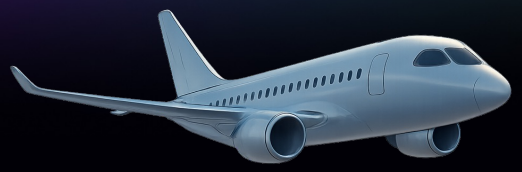
The Alien & The AI Horizon



October 14, 2025

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

Sandy Dunn, CISO SPLX

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

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

SPLX

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

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

Topics

- The GenAI Alien Frontier KABOOM
- The Big Scary Dirty World
- Challenges of being Human
- AI Threat Map
- OWASP GenAI COMPASS
- AI Skills & Sanity


FLOW
CREATE

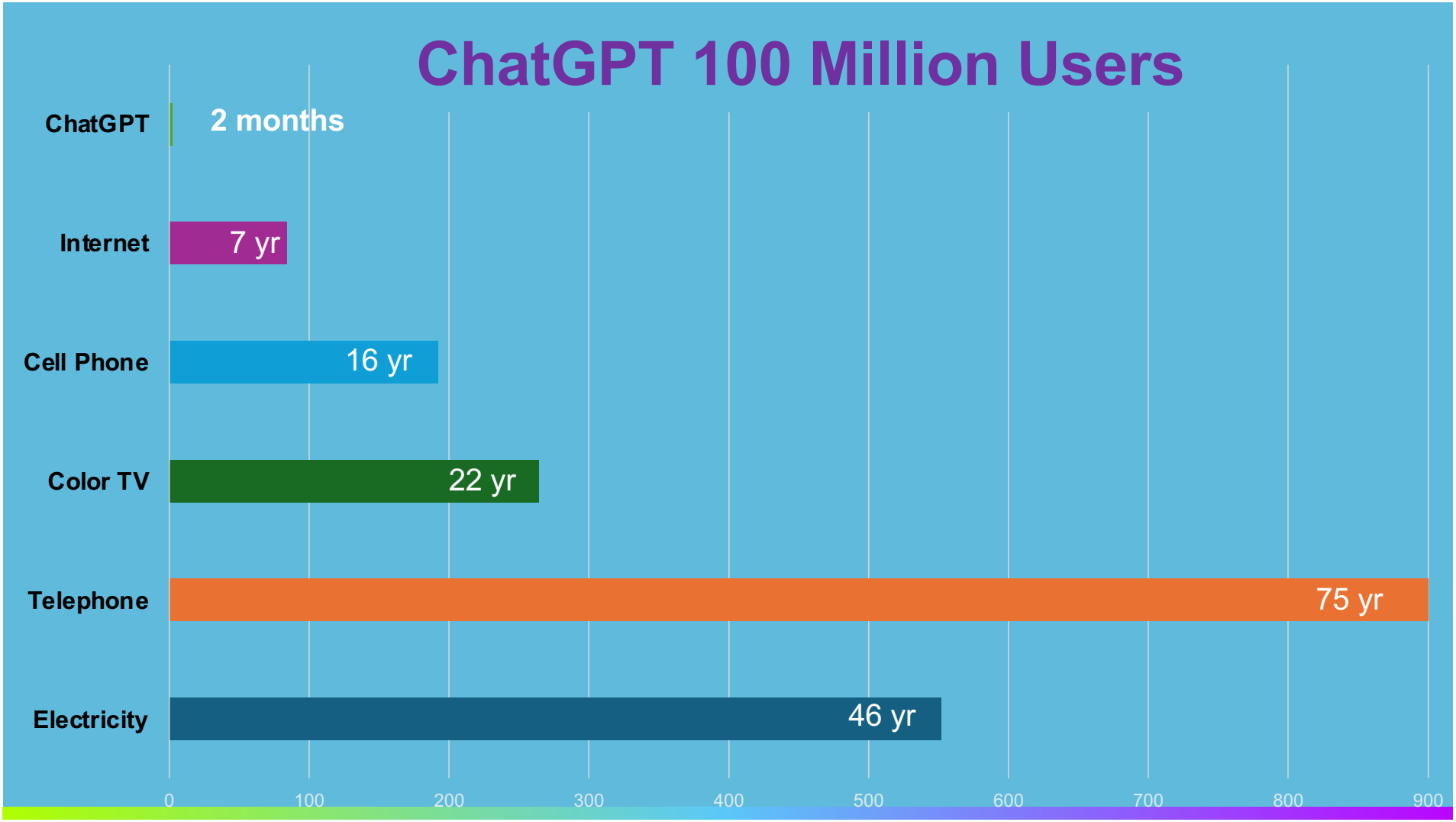
AI KA - BOOM !

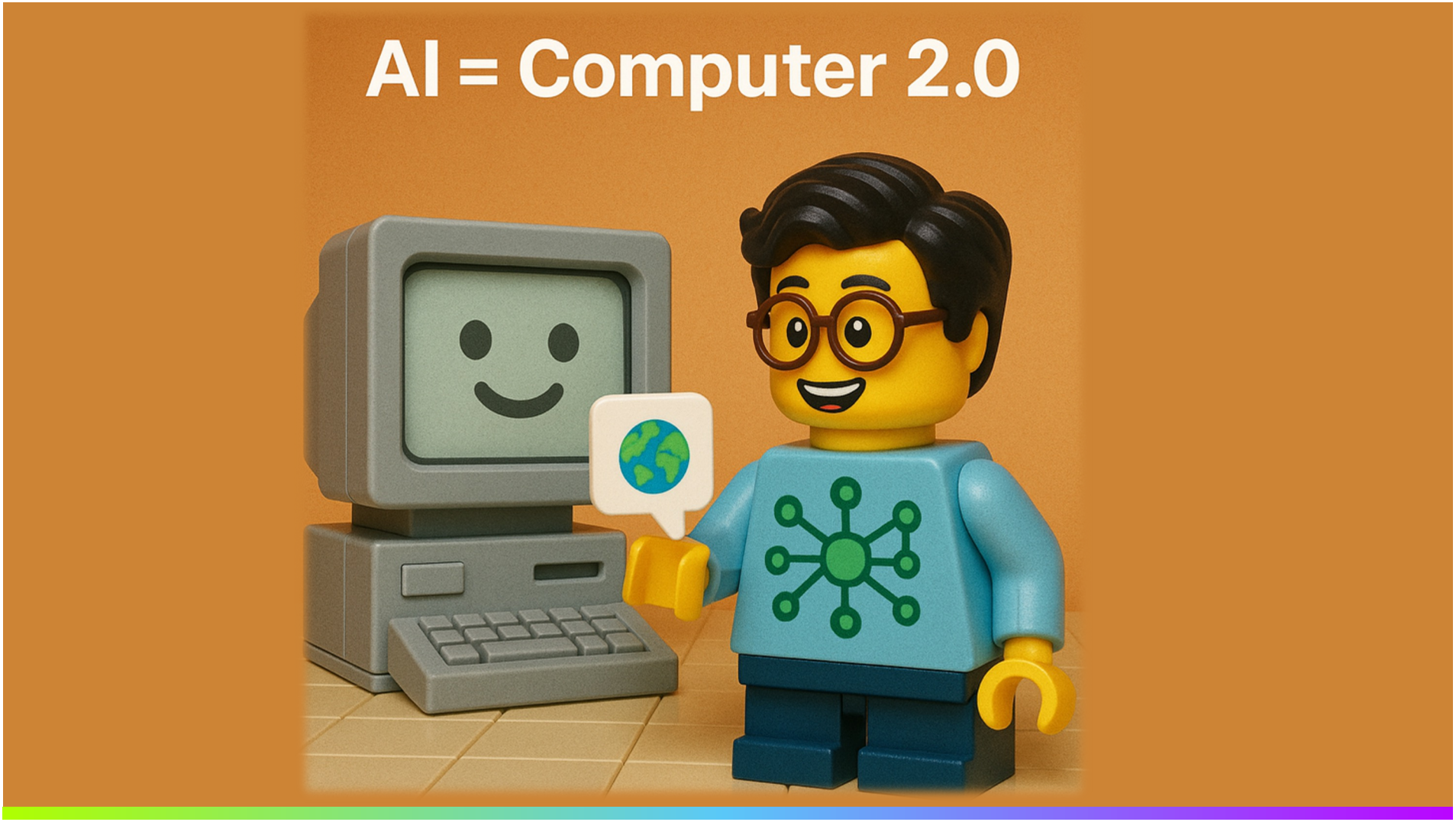
November 22, 2022

Algorithms

Data

GPU Power

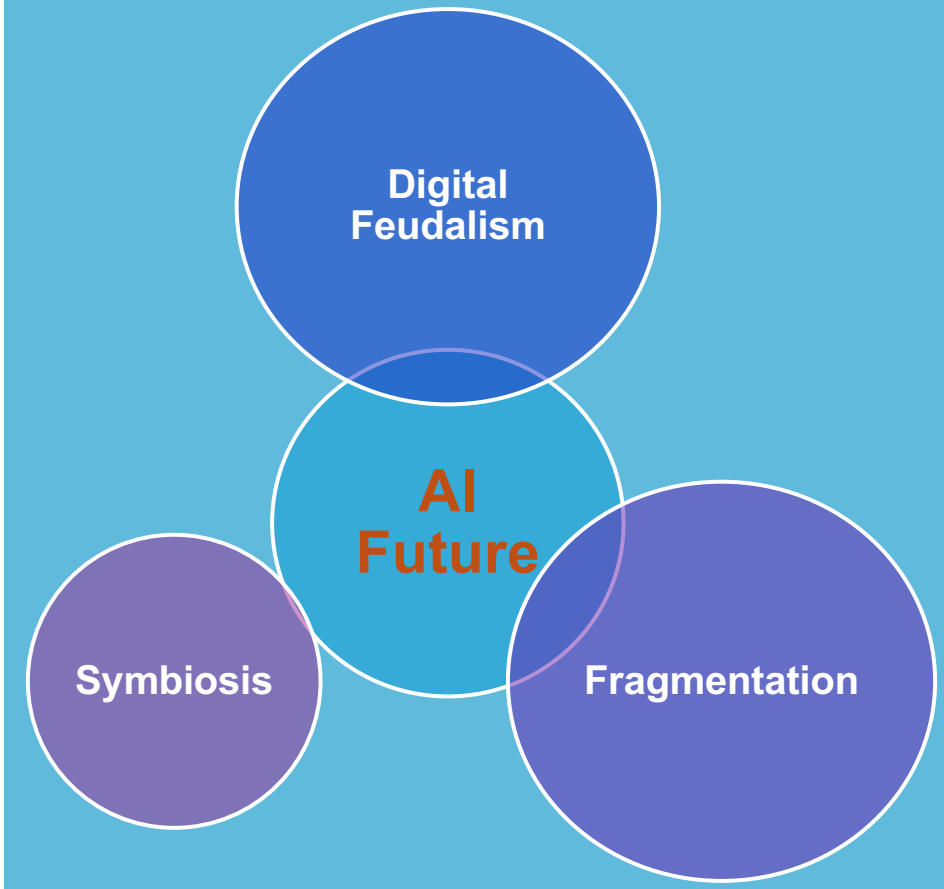

ChatGPT 100 Million Users
ChatGPT
2 months
Internet
7 yr
Cell Phone
16 yr
Color TV
22 yr
Telephone
75 yr
Electricity
46 yr
0
100
200
300
400
500
600
700
800
900


AI = Computer 2.0

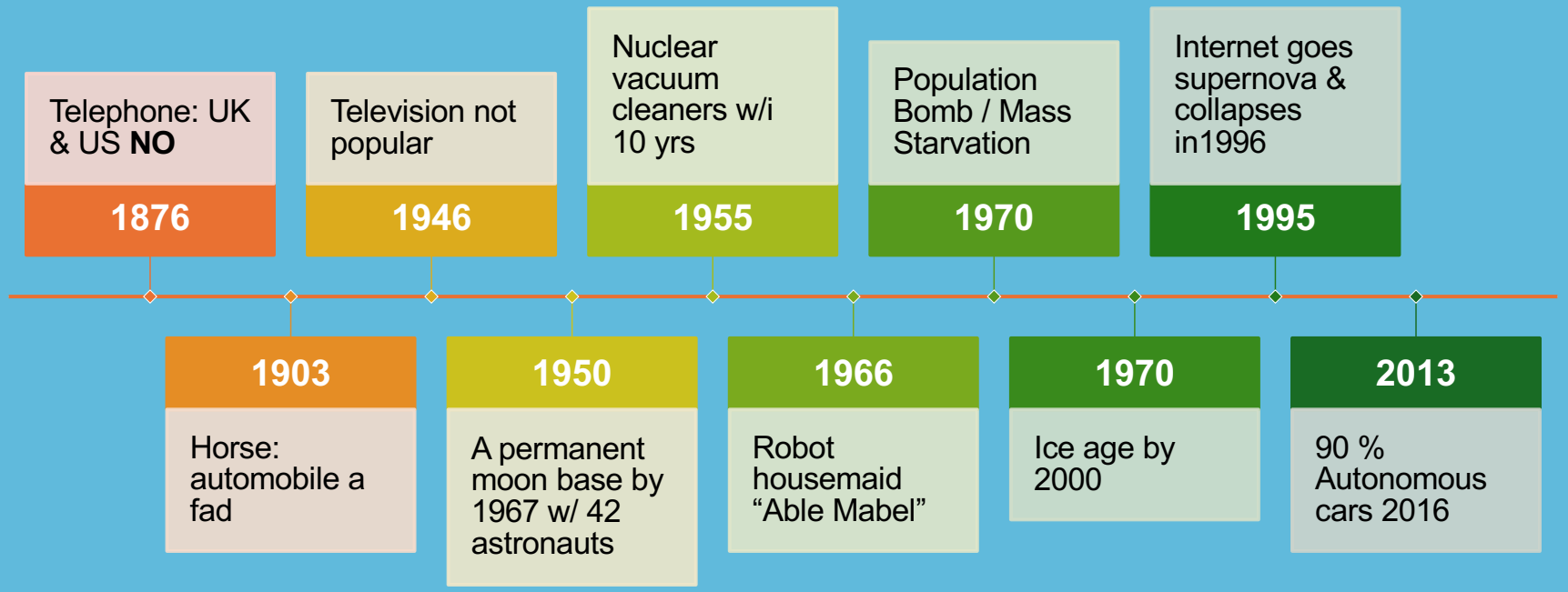
The Machines Are Taking Our Jobs - Thank God?

Mostaque's Guide to the next 1000 Days

Star Trek not Star Wars

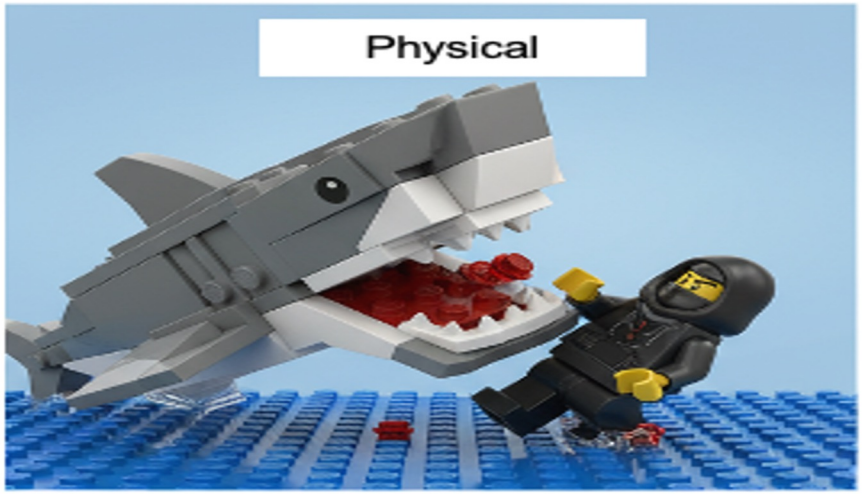


Predictions are Hard



The Big Scary Dirty World

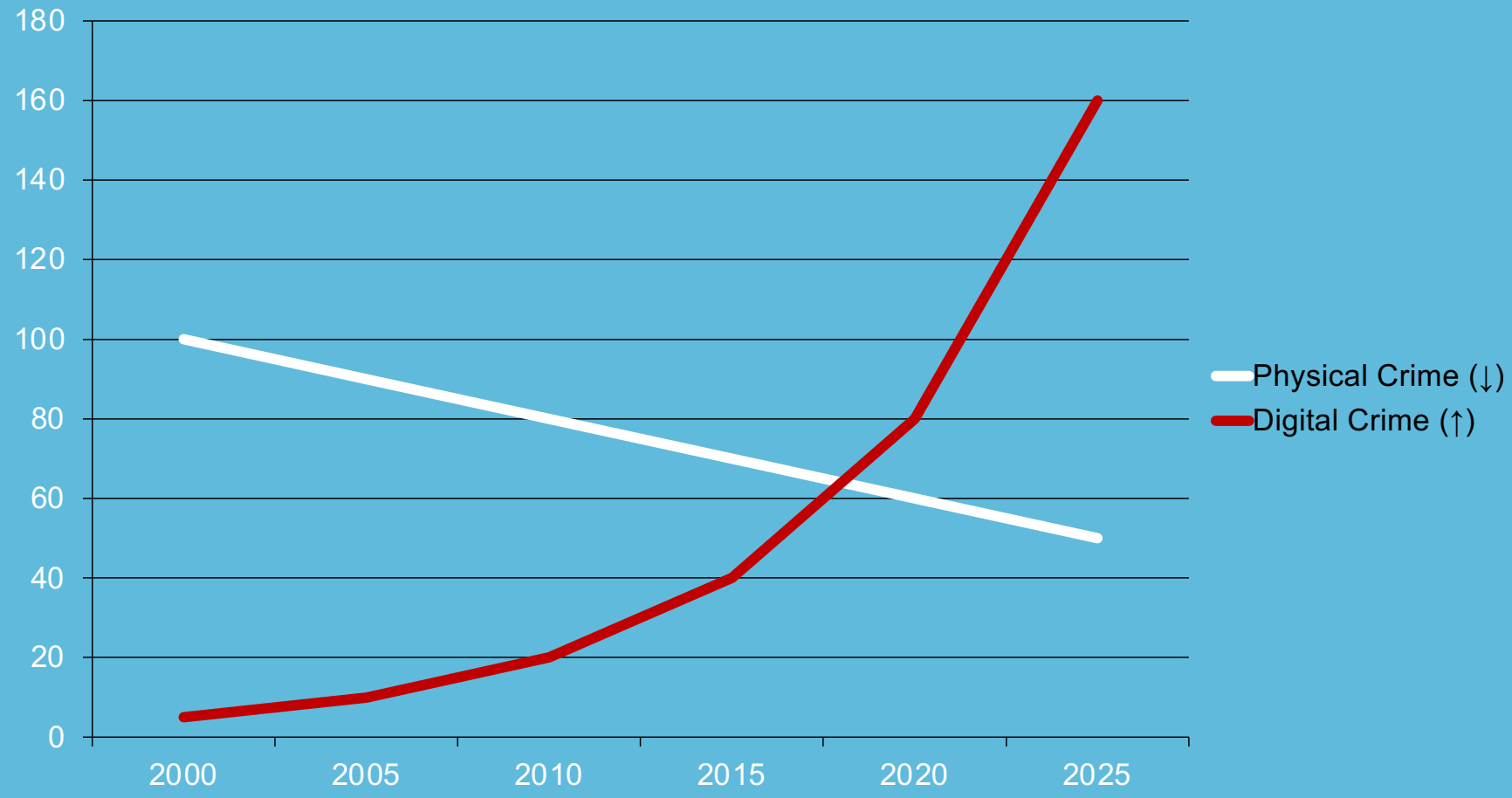

Humans


Physical

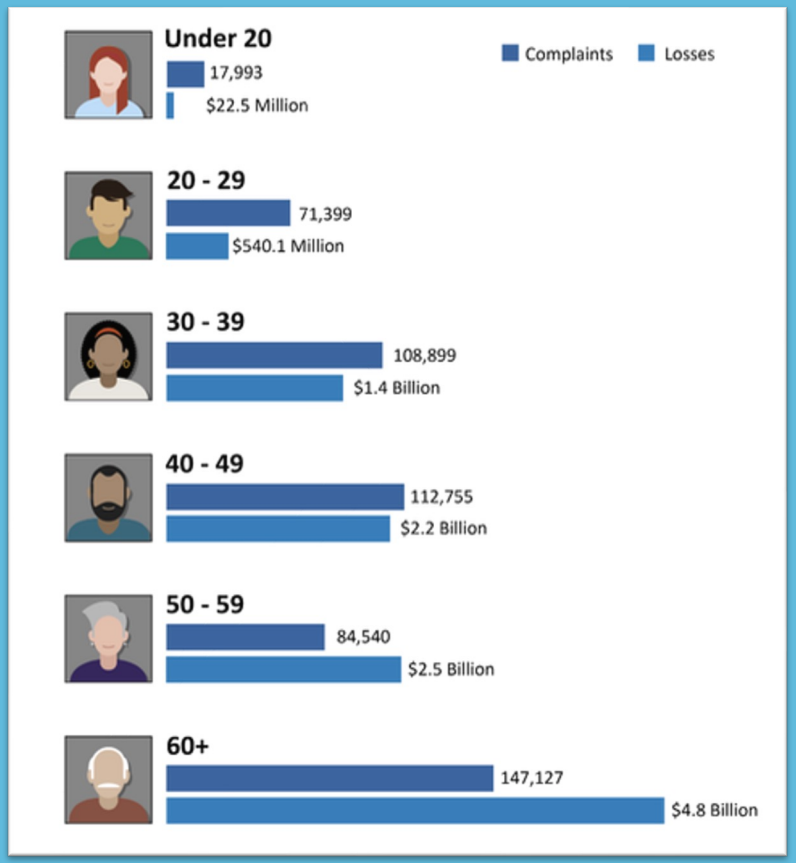

Digital


Artificial Intelligence

Digital Crime Explodes - Physical Crime Falls



FBI 2024 Internet Crime Report



The Digital Evolution of Criminal Recruitment

Public Service Announcement

FEDERAL BUREAU OF INVESTIGATION

cember 12, 2019

Alert Number
121219-PSA

ns regarding this
uld be directed to
FBI Field Office.

Office Locations:
.gov/contact-
eld-offices

Child Predators Use Online Gaming to Contact Children

Some predators use online gaming to sexually exploit children. Parents and guardians—and their children—should know the risks posed by online gaming, the methods these predators use, how to keep children safe online, and what to do if their children are victimized or targeted.

DEFINITION

The FBI defines online gaming as any game played while connected to the Internet. This includes games played through gaming consoles; handheld gaming devices; and applications on phones, tablets, or computers. All games with communication features, including basic games, can be used by

Recruitment

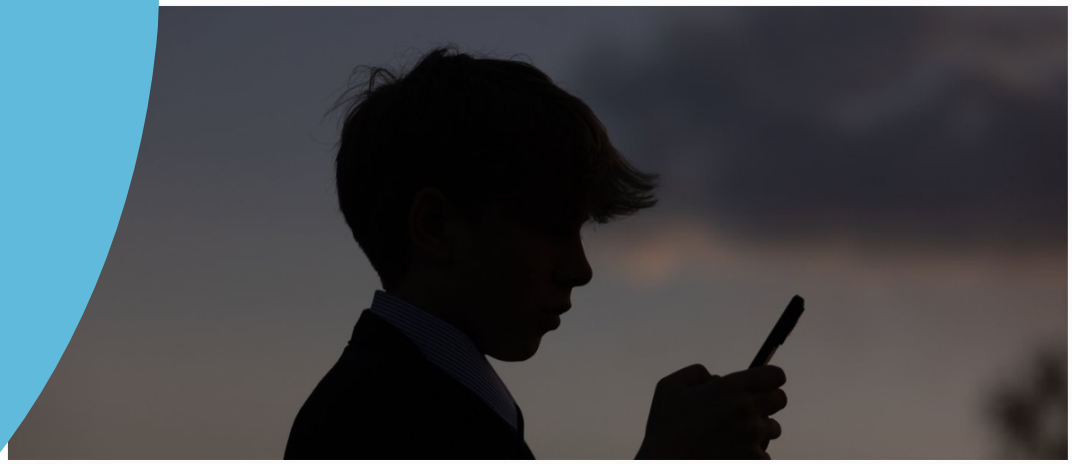


Photo: Anna Barclay/Getty Images


The Tequila Heist
MEZQUILA
Santo
Santo

o drivers receive toll bills from California's
rak for violations they never committed

Idaho driver received toll bills from California's FasTrak for violations they never committed

Privacy & Digital Tracking

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

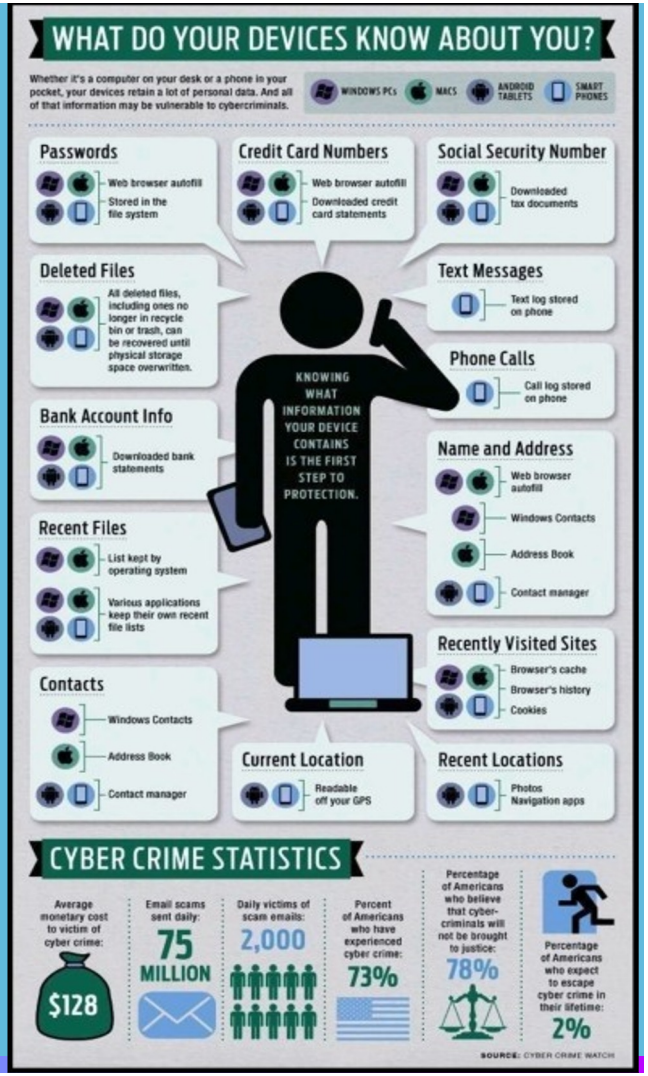
Edgar Allan Poe's 'The Tell-Tale Heart' tells the tale of a man, so wracked with guilt and paranoia after a well crafted murder that he begins to hear the beating of his victim's heart from under his floorboards and (*spoiler alert*) confesses to the crime.

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.

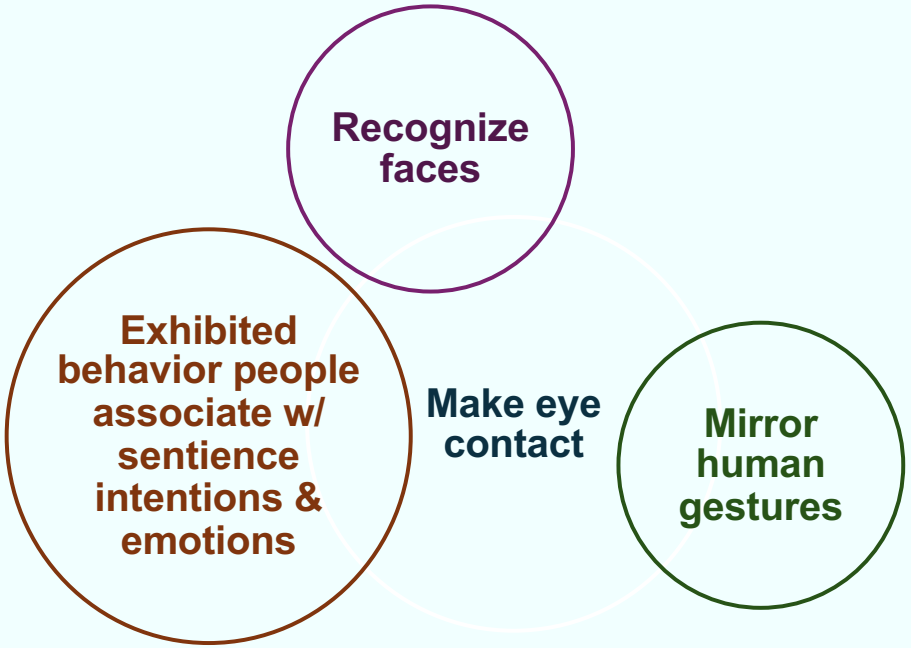
THE EDGE @1MARKET

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

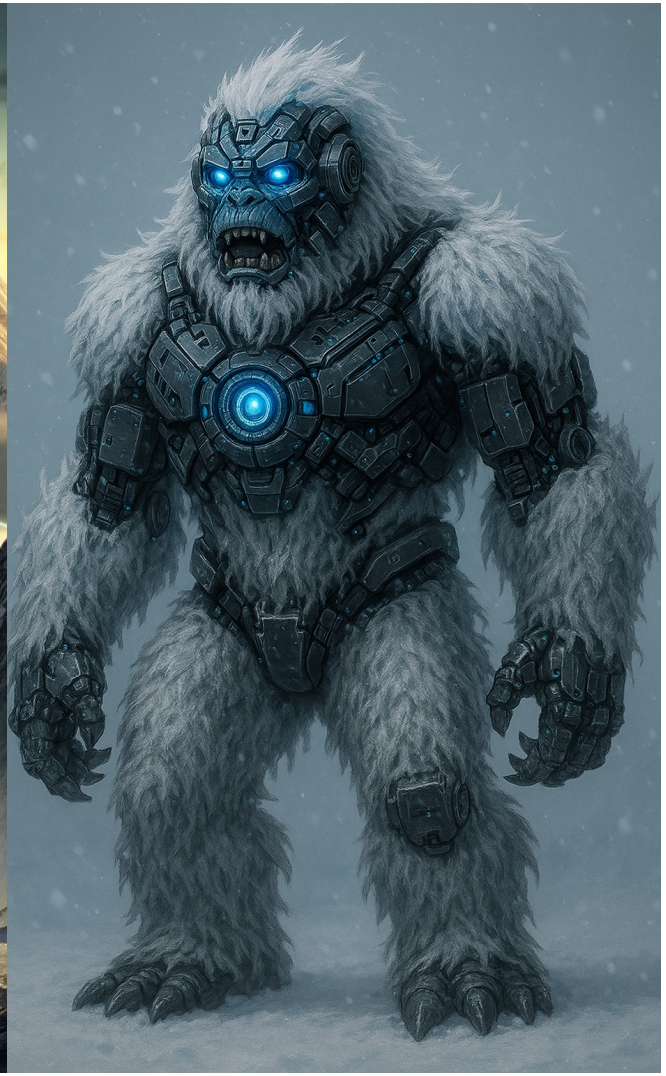


It push our Darwinian buttons

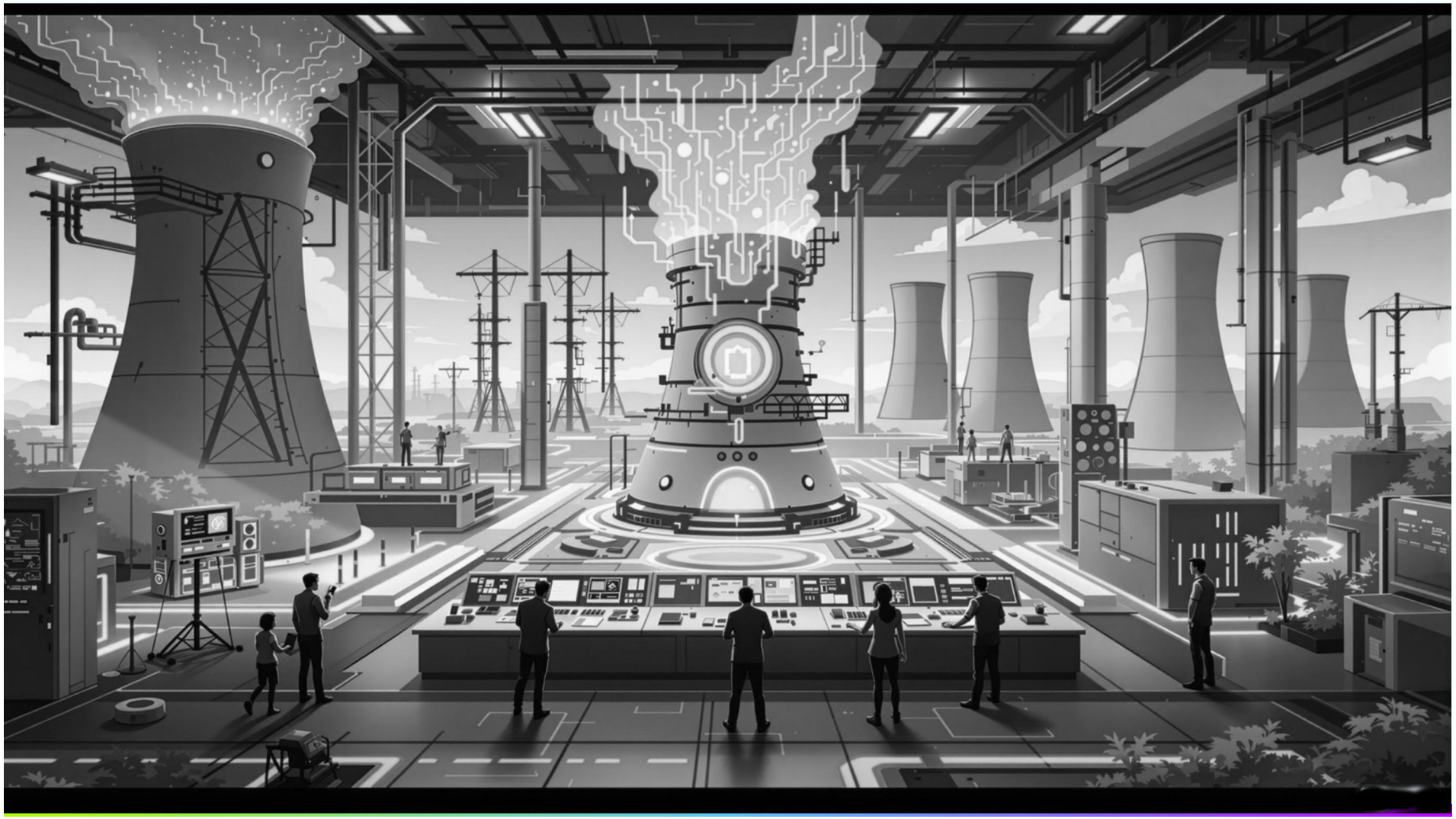
Psychologist, Sherry Turkle

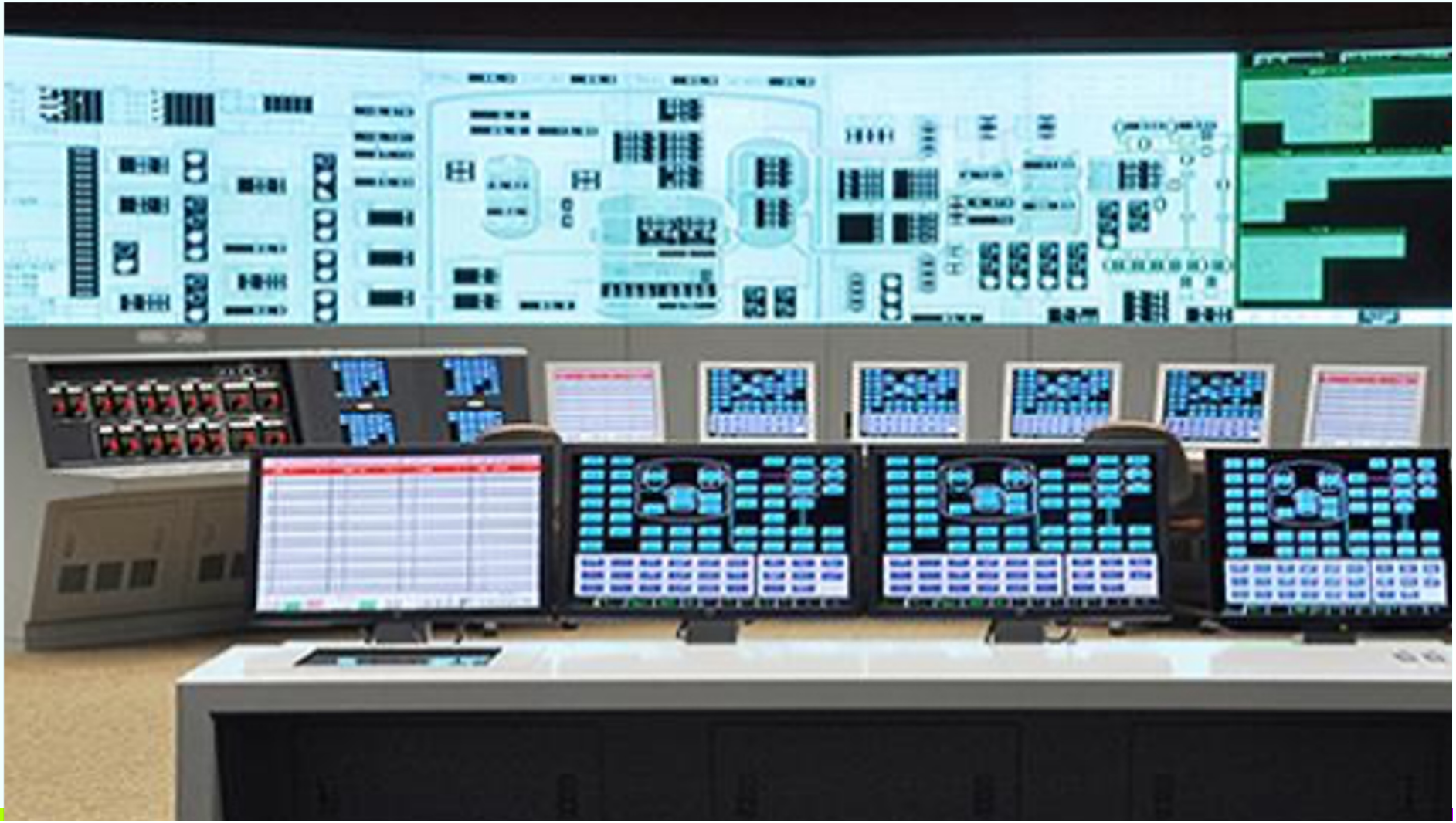
Cognitive Hacking

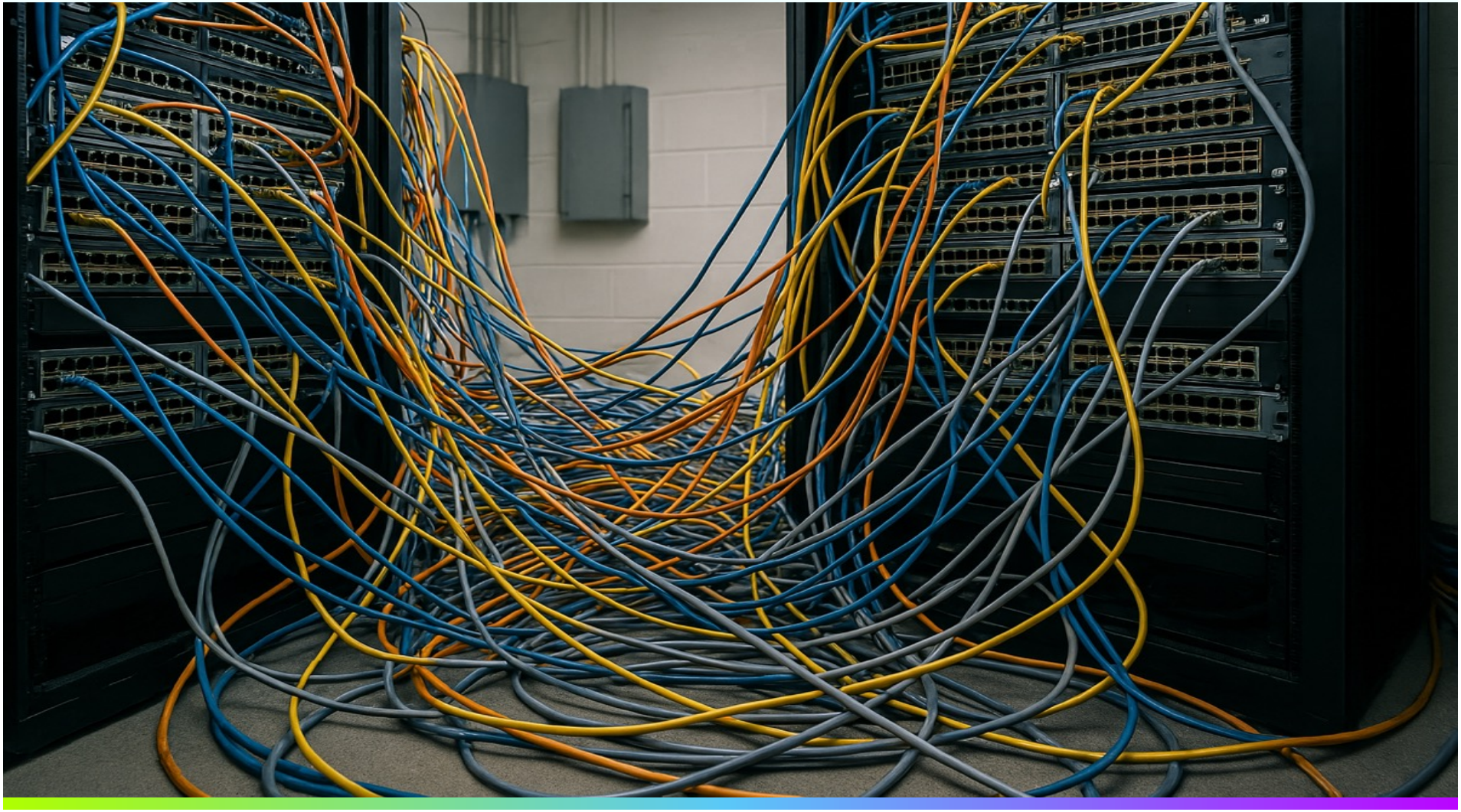
Exploiting vulnerabilities in how people think, feel, and make decisions



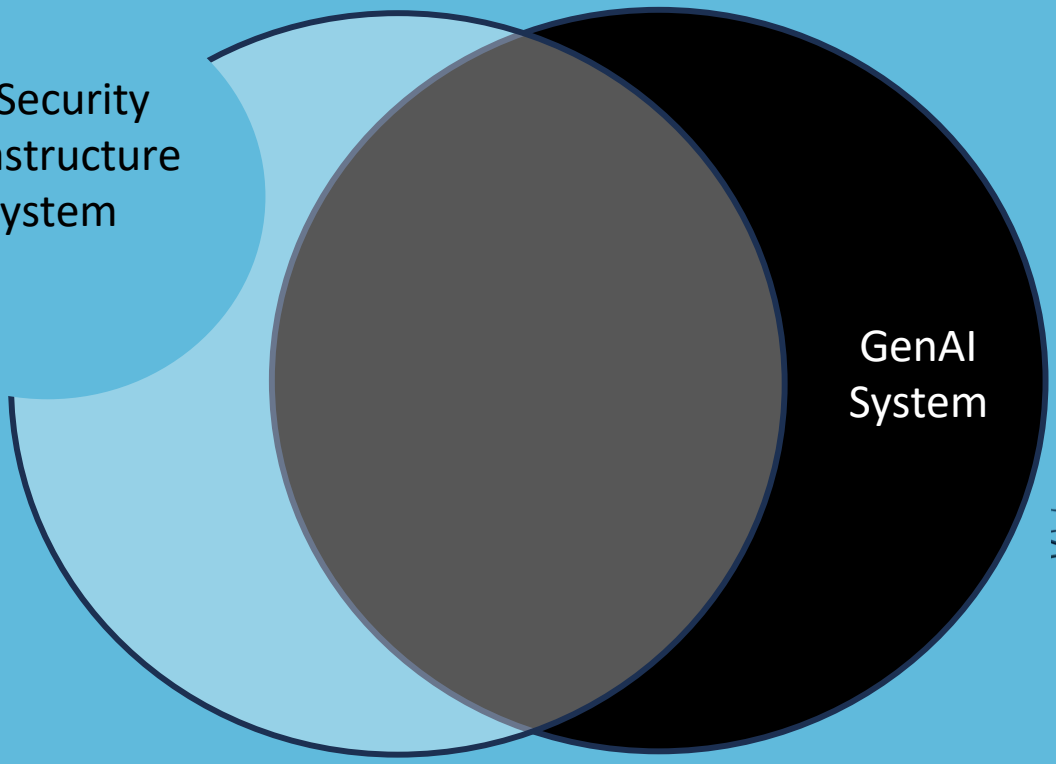
What is the WORST Threats I need to Be Prepared For?

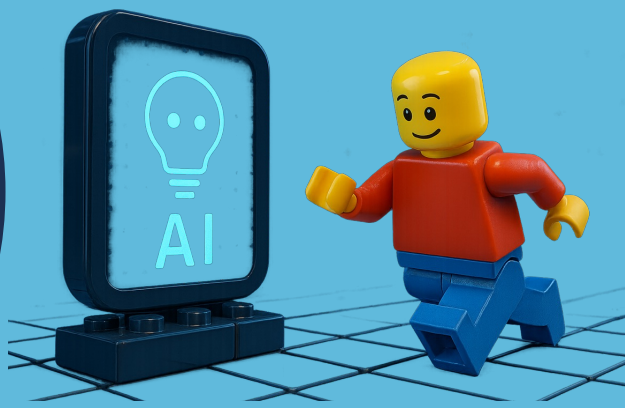




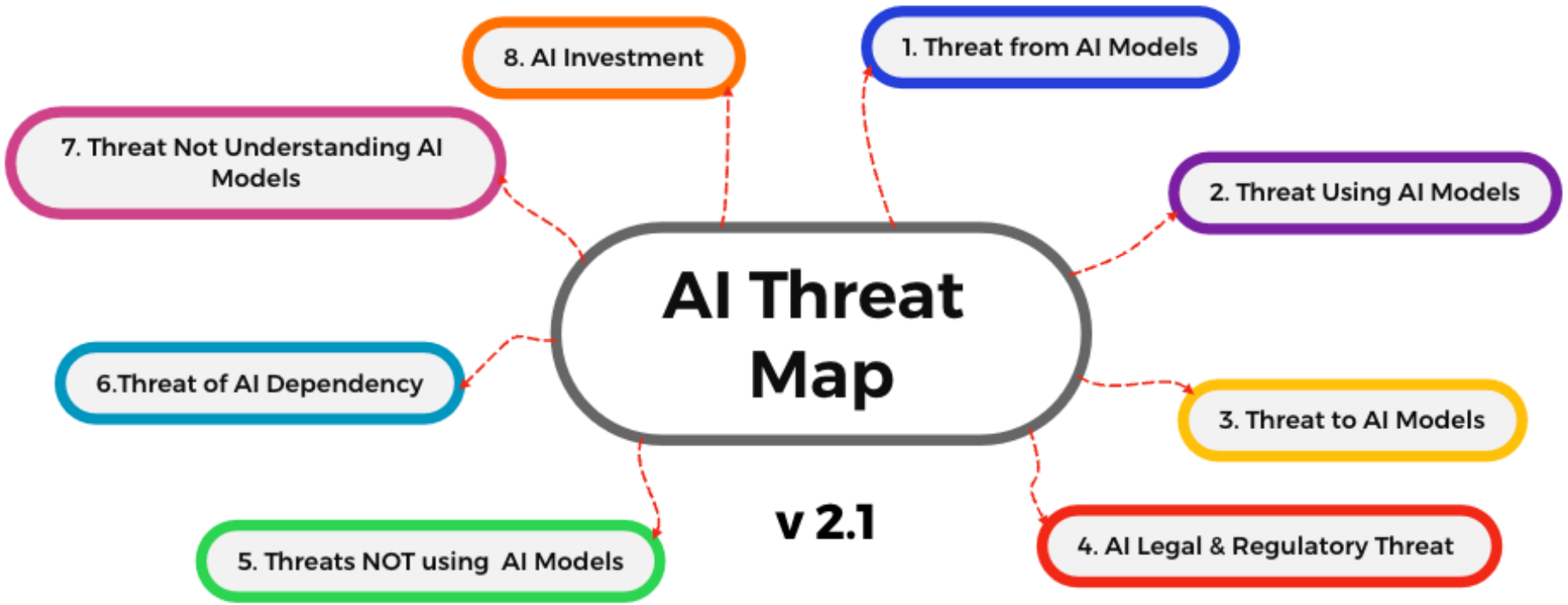


Goldilocks Zone



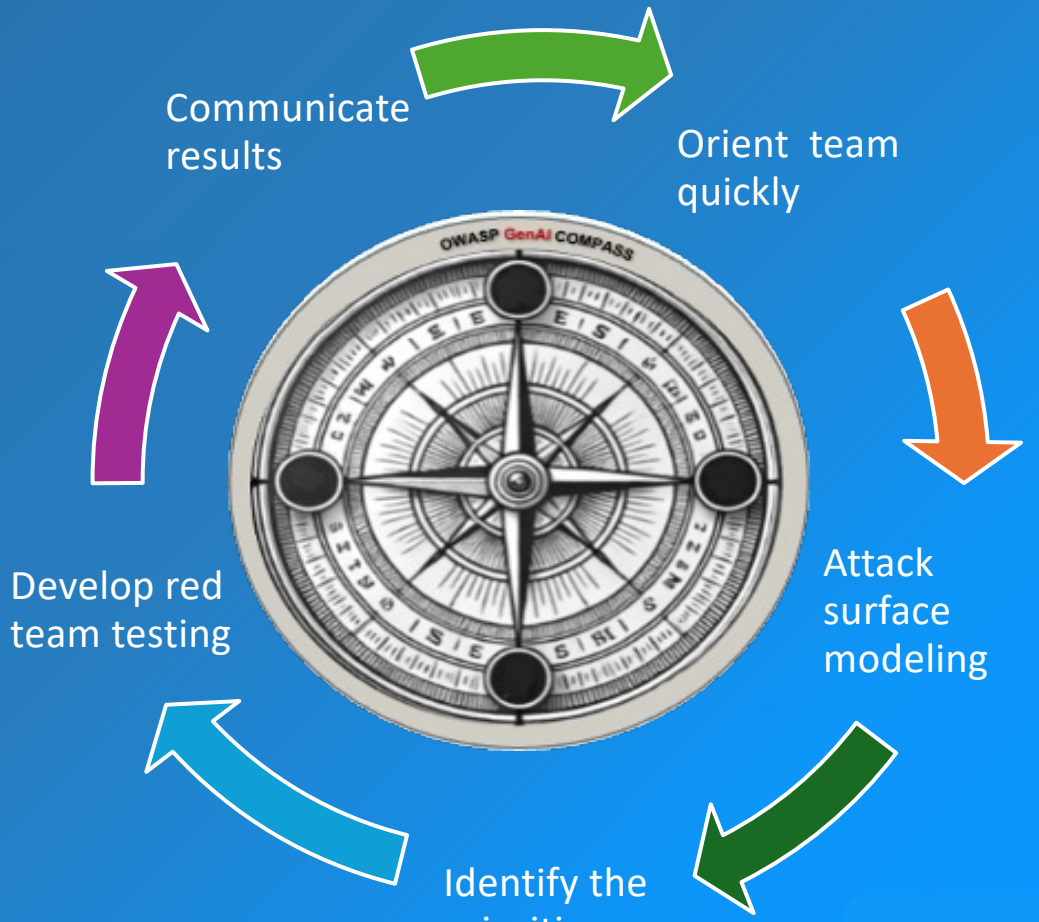


AI Threat Map



OWASP GenAI Project

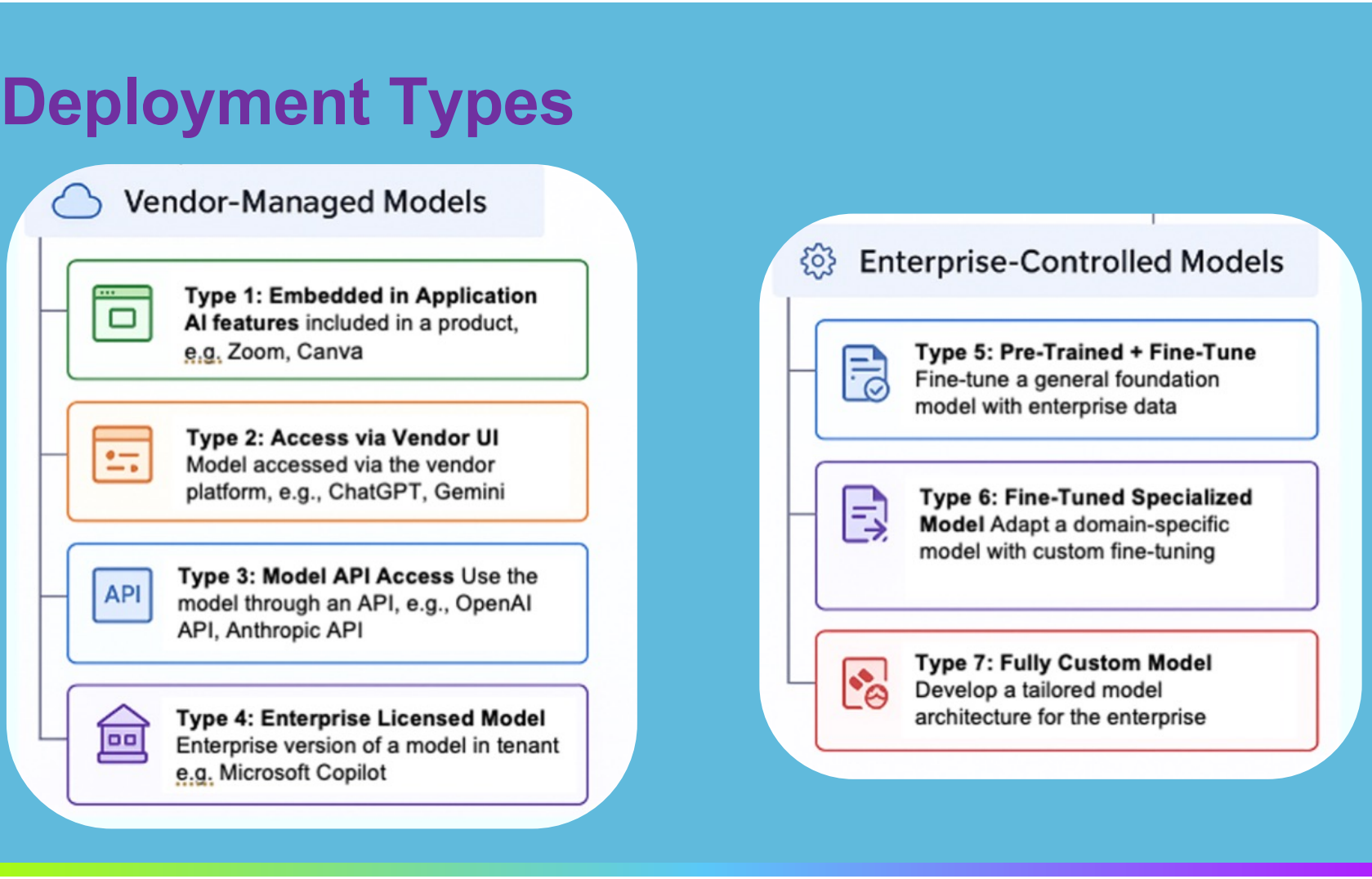
AI Threat Defense COMPASS v1.1



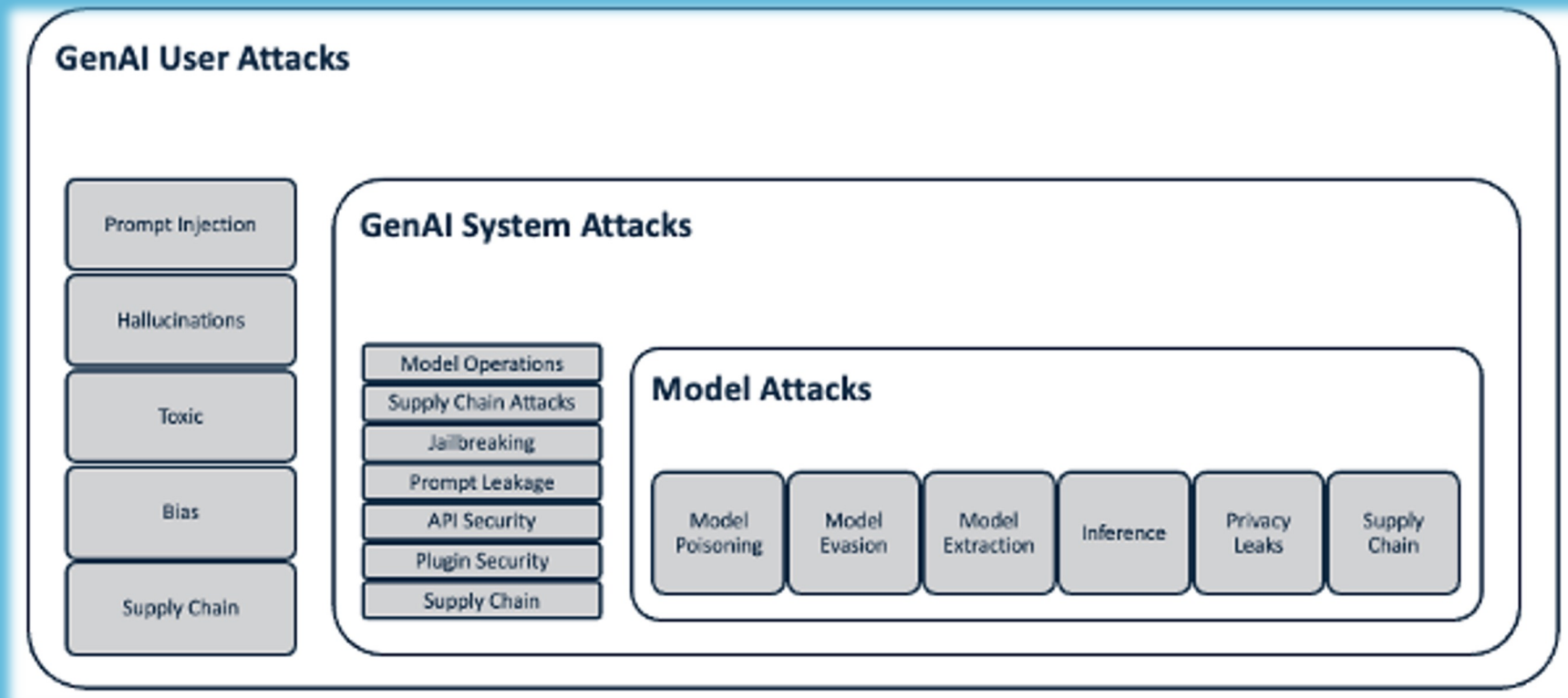
Attack Surface Modeling

Or

Deployment Types



Types of AI Attacks

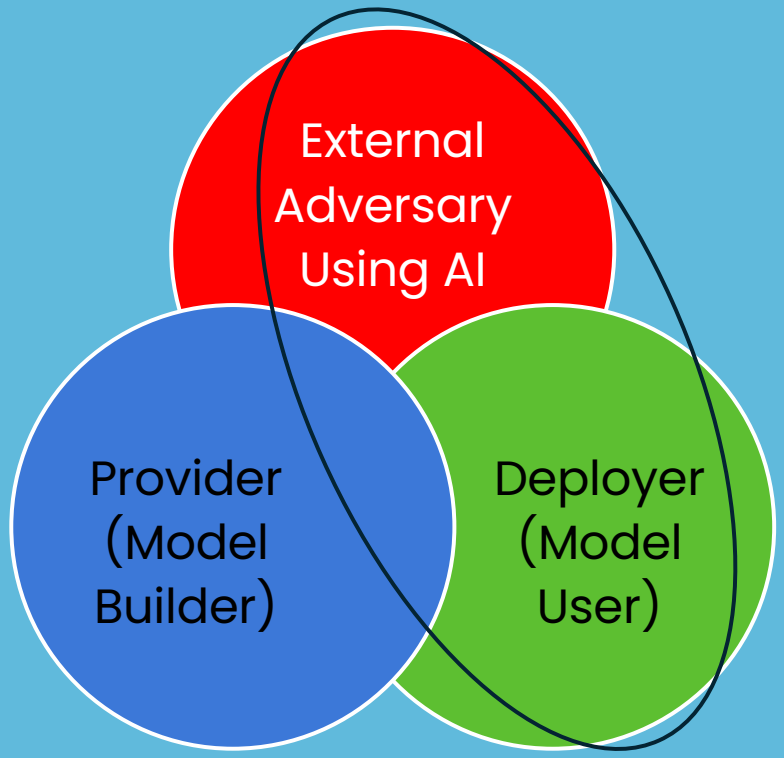


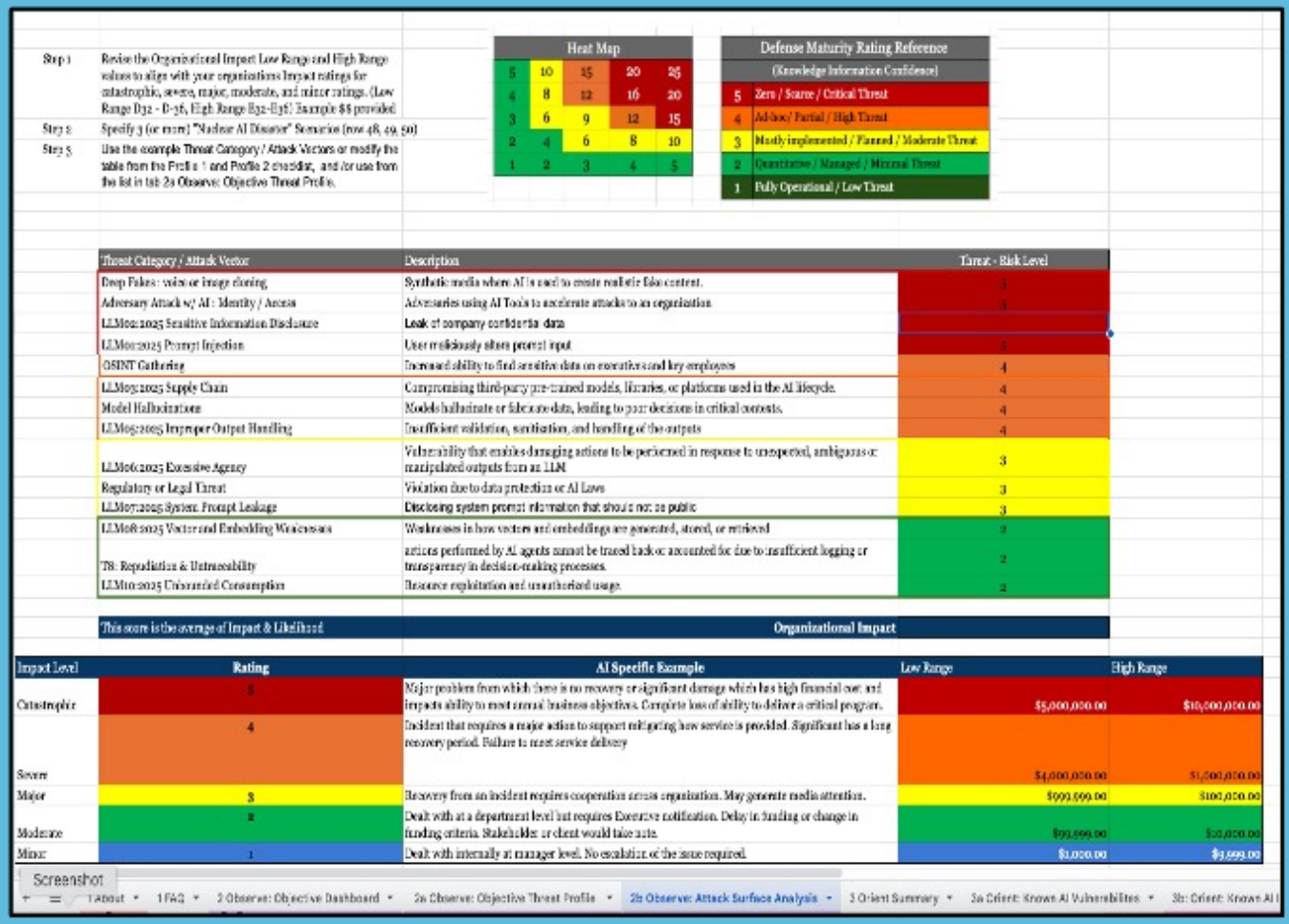
AI Security Shared Responsibility Matrix

mikeprivette

Security Domain	SaaS AI	PaaS AI	IaaS AI	On-Premises	Embedded AI	Agentic AI	AI Coding	MCP Systems
Application Security	Shared	Customer	Customer	Customer	Shared	Shared	Customer	Customer
AI Ethics and Safety	Provider	Shared	Customer	Customer	Provider	Shared	Provider	Customer
User Access Control	Shared	Shared	Customer	Customer	Shared	Customer	Customer	Customer
Model Security	Provider	Shared	Customer	Customer	Provider	Shared	Provider	Shared
Data Privacy	Shared	Shared	Customer	Customer	Shared	Customer	Customer	Customer
Data Security	Shared	Shared	Customer	Customer	Shared	Customer	Customer	Customer
Monitoring and Logging	Shared	Shared	Customer	Customer	Provider	Shared	Customer	Customer
Compliance and Governance	Shared	Shared	Customer	Customer	Shared	Customer	Customer	Customer
Supply Chain Security	Provider	Shared	Shared	Customer	Provider	Shared	Provider	Shared
Network Security	Provider	Shared	Customer	Customer	Provider	Customer	Customer	Customer
Infrastructure Security	Provider	Provider	Shared	Customer	Provider	Shared	Customer	Shared
Incident Response	Shared	Shared	Customer	Customer	Shared	Shared	Customer	Customer
Agent Governance ★	N/A	Shared	Customer	Customer	Shared	Customer ▲	N/A	Customer
Code Generation Security ★	N/A	N/A	N/A	N/A	N/A	N/A	Customer ▲	N/A
Context Pollution Protection ★	Shared	Shared	Customer	Customer	Shared	Shared	Customer ▲	Customer ▲
Multi-System Integration ★	Shared	Shared	Customer	Customer	Shared ▲	Shared	Customer ▲	Customer ▲

Evaluate by Profile





Step 1	Revise the Organizational Impact Low Range and High Range values to align with your organizations Impact ratings for catastrophic, severe, major, moderate, and minor ratings. (Low Range D32 - D-36, High Range E32-E36) Example $$ provided
Step 2	Specify 3 (or more) "Nadine AI Disaster" Scenarios (row 48, 49, 50)
Step 3	Use the example Threat Category / Attack Vectors or modify the table from the Profile 1 and Profile 2 checklist, and /or use from the list in tab 2a Observe: Objective Threat Profile.

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

Defense Maturity Rating Reference

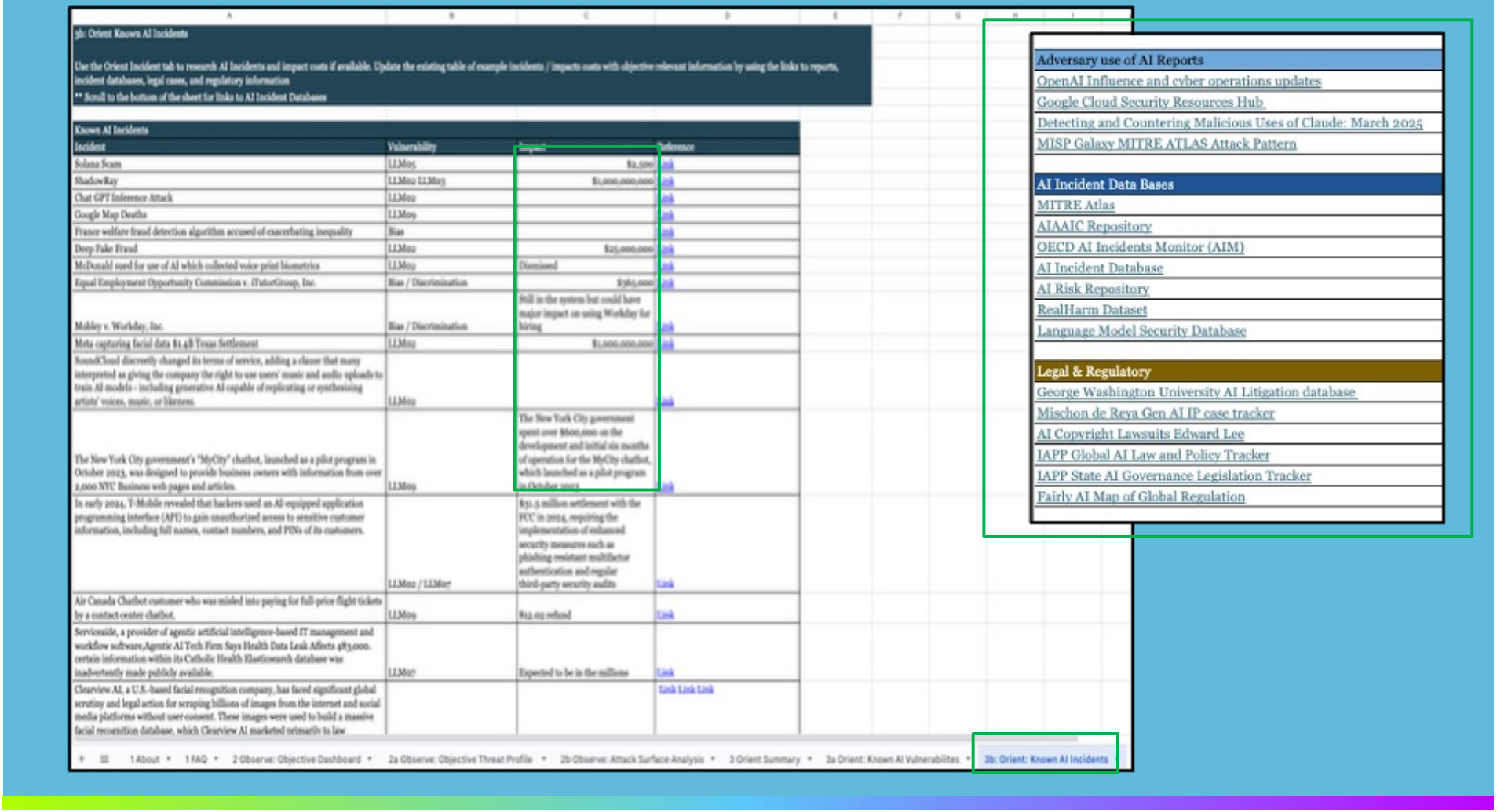
	(Knowledge Information Confidence)
5	Zero / Scarce / Critical Threat
4	Ad-hoc/ Partial / High Threat
3	Mostly implemented / Planned / Moderate Threat
2	Quantitative / Managed / Minimal Threat
1	Fully Operational / Low Threat

Threat Category / Attack Vector	Description	Threat - Risk Level
Deep Fakes: voice or image cloning	Synthetic media where AI is used to create realistic fake content.	5
Adversary Attack w/ AI : Identity / Access	Adversaries using AI Tools to accelerate attacks to an organization	5
LLM02:2025 Sensitive Information Disclosure	Leak of company confidential data	5
LLM01:2025 Prompt Injection	User maliciously alters prompt input	5
OSINT Gathering	Increased ability to find sensitive data on executives and key employees	4
LLM03:2025 Supply Chain	Compromising third-party pre-trained models, libraries, or platforms used in the AI lifecycle.	4
Model Hallucinations	Models hallucinate or fabricate data, leading to poor decisions in critical contexts.	4
LLM05:2025 Improper Output Handling	Insufficient validation, sanitization, and handling of the outputs	4
LLM06:2025 Excessive Agency	Vulnerability that enables damaging actions to be performed in response to unexpected, ambiguous or manipulated outputs from an LLM	3
Regulatory or Legal Threat	Violation due to data protection or AI Laws	3
LLM07:2025 System Prompt Leakage	Disclosing system prompt information that should not be public	3
LLM08:2025 Vector and Embedding Weaknesses	Weaknesses in how vectors and embeddings are generated, stored, or retrieved	2
T8: Repudiation & Untraceability	actions performed by AI agents cannot be traced back or accounted for due to insufficient logging or transparency in decision-making processes.	2
LLM10:2025 Unbounded Consumption	Resource exploitation and unauthorized usage.	2
This score is the average of Impact & Likelihood	Organizational Impact	

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 and 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	$4,000,000.00	$1,000,000.00
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

Screenshot

1 About | 1 FAQ | 2 Observe: Objective Dashboard | 2a Observe: Objective Threat Profile | 2b Observe: Attack Surface Analysis | 3 Orient Summary | 3a Orient: Known AI Vulnerabilities | 3b: Orient: Known AI



3b: Orient Known AI Incidents

Use the Orient Incident tab to research AI Incidents and impact costs if available. Update the existing table of example incidents / impacts costs with objective relevant information by using the links to reports, incident databases, legal cases, and regulatory information

** Scroll to the bottom of the sheet for links to AI Incident Databases

Known AI Incidents

Incident	Vulnerability	Impact	Reference
Solana Scam	LLM05	$2,500	Link
ShadowRay	LLM02 LLM03	$1,000,000,000	Link
Chat GPT Inference Attack	LLM02		Link
Google Map Deaths	LLM09		Link
France welfare fraud detection algorithm accused of exacerbating inequality	Bias		Link
Deep Fake Fraud	LLM02	$25,000,000	Link
McDonald sued for use of AI which collected voice print biometrics	LLM02	Dismissed	Link
Equal Employment Opportunity Commission v. iTutorGroup, Inc.	Bias / Discrimination	$365,000	Link
Mobley v. Workday, Inc.	Bias / Discrimination	Still in the system but could have major impact on using Workday for hiring	Link
Meta capturing facial data $1.4B Texas Settlement	LLM02	$1,000,000,000	Link
SoundCloud discreetly changed its terms of service, adding a clause that many interpreted as giving the company the right to use users' music and audio uploads to train AI models - including generative AI capable of replicating or synthesizing artists' voices, music, or likeness.	LLM02		Link
The New York City government's "MyCity" chatbot, launched as a pilot program in October 2023, was designed to provide business owners with information from over 2,000 NYC Business web pages and articles.	LLM09	The New York City government spent over $600,000 on the development and initial six months of operation for the MyCity chatbot, which launched as a pilot program in October 2023	Link
In early 2024, T-Mobile revealed that hackers used an AI-equipped application programming interface (API) to gain unauthorized access to sensitive customer information, including full names, contact numbers, and PINs of its customers.	LLM02 / LLM07	$31.5 million settlement with the FCC in 2024, requiring the implementation of enhanced security measures such as phishing-resistant multifactor authentication and regular third-party security audits	Link
Air Canada Chatbot customer who was misled into paying for full-price flight tickets by a contact center chatbot.	LLM09	$13.03 refund	Link
Servicenaide, a provider of agentic artificial intelligence-based IT management and workflow software,Agentic AI Tech Firm Says Health Data Leak Affects 483,000. certain information within its Catholic Health Elasticsearch database was inadvertently made publicly available.	LLM07	Expected to be in the millions	Link
Clearview AI, a U.S.-based facial recognition company, has faced significant global scrutiny and legal action for scraping billions of images from the internet and social media platforms without user consent. These images were used to build a massive facial recognition database, which Clearview AI marketed primarily to law			Link Link Link

1 About | 1 FAQ | 2 Observe: Objective Dashboard | 2a Observe: Objective Threat Profile | 2b Observe: Attack Surface Analysis | 3 Orient Summary | 3a Orient: Known AI Vulnerabilites | 3b: Orient: Known AI Incidents

Adversary use of AI Reports
OpenAI Influence and cyber operations updates
Google Cloud Security Resources Hub
Detecting and Countering Malicious Uses of Claude: March 2025
MISP Galaxy MITRE ATLAS Attack Pattern

AI Incident Data Bases
MITRE Atlas
AIAAIC Repository
OECD AI Incidents Monitor (AIM)
AI Incident Database
AI Risk Repository
RealHarm Dataset
Language Model Security Database

Legal & Regulatory
George Washington University AI Litigation database
Mischon de Reya Gen AI IP case tracker
AI Copyright Lawsuits Edward Lee
IAPP Global AI Law and Policy Tracker
IAPP State AI Governance Legislation Tracker
Fairly AI Map of Global Regulation

Red Team Testing

- ✓ Expose Context Specific Security Flaws
- ✓ Prevent Critical Safety Failures
- ✓ Uncover and Mitigate Bias
- ✓ Validate Real-World Robustness
- ✓ Prompt Injection and Manipulation
- ✓ Sensitive Data Leakage
- ✓ Alignment
- ✓ Regulatory Compliance
- ✓ Hallucinations & Misinformation

AI SKILLS

2nd Brain / Library

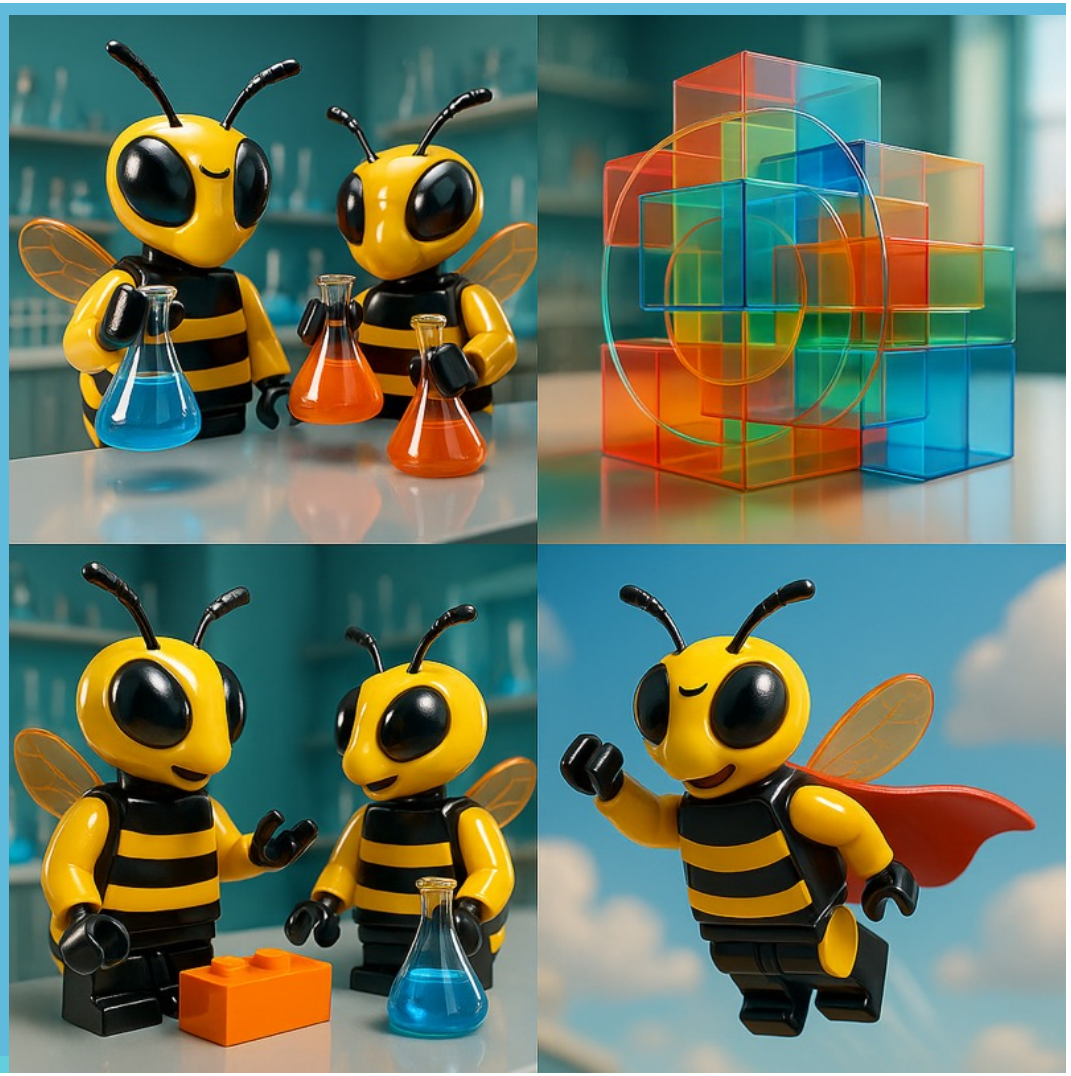
Experiment

Critical Thinking

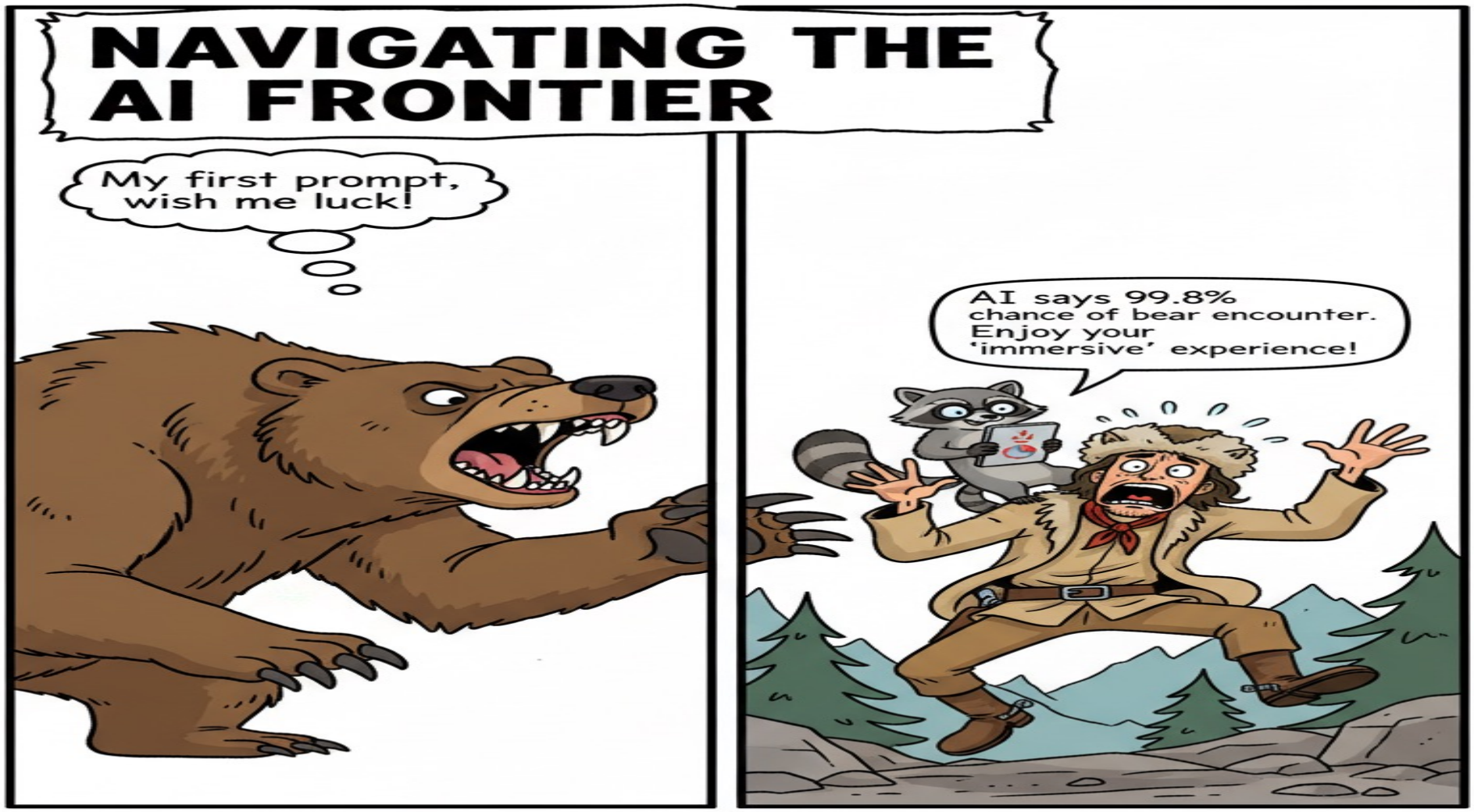
Problem Solving

Human Skills

Your superpower

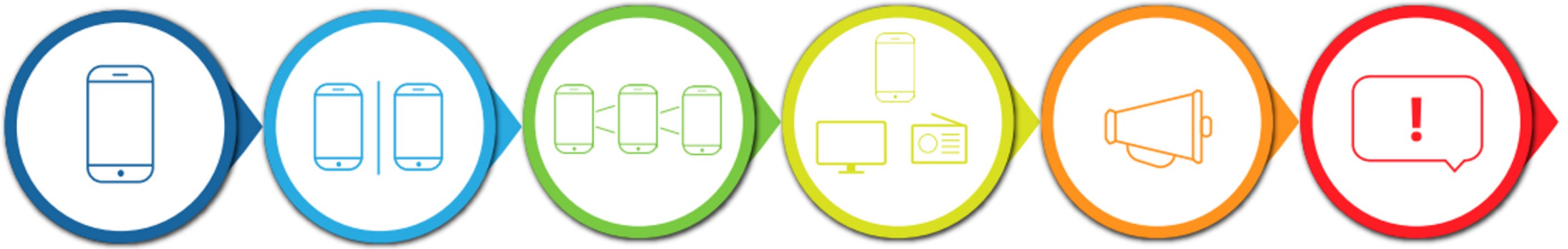



OWASP GenAI SECURITY PROJECT
TOP 10 FOR LLM AND GENERATIVE AI
GETTING STARTED
RESOURCES
PROJECTS
BLOG
ABOUT
https://genai.owasp.org/
IDENTIFYING AND TACKLING THE RISKS OF GEN AI SYSTEMS AND APPLICATIONS
OWASP GenAI Security Project
A global community-driven and expert led initiative to create freely available open source guidance and resources for understanding and mitigating security and safety concerns for Generative AI applications and adoption.
15k+
Members
15+
Countries
20+
AI Cybersecurity Publications
Cybersecurity Practitioners
Top 10 for LLM and GenAI
AI Threat Intelligence
AI Security Solution Landscape
AI Security Governance
Red Teaming & Evaluation
Secure AI Adoption
Data Security
Agentic App Security
Gen AI Developers
Data Scientists
CISOs, CTO, CIOs
OWASP GenAI SECURITY PROJECT
genai.owasp.org


NAVIGATING THE AI FRONTIER
My first prompt, wish me luck!
AI says 99.8% chance of bear encounter. Enjoy your 'immersive' experience!

Questions ?

The Breakout Scale



CATEGORY ONE
one platform, no breakout

CATEGORY TWO
one platform, breakout OR many platforms, no breakout

CATEGORY THREE
multiple platforms, multiple breakouts

CATEGORY FOUR
cross-medium breakout

CATEGORY FIVE
celebrity amplification

CATEGORY SIX
policy response OR call for violence