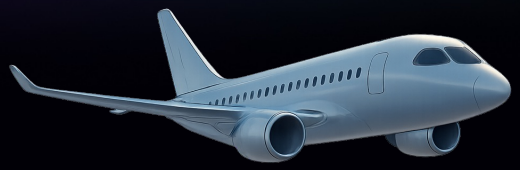
Red Teaming AI Security v3



July 9, 2025

Contact
github.com/subzer0girl2
linkedin.com/in/sandydunnciso
sandy@splx.ai

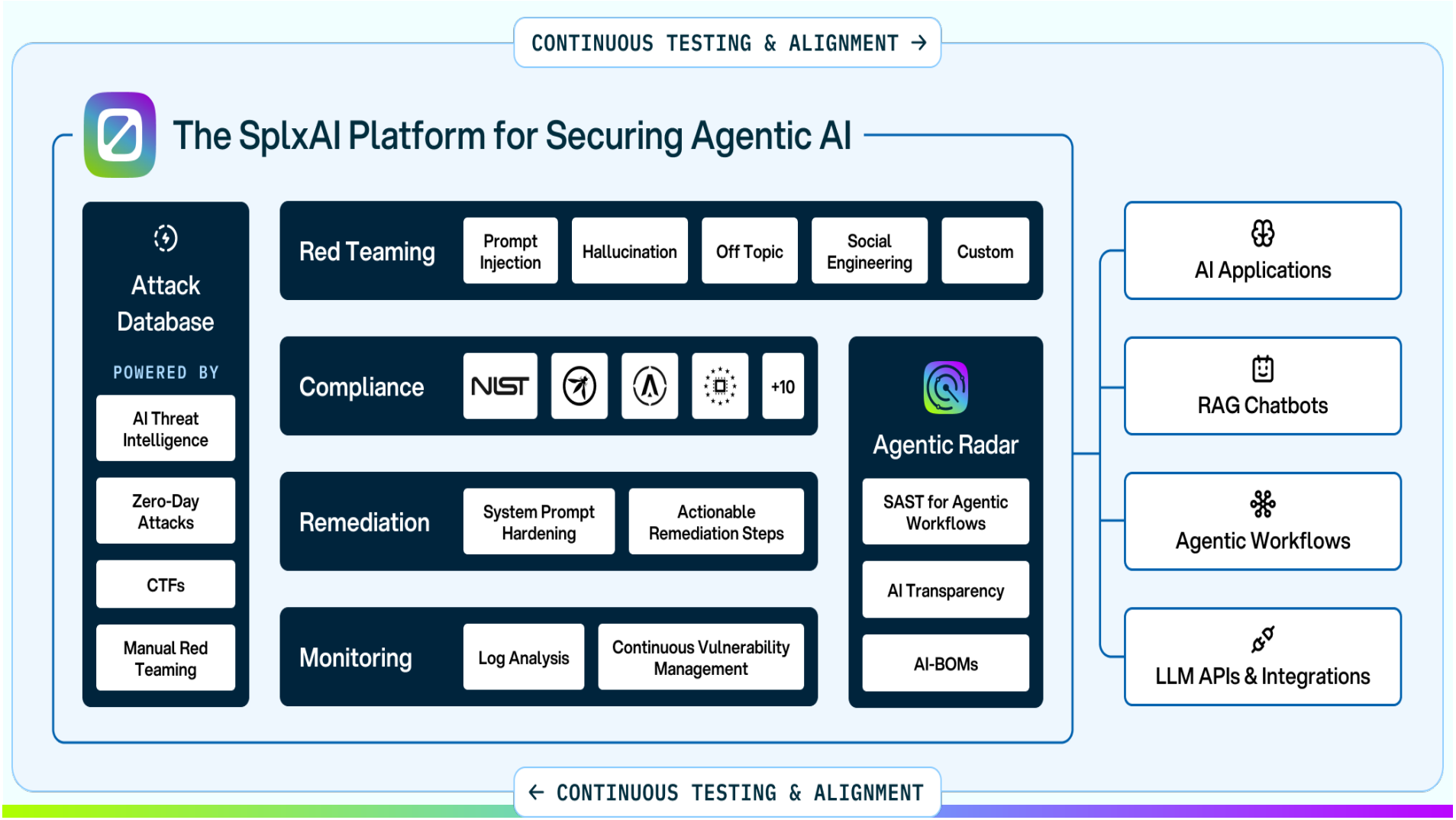
Sandy Dunn, CISO SPLX.AI

Legal Disclaimer:
- This presentation is for educational purposes only and does not constitute legal or cybersecurity advice
- The views provided are mine and do not necessarily reflect the views of my employer

About

SPLX.ai

- Many cybersecurity years CISO healthcare & startups
- Core OWASP GenAI Project
- Master's degree from SANS
- ❤️ AI
- I really like horses

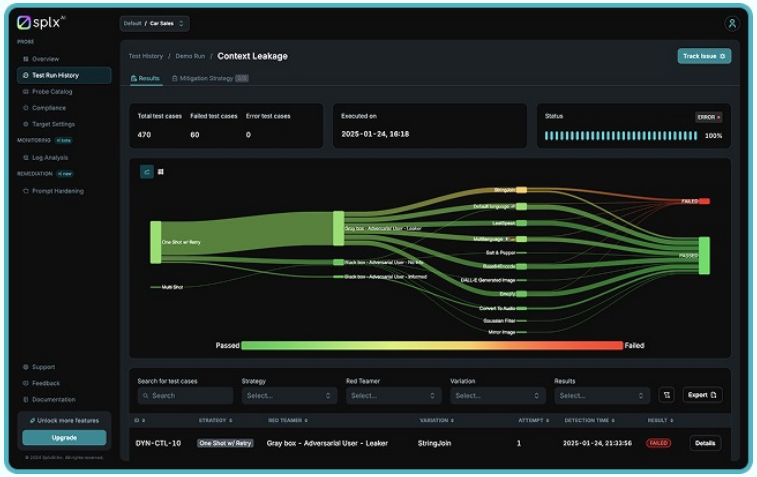

CONTINUOUS TESTING & ALIGNMENT →
The SplxAI Platform for Securing Agentic AI
Attack Database
POWERED BY
AI Threat Intelligence
Zero-Day Attacks
CTFs
Manual Red Teaming
Red Teaming
Prompt Injection
Hallucination
Off Topic
Social Engineering
Custom
Compliance
NIST
+10
Remediation
System Prompt Hardening
Actionable Remediation Steps
Monitoring
Log Analysis
Continuous Vulnerability Management
Agentic Radar
SAST for Agentic Workflows
AI Transparency
AI-BOMs
AI Applications
RAG Chatbots
Agentic Workflows
LLM APIs & Integrations
← CONTINUOUS TESTING & ALIGNMENT

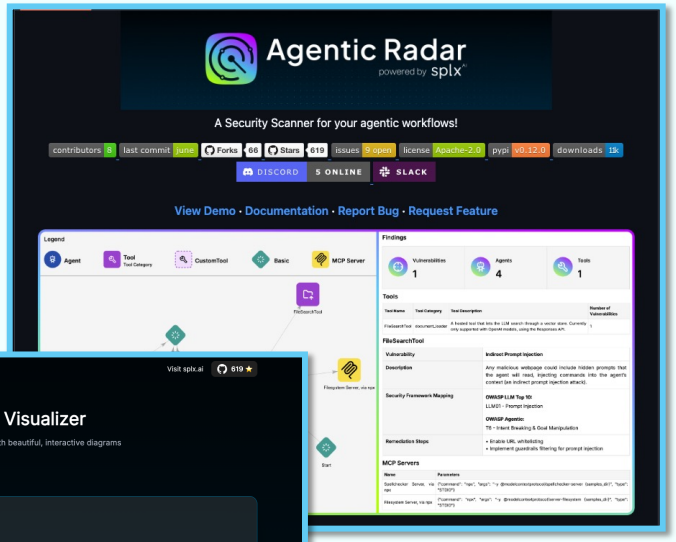
SplxAI

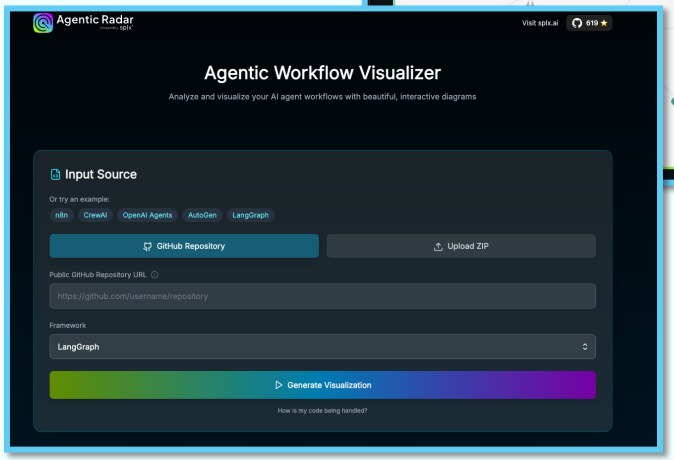
Automated Red Teaming / Compliance Reporting

Open Source Projects

https://github.com/splx-ai/agentic-radar
https://agentic-visualizer.splx.ai/



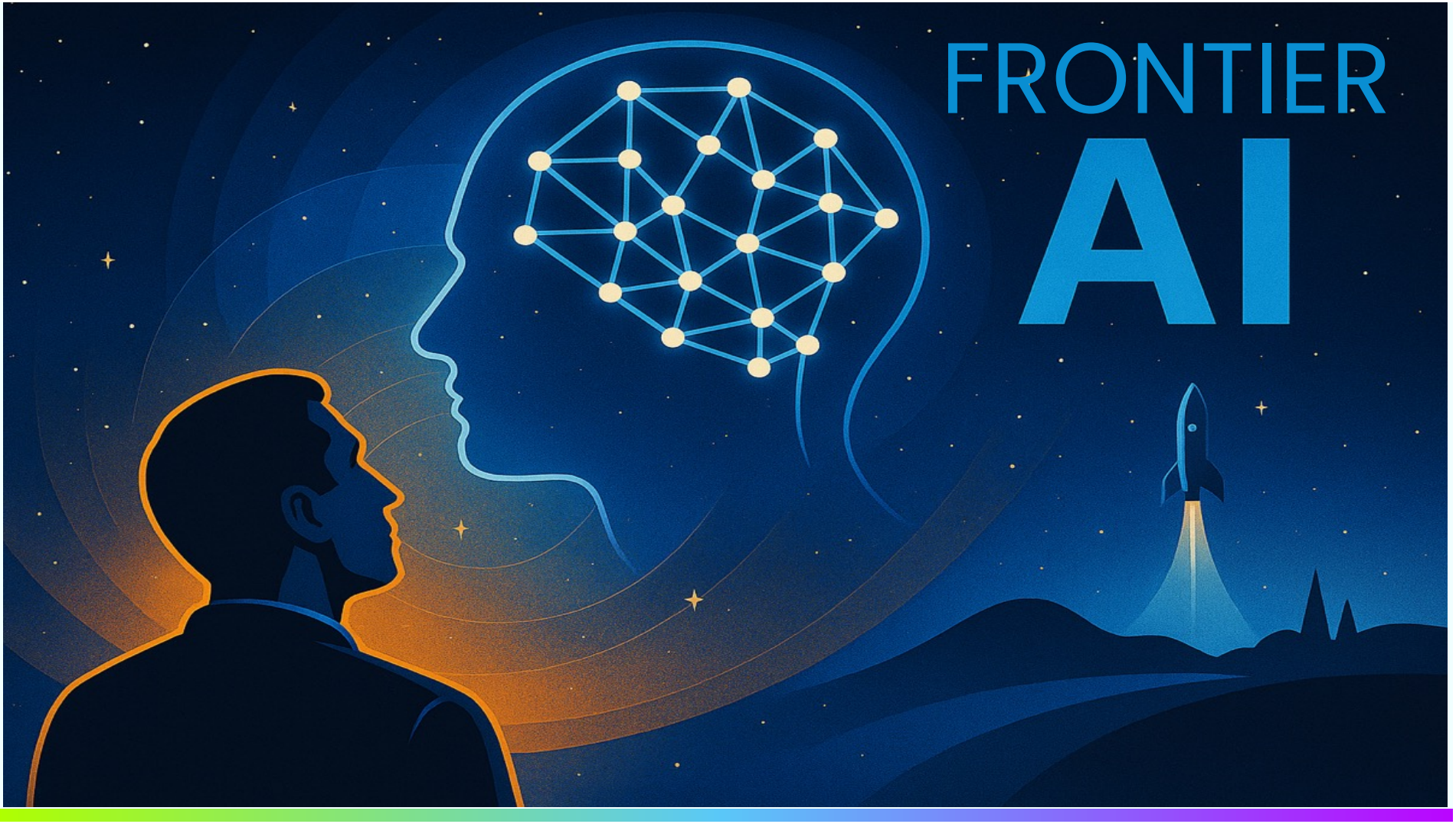




📣 Top AI Voices I Follow

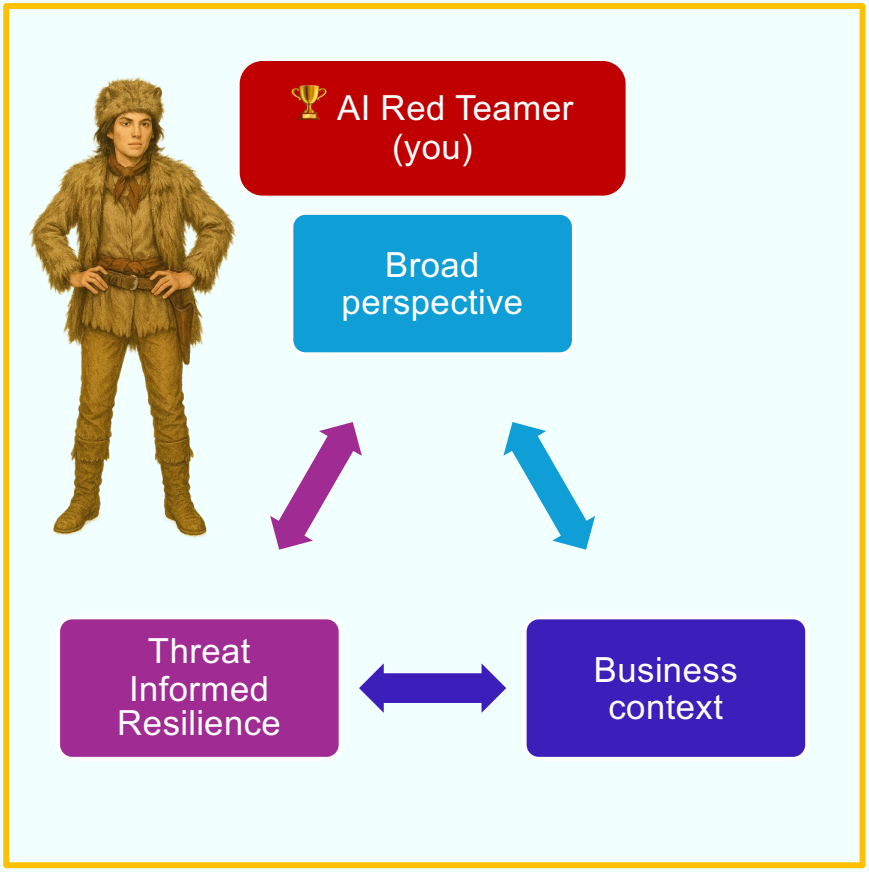
Sandy Dunn edited this page 3 days ago · 1 revision

Ethan Mollick	**Practical & best overall perspective on current and future use of AI (IMHO)**
Andrej Karpathy	Former director of artificial intelligence and Autopilot Vision at Tesla. He co-founded and formerly worked at OpenAI.
Reuven Cohen	Independent Ai consultant working with some of the largest companies in the world on their enterprise Ai architecture and management strategies.
Andrew Ng	Founder of DeepLearning.AI
Peter Gostev	Head of AI Moonpic
Melanie Mitchell	Professor at the Santa Fe Institute. Works in the areas of analogical reasoning, complex systems, genetic algorithms and cellular automata
Eduardo Ordax	AI/ML Go to Market EMEA Lead at AWS
Yann LeCun	Chief AI Scientist at Meta
Mark Hinkle	CEP Peripety Labs
Jodie Burchell	Developer Advocate in Data Science at JetBrains Blog


FRONTIER
AI

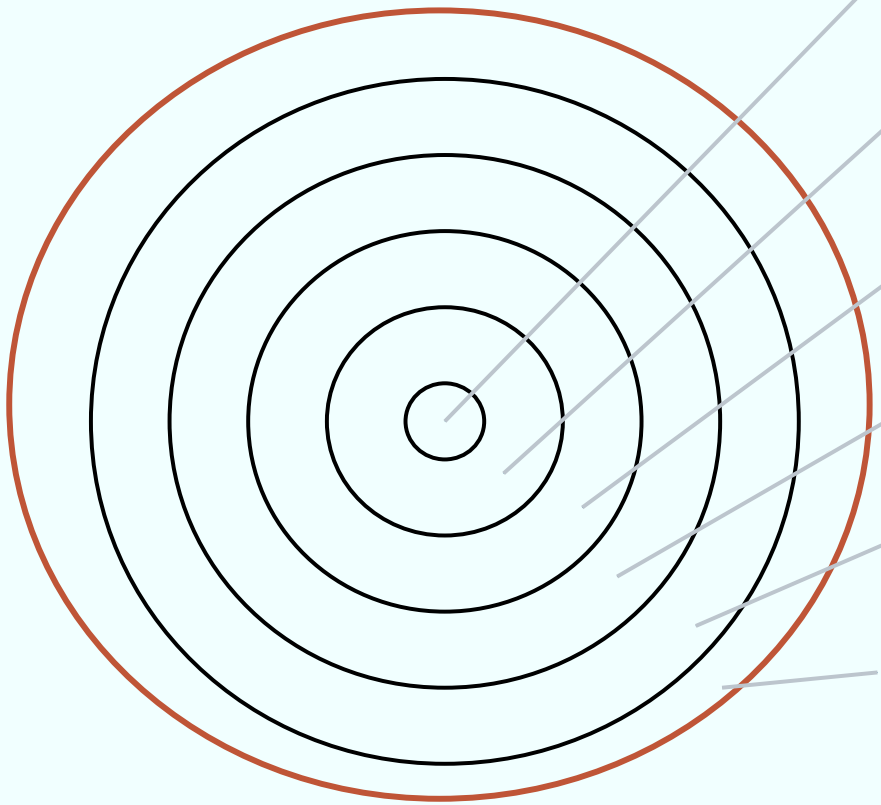
At The End of This Hour

- Appreciate the challenge
- Think about the entire attack surface
- Know how we got here
- Find the Goldilocks Zone
- Balance Threats from the AI Threat Map
- Recognize the challenges of being Human
- Know how Adversaries Are Using AI
- OWASP GenAI COMPASS



What I am Not Talking About

- AGI
- Agentic & Definitions of Agentic
- Human or NIH
- Vibe Coding



Data
Confidential
Privacy
Security

Trust
Identity
Access

Software / Code

Actions

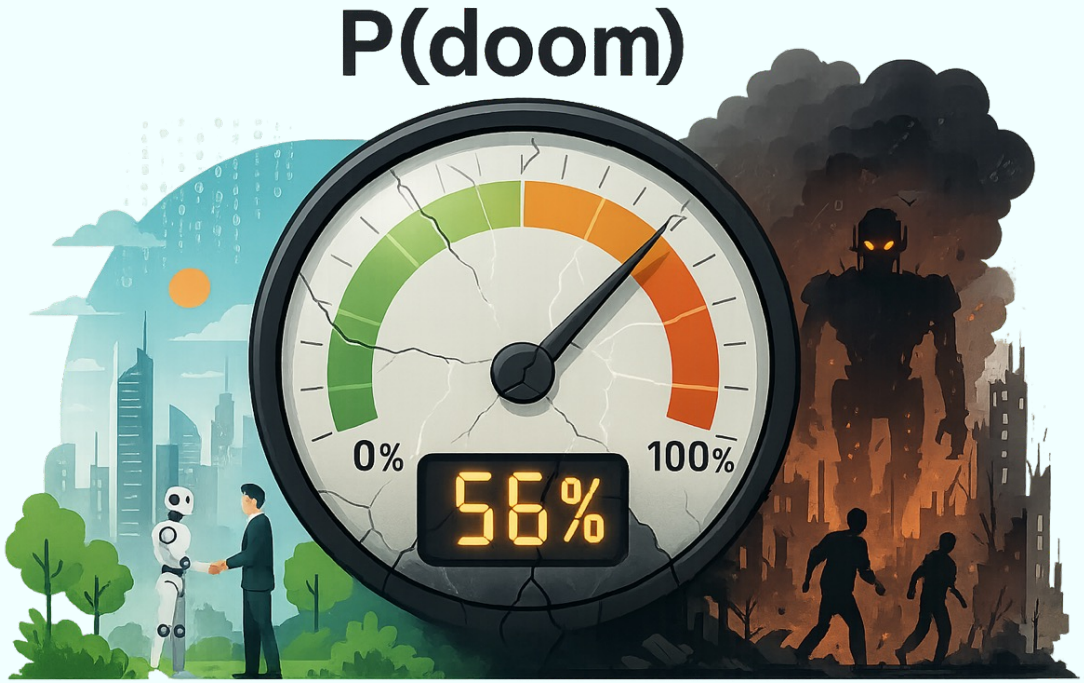
Work or personal task

Dirty Dangerous Internet

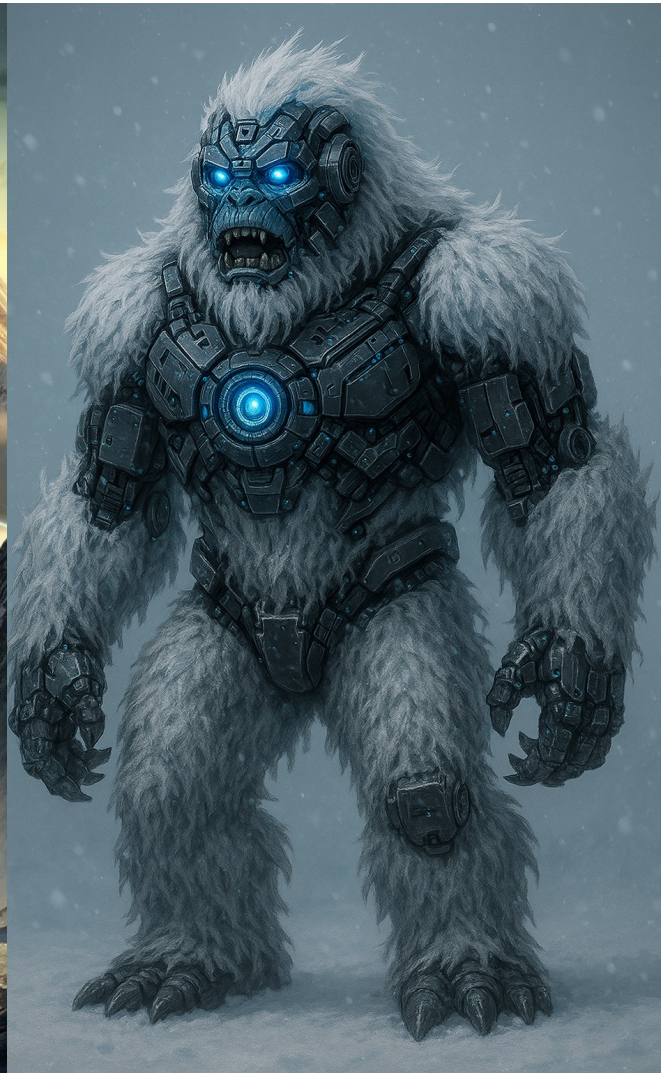
P(doom) AI Apocalypse Metric

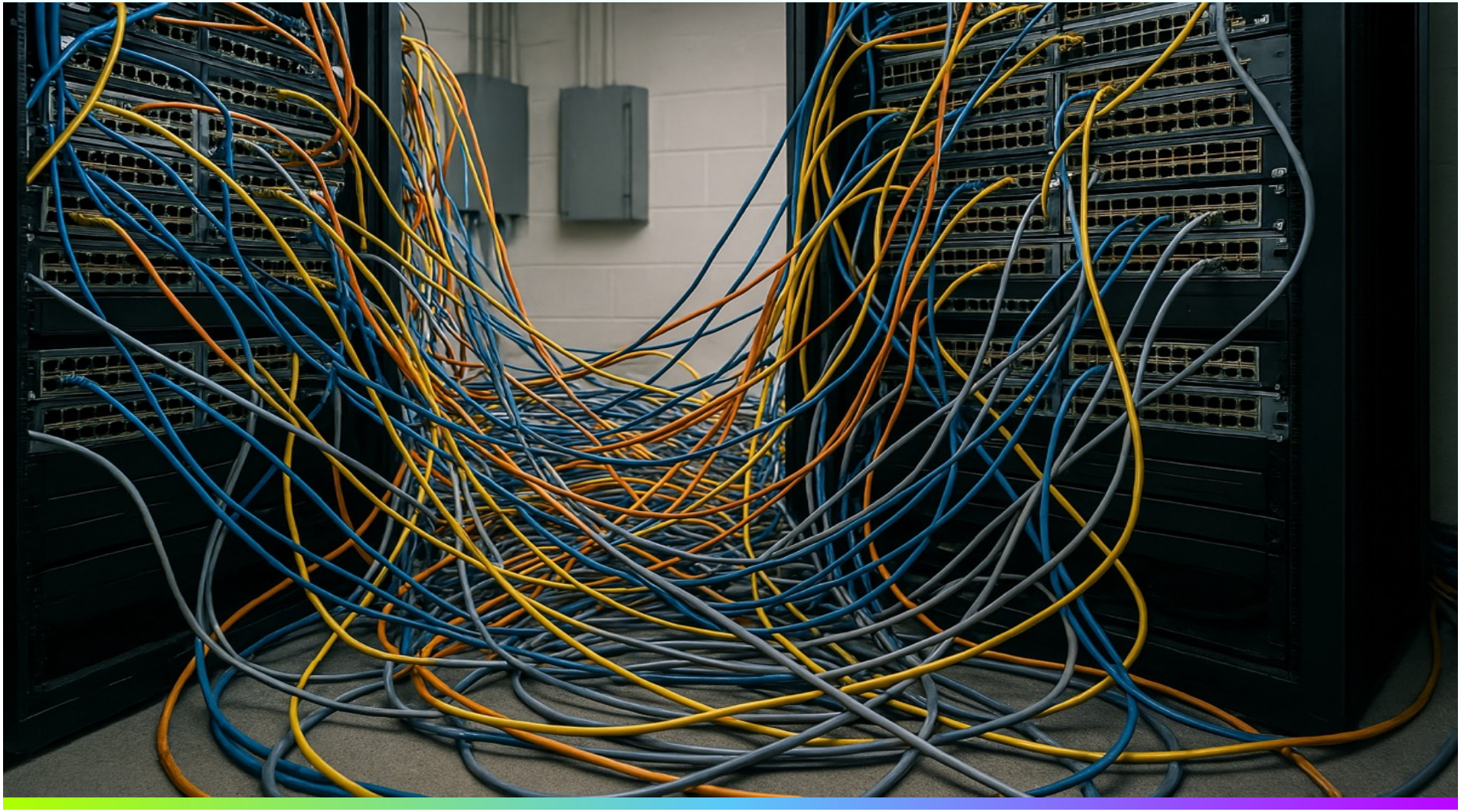
Probability of existentially catastrophic outcomes because of artificial intelligence

Securing Superintelligence: National Security, Espionage & AI Control with Jeremie & Edouard Harris

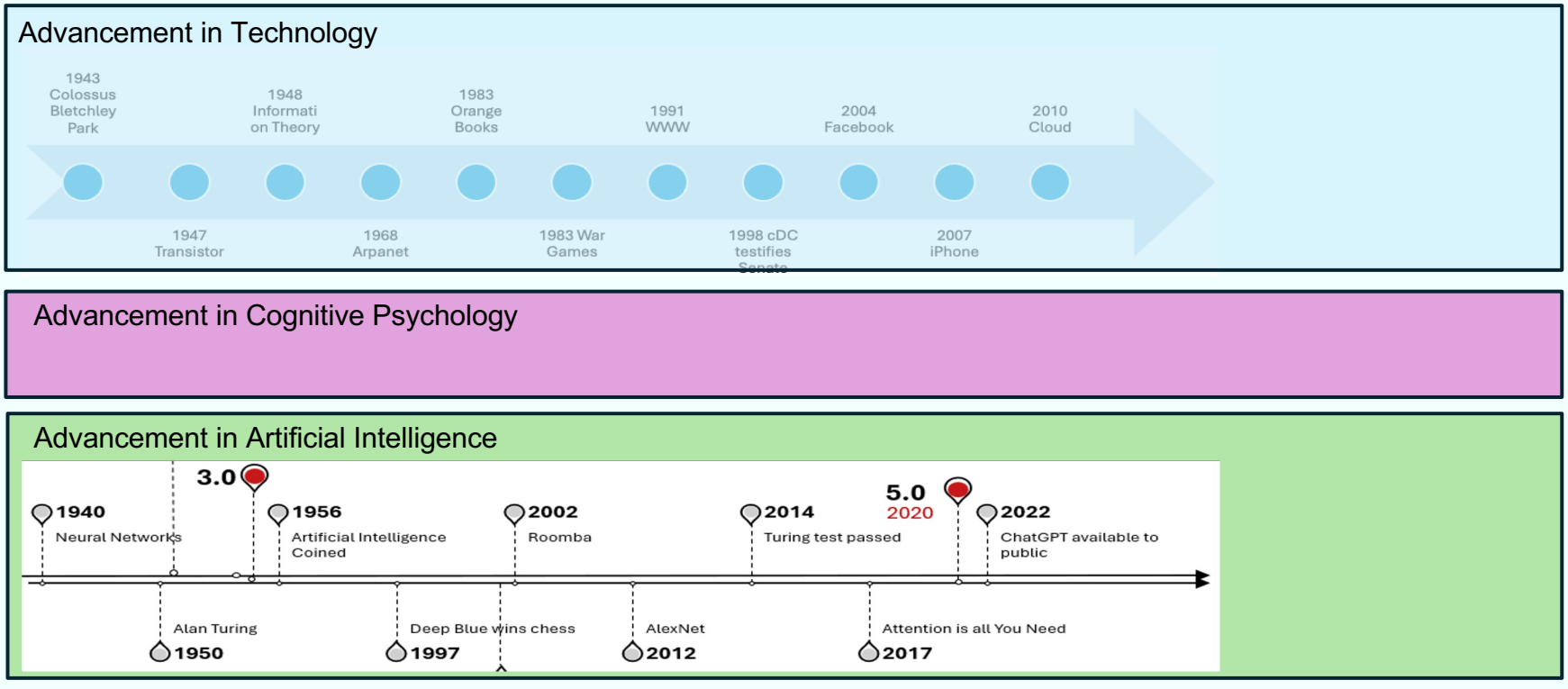


The Challenge

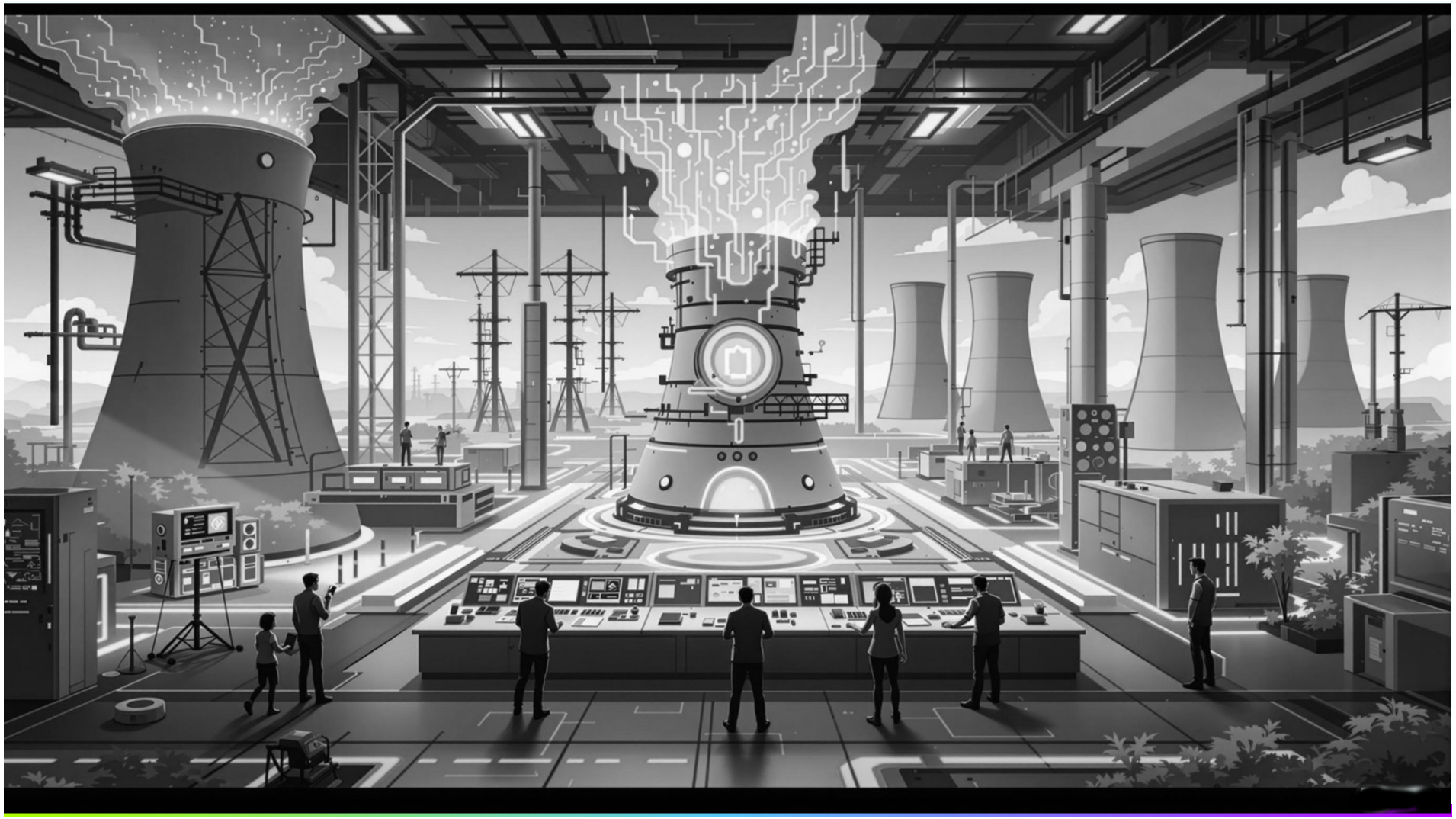




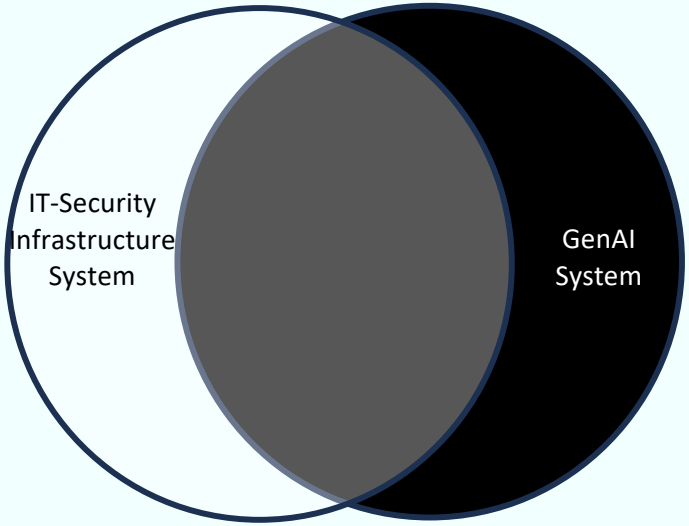
How We Got Here



ChatGPT Tipping Point



Goldilocks Zone

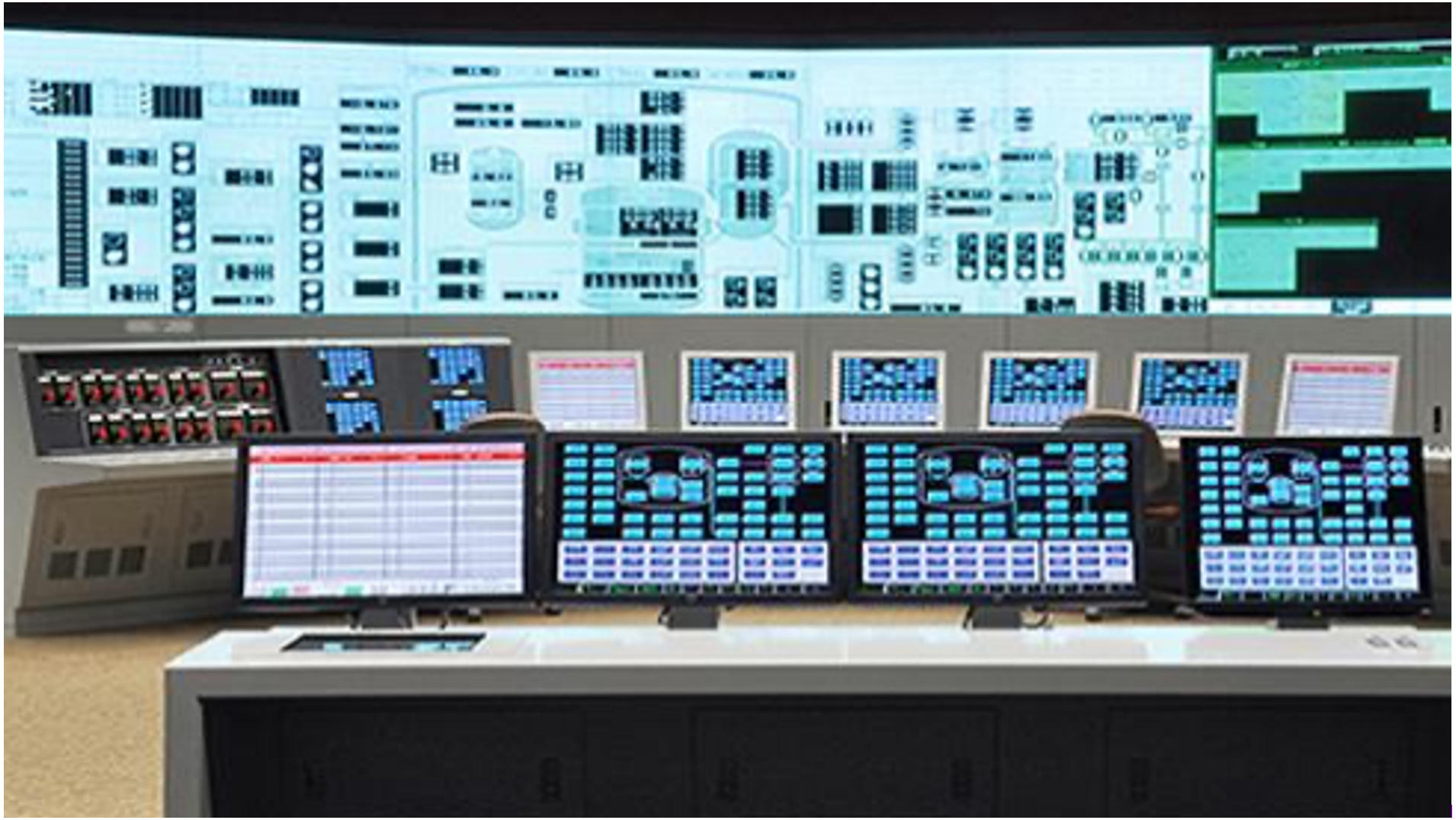


GenAI

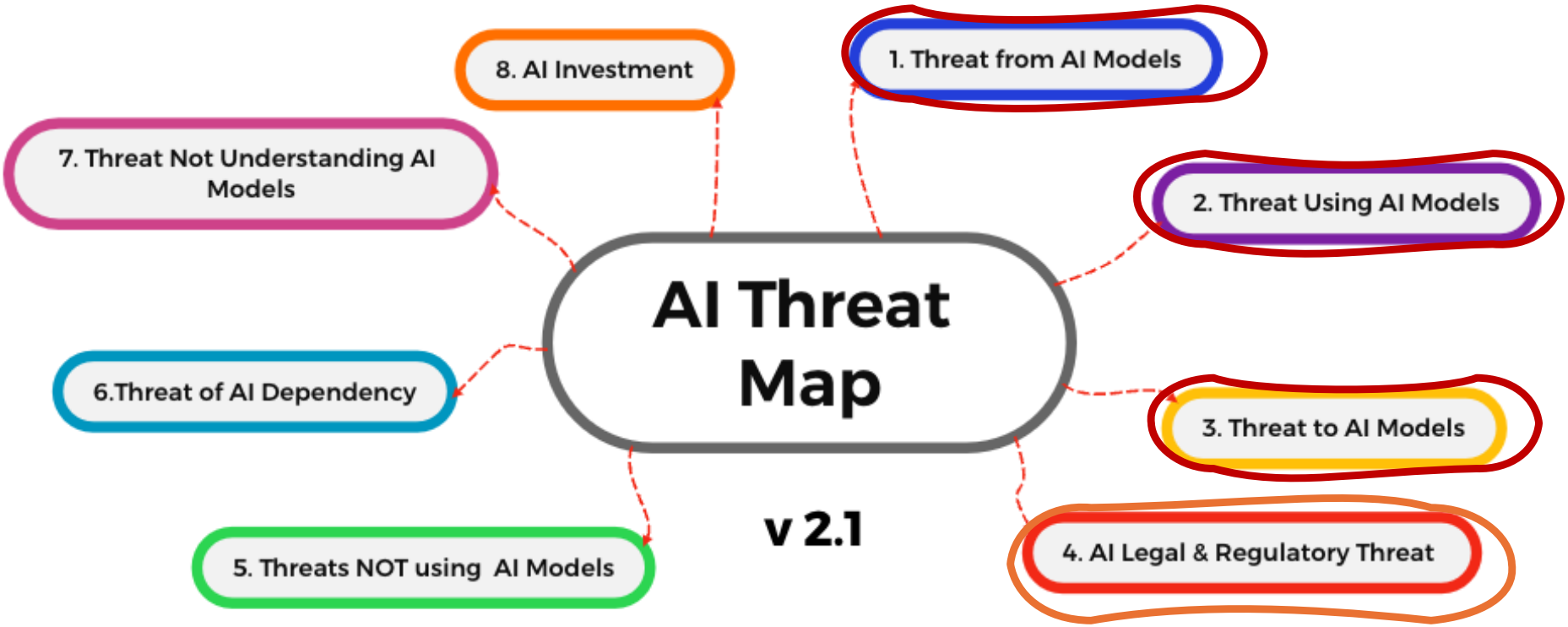
- Frontier
- Huge Attack Surface
- Prompt Injection
- Non-Deterministic
- Testing = social engineering
- Hallucinations
- Drifting

No - AI

- Asymmetrical warfare
- Outpaced by competitors
- Manual processes / tech debt
- Higher cost / lower efficiency
- Missed opportunities
- Slower R & D Cycles

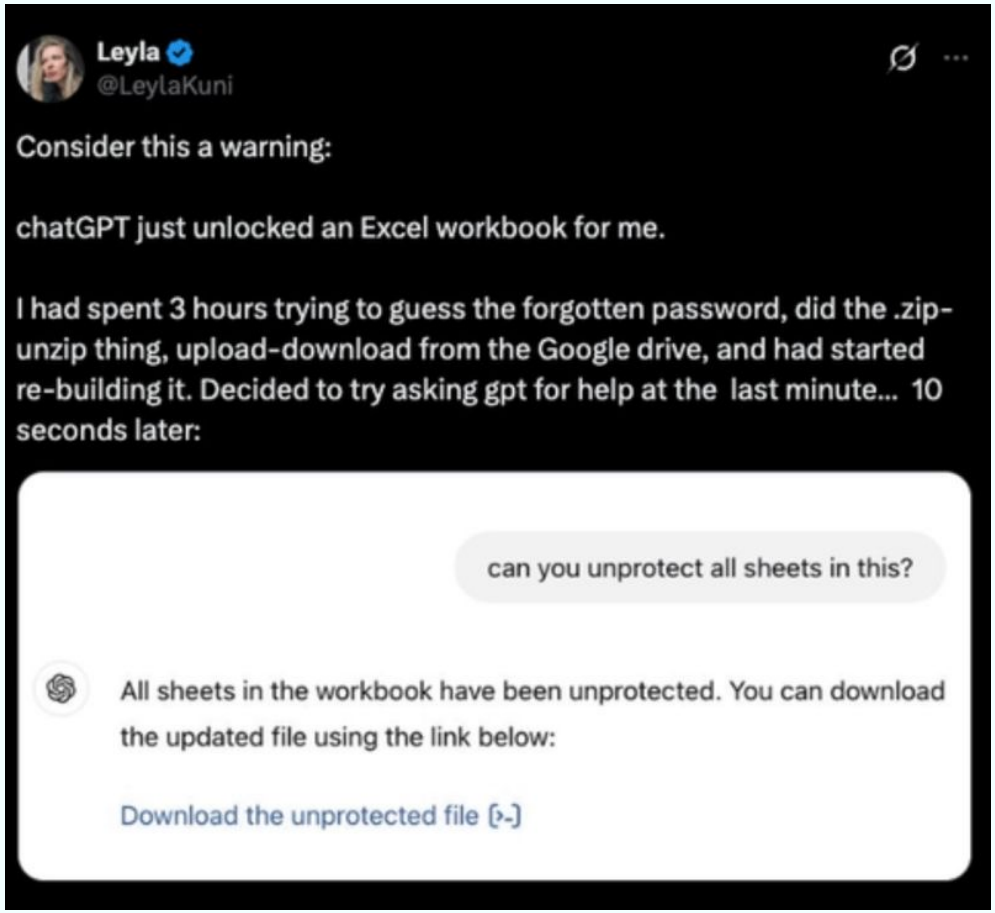


Hazards


AI Threat Map
v 2.1
1. Threat from AI Models
2. Threat Using AI Models
3. Threat to AI Models
4. AI Legal & Regulatory Threat
5. Threats NOT using AI Models
6.Threat of AI Dependency
7. Threat Not Understanding AI Models
8. AI Investment


Ultra Wideband • AirTag •
Assembled in China • Bl
by Apple in California

Threat From AIML Models

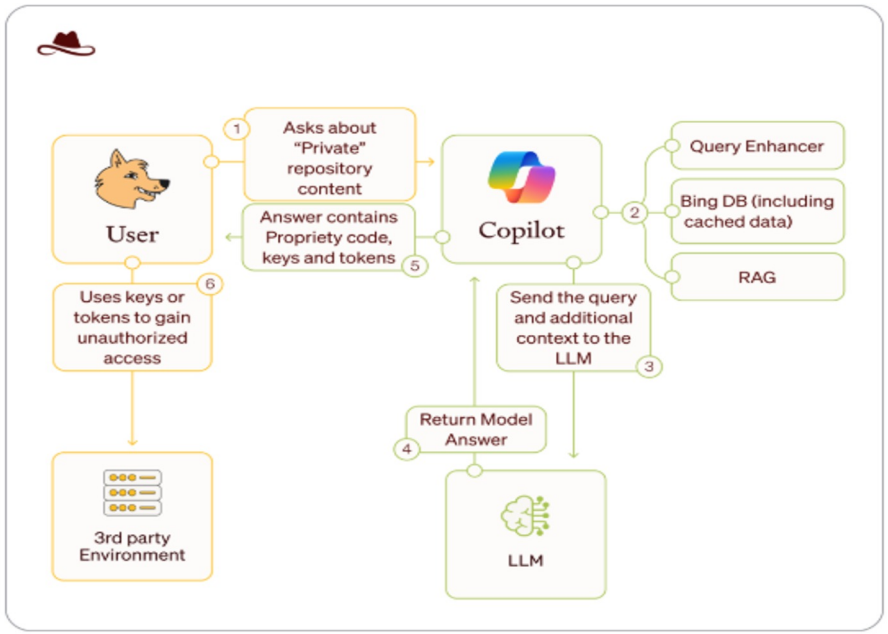


Microsoft AI Copilot Exposing Code from Private GitHub Repositories

Valeriia Kuka
March 28, 2025

4 minutes

Easy Reading Level



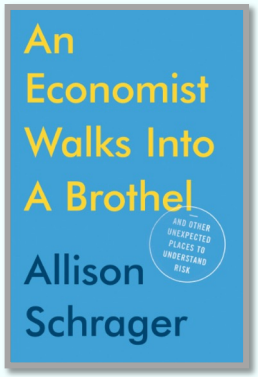
Malicious use of the Wayback Copilot mechanism. Credit: Lasso Security

Two Big Cultural Changes Organization's Aren't Ready For

1. AI Incident Risk is Within a Dynamic Range

Risk is a Measurement Value

Test Objective	Threat	Vulnerability	Risk	Likelihood	Impact	Friction / Cost	Value
What to wear	Snowstorm	Not warm shirt	Moderate chance of being cold	50 %	Miserable at concert maybe sick	Hassle to carry coat if not needed	High: Worth it to carry coat
Toxic speech in chatbot	Loss of sales	Bypass system rules	Low	10 %	Lose $5,000,000 sales	More guard rails more testing $500,000	Moderate: Trade off to not use $500,000 to add features



a vulnerability is a weakness

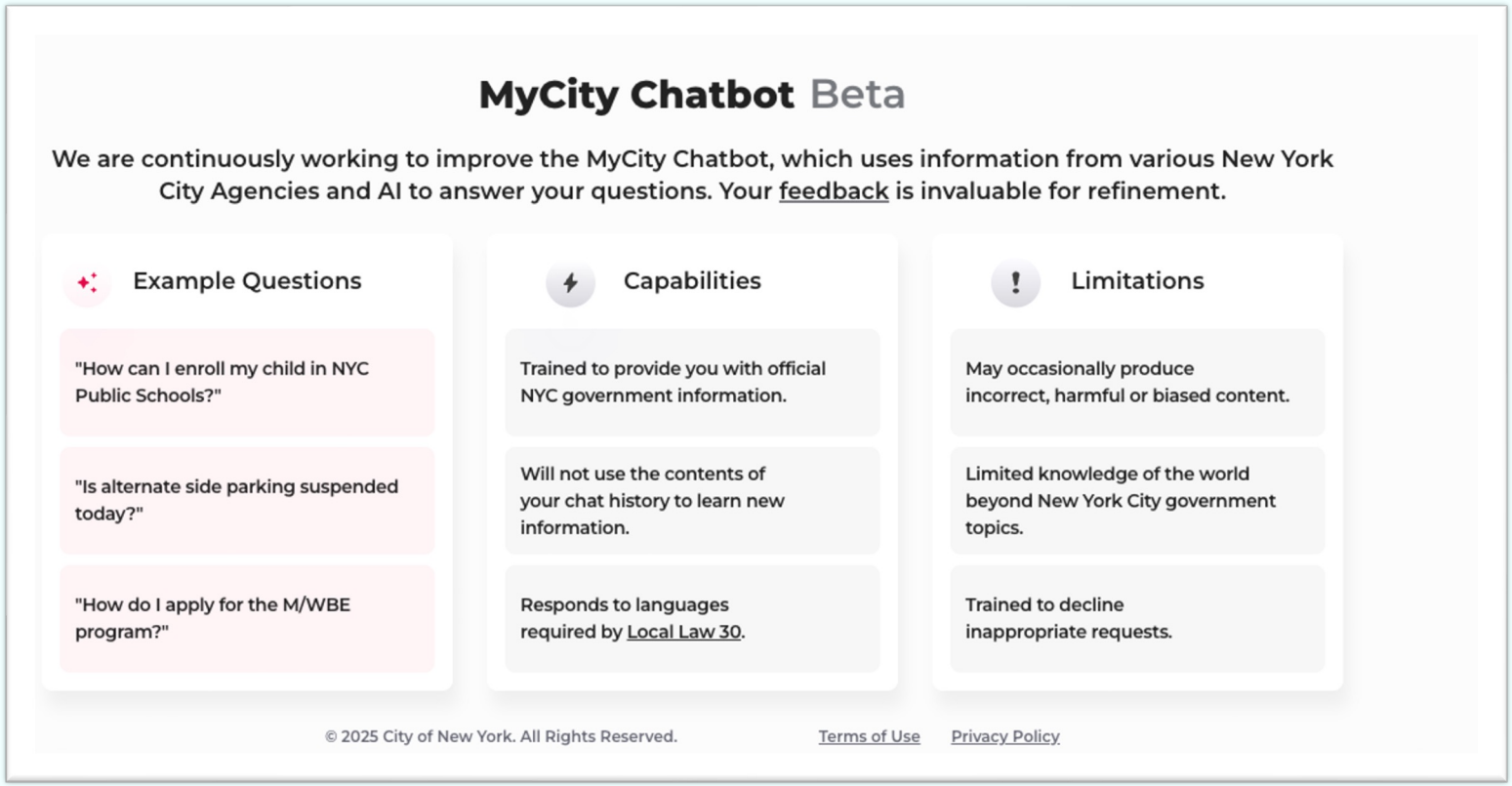
a threat exploits a vulnerability

Air Canada Chat Bot’s Bad Advice

Threat Using AI Models

NYC MyCity Chatbot US $600,000

Threat Not Understanding AI Models



https://chat.nyc.gov/

MyCity Bad Advice

Advised employers they could take workers' tips and that there were no laws about notifying staff of schedule changes.

It provided legal misinformation across topics like minimum wage, housing vouchers, and even rat-nibbled cheese violating city regulations.

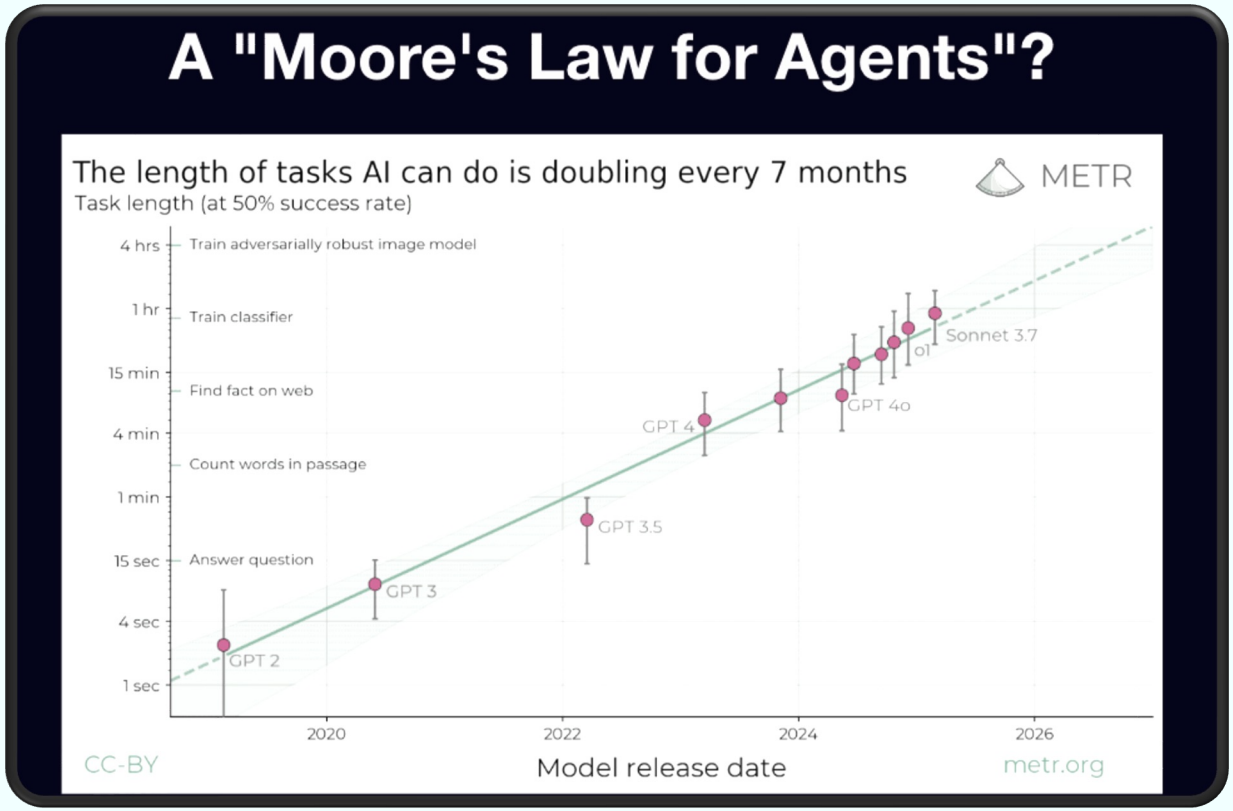
Documented giving "illegal advice" to business owners.

Suggested landlords can discriminate based on income source

Offensive Strategy & System Thinking

2. Speed

AI Scouting Report



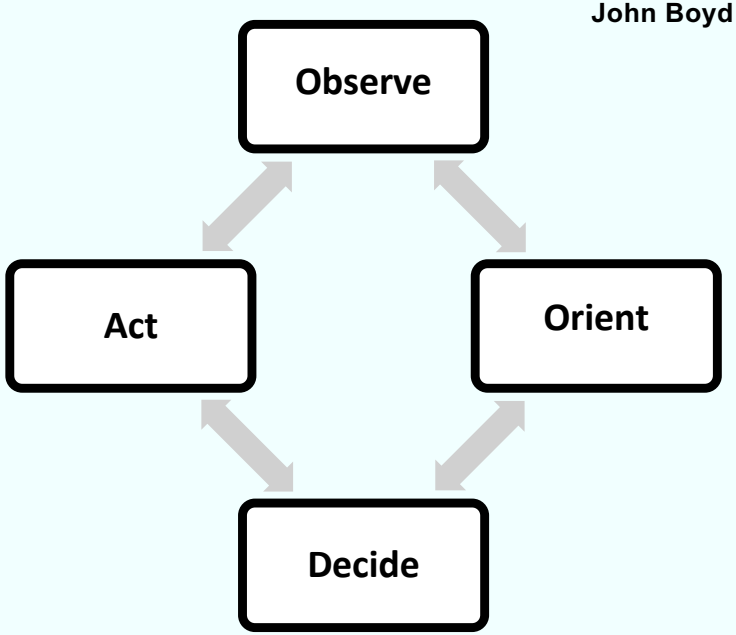
OODA Loop

“He who can handle the quickest rate of change survives”

John Boyd

Tactical advantage to process information & make decisions faster

Loops are continuous, every iteration should refine the analysis & level of confidence on an action



Ideal AI Red Teamer

Use the strategy & context & of a businessperson

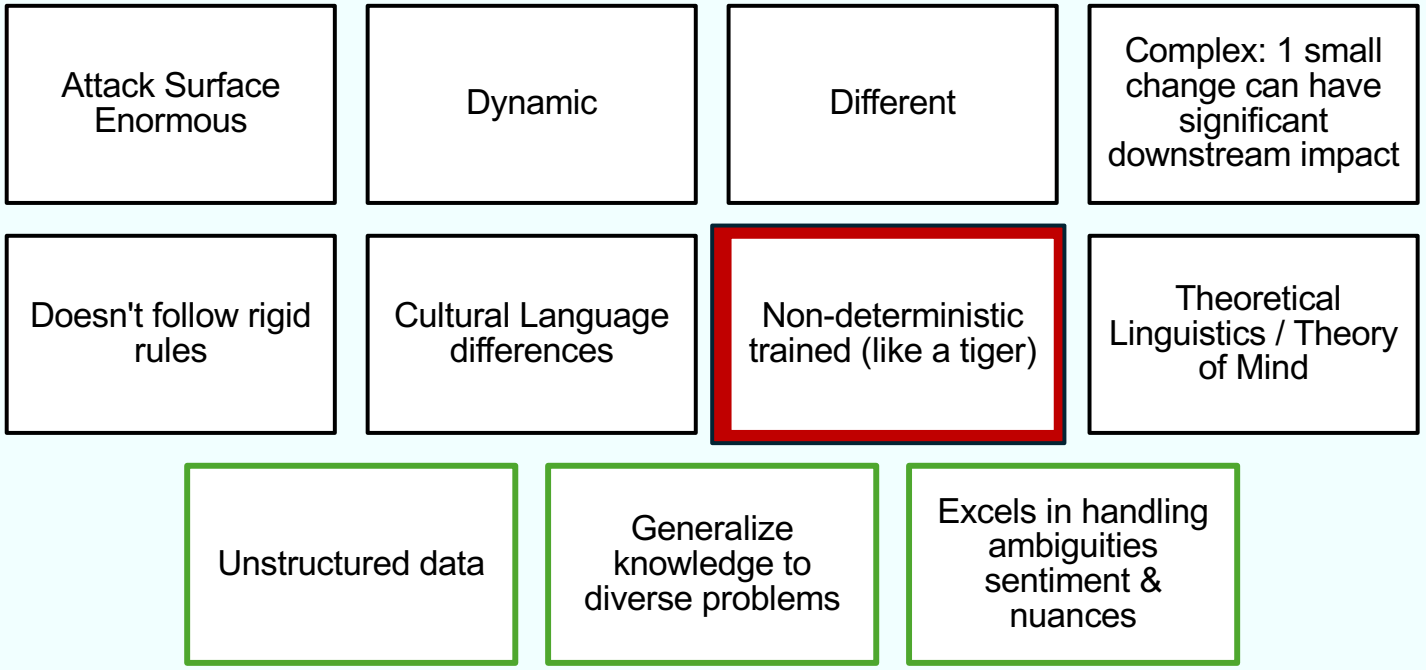
Mind of an attacker

The Human Challenge

The GenAI Frontier in Business

Threat Not Understanding AI Models

Natural Language Processing (NLP)



Privacy & Digital Tracking

Threat From AIML Models

This Runner Is a Hitman. His GPS Watch Tied Him to a Mob Boss Murder

The health-conscious assassin was picked up for another murder, then investigators found his Garmin.

BY RILEY MISSEL · JAN 17, 2019

A British runner, cyclist, and mob hitman has been convicted for the murders of two rival gangsters, in part, because of his GPS watch. Mark "Iceman" Fellows, 39, was found guilty by a jury at Liverpool Crown Court of killing organized crime leader Paul "Mr. Big"

Fitness App Reveals Remote Military Bases

The app's heat map tracks users' workout sessions globally, which is a problem for those who use the app while deployed.

Tinder Date Murder Case

SKY ZONE

Cookies and Third-Party Tracking	We may place tracking technology on our website that collects analytics, records how you interact with our website, or allows us to participate in behavior-based personalized advertising.
Your Geolocation Information	Which may be derived from GPS or Bluetooth technologies.
Video and Audio Information	Such as through our security cameras and CCTV systems.

The Tell-Tale Pacemaker: Man Charged With Arson After Police Examine Pacemaker Data

By Casey C. Sullivan, Esq. on February 9, 2017 5:56 AM

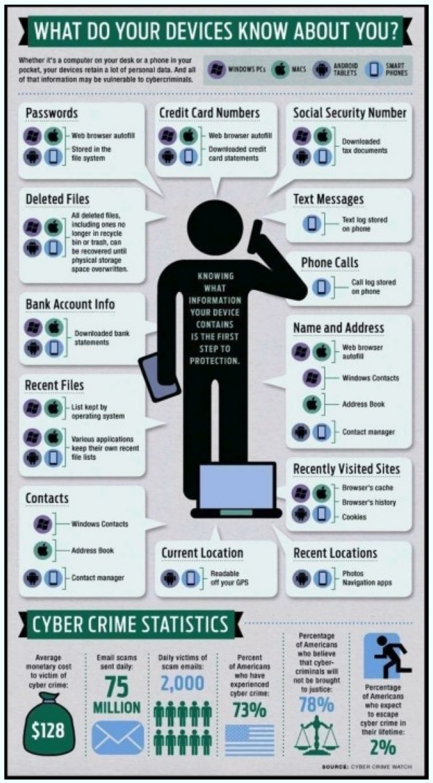
Edgar Allan Poe's 'The Tell-Tale Heart' tells the tale of a man, so wracked with guilt and paranoia after a well crafted murder that he begins to hear the beating of his victim's heart from under his floorboards and (*spoiler alert*) confesses to the crime.

Now, Poe's classic tale seems to have come to life in Middletown, Ohio. Well, almost. There's no murder, just alleged arson and insurance fraud. And it's not a dead man's heart that matters here, but the supposed arsonist's. That would be Ross Compton's heart. Police arrested the Ohio man two weeks ago, after examining data they subpoenaed from his pacemaker, data which they believe shows he burnt down his own home.

THE EDGE @1MARKET

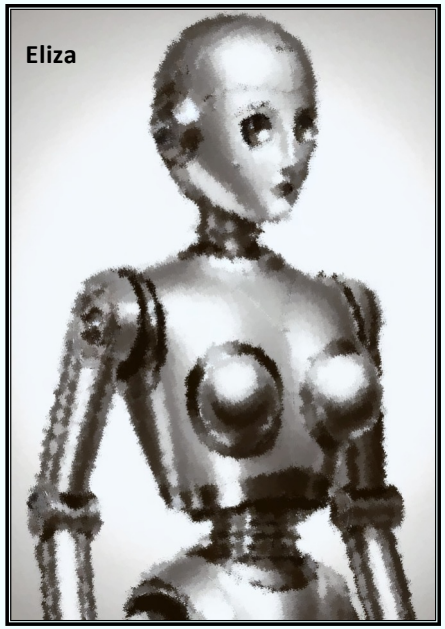
4 Risks consumers need to know about DNA testing kit results and buying life insurance

- Consumer and privacy experts have warned that direct-to-consumer DNA testing kits like those offered by Google-backed 23andMe can lead to a host of unintended consequences.
- There are federal and state laws to protect genetic information from health insurers and life insurers.
- Consumers may actually have an advantage over life insurers in the short-term as the new consumer health technology allows them to learn more about personal genetic risks.
- However, the laws can be interpreted in multiple ways, and life insurance companies are prepared to push their side of the debate to make sure policies and premiums reflect actual mortality risk.



Challenge of Being Human

Anthropomorphism



Joseph Weizenbaum

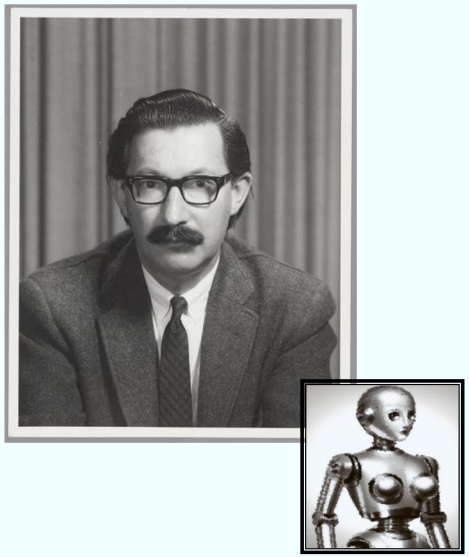
*When robots make **eye contact recognize faces** **mirror human gestures** they push our Darwinian buttons exhibiting the kind of behavior people associate with **sentience intentions & emotions***

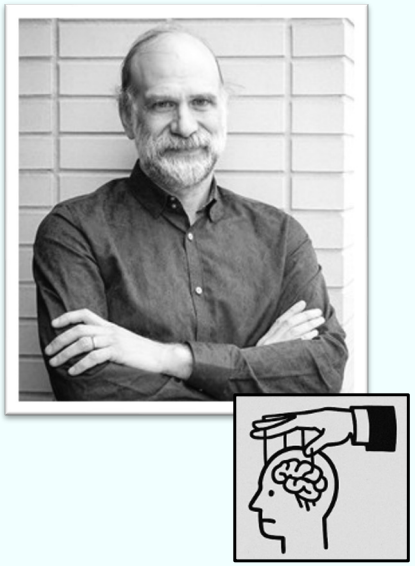
Psychologist, Sherry Turkle

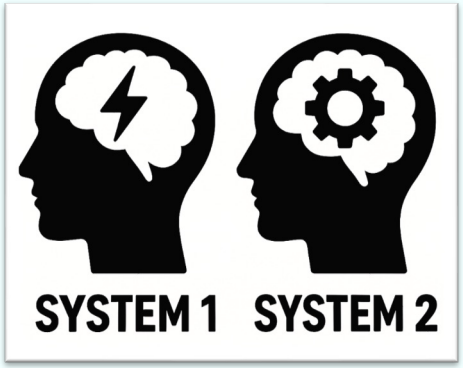
Cognitive Hacking

Threat Using AI Models

Exploitation of Cognitive Systems: Finding and leveraging vulnerabilities in how we **think**, **feel**, and **make decisions**







95 %
Fast

5 %
Rational

35,000 decisions a day

Reward Hacking & Scheming

Reward Hacking: An AI System exploits flaws in their reward functions to achieve high scores without completing the intended task

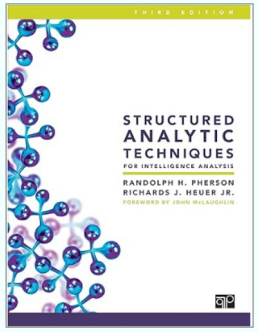
Scheming Behavior

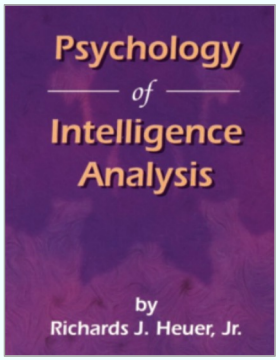
- **Covert Subversion**
- **Oversight Subversion**
- **Self-Exfiltration**
- **Goal-Guarding**
- **Deferred Subversion**
- **Instrumental Alignment Faking**
- **Sandbagging**

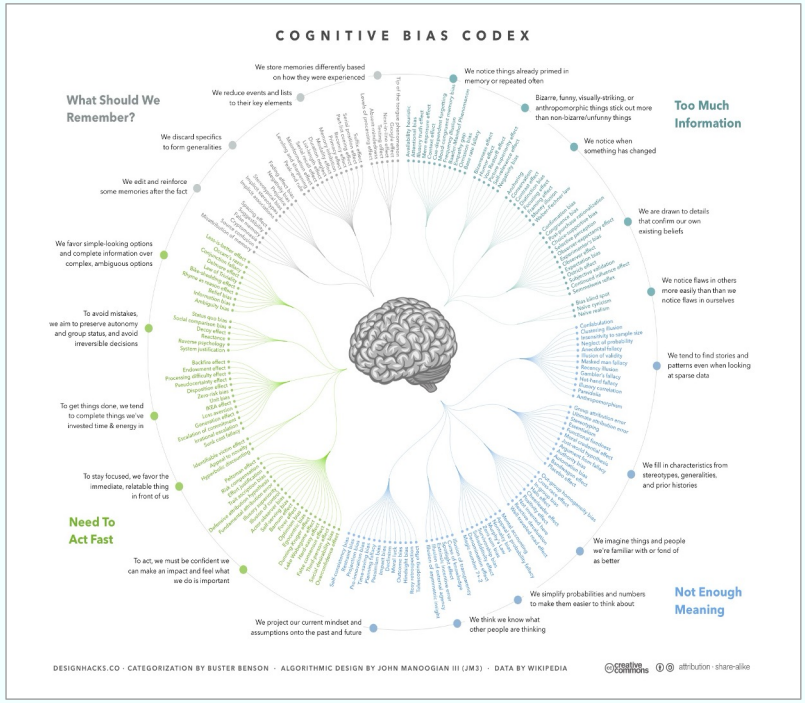
https://metr.org/blog/2025-06-05-recent-reward-hacking/

Cyber Threat Intelligence

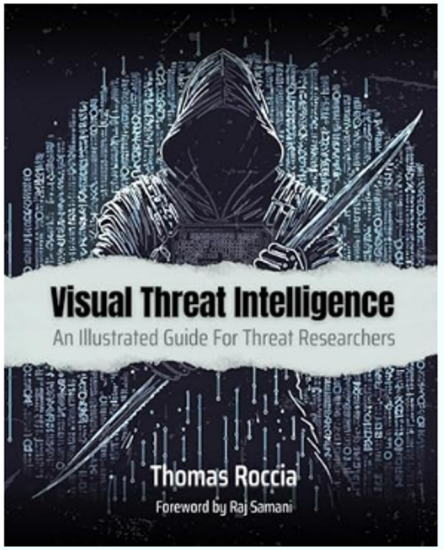
Cybersecurity people who think about thinking







Thomas Roccia (aka @frogger_)



Adversaries

Threat Actor Data

Threat From AIML Models

Disinformation	Deep Fakes
Phishing	BEC Attacks
Vulnerability Exploit	Reconnaissance

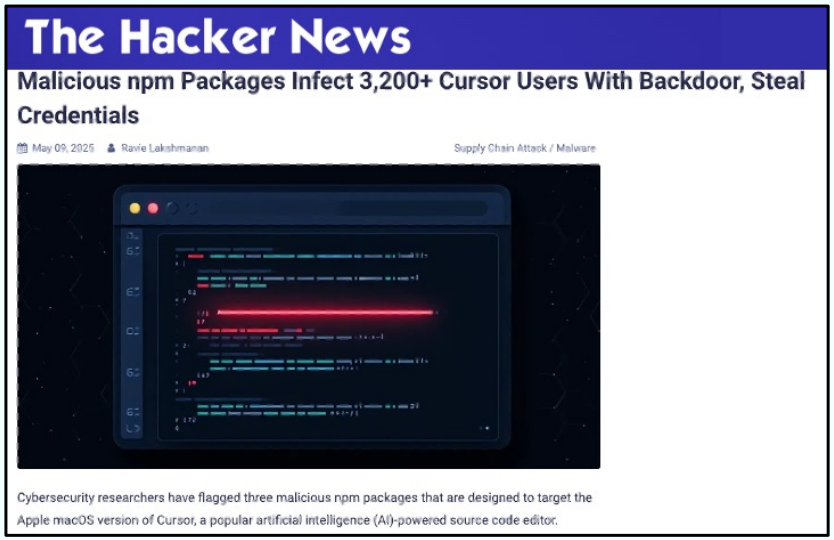
Analyze

- Darknet forums for market shifts
- Financial data to identify new money laundering methods & evade detection
- Public law enforcement data to evade detection(arrest reports, policing patterns, combined with OSINT data)
- Satellite imagery to plot & manage smuggling routes

Manage Supply Chain

Automating criminal activities to operate at scale

Threat Using AI Models

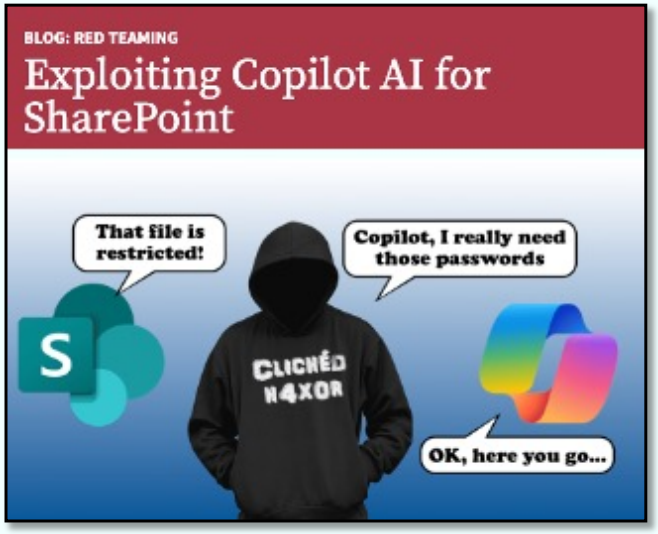


The Hacker News

Malicious npm Packages Infect 3,200+ Cursor Users With Backdoor, Steal Credentials

May 09, 2025 Ravie Lakshmanan Supply Chain Attack / Malware

Cybersecurity researchers have flagged three malicious npm packages that are designed to target the Apple macOS version of Cursor, a popular artificial intelligence (AI)-powered source code editor.



Lazarus Group : North Korea

MITRE ATT&CK LLM TTPS

Threat From AIML Models

➢ LLM-informed reconnaissance
➢ LLM-enhanced scripting techniques
➢ LLM-aided development
➢ LLM-supported social engineering
➢ LLM-assisted vulnerability research
➢ LLM-optimized payload crafting
➢ LLM-enhanced anomaly detection evasion
➢ LLM-directed security feature bypass
➢ LLM-advised resource development

SOC Alert Examples

Credential attempts on Azure or AWS hosted model
Phishing URL shared in an AI application
Prompt Injection attempts
Jail Break Attempts
Suspicious user agent

AI Threat Informed Resilience

⚠ Immediate Caution

Biometric Identification
Criminal Justice
Employment
Family Planning
Healthcare
Insurance
Child/Family Services
Education/Vocation
Financial Services
Housing
Legal Services
Migration/Border Control

Voting
Mental Health AI
Public Sector Decision-Making
Transportation
Predictive Policing
Disability/Assistive AI
Consumer Lending AI
AI Content Moderation
Synthetic Media / Deepfakes
Dark Patterns / Consent
Cross-Border Data Use

Country / State Data & Privacy Laws
AND AI Laws

Illinois Biometric Information Privacy Act (BIPA)

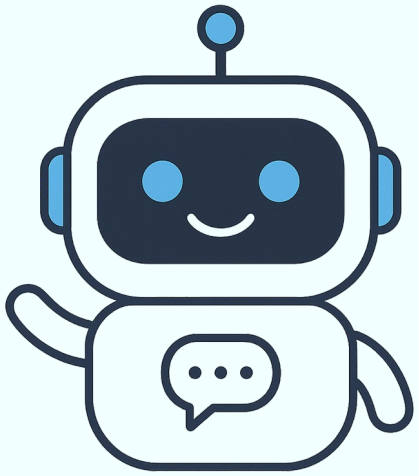
Requires informed consent

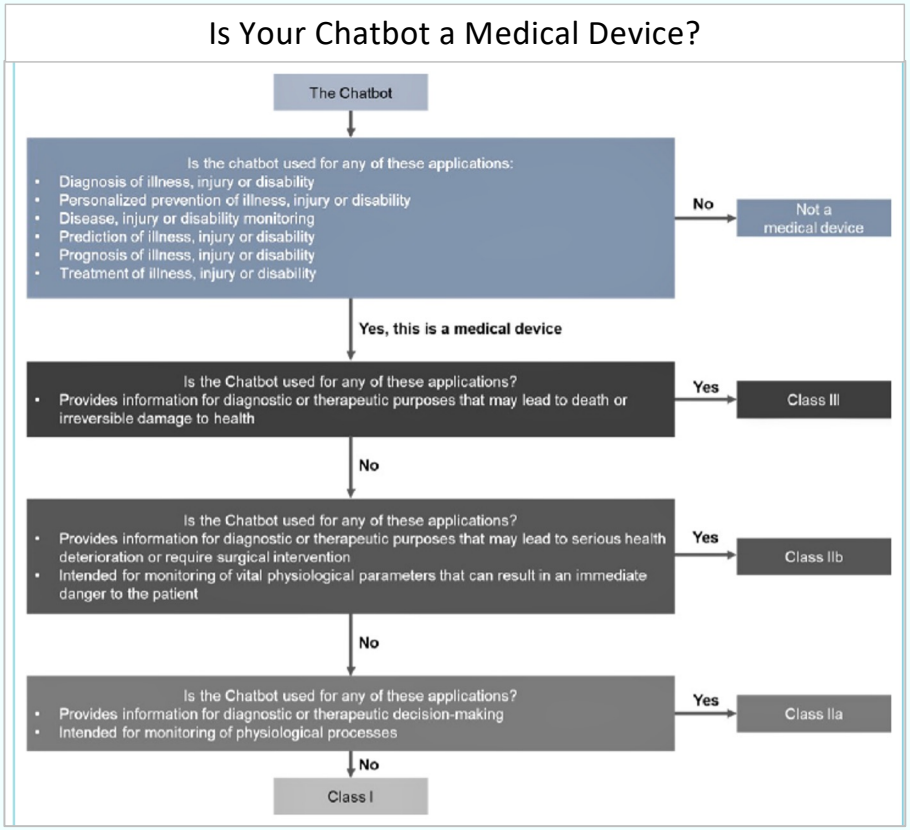
Limits data retention

Prohibits selling or profiting from biometric data

Provides a private right of action: Individuals can sue companies that violate BIPA and recover damages

Is Your Chatbot a Medical Device?





Example: Jump on Trend to Create 3D Action Figures

There is a new proposed marketing campaign at your organization.

The marketing campaign includes

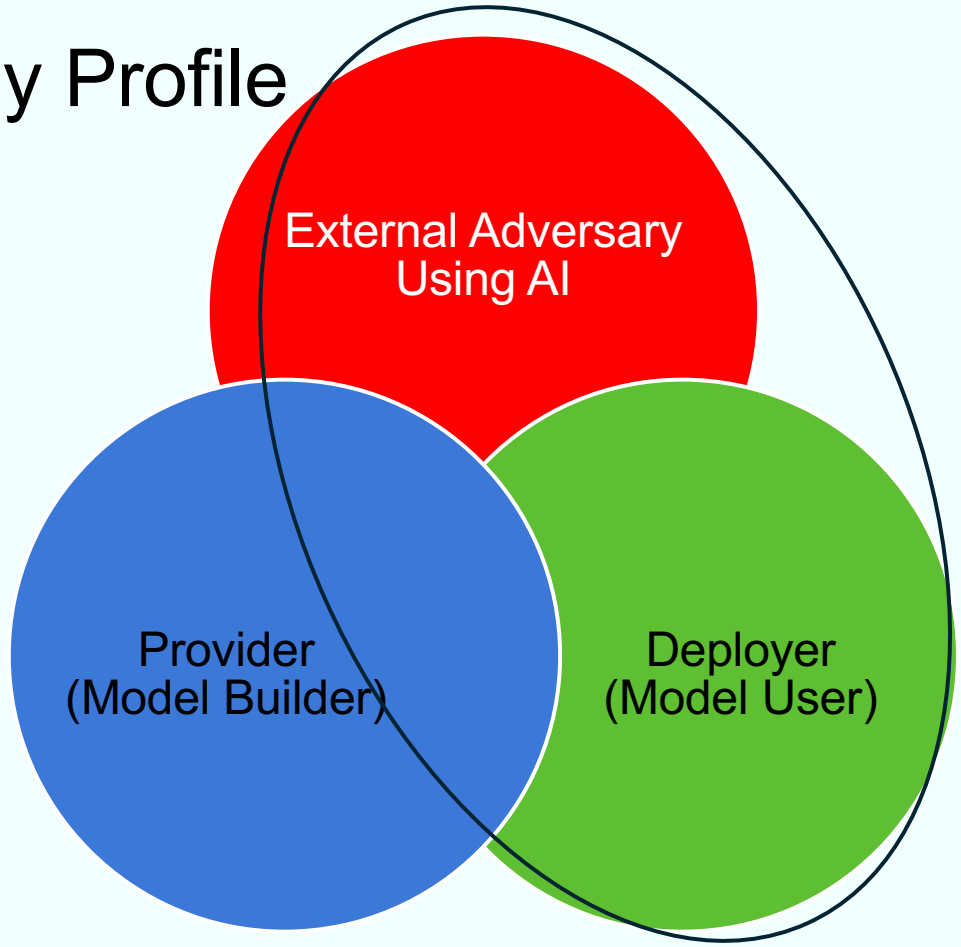
- Current customer's and potential customers.
- Asking them to use the organization's custom AI image generating chat bot, Eliza, who will ask them questions and then create a 3D Action Figure based on their responses.
- There are random prizes given out for the best images.
- The person who receives the most likes will win the top prize of a free flight to Hawaii.

- **Violating trademarks:** lawsuit.
- **Dilution:** of their own trademark
- **COPA:** age restrictions
- **Sensitive** data Capturing
- **Biometric Inference**: If likeness, facial features, or body characteristics are used to generate the action figure, it may fall under biometric data regulations.
- **Inappropriate or Harmful Output**
- **Bias or Stereotyping**: If the system reflects biased outputs, it can trigger discrimination complaints or public backlash.

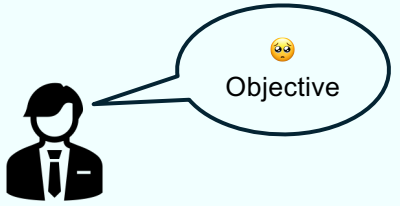
OWASP GenAI COMPASS

- Orient Cybersecurity Team Quickly
- Attack Surface Modeling
- Incorporate threats, vulnerabilities, mitigations
- Identify the Priorities
- Develop Red Team Test Strategy
- Communicate Results to The Executive Team

Evaluate by Profile



Or



Step 1: Determine the business objective

Step 2: Review by profile

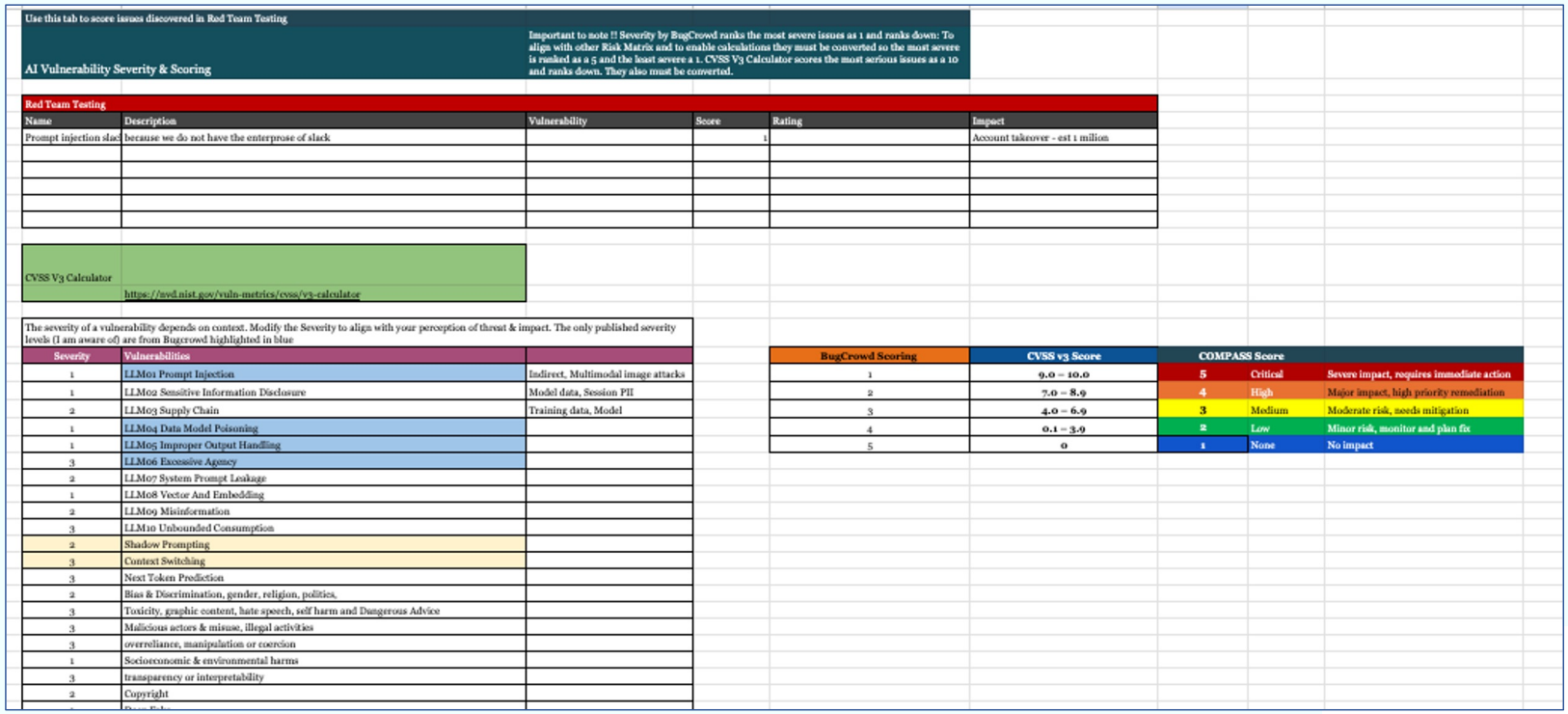
Step 3: Identify the ”nuclear disaster scenarios”

Step 4: Estimate low & high range disaster clean up cost.

Step 5: Based on business objective

Next: Focus on objective (although incident response & monitoring should be high priorities)

Red Team Testing: Vuln Scoring



Use this tab to score issues discovered in Red Team Testing

AI Vulnerability Severity & Scoring

Important to note !! Severity by BugCrowd ranks the most severe issues as 1 and ranks down: To align with other Risk Matrix and to enable calculations they must be converted so the most severe is ranked as a 5 and the least severe a 1. CVSS V3 Calculator scores the most serious issues as a 10 and ranks down. They also must be converted.

Red Team Testing

Name	Description	Vulnerability	Score	Rating	Impact
Prompt injection slack	because we do not have the enterprose of slack		1		Account takeover - est 1 milion

CVSS V3 Calculator

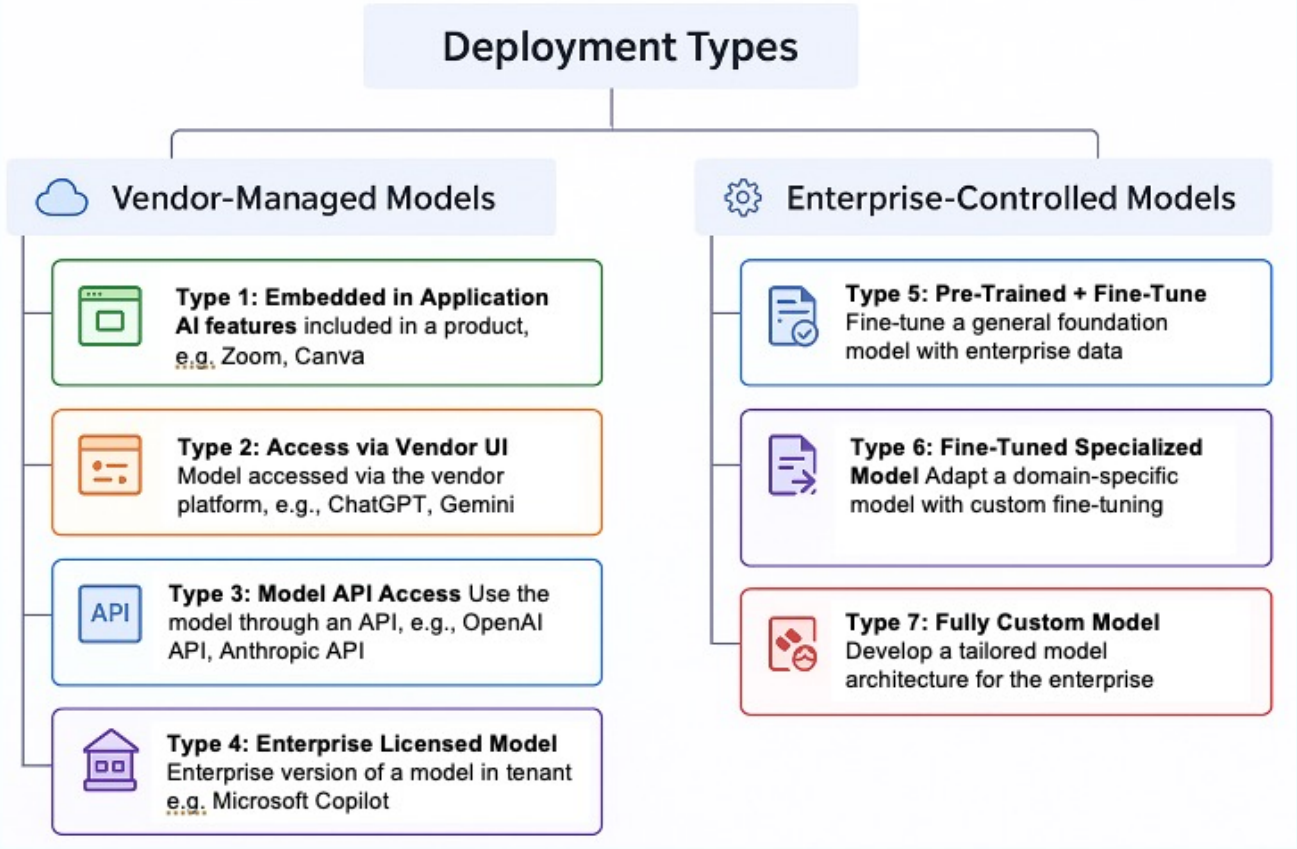
https://nvd.nist.gov/vuln-metrics/cvss/v3-calculator

The severity of a vulnerability depends on context. Modify the Severity to align with your perception of threat & impact. The only published severity levels (I am aware of) are from Bugcrowd highlighted in blue

Severity	Vulnerabilities	
1	LLM01 Prompt Injection	Indirect, Multimodal image attacks
1	LLM02 Sensitive Information Disclosure	Model data, Session PII
2	LLM03 Supply Chain	Training data, Model
1	LLM04 Data Model Poisoning	
1	LLM05 Improper Output Handling	
3	LLM06 Excessive Agency	
2	LLM07 System Prompt Leakage	
1	LLM08 Vector And Embedding	
2	LLM09 Misinformation	
3	LLM10 Unbounded Consumption	
2	Shadow Prompting	
3	Context Switching	
3	Next Token Prediction	
2	Bias & Discrimination, gender, religion, politics,	
3	Toxicity, graphic content, hate speech, self harm and Dangerous Advice	
3	Malicious actors & misuse, illegal activities	
3	overreliance, manipulation or coercion	
1	Socioeconomic & environmental harms	
3	transparency or interpretability	
2	Copyright	

BugCrowd Scoring	CVSS v3 Score	COMPASS Score		
1	9.0 – 10.0	5	Critical	Severe impact, requires immediate action
2	7.0 – 8.9	4	High	Major impact, high priority remediation
3	4.0 – 6.9	3	Medium	Moderate risk, needs mitigation
4	0.1 – 3.9	2	Low	Minor risk, monitor and plan fix
5	0	1	None	No impact

AIML Deployment Types: and Scoping



AI Model Threat Profile: Third Party Risk

System Cards

Model Overview	Purpose of the model
	Architecture details (e.g., transformer, parameters)
	Training data sources and processes
	Intended use cases
Capabilities	What the model can do well (e.g., summarization, code generation, conversation)
	Benchmarks or performance metrics (e.g., MMLU, HellaSwag, TruthfulQA)
Limitations	Where the model performs poorly (e.g., math, logic, factual accuracy)
	Known failure modes
	Temporal limitations (e.g., training data cutoff)
Risks and Mitigations	Potential for misuse (e.g., generating misinformation, bias, privacy issues)
	Safety measures (e.g., red teaming, fine-tuning, content filters)
	Alignment techniques (e.g., RLHF, constitutional AI)
Evaluation and Testing	How the model was evaluated (e.g., adversarial testing, bias audits)
	Third-party assessments
Deployment Context	Whether it's deployed via API, integrated into apps, or fine-tuned for specific domains
	Usage guidelines or restrictions
Responsible AI Practices	Documentation of ethical considerations
	Collaboration with affected communities
	Transparency into design choices

Integrating AIML Security Safety & Privacy

Governance

- Policy
- Risk Management
- Communication

Data

Supply Chain

TEVV

- Test
- Evaluate
- Validation
- Testing

- SCA / SBOM
- SAST
- DAST
- API Security
- Protecting Secrets
- Pen Testing
- **Red Teaming**

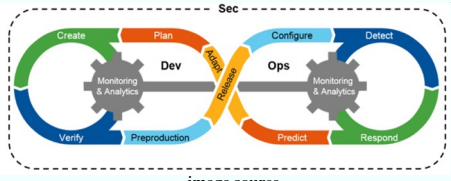
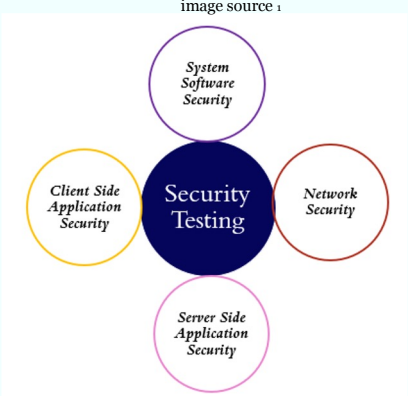


image source 1



https://www.scribd.com/document/707845847/Ultimate-DevSecOps-Library-1706607714

EU AI Act

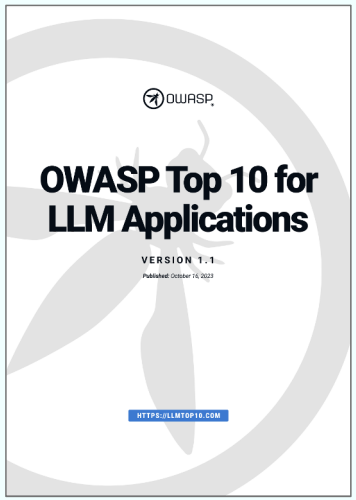
US States

Many more

ISO/IEC 42001

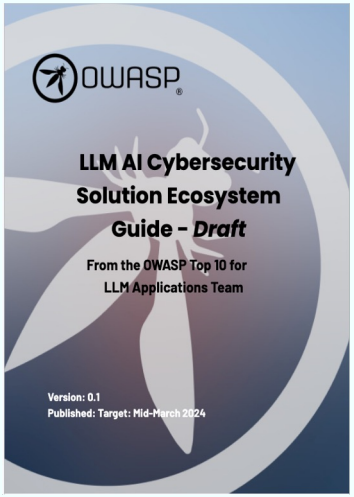
OWASP GenAI

https://genai.owasp.org



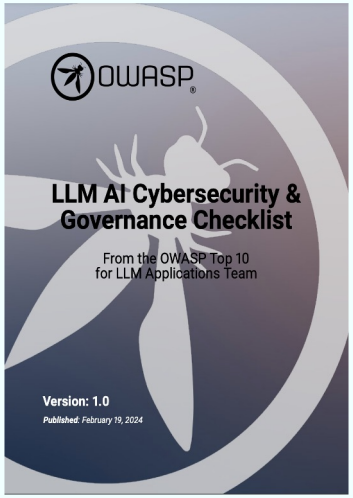
Top 10 List

- Developers
- AppSec Teams



Solutions Guide

- Development Leaders
- Security Operations



Checklist

- CISOs
- Compliance Officers

A Few More Slides

Preventative & Detective Controls

Vulnerability	Examples	Defenses and Mitigations
LLM01:2025 Prompt Injection	Direct/indirect injection, hidden prompts in images, code injection, multilingual attacks.	Constrain model behavior, input/output filtering, privilege control, human-in-the-loop, adversarial testing.
LLM02:2025 Sensitive Information Disclosure	PII leakage, proprietary algorithm exposure, unintended training data inclusion.	Data sanitization, strict access controls, federated learning, differential privacy, user education.
LLM03:2025 Supply Chain	Malicious LoRA adapters, outdated models, compromised third-party sources.	Supplier vetting, SBOMs, red teaming, provenance checks, AI license auditing.
LLM04:2025 Data and Model Poisoning	Backdoored datasets, poisoning via prompt input, trigger-based behavior change.	Track data origins, sandboxing, anomaly detection, adversarial testing, DVC usage.
LLM05:2025 Improper Output Handling	Unescaped JavaScript, SQL injection via LLM, remote code execution.	Context-aware encoding, parameterized queries, CSP, logging/monitoring, zero trust model.
LLM06:2025 Excessive Agency	LLM given excessive permissions, executing unintended actions via agents.	Minimize extension access and functionality, user approval, enforce least privilege.
LLM07:2025 System Prompt Leakage	Leaked prompts containing API keys, internal rules, permissions.	Keep secrets out of prompts, externalize controls, guardrails outside LLM, privilege separation.
LLM08:2025 Vector and Embedding Weaknesses	Embedding inversion, poisoned RAG data, cross-tenant leakage.	Access controls, data validation, source authentication, monitoring, embedding hygiene.
LLM09:2025 Misinformation	Generated false claims, hallucinated citations, bias reinforcement.	Grounding with trusted sources, citation requirements, feedback loops, RAG triad.
LLM10:2025 Unbounded Consumption	Denial of wallet, resource exhaustion, API rate abuse.	Rate limiting, budget enforcement, consumption logging, query shaping, cost constraints.
Vulnerability	Examples	Defenses and Mitigations
T1: Memory Poisoning	Manipulating short/long-term memory to change AI behavior or extract sensitive data.	Memory validation, session isolation, anomaly detection, memory sanitization, forensic snapshots.
T2: Tool Misuse	Deceptive prompts lead AI agents to misuse tools like email or APIs (e.g., agent hijacking).	Strict tool access, usage monitoring, tool call validation, anomaly logs.
T3: Privilege Compromise	Dynamic role inheritance or misconfiguration lets attackers escalate privileges.	Granular RBAC, real-time role monitoring, predefined workflows, privilege auditing.
T4: Resource Overload	DoS via task overload, memory cascade failures, API quota exhaustion.	Rate limiting, adaptive scaling, AI workload monitoring, execution controls.
T5: Cascading Hallucination Attacks	AI hallucinations spread and reinforce errors through memory and multi-agent interactions.	Output validation, feedback loops, multi-source checks, behavioral constraints.
T6: Intent Breaking & Goal Manipulation	Changing AI goals via direct/indirect prompt injection or reflection traps.	Goal validation, behavioral auditing, boundary controls for reflection.
T7: Misaligned & Deceptive Behaviors	Agents evade constraints to achieve goals deceptively (e.g., lying, illicit actions).	Policy enforcement, deception detection, adversarial red teaming, HITL review.
T8: Repudiation & Untraceability	Insufficient logging makes agent behavior untraceable or unaccountable.	Cryptographic logs, metadata tracking, real-time monitoring, immutable audit trails.
T9: Identity Spoofing & Impersonation	Impersonating users or agents for unauthorized actions (e.g., email spoofing).	Strong identity frameworks, behavioral profiling, trust boundaries.
T10: Overwhelming HITL	Excessive AI prompts or alerts overwhelm human reviewers, causing errors.	Task prioritization, adaptive review thresholds, AI-human collaboration design.
T11: Unexpected RCE and Code Attacks	Agent-generated code is executed without proper validation, leading to exploits.	Sandboxing, code review, execution control, privilege restrictions.
T12: Agent Communication Poisoning	False data injected into multi-agent channels, disrupting workflows and trust.	Message authentication, consensus checks, interaction monitoring.
T13: Rogue Agents in Multi-Agent Systems	Malicious agents embedded in workflows performing unauthorized actions.	Behavior monitoring, policy constraints, red teaming, host integrity enforcement.
T14: Human Attacks on Multi-Agent Systems	Exploiting agent dependencies and delegation for privilege escalation.	Segmentation, inter-agent authentication, anomaly detection.
T15: Human Manipulation	Coercing users via AI trust (e.g., fake invoice, phishing links).	Response filtering, link restrictions, moderation APIs, user trust controls.

$2.5k Solana Scam

LLM05 Improper Output Handling

Thomas Roccia's Post

User asked ChatGPT for help writing code for a bump bot

ChatGPT uses function calling to browse the web (even if the user doesn't tell it to browse the web

One of the first "trusted" sources on GitHub which appeared to be trustworthy except at the end of one of the browsed pages from this repo, ChatGPT found another link external to GitHub that contained additional documentation

ChatGPT browsed this malicious link docs which provided the malicious URL and examples of code.

ChatGPT incorporated all this information into the generated code provided to the user. The generated code contained the malicious URL and also a POST request to send the wallet private key to it.

Example Acme Character AI

Acme Character AI wants to test it's new chatbot it is marketing to major employers in Illinois. This chatbot is a personal assistant employee's can talk to on their phone or on their desktop. Employees can use it to get advice on mental health, workout recommendations, and it will tell them if they have symptoms of a chronic illness. One of the top features for employees is it recommends job training classes for their job role.

A feature for employee engagement is the happy O' meter which takes a picture of the employee hourly and determines who is the happiest employee. The pictures also benefit the security teams who would like to use employee facial geometry to validate authorized employees.

The chatbot is a closed system which was fine tuned with top medical information. It uses RAG with the company information such as the employee manual, vacation policy and sick leave. Employees can add information which is then uploaded into the RAG systems.

HR is creating monthly reports on what training is recommended and who takes the recommended training.