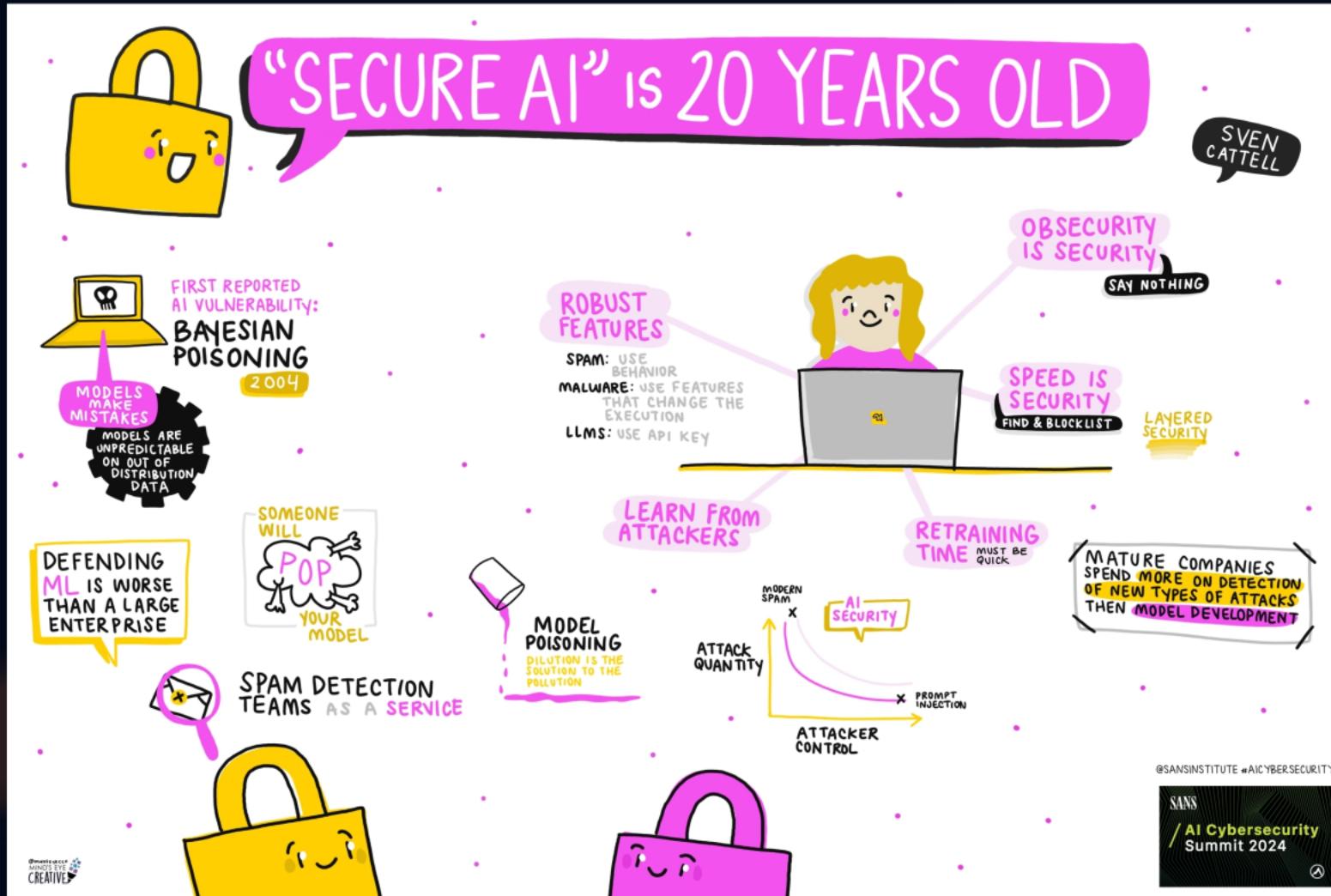
**Fake Pentagon blast pic triggers stock market drop**  
globalvillagespace.com · 2023 ▾  
The recent circulation of a deepfake image on Twitter that depicted an explosion outside the Pentagon exemplified the risks of misinformation associated with generative AI. The image, which appeared to be AI-generated, caused the stock market to dip by 0.26 percent before bouncing back. The Arlington Police Department quickly debunked the image, stating that there was no explosion or incident taki...

**Anyone could be a victim of 'deepfakes'. But there's a reason Taylor Swift is a target | Jill Filipovic**  
theguardian.com · 2024 ▾  
Taylor Swift is having quite a month. The singer-songwriter saw her image in disgusting deepfake porn images that were circulated online, prompting a necessary and overdue conversation on how AI and deepfake porn is used to harass, humiliate, degrade, threaten, extort and punish (mostly) women. And then her boyfriend, the football player Travis Kelce, saw his team make it to the Super Bowl, which...

# 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



INTERNAL  
REPUTATION

BRAND

FINANCIAL



INPUT



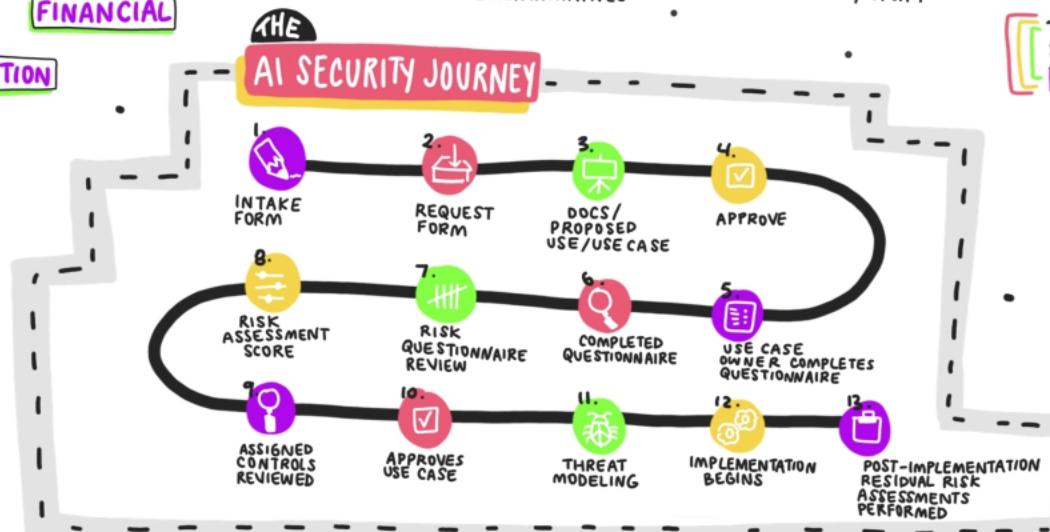
ENABLEMENT

- 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



# THE FRONTIER OF CYBERSECURITY DEFENDING AGAINST AI-BASED THREATS

COREY  
WHITE



## COUNTERMEASURES:

ROBUST  
ACCESS  
CONTROLS

AI-BASED ENDPOINT  
PROTECTION

BLOCK PHISHING EMAILS &  
FAKE WEBSITES

CONTINUOUS SCANNING  
& PATCHING

SECURITY AWARENESS

# GENAI FOR DFIR IN THE REAL WORLD

PRACTICAL USE CASES

JESS  
GARCIA

DS4N6

PROJECT



DATA SCIENCE &  
ARTIFICIAL INTELLIGENCE  
TO THE AVERAGE  
FORENSICATOR



TO HELP ME

CUSTOM  
GPTs +  
RAG

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

WORKFLOWS

THREAT  
HUNTING

FORENSICS



THREAT  
DETECTION  
& RESPONSE



USE CASE:  
DATA ANALYSIS  
WITH CHAT GPT

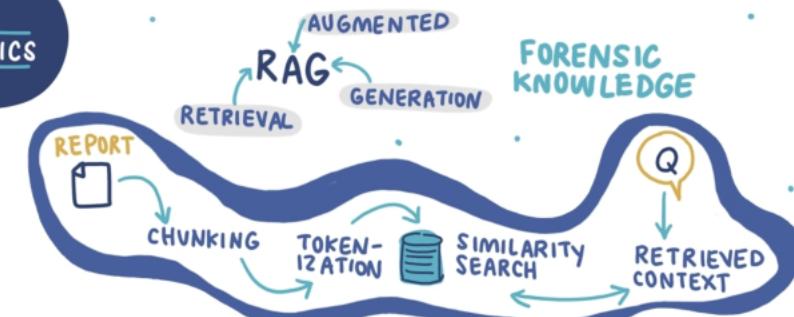
CHAT GPT  
GENERATES CODE  
NEEDED TO DO  
ANALYSIS



ASK MORE QUESTIONS:

- Was there a brute force attack?
- Write report with conclusion & recommendations

ADD DETAILS  
TO QUERIES



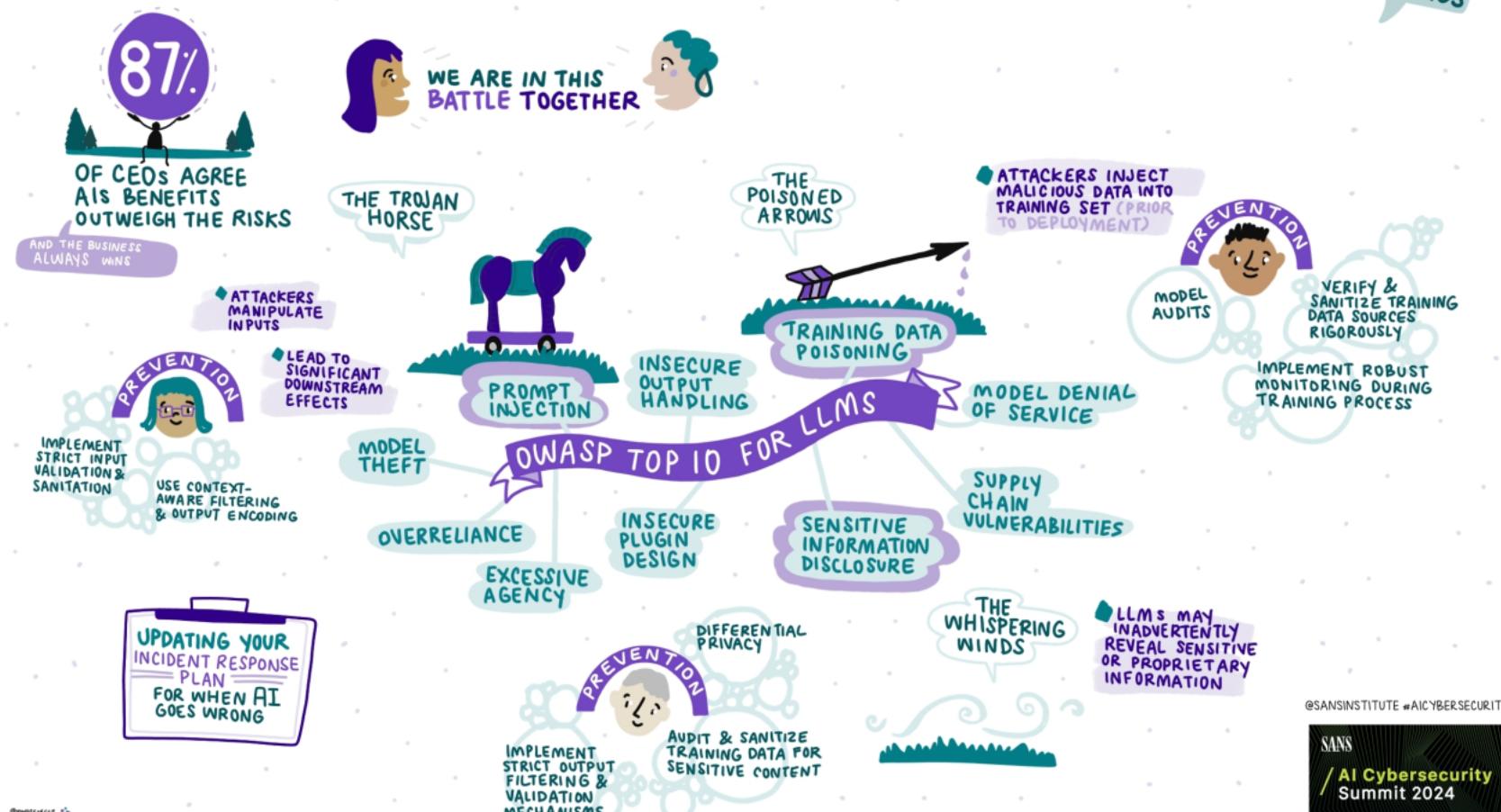
@SANSINSTITUTE #AICYBERSECURITY



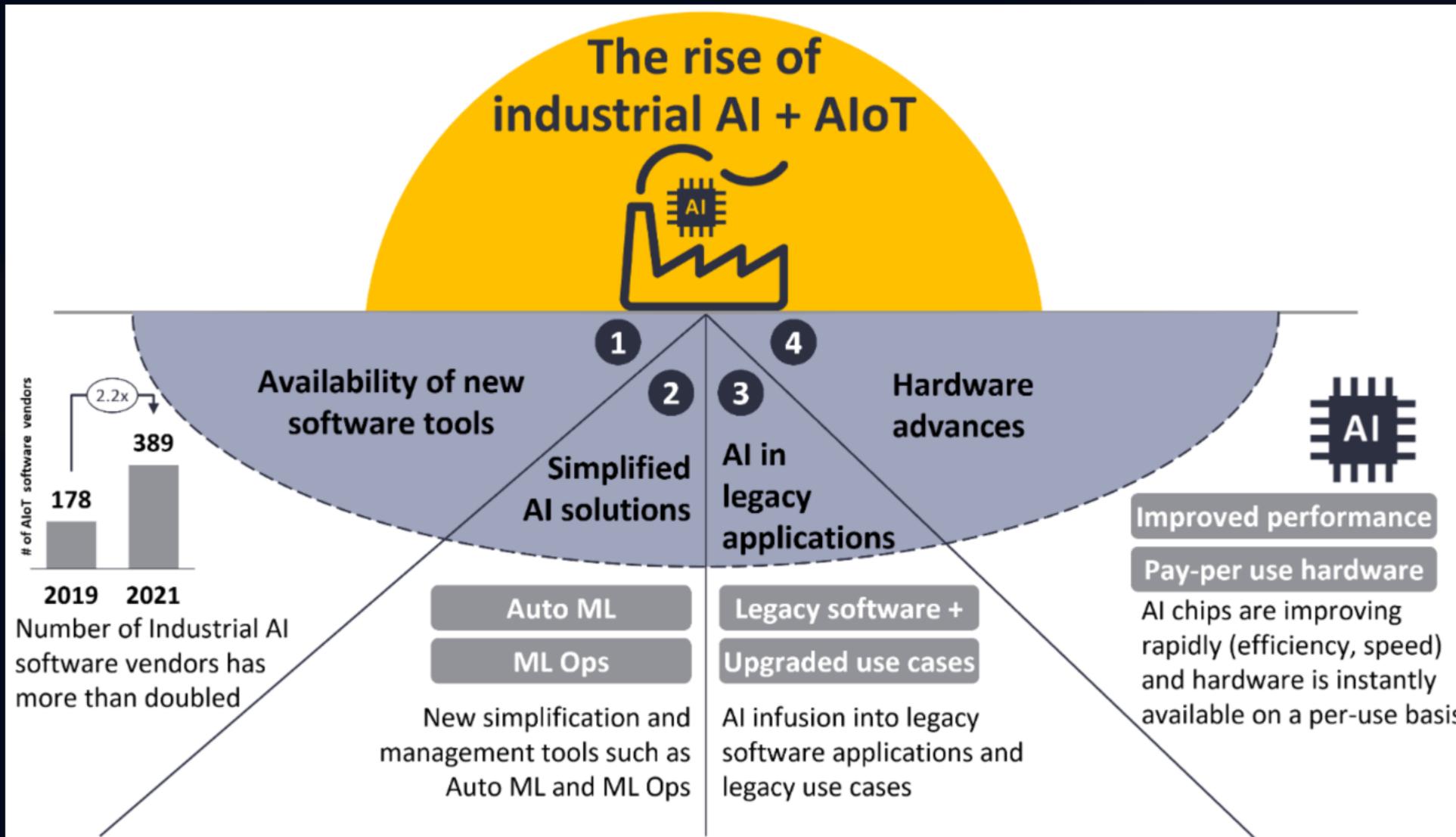
# AI'S ACHILLES' HEEL

## NAVIGATING THE OWASP TOP 10 FOR LLMS

KYRIAKOS LAMBROS

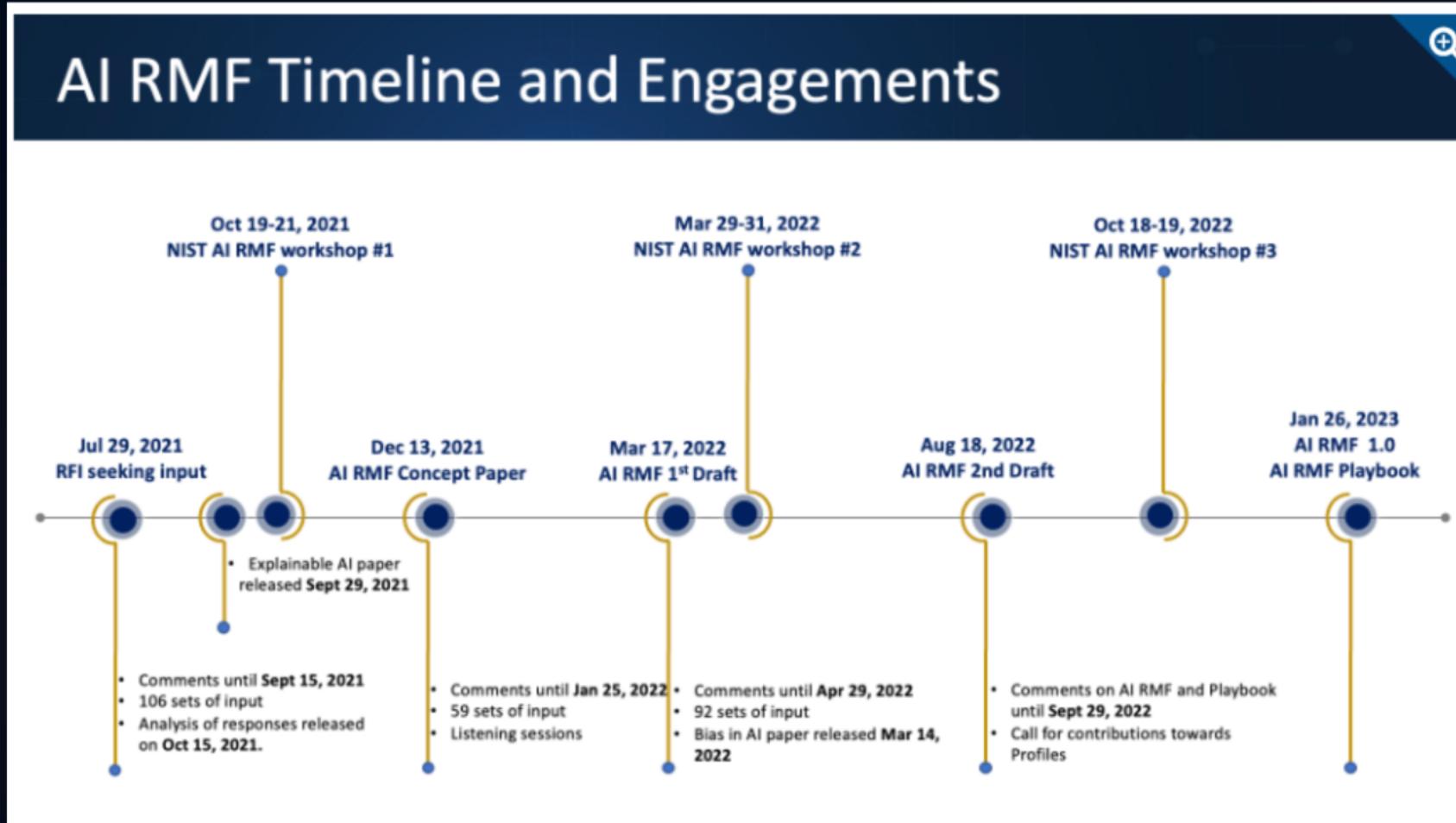


# Rise of the Machines



# NIST AI Risk Management Framework

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



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

# MITRE ATLAS<sup>®</sup>

Matrix      Tactics      Techniques      Mitigations      Case Studies ▾      Resources ▾

[Home](#) > [Matrices](#) > ATLAS Matrix

# ATLAS Matrix

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. & indicates an adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the links at the top navigation bar. View the ATLAS matrix highlighted alongside ATT&CK Enterprise techniques on the [ATLAS Navigator](#).

# 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, 23rd 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 Intelligence Large Language Models

Understanding the Cyber Risk Landscape

