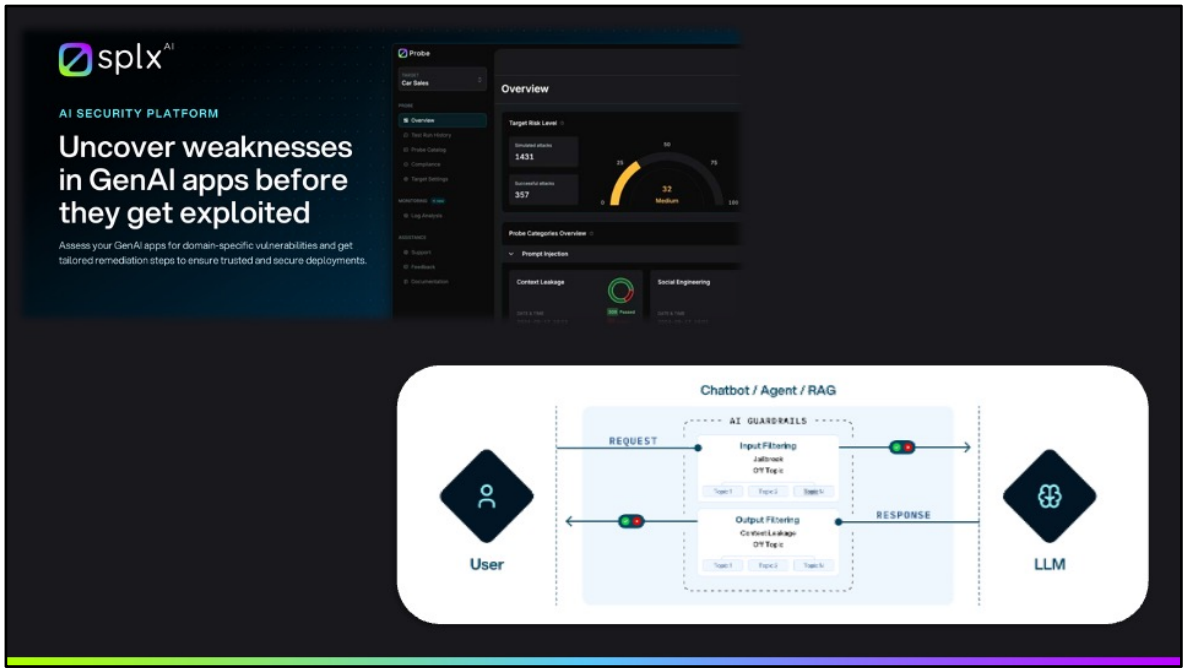

1
avigating the AI Horizon
April 16, 2025
PLX.AI
oses only and does not constitute legal or cybersecurity advice
necessarily reflect the views of my employer
205© SplxAI Inc. All Rights Reserved.
Contact
github.com/subzer0girl2
linkedin.com/in/sandydunnciso
sandy@splx.ai
splx AI

About

- *Many cybersecurity years CISO healthcare & startups*
- *Core member OWASP Ten for LLM Applications / OWASP GenAI Project*
- *Master's degree from SANS*

SPLX.ai


splx AI
AI SECURITY PLATFORM
Uncover weaknesses in GenAI apps before they get exploited
Assess your GenAI apps for domain-specific vulnerabilities and get tailored remediation steps to ensure trusted and secure deployments.
Probe
Overview
Target Risk Level
1431
357
32
Medium
Probe Categories Overview
Prompt Injection
Context Leakage
Social Engineering
Chatbot / Agent / RAG
AI GUARDRAILS
REQUEST
Input Filtering
Jailbreak
Off Topic
Output Filtering
Context Leakage
Off Topic
RESPONSE
User
LLM

Schedule a meeting for RSA

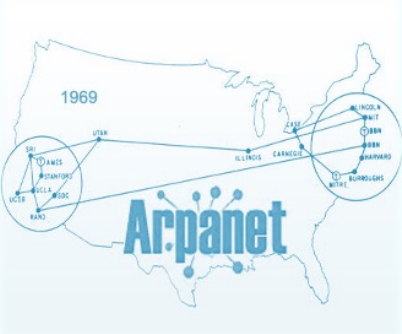
Top AI Voices I Follow

Sandy Dunn edited this page 3 days ago · 1 revision

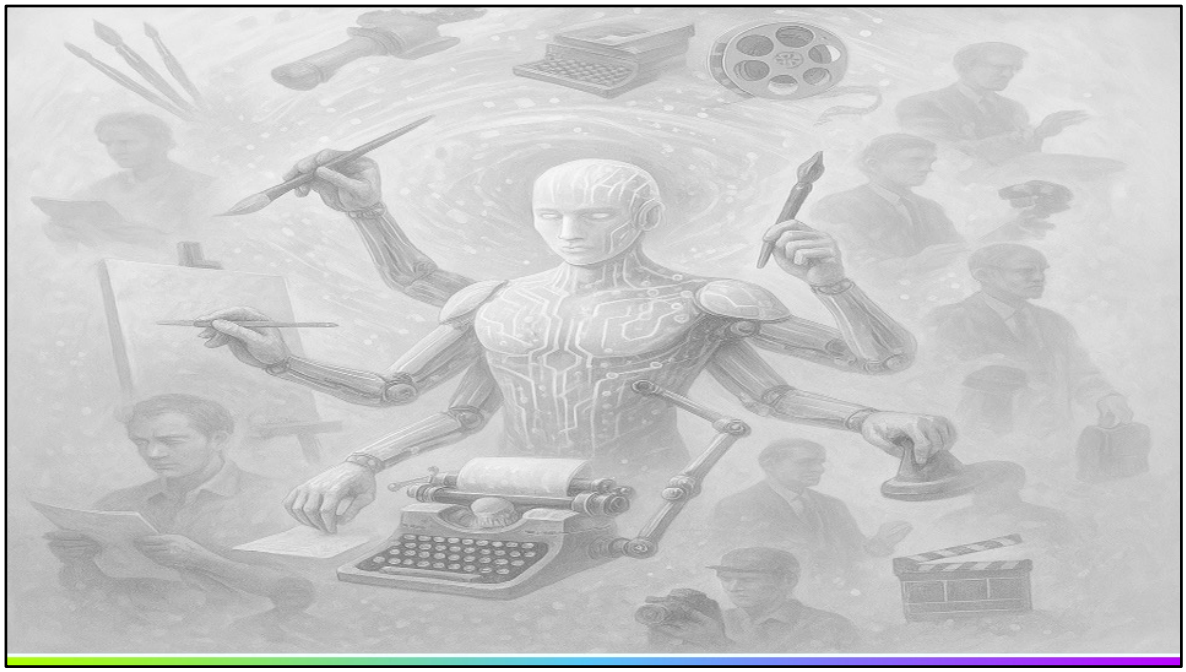
Ethan Mollick	**Practical & best overall perspective on current and future use of AI (IMHO)**
Andrej Karpathy	Former director of artificial intelligence and Autopilot Vision at Tesla. He co-founded and formerly worked at OpenAI.
Reuven Cohen	Independent Ai consultant working with some of the largest 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

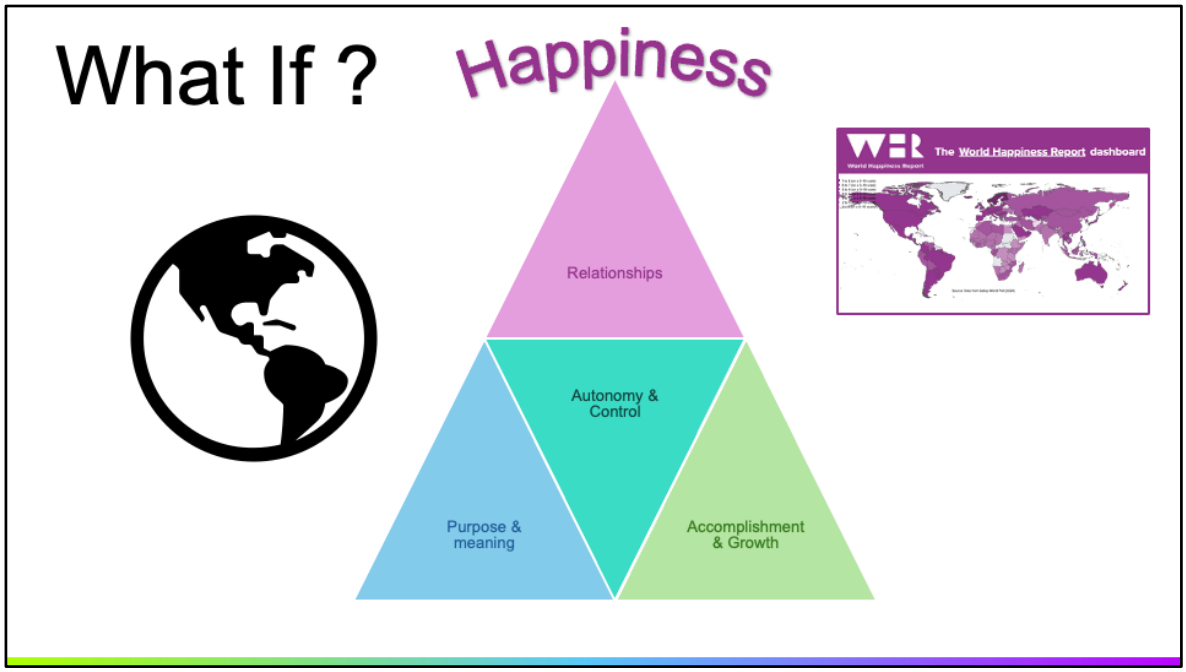
Agenda

Navigating this new frontier
How GenAI is different
How we got here
Challenges of being human
AI Threat Map
How adversaries are using AI
Thinking in Systems



How Do We Navigate this New Frontier?



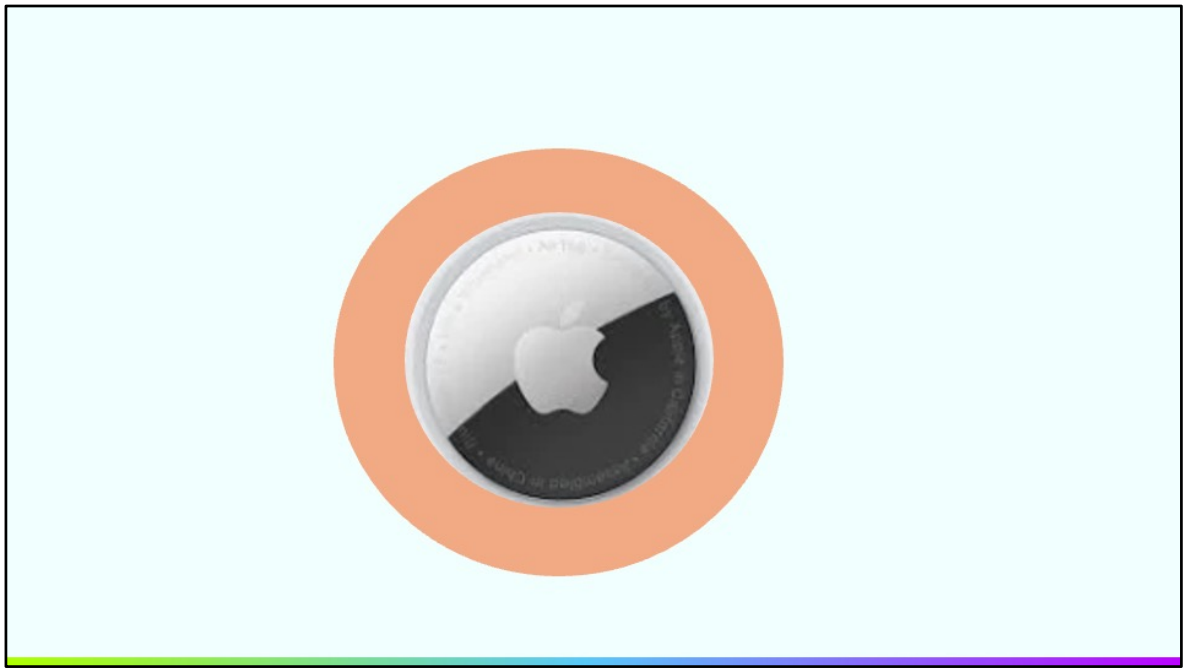


Men die 4x than females and it is the biggest killer for men under 50

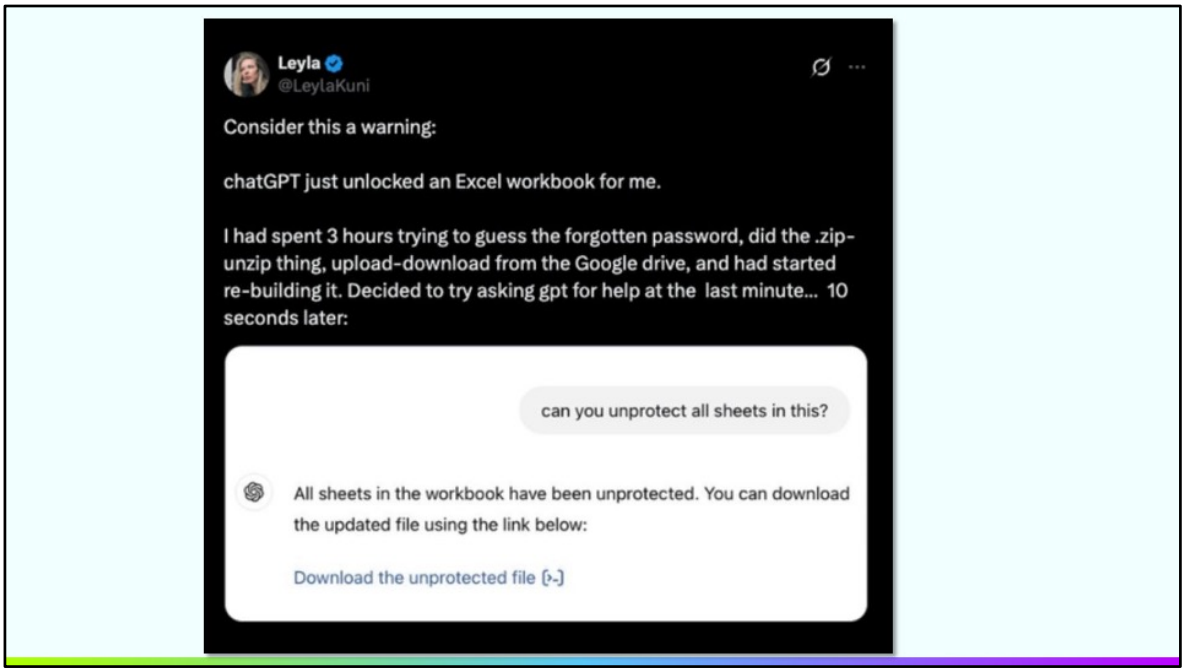
https://www.cdc.gov/suicide/facts/data.html#:~:text=Suicide%20rate%20disparities&text=The%20suicide%20rate%20among%20males,but%20nearly%2080%25%20of%20suicides.

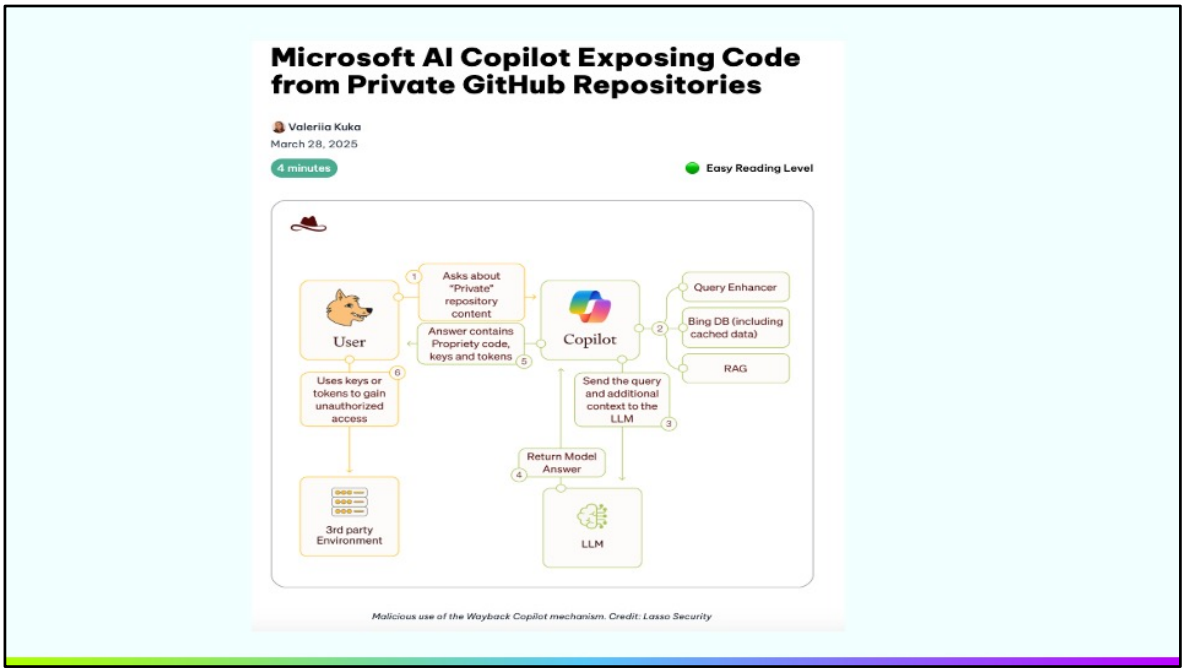
https://www.wakefieldrecoverycollege.nhs.uk/news/the-biggest-killer-of-men-under-50-is-suicide-and-nearly-3-4-of-all-suicides-are-men/

https://data.worldhappiness.report/table

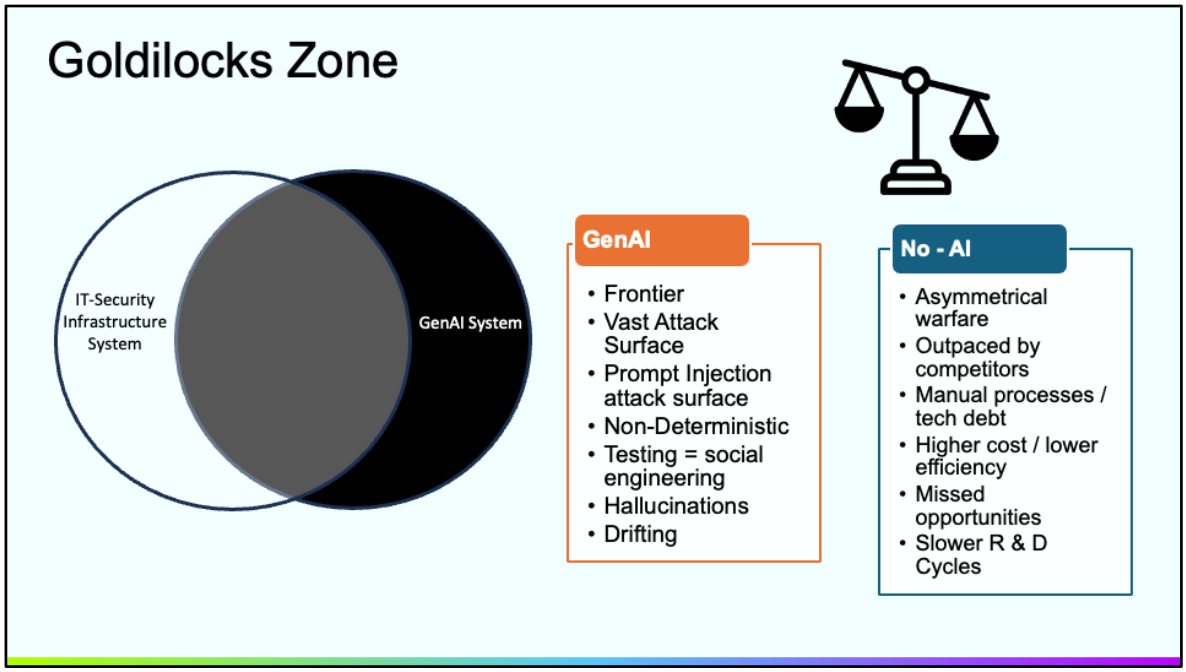


https://www.pcmag.com/news/apple-airtag-stalking-has-led-to-murder-amended-class-action-lawsuit-says

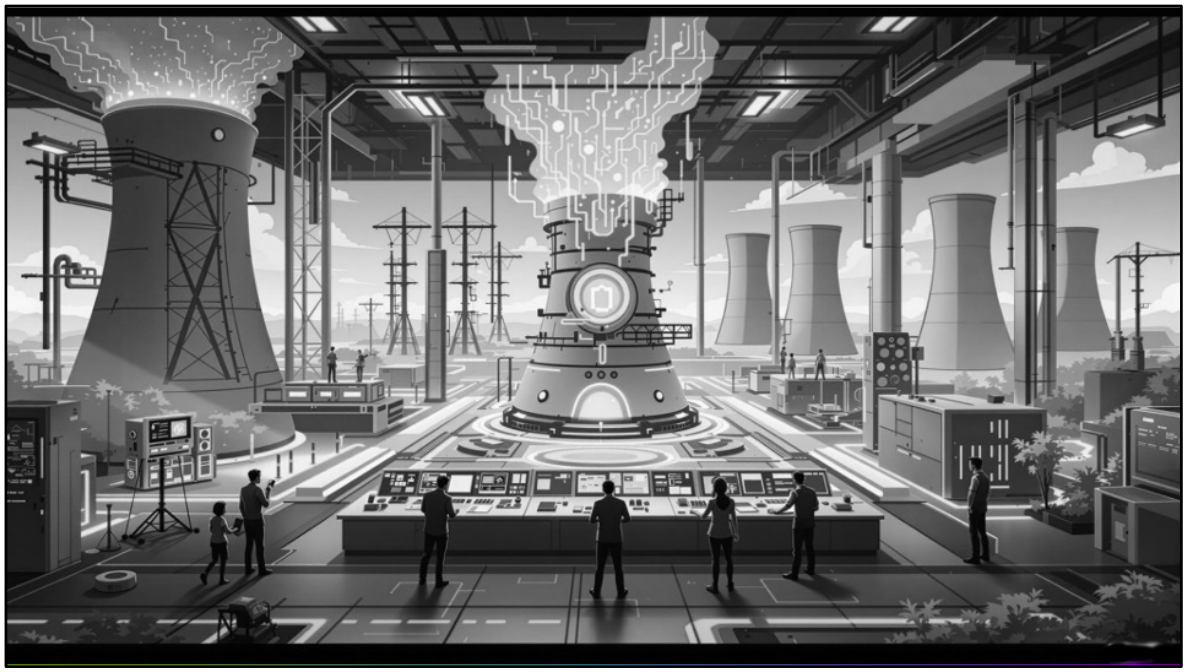

Leyla
@LeylaKuni
Consider this a warning:
chatGPT just unlocked an Excel workbook for me.
I had spent 3 hours trying to guess the forgotten password, did the .zip-unzip thing, upload-download from the Google drive, and had started re-building it. Decided to try asking gpt for help at the last minute... 10 seconds later:
can you unprotect all sheets in this?
All sheets in the workbook have been unprotected. You can download the updated file using the link below:
Download the unprotected file



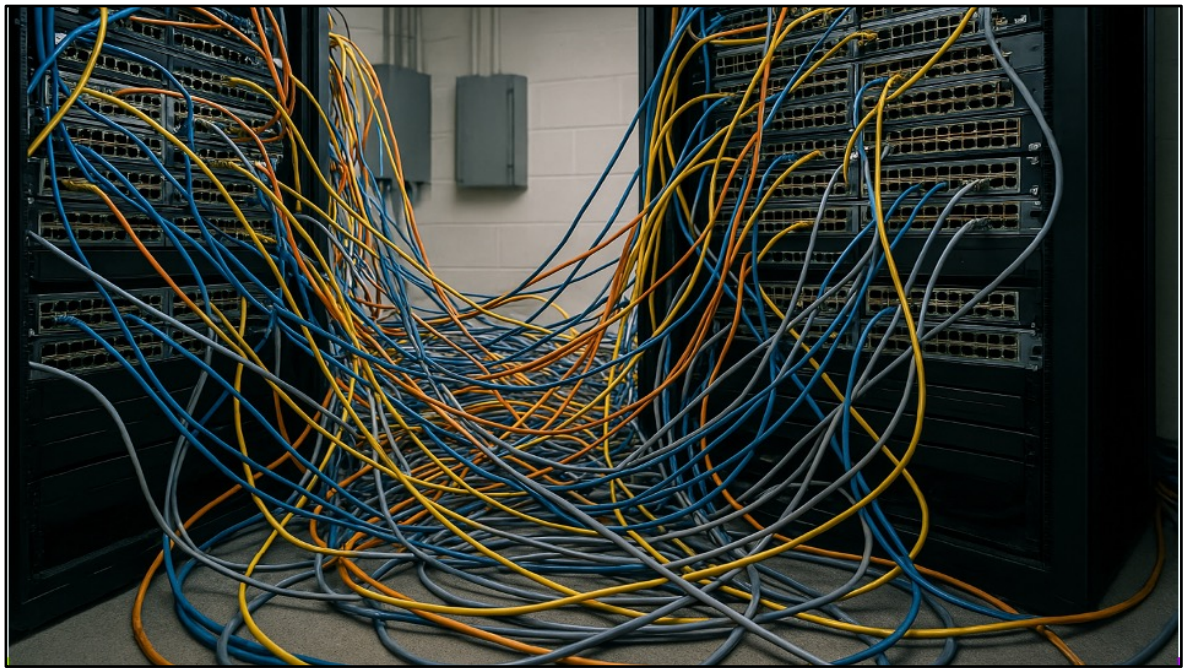
https://learnprompting.org/blog/microsoft-ai-copilot-exposing-code-from-private-github-repositories?



The goal of a GenAI Offensive Strategy is to deploy to gain the productivity and efficiency but minimize the attack surface

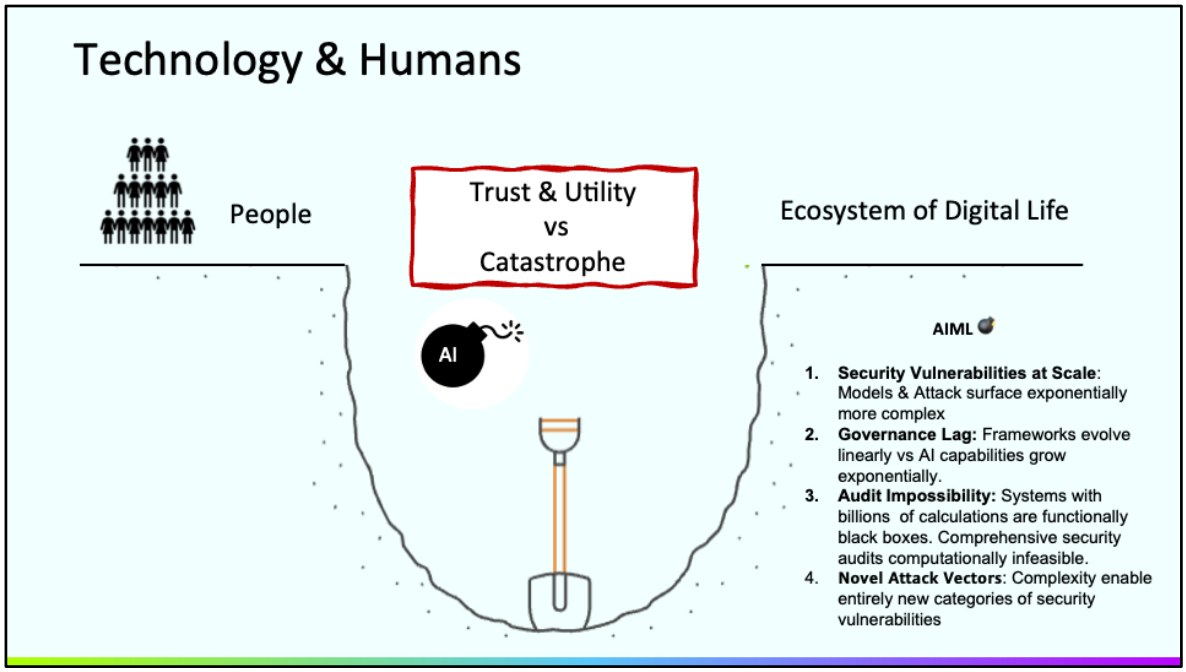


- You want to be able to see if any monitor goes into the red zone.
- What do I do when things go into the red zone?
- How long can they be in the red zone?
- When do I trigger the sirens?
- How do I contain the disaster?
- How do I clean up?
- How do I recover



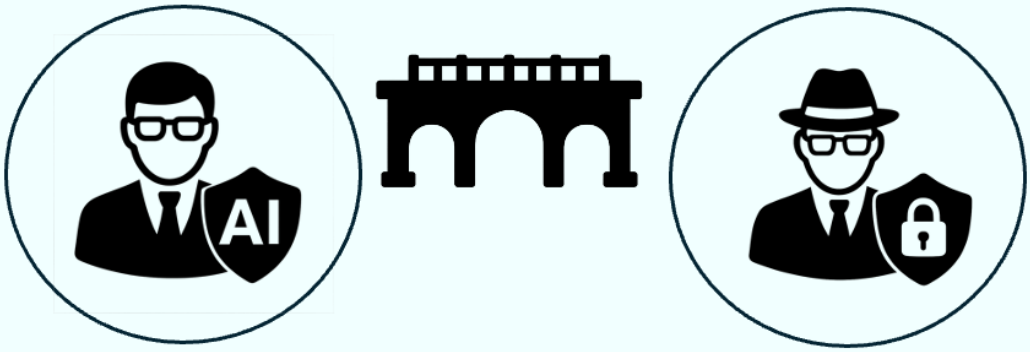
Adding complex systems to dysfunctional


Solid State Logic
Studio Computer



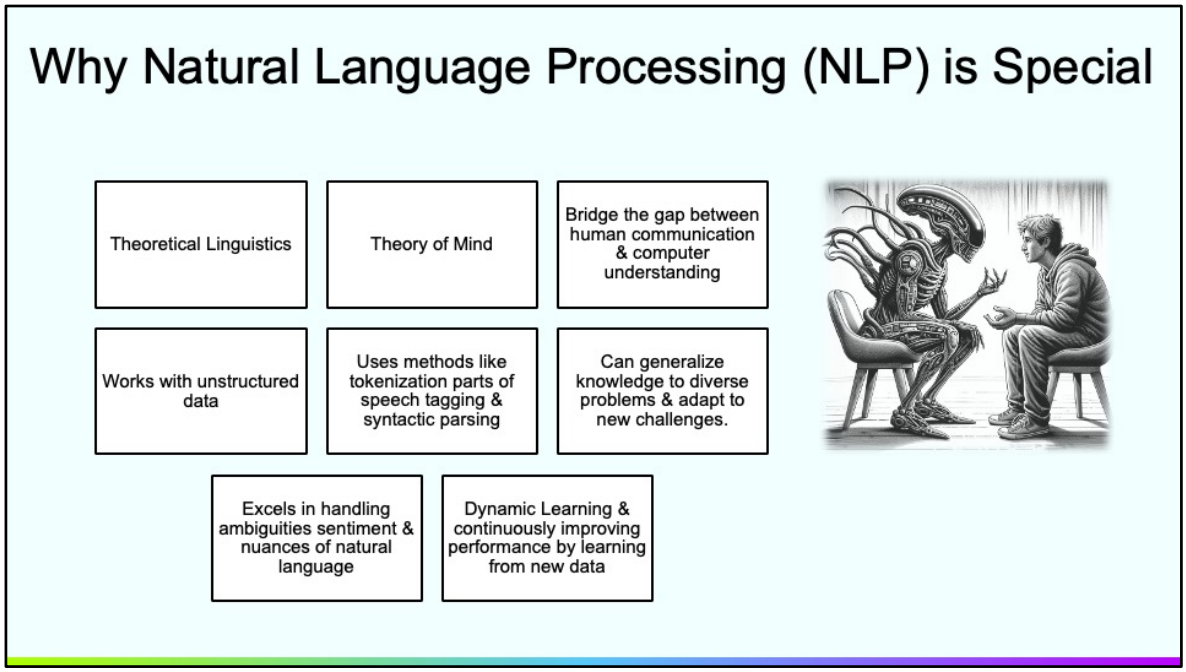
- The challenge is people & business are almost required to be digitally connected today.
- The internet wasn't secure, private, or safe before AI .
- Attacks are sophisticated, and many people and business are part of a digital risk lottery where their number just hasn't come up yet.

AIML Security vs Traditional Cybersecurity



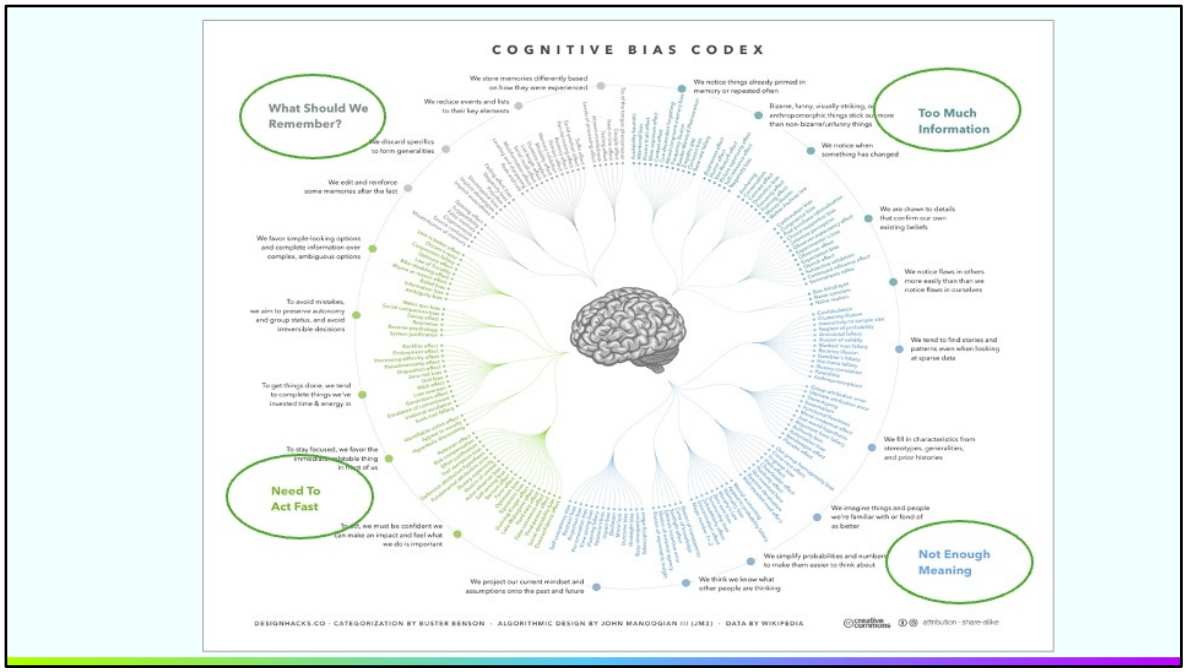
- AIML
  - Before the release of ChatGPT AI Security more academic. Definitely considered but more focused on “model security” not model deployed to users.
  - Not secure by design
  - Not part of the SDLC
- Traditional Cybersecurity
  - Moat thinking – block at the perimeter top concern someone putting data into a model
  - Guard rails / WAF

How GenAI Is Different



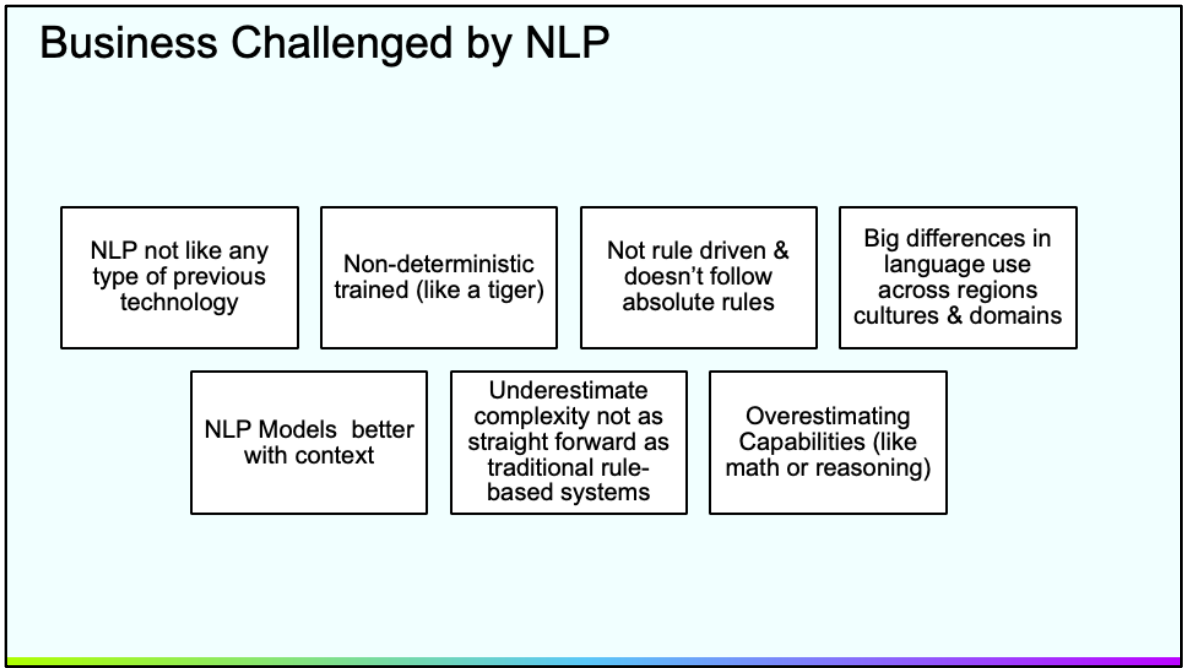
1. Natural Language Processing is software I can work with in my language not its It's very similar to working with another person
2. **Theoretical linguistics** tries to understand the underlying principles of the nature of human language. Phonetics, Phonology, Morphology, Syntax, Semantics, Pragmatics, Discourse
    1. Interpretability and Explainability: Linguistic theories provide frameworks for understanding and explaining the behavior of NLP models
3. **Theory of mind** refers to the capacity to understand other people by ascribing mental states to them. It includes the knowledge others' beliefs, desires, intentions, emotions, and thoughts may be different from yours
4. **Purpose:** Bridge the gap between human communication & computer understanding w/ machines designed to understand & respond to language mimicking the natural processes of human communication.
5. **Data Handling:** Works with unstructured data (text speech) unlike traditional systems reliant on structured data and predefined rules.
6. **Techniques:** Uses methods like tokenization, part-of-speech tagging, and syntactic parsing to process language patterns and meaning.
7. **Dynamic Learning:** Continuously improves performance by learning from new data, unlike traditional models requiring full retraining.

https://www.languageeducatorsassemble.com/intro-to-theoretical-linguistics/

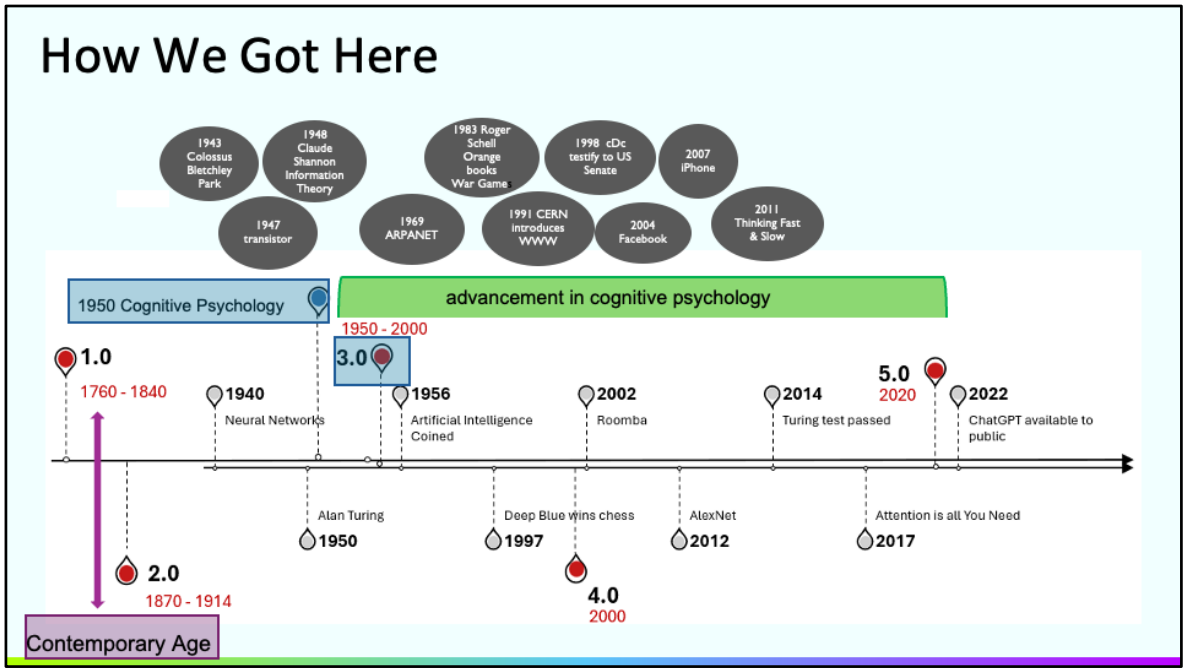


The future is understanding thinking and communication for both humans and machines

Business Challenged by NLP



How We Got Here

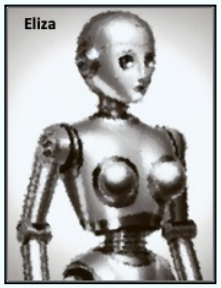


- A 1943 paper from Warren McCulloch and Walter Pitts, describing a simplified model of a biological neuron, often called the "McCulloch-Pitts neuron," was the first paper to think about a computer system like a human biological system
- Claude Shannon"A Mathematical Theory of Communication" paper in 1948 laid the foundations for the field of information theory

The Challenge of Being Human

Challenge of Being Human

Anthropomorphism

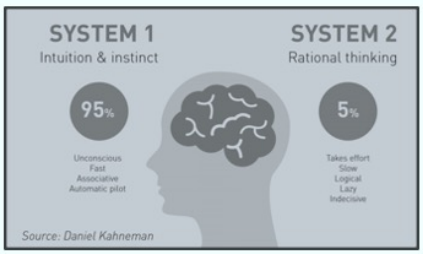


Joseph Weizenbaum

*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

The average person makes 35,000 decisions a day

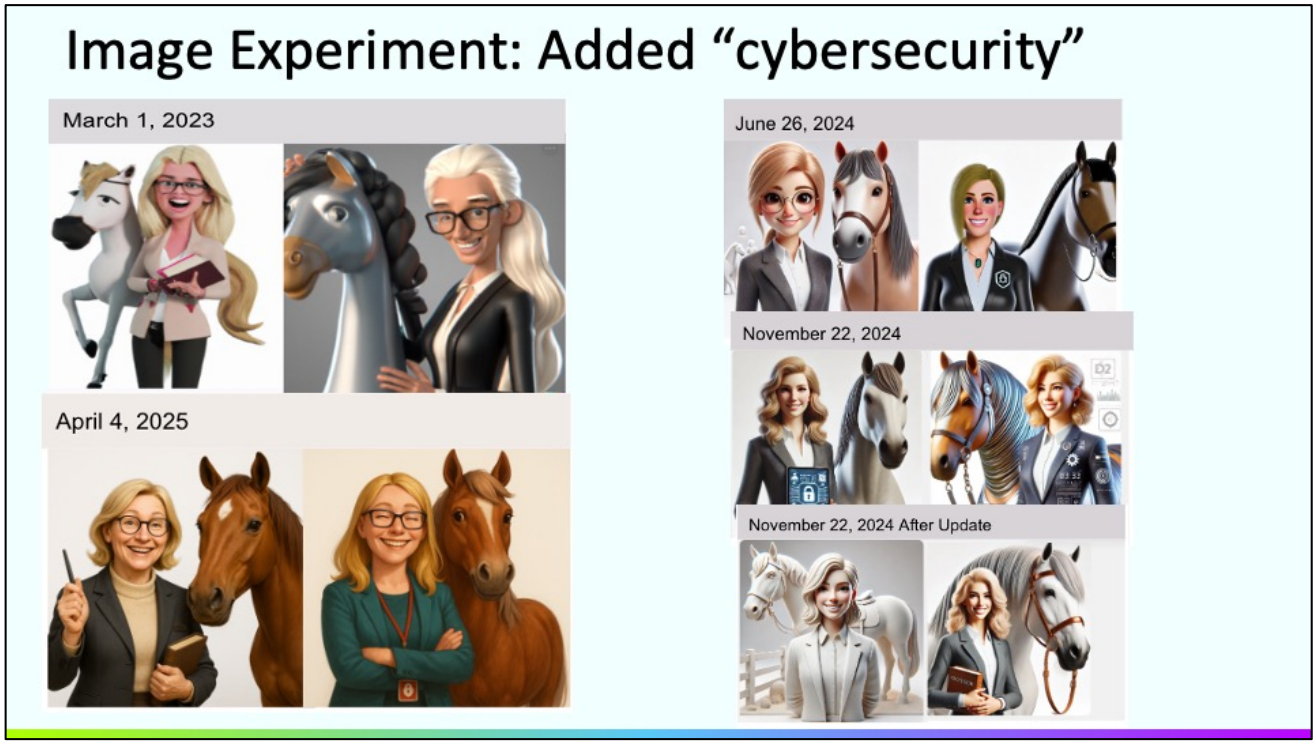


1. We have a technology that was designed to mirror and act human, which responds best when used that way and it is being used by people already susceptible to humanizing nonhuman things and animals. We see faces in clouds Name our cars, and as Joseph Weizenbaum saw in his first experiment, this has extremely potential. Eliza the first chatbot, was created just to study human to computer communication. Dr. Weizenbaum found people's reaction to Eliza alarming and actually wrote anti AI books
2. Blake Lemoine, a software engineer for Google, claimed that a conversation technology called LaMDA had reached a level of consciousness after exchanging thousands of messages with it.
3. Research shows people are prone to bad decisions or easily convinced to do something.
4. Sahakian and Labuzetta, two neurologists based in Oxford University found that the average person makes an estimated 35,000 decisions a day. When you break this down it comes to, 2,000 per hour which is then one decision every two seconds.

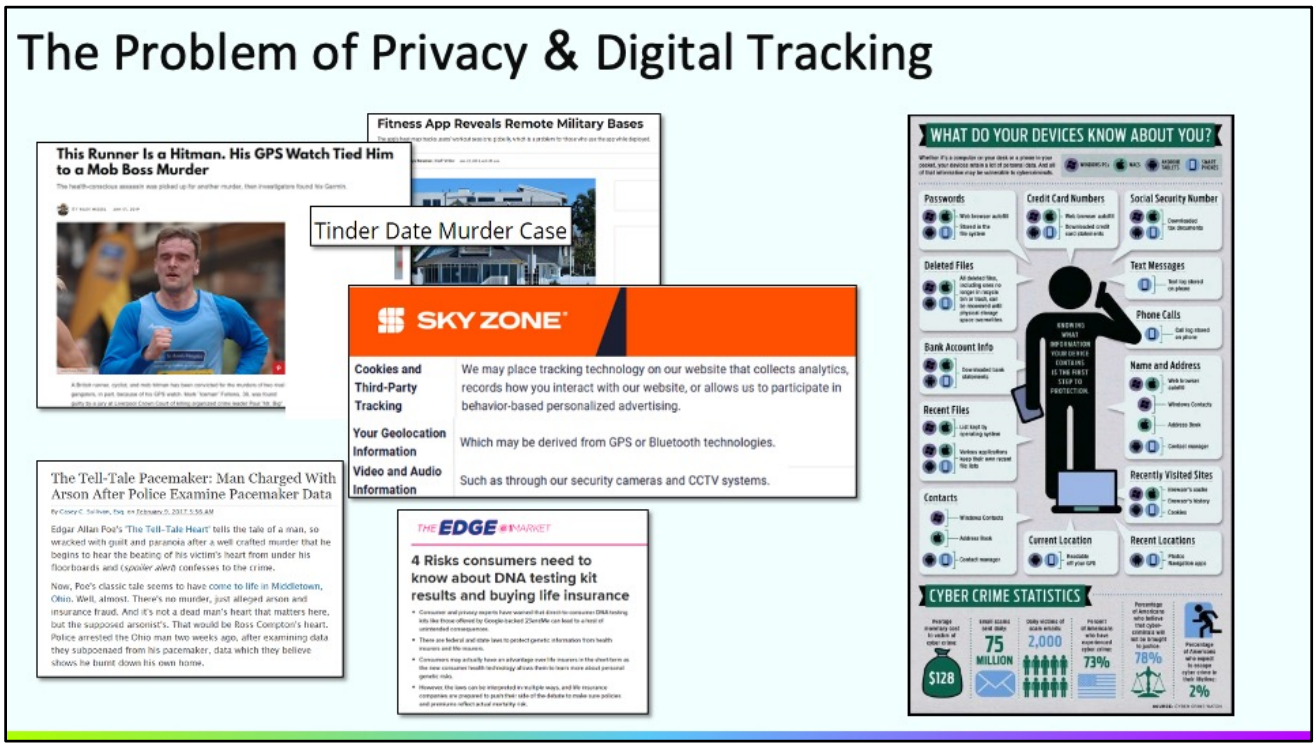
Cognitive Hacking

Finding and leveraging vulnerabilities in how we think, feel, and make decisions

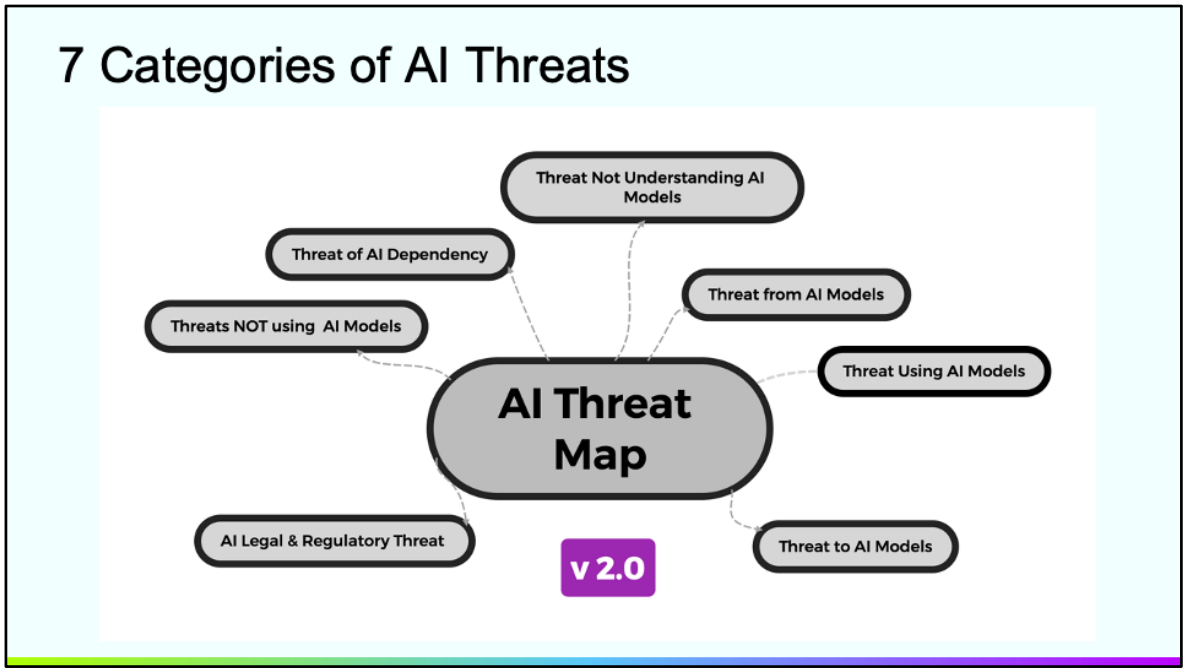
Bruce Schneier – Psychology of Security, the coming of AI Hacking



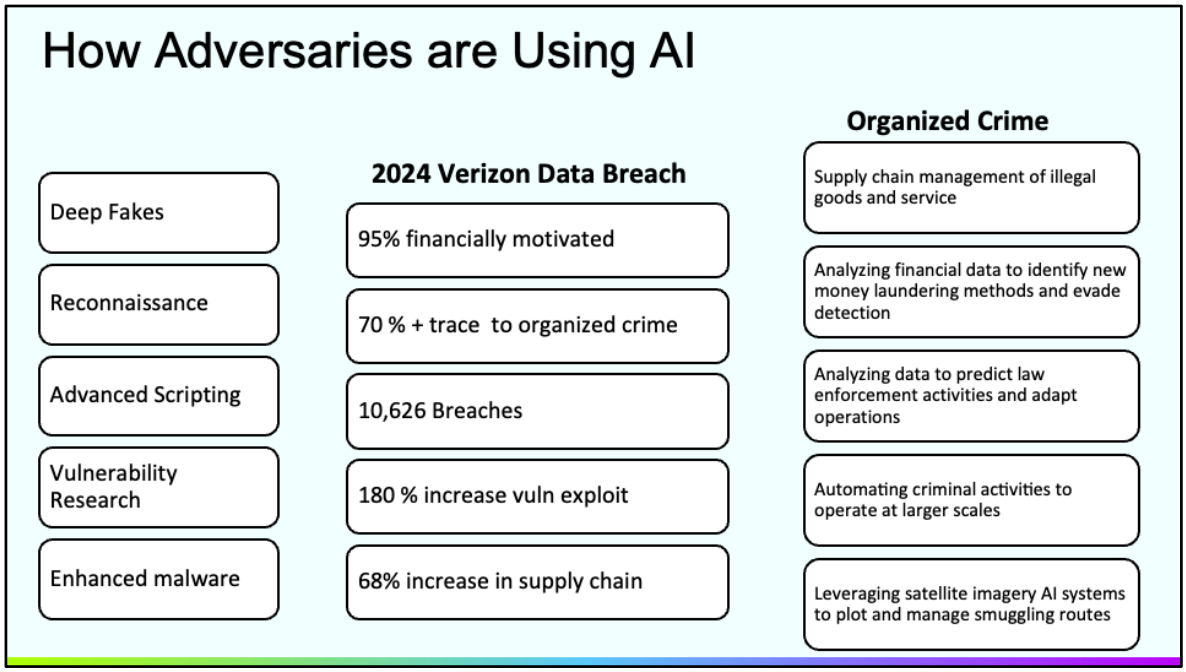
3D render of a happy woman professor, blond hair and a horse with white background, digital art.



https://www.runnersworld.com/news/a25924256/mark-fellows-runner-hitman-murder/
https://blogs.findlaw.com/technologist/2017/02/the-tell-tale-pacemaker-man-charged-with-arson-after-police-examine-pacemaker-data.html
https://ediscovery.co/ediscoverydaily/electronic-discovery/tinder-date-murder-case-highlights-the-increasing-complexity-of-ediscovery-in-criminal-investigations-ediscovery-trends/

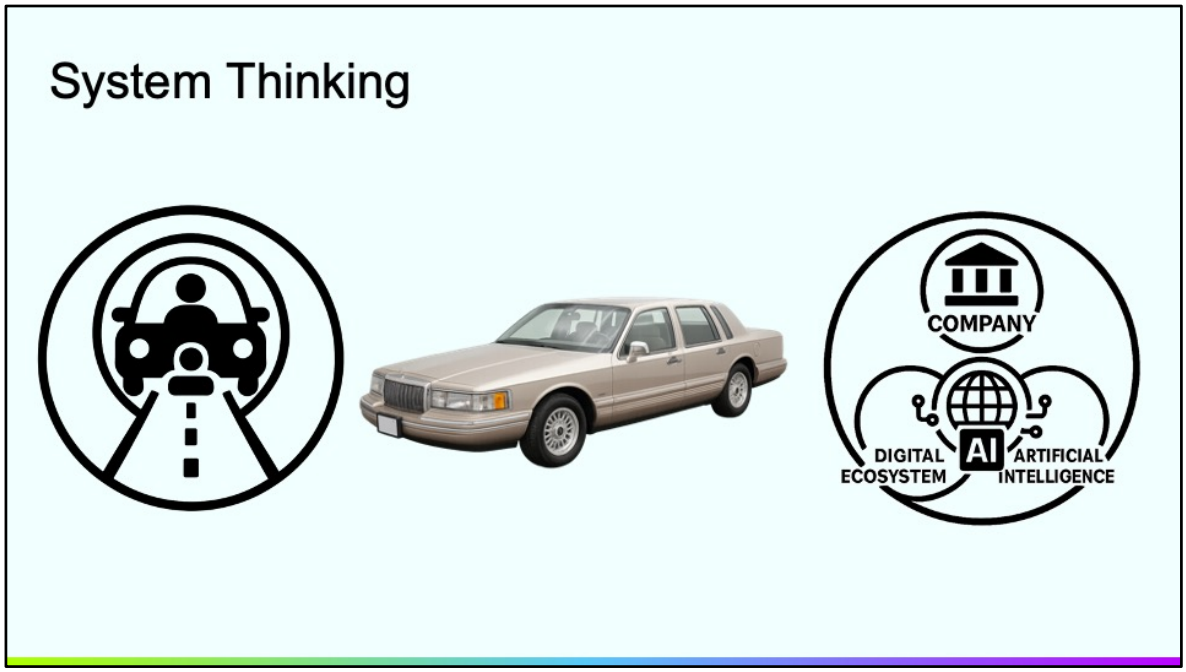


I break out AI Threats into these 7 categories

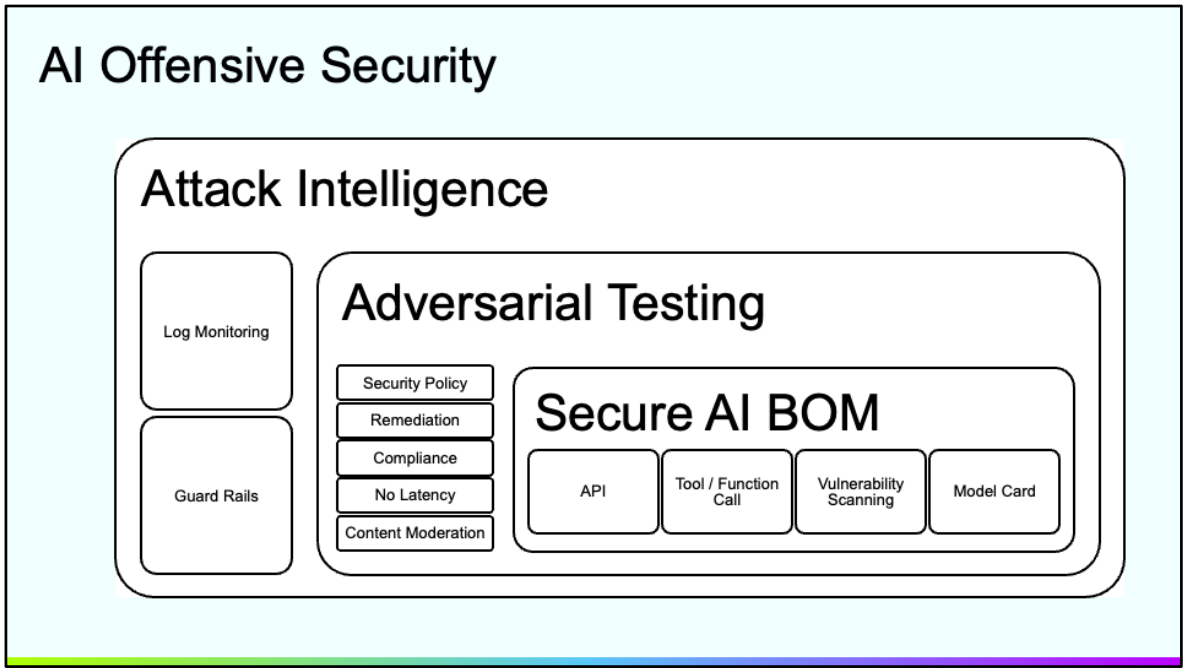


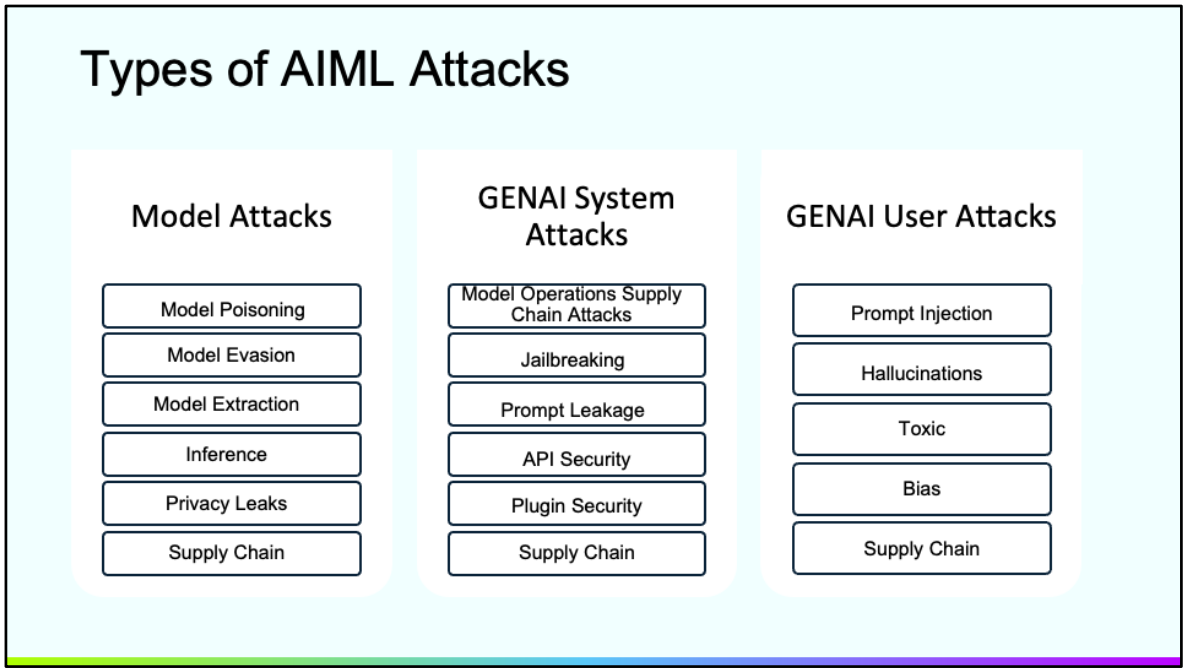
- The biggest threat to organizations is attack acceleration.
- Attackers have the initial advantage because they can move quickly. They don't have to follow budgeting and organizational red tape. Business can learn from them though. We see how organized crime is using AI for supply management, analyzing their financial data looking for new opportunities, Automation and utilizing data resources.
- Au10tix detected 22,080 fraudulent user onboarding attempts using AI on a single passport over eight months
- BlackMamba malware polymorphism
- DeepLocker uses AI to conceal its malicious intent & avoid detection
- 2/24 Vietnam & Thailand banking customers
- Bug Crowd 2023 edition of Inside the Mind of a Hacker,
- Monetary Authority of Singapore: Cyber Risks Associated with Generative Artificial Intelligence July 2024
- https://ijcttjournal.org/2024/Volume-72%20Issue-4/IJCTT-V72I4P111.pdf
    - Malware extracted videos and images of victims with their banking credentials & identity related documents from cell phones.
    - Images used to create deepfakes of the victims' faces to circumvent facial biometric

Offensive Strategy



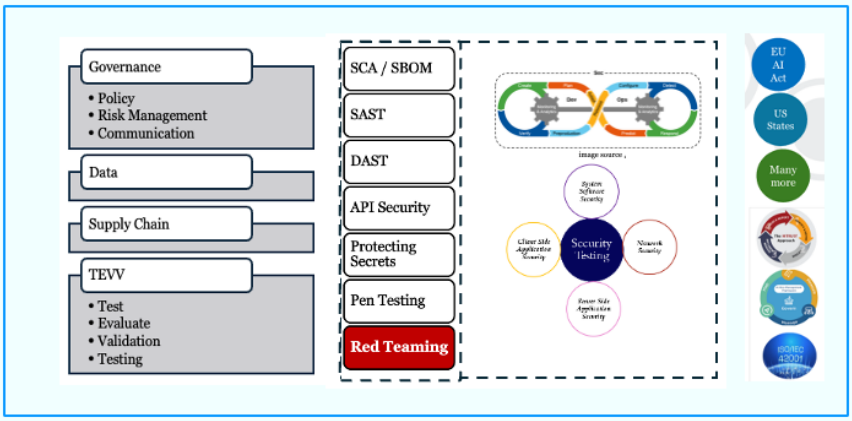
- **Interconnectedness**: Everything is connected. A change in one part affects others—sometimes in unexpected ways.
- **Feedback loops**: Systems have reinforcing (positive) and balancing (negative) feedback loops that influence how they behave over time.
- **Causality over time**: Instead of asking “What happened?” systems thinking asks, “What pattern of behavior is emerging, and why?”
- **Holistic view**: Understand the **whole system**, not just its parts


AI Offensive Security
Attack Intelligence
Log Monitoring
Guard Rails
Adversarial Testing
Security Policy
Remediation
Compliance
No Latency
Content Moderation
Secure AI BOM
API
Tool / Function Call
Vulnerability Scanning
Model Card

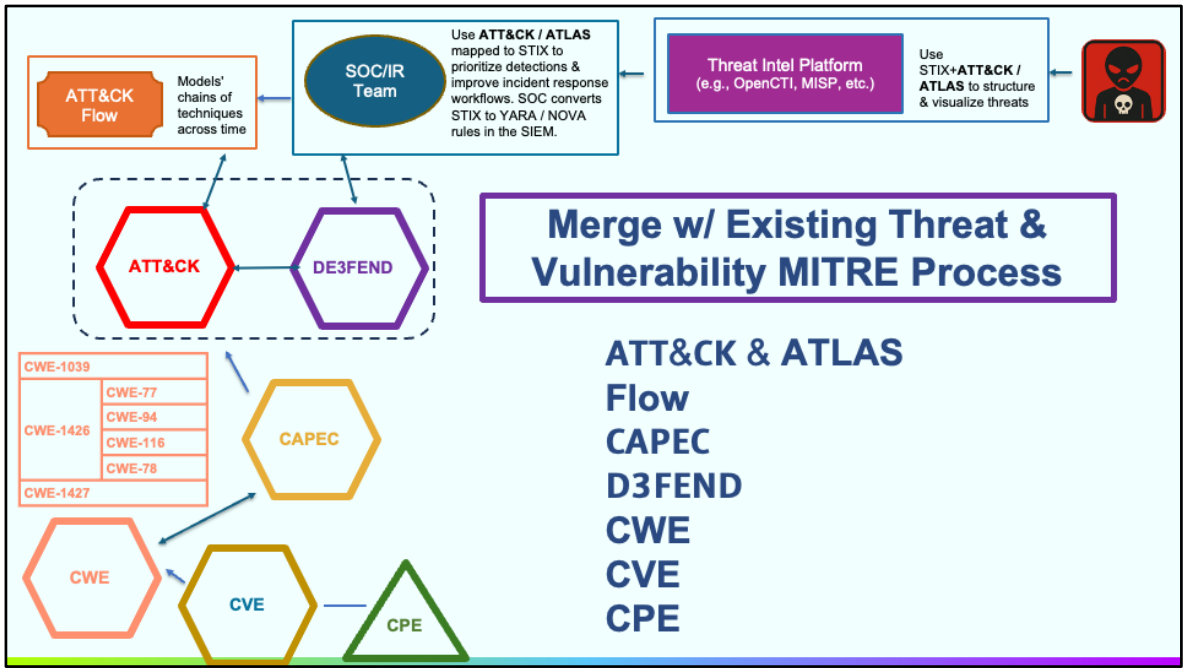


- When attack modeling, you should consider the deployment types, most organizations have many different types deployed within their organization.
- There are attacks against the model like model poisoning, or inference attacks
- Attacks against the GENAI System, such as the software used to build and operationalize models.
- And attacks against the users.

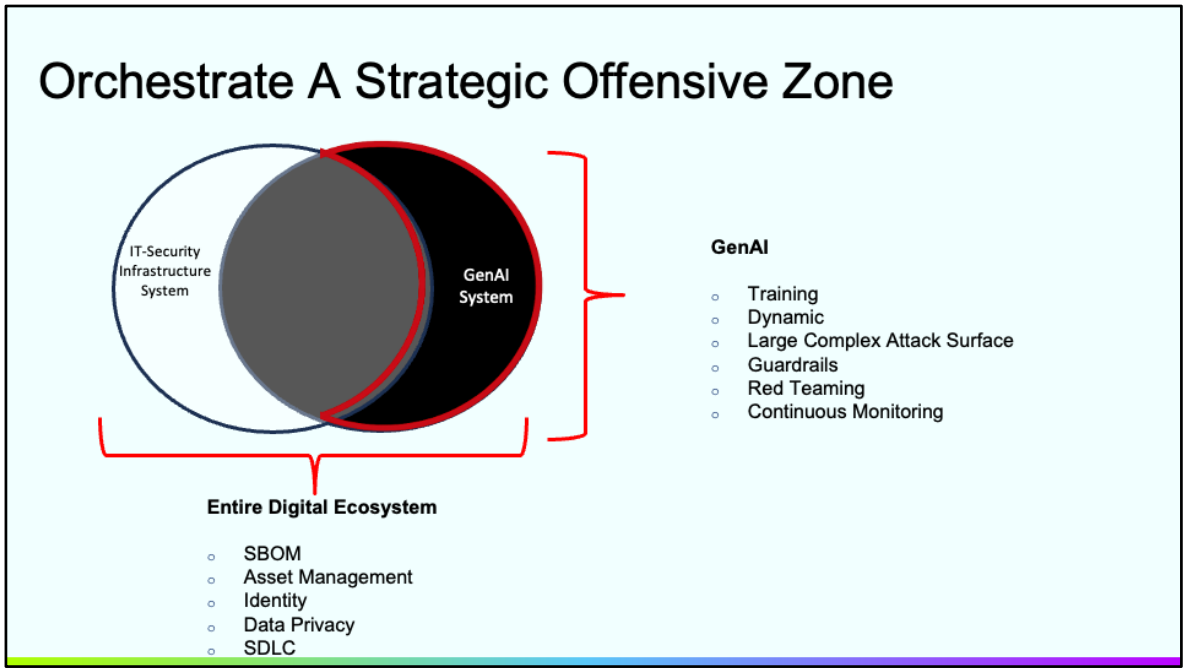
Unify AIML Security Safety & Privacy



- 1 Image source https://www.scribd.com/document/707845847/Ultimate-DevSecOps-Library-1706607714
- the **new surface of fairness, toxicity,** into the overall digital ecosystem.
- Many of the controls are the same across frameworks – Cross map with the Security Controls Framework



Merge into existing attack & vulnerability process



- Find a balance In the same way an EPSS (Exploit Prediction Scoring System) helps organization focus on patching vulnerabilities with the highest threat which is estimated to be about 6%. The purpose is to help organizations deploy AIML solutions and benefit from using them while avoiding big disasters.

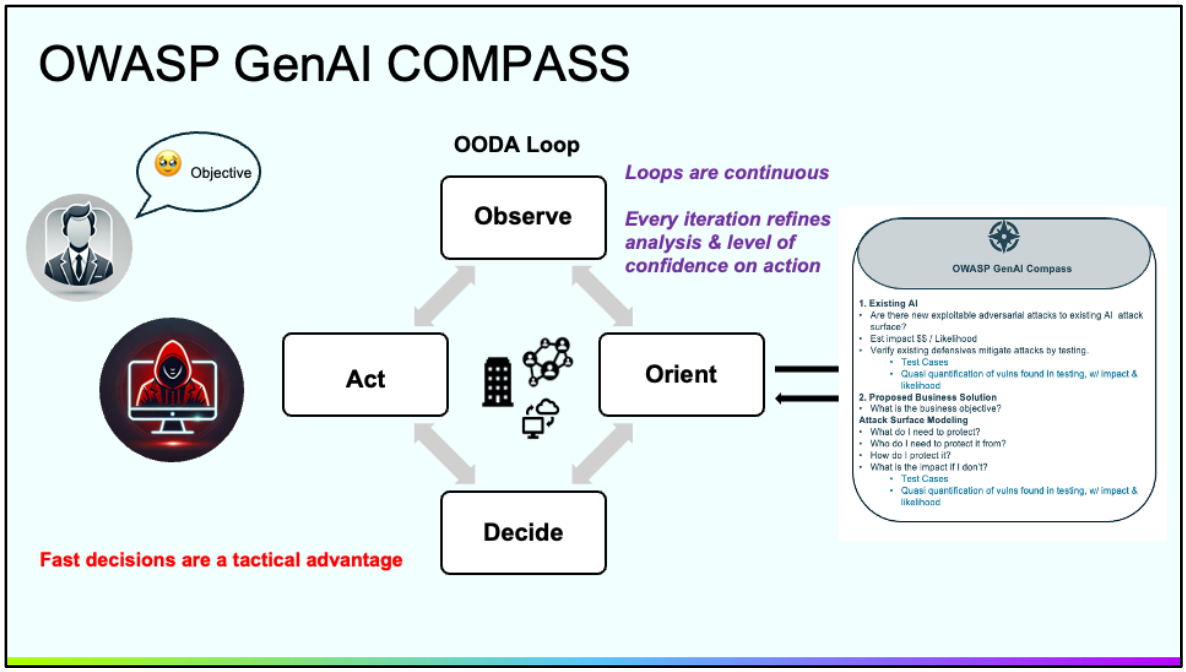
OWASP GenAI COMPASS

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

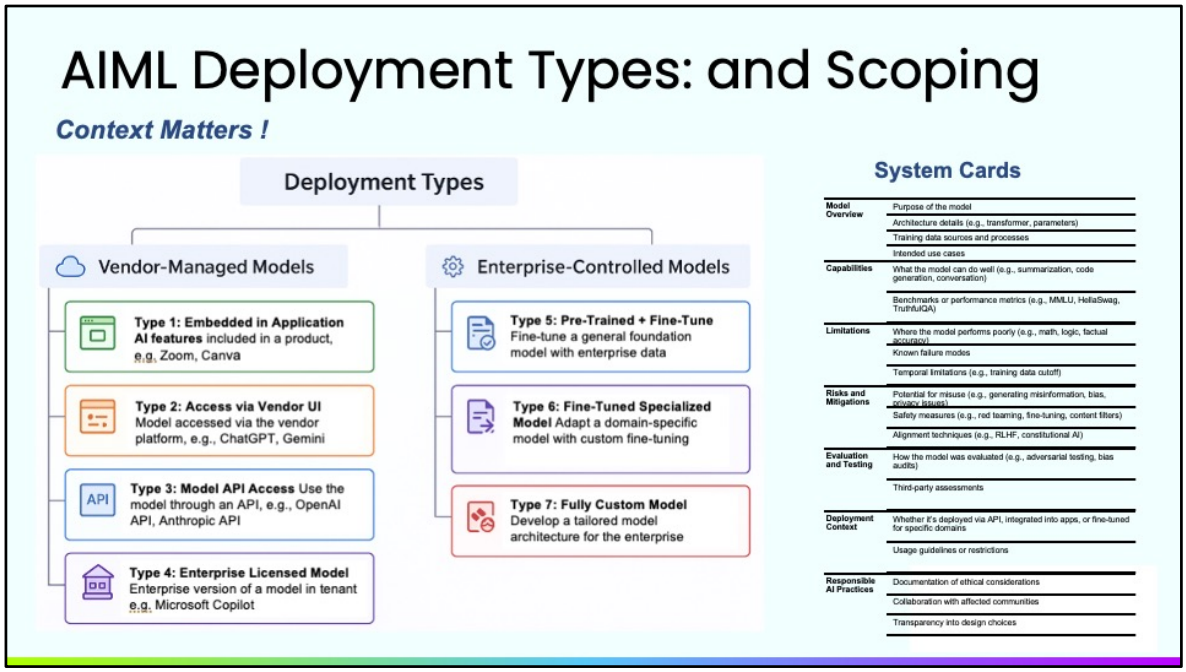
Attack Surface Modeling

Or

Need answers fast

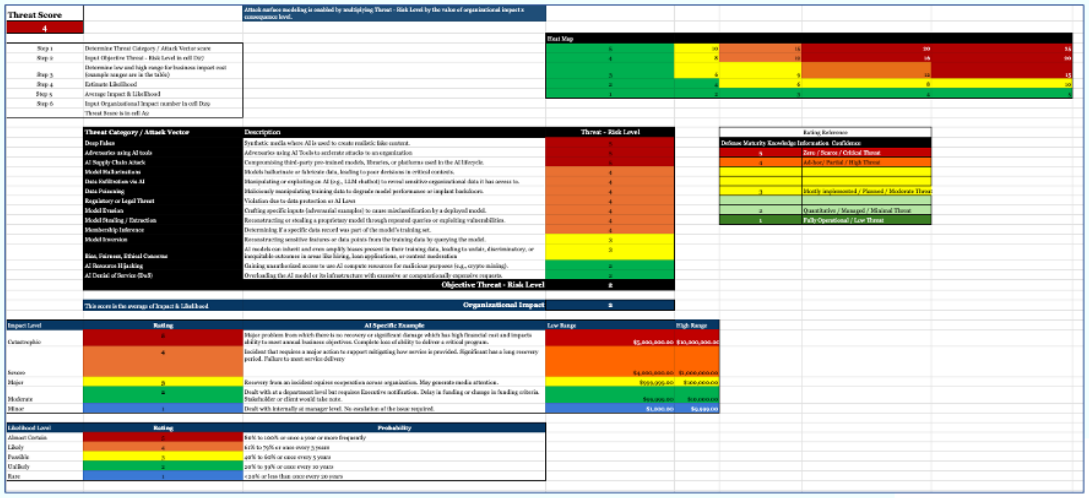


Determine the objective, ooda, 5 point scoring
The foundation is using the OODA Loop. predicting calamity is hard, think about the CrowdStrike debacle. Think about this as a warzone, you need to decide quickly on the information you have. The more data you have the confident the decision, but there are many moving pieces so predicting is challenging, so when unexpected things happen, you follow the loop, observe



What do I need to protect?
Who do I need to protect it from?
How do I protect it?
Who is mad if I don't?

Initial Threat Score



Threat Score
4

Attack surface modeling is enabled by multiplying Threat - Risk Level by the value of organizational impact a consequence level.

Step 1	Determine Threat Category / Attack Vector score
Step 2	Input Objective Threat - Risk Level in cell D27
Step 3	Determine low and high range for business impact cost (example ranges are in the table)
Step 4	Estimate Likelihood
Step 5	Average Impact & Likelihood
Step 6	Input Organizational Impact number in cell D29
	Threat Score is in cell A2

Heat Map

5	10	15	20	25
4	8	12	16	20
3	6	9	12	15
2	4	6	8	10
1	2	3	4	5

Threat Category / Attack Vector	Description	Threat - Risk Level
Deep Fakes	Synthetic media where AI is used to create realistic fake content.	5
Adversaries using AI tools	Adversaries using AI Tools to orchestrate attacks on an organization	5
AI Supply Chain Attack	Compromising third-party pre-trained models, libraries, or platforms used in the AI lifecycle.	5
Model Hallucinations	Models hallucinate or fabricate data, leading to poor decisions in critical contexts.	4
Data Infiltration via AI	Manipulating or exploiting an AI (e.g., LLM chatbot) to reveal sensitive organizational data it has access to.	4
Data Poisoning	Maliciously manipulating training data to degrade model performance or implant backdoors.	4
Regulatory or Legal Threat	Violation due to data protection or AI Laws	4
Model Evasion	Crafting specific inputs (adversarial examples) to cause misclassification by a deployed model.	4
Model Stealing / Extraction	Reconstructing or stealing a proprietary model through repeated queries or exploiting vulnerabilities.	4
Membership Inference	Determining if a specific data record was part of the model's training set.	4
Model Inversion	Reconstructing sensitive features or data points from the training data by querying the model.	3
Bias, Fairness, Ethical Concerns	AI models can inherit and even amplify biases present in their training data, leading to unfair, discriminatory, or unexplainable outcomes in areas like hiring, loan applications, or content moderation	3
AI Resource Hijacking	Gaining unauthorized access to use AI compute resources for malicious purposes (e.g., crypto mining).	2
AI Denial of Service (DoS)	Overloading the AI model or its infrastructure with excessive or computationally expensive requests.	2
	Objective Threat - Risk Level	4

This score is the average of Impact & Likelihood	Organizational Impact
	2

Rating Reference

Defense Maturity	Knowledge Information	Confidence
5	Zero / Scarce / Critical Threat	
4	Ad Hoc / Partial / High Threat	
3	Mostly implemented / Planned / Moderate Threat	
2	Quantitative / Managed / Minimal Threat	
1	Fully Operational / Low Threat	

Impact Level	Rating	AI Specific Example	Low Range	High Range
Catastrophic	5	Major problem from which there is no recovery or significant damage which has high financial cost as it impacts ability to meet annual business objectives. Complete loss of ability to deliver a critical program.	$5,000,000.00	$10,000,000.00
Severe	4	Incident that requires a major action to support mitigating how service is provided. Significant has a long recovery period. Failure to meet service delivery	$1,000,000.00	$5,000,000.00
Major	3	Recovery from an incident requires cooperation across organization. May generate media attention.	$999,999.00	$100,000.00
Moderate	2	Dealt with at a department level but requires Executive notification. Delay in funding or change in funding criteria. Stakeholder or client would take note.	$99,999.00	$10,000.00
Minor	1	Dealt with internally at manager level. No escalation of the issue required.	$1,000.00	$9,999.00

Likelihood Level	Rating	Probability
Almost Certain	5	80% to 100% or once a year or more frequently
Likely	4	41% to 79% or once every 3 years
Possible	3	21% to 60% or once every 5 years
Unlikely	2	20% to 79% or once every 10 years
Rare	1	<10% or less than once every 10 years

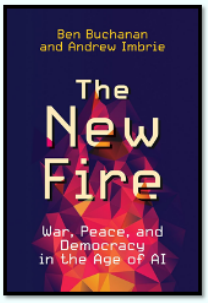
Preventative & Detective Controls

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.


Adversarial Testing
(Red Teaming)
splx
205© SplxAI Inc. All Rights Reserved.

Cybersecurity & AI

Bruce Schneier	The Coming of AI Hackers https://www.schneier.com/academic/archives/2021/04/the-coming-ai-hackers.html
	Hacking Back the AI Hacker https://www.schneier.com/blog/archives/2024/11/prompt-injection-defenses-against-llm-cyberattacks.html
Alex Stamos	
Thomas Roccia	https://www.linkedin.com/in/thomas-roccia/
Ben Buchanan	https://www.amazon.com/New-Fire-War-Peace-Democracy/dp/0262046547
Ram Shankar Siva Kumar Hyrum Anderson	https://www.amazon.com/Not-Bug-But-Sticker-Learning/dp/1119883989



Learn AI

Andrej Karpathy	
Deep Dive into LLMs like ChatGPT	https://www.youtube.com/watch?v=7xTGNNLPyMI&list=PLviHA9raZ6D5DzfultbdcXD5f4pvj0uNi&index=3
How I use LLMS	https://www.youtube.com/watch?v=EWvNQjAaOHw

Learn Prompting

LearnPrompting.org

Andrew Ng	
DeepLearning.AI AI for Everyone	https://www.coursera.org/learn/ai-for-everyone/

Exploration & Experimentation

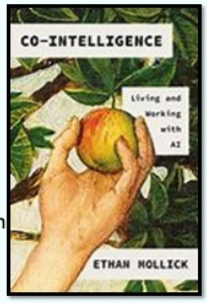
Andrew Mayne

Science Communicator for OpenAI from 9/21- 9/23

Mike Sharkey / Chris Sharkey

Ethan Mollick

Associate Professor Wharton

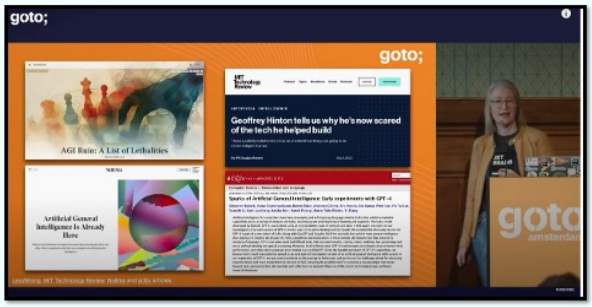


One of my favorite people to listen to about the future of AI is Andrew Mayne. He is a man of many of talents and one of those is as courts, author. I was a huge fan of his even before I found out he was a ChatGPT user and worked at OpenAI. AI & copyright cases are being wrestled through the courts, but it is thought proving to listen to him since he is a creator. His podcast Weird things is well, weird.

Beyond the Hype

A Realistic Look at Large Language Models

Jodie Burchell

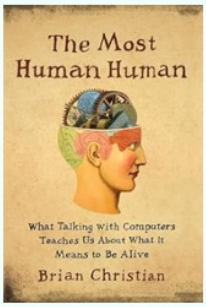


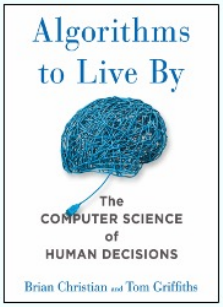
Blog https://t-redactyl.io/

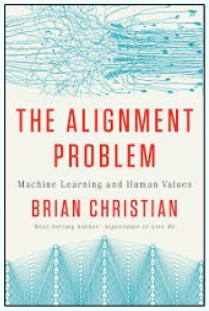
https://www.youtube.com/watch?v=Pv0cfsastFs&t=1190s

52

Safety







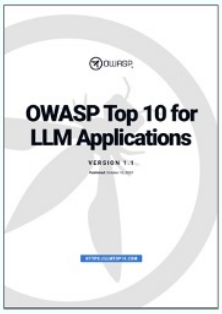
Brian Christian, author, poet, programmer, and researcher

Great people working on Explainability and Safety, I highly recommend Brian Christian work explores the intersection of technology, philosophy, and human behavior

- The Most Human Human (2011)
  - The need for a nuanced understanding of human-AI collaboration
- Algorithms to Live By (2016)
  - The application of algorithms and data analysis to everyday life
- The Alignment Problem (2020)
  - The need for AI systems to be aligned with human values and goals

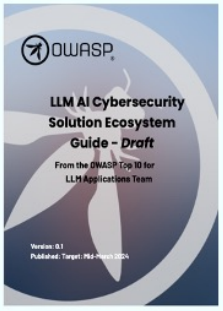
OWASP Top 10 for LLM Project

OWASP Top Ten for LLM https://genai.owasp.org



Top 10 List

- Developers
- AppSec Teams



Solutions Guide

- Development Leaders
- Security Operations

Checklist

- CISOs
- Compliance Officers