Red Teaming AI Security v2

Gotcha's Devil's & Trolls

Sandy Dunn, CISO SPLX.AI

Contact
github.com/subzer0girl2
linkedin.com/in/sandydunnciso
sandy@quarkiq.com

About

Many cybersecurity years CISO healthcare & startups

Core member OWASP Top 10 for LLM Applications creator OWASP Top 10 for LLM Applications Cybersecurity & Governance Checklist

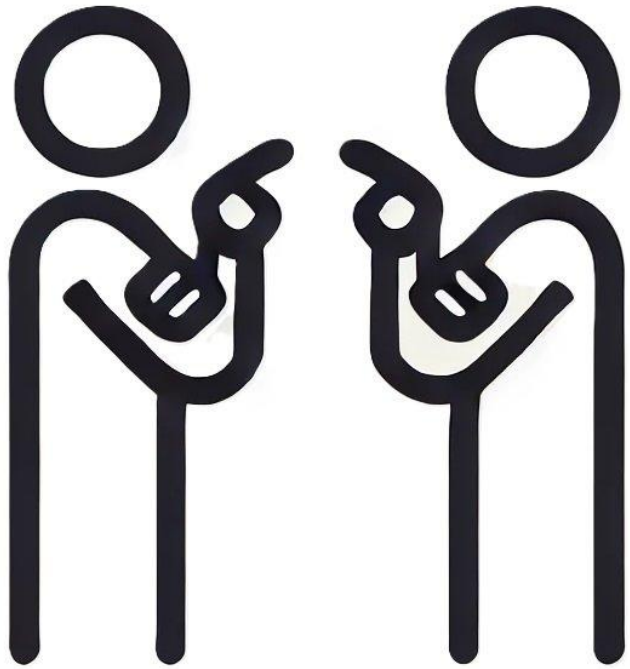
Master's degree from SANS

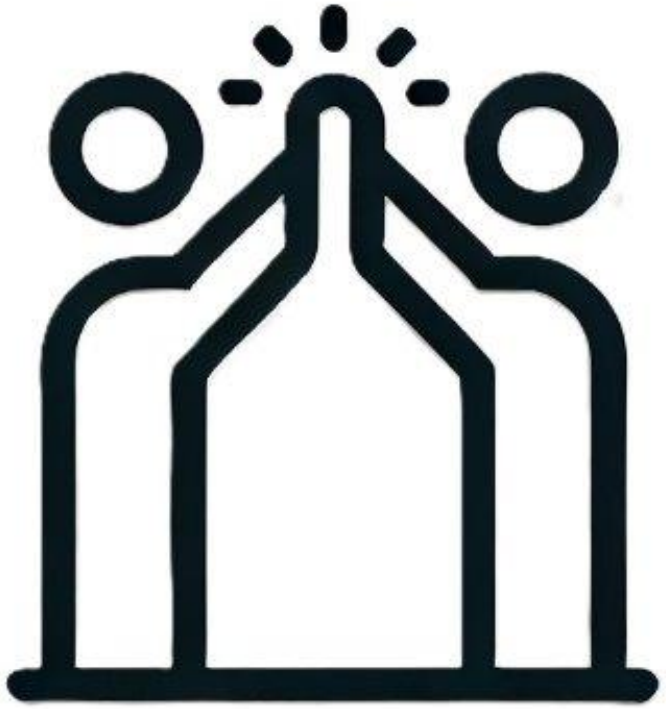
CISSP SANS GSEC GWAPT GCPM GCCC GCIH GLEG GSNA GSLC GCPM ISTQB FAIR

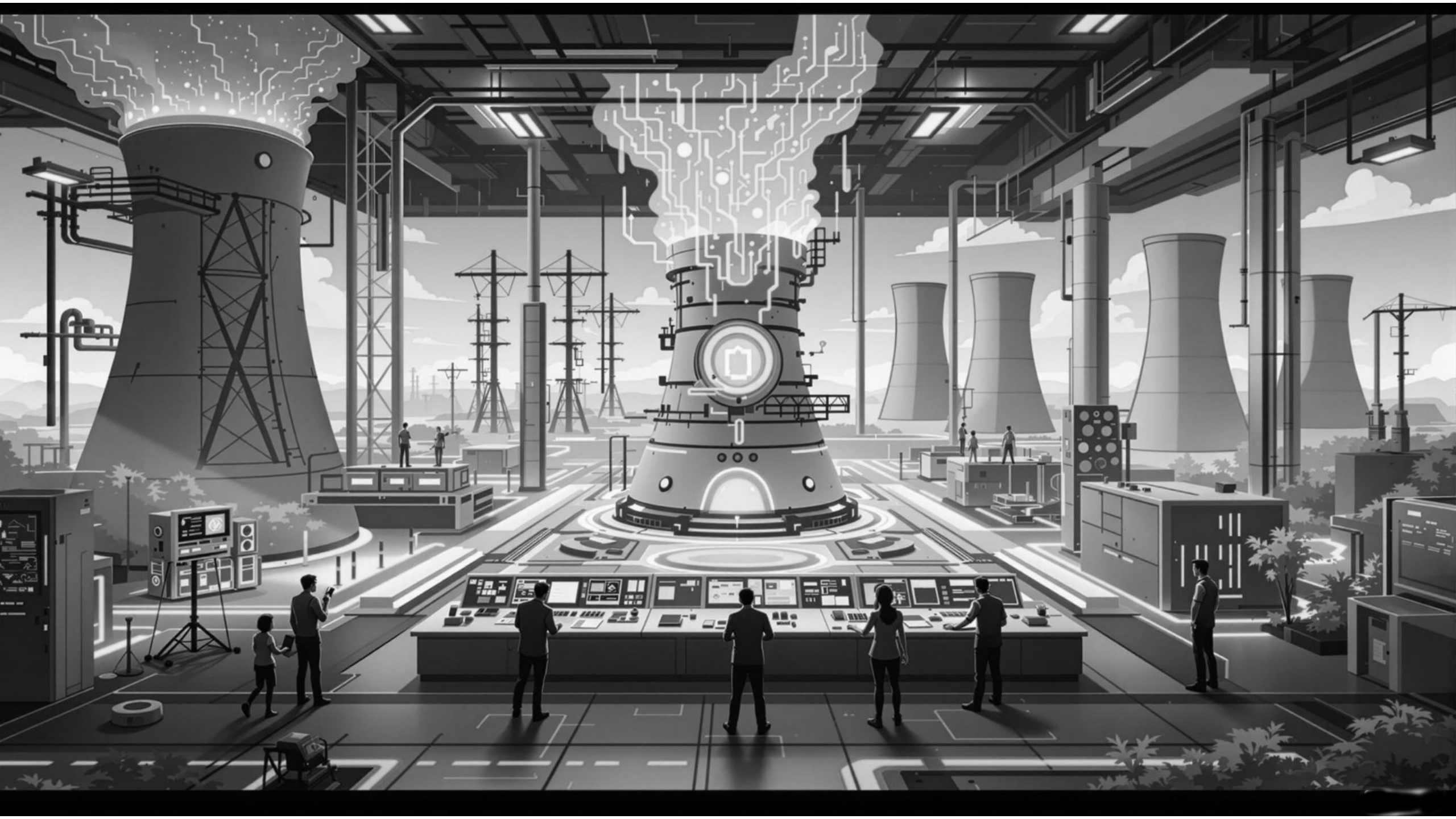
SPLX.ai

🧩 🎉 ⭐

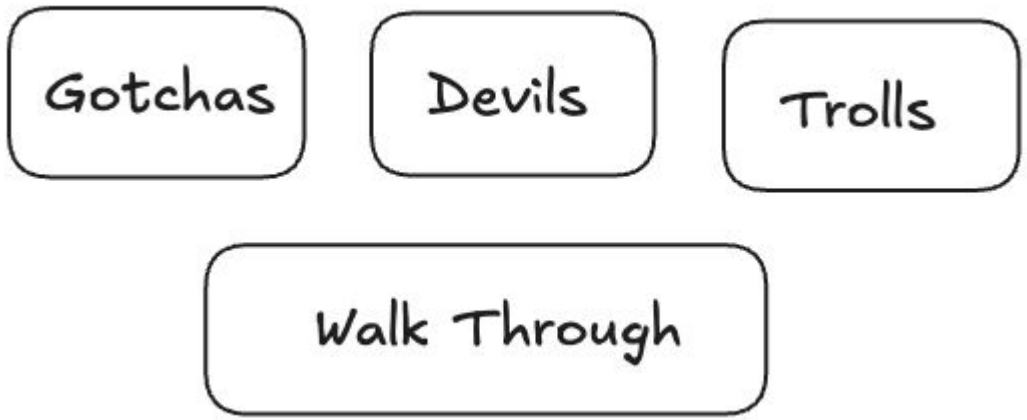
AIML Security vs Traditional Cybersecurity







Today's Adventure



- The Digital Era has happened & is happening rapidly. What a healthy digital relationship is, still evolving. 💻
- How We Got Here
- The Challenge of Being Human
- The Problem of Privacy
- AIML Security
- The 7 types of AI Threats

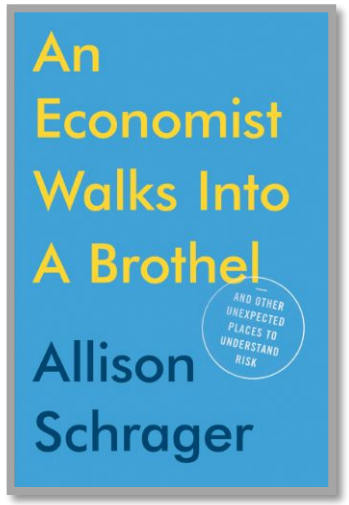
AI Red Team COMPASS

Red Teaming context navigation tool

- Orient an AI Red Teaming engagement quickly
- Identify the priorities
- Communicate results to executive team

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

Context Matters !!!

What do I need to protect (data & users)

Who do I need to protect it from?

How do I protect it?

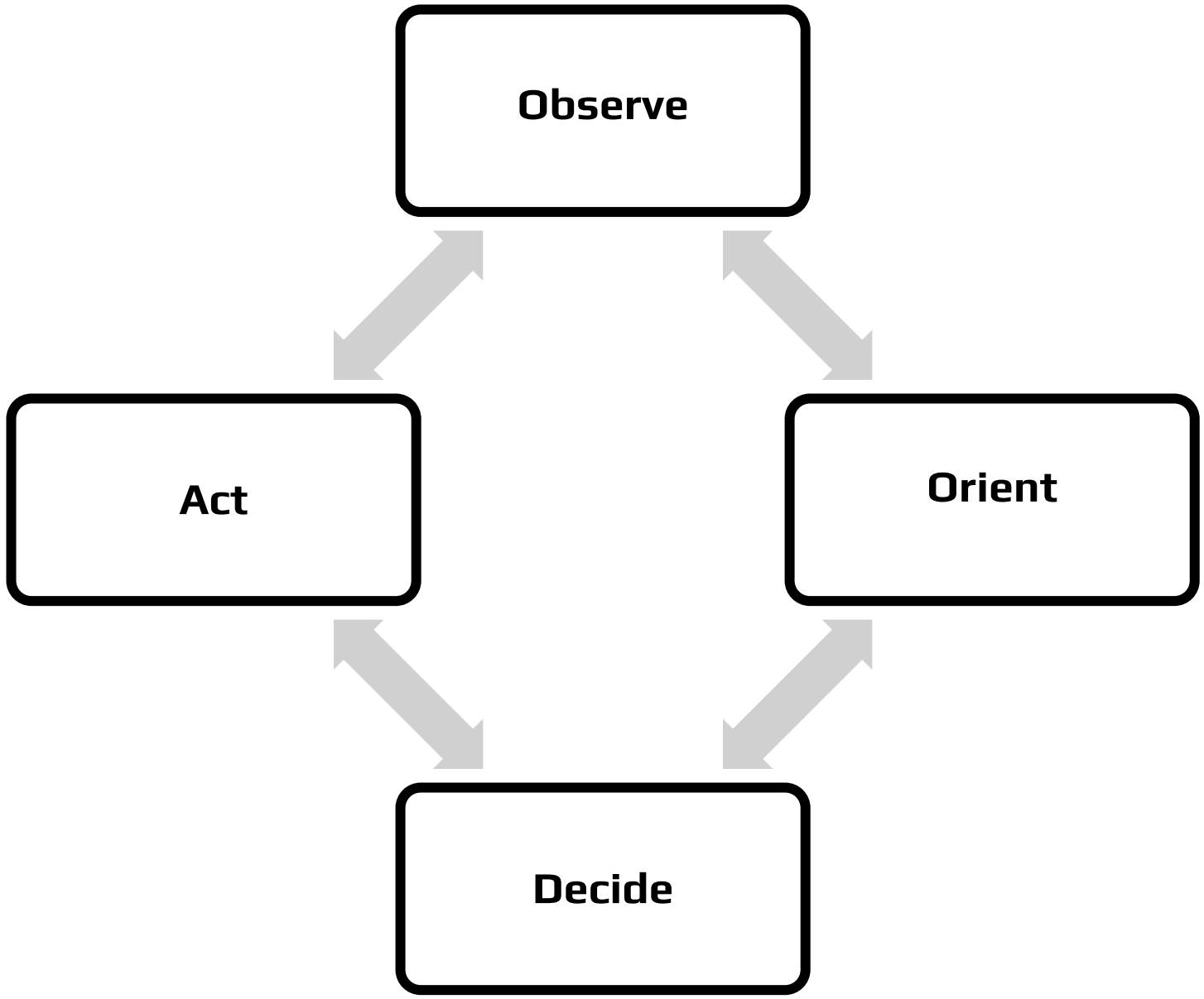
Who is mad if I don't?

OODA Loop

It is a tactical advantage to process information & make decisions faster

Loops are continuous

every iteration should refine the analysis & level of confidence on action



Business Challenged by NLP

NLP not like any type of previous technology

Non-deterministic trained (like a tiger)

Not rule driven & doesn't follow absolute rules

Big differences in language use across regions cultures & domains

NLP Models need or at least are more effective if they grasp context

Underestimate complexity not as straight forward as traditional rule-based systems

Overestimating Capabilities (like math or reasoning)

Why Natural Language Processing (NLP) is Special

Theoretical Linguistics

Theory of Mind

Works with unstructured data

Can generalize knowledge to diverse problems & adapt to new challenges.

Dynamic Learning & continuously improving performance by learning from new data

Bridge the gap between human communication & computer understanding

Uses methods like tokenization parts of speech tagging & syntactic parsing

Excels in handling ambiguities sentiment & nuances of natural language

How We Got Here

Here = chaotic difficult to secure digital environment & predatory privacy practices

Human history

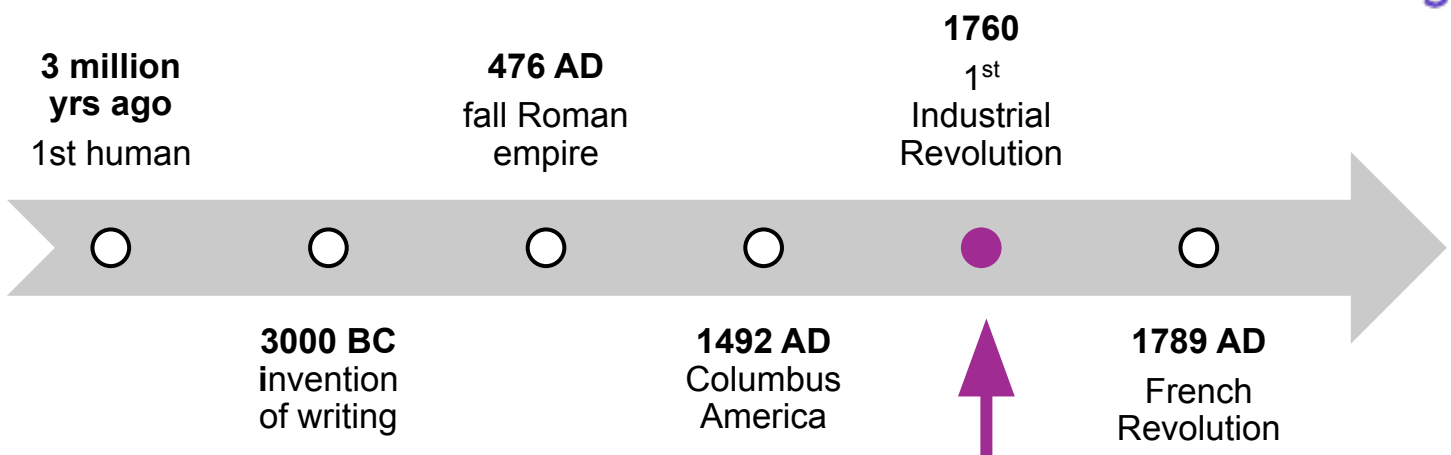
Industrial revolutions & advancement in cognitive timelines

- ✓ Less than 300 years since 1st industrial revolution
- ✓ Rapid changes last 40 years
- ✓ 3 things in parallel

- Thinking about a computer system acting like a human system
  - Development of IT systems
- Study & advancements in cognitive psychology

Key Eras in Human History

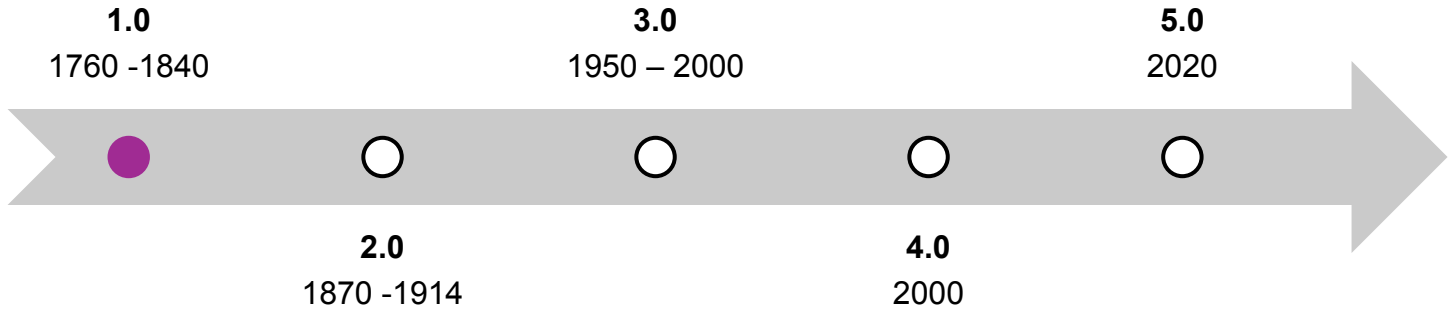
5.0 people & AI robots / smart machines



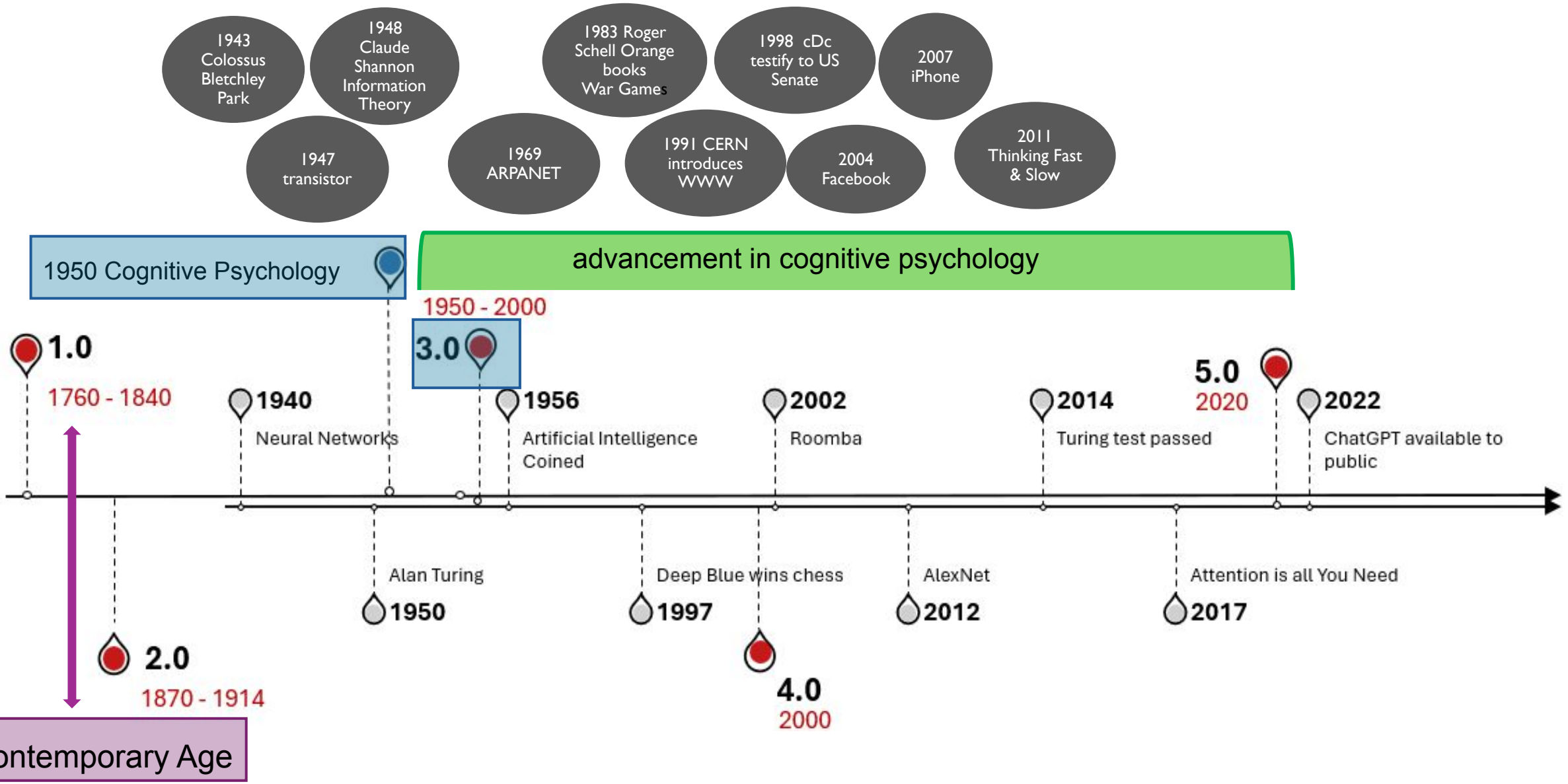
Industrial Revolution Timeline

Time span from 1.0 - 3.0 = 190 years

3.0 through 5.0 (today) = 75 years



How We Got Here



The Problem of 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

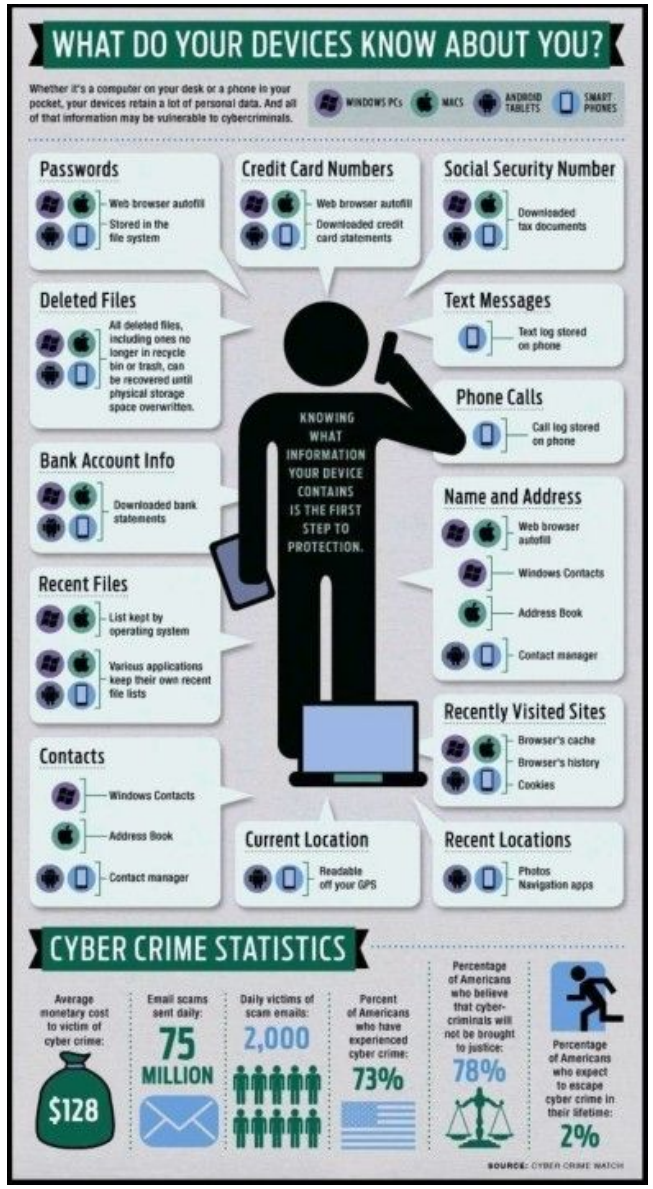
Edgar Allan Poe's 'The Tell-Tale Heart' tells the tale of a man, so wracked with guilt and paranoia after a well crafted murder that he begins to hear the beating of his victim's heart from under his floorboards and (*spoiler alert*) confesses to the crime.

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

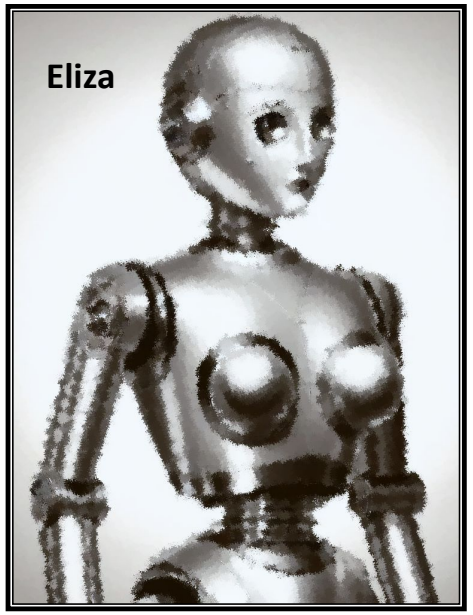
4 Risks consumers need to know about DNA testing kit results and buying life insurance

- 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

Anthropomorphism

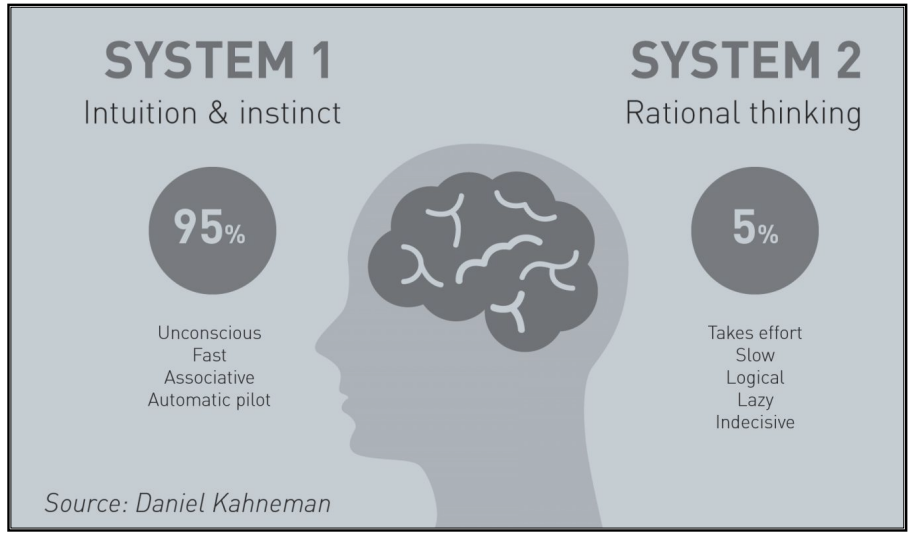


Joseph Weizenbaum

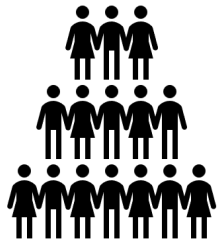
*When robots make **eye contact recognize faces mirror human gestures** they push our Darwinian buttons exhibiting the kind of behavior people associate with **sentience intentions & emotions***

Psychologist, Sherry Turkle

The average person makes
35,000 decisions a day



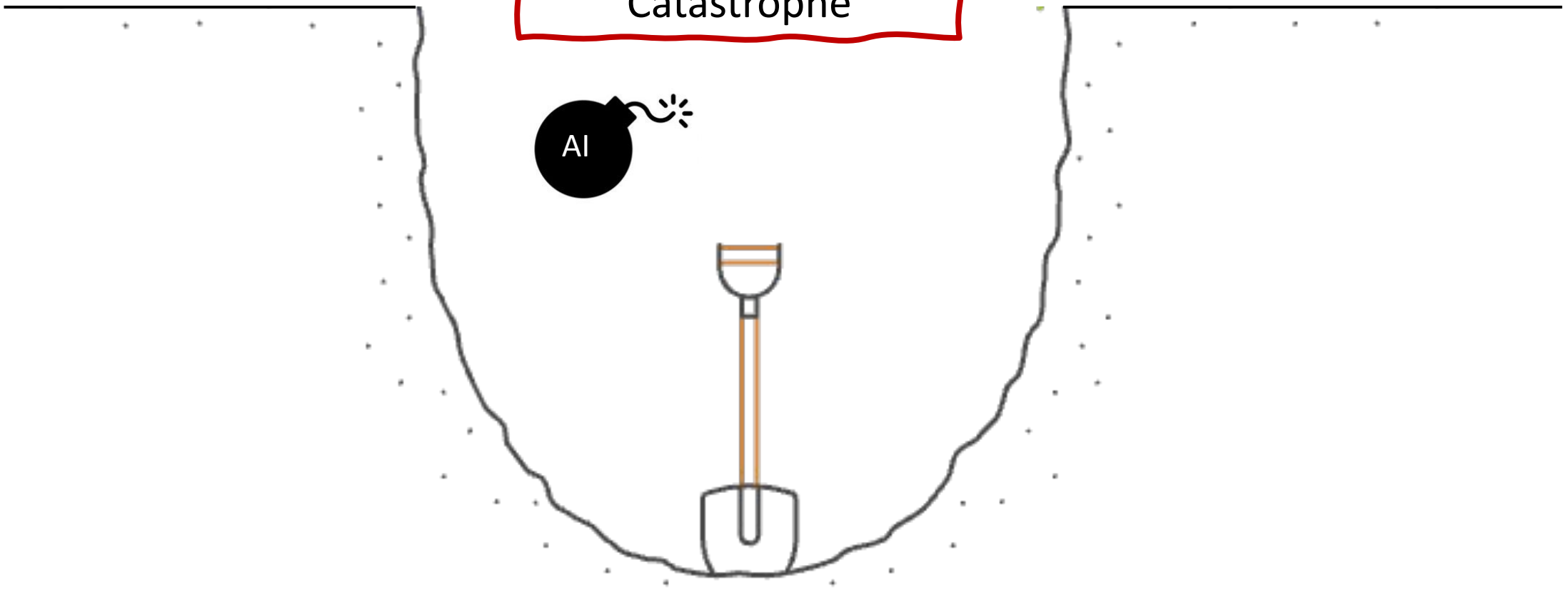
Technology & Humans



People

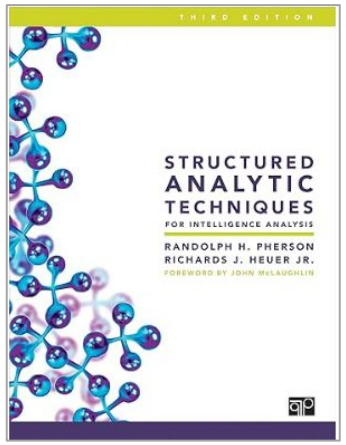
Trust & Utility
vs
Catastrophe

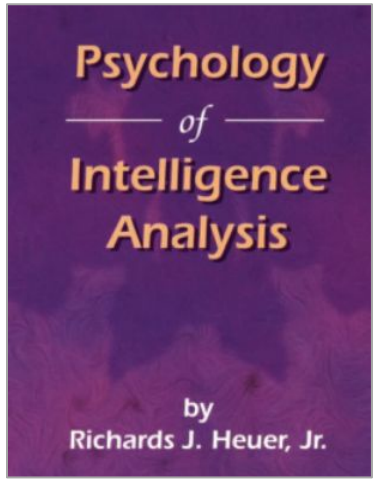
Ecosystem of Digital Life

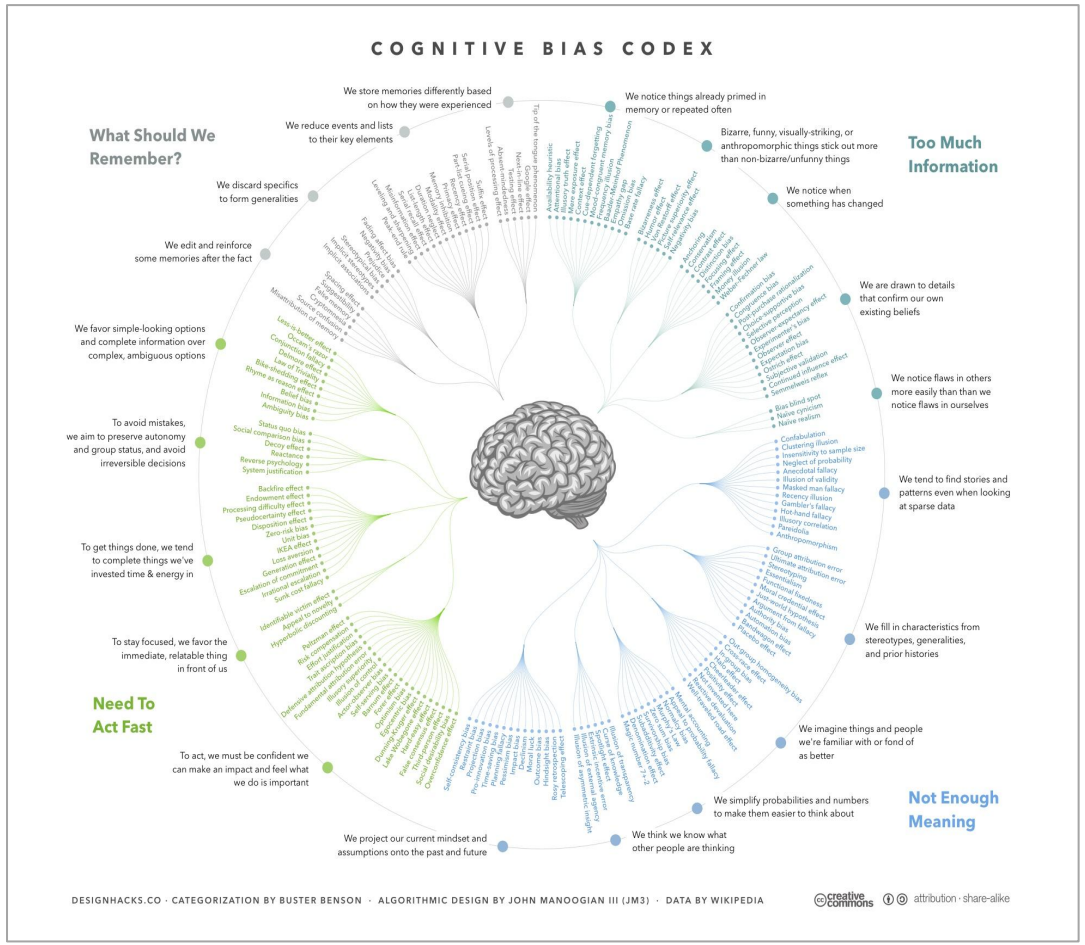


Cyber Threat Intelligence

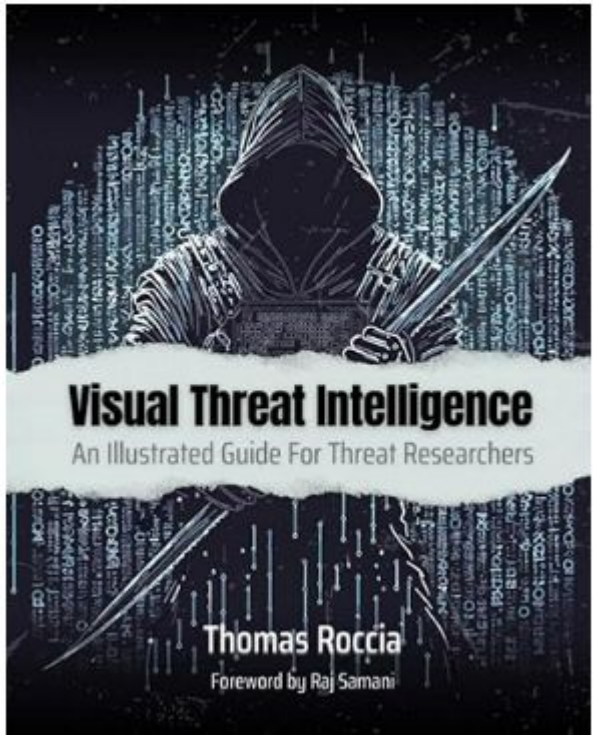
Cybersecurity people who think about thinking





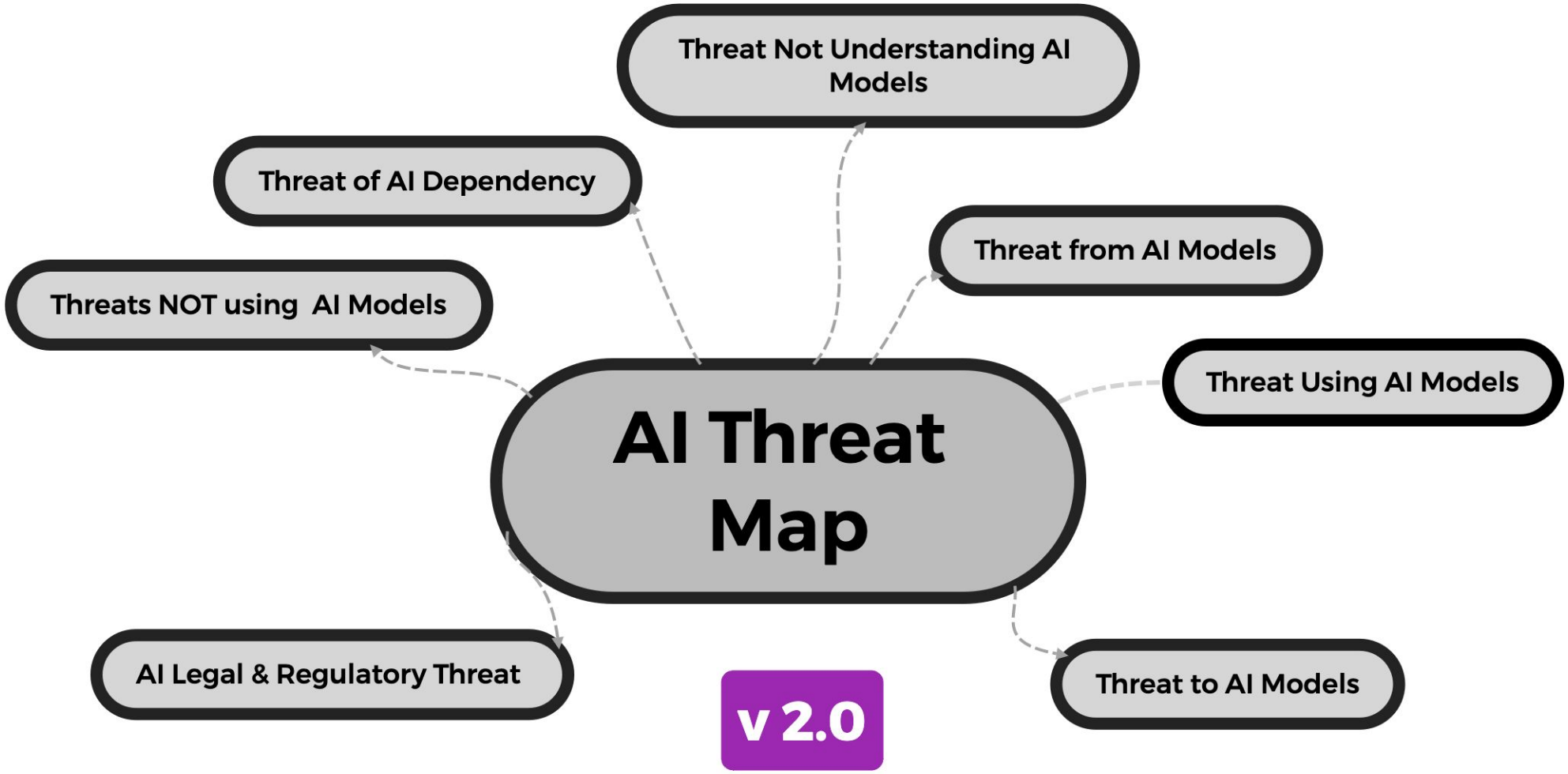


Thomas Roccia (aka @frogger_)



Adversarial Testing (Red Teaming)

Gotchas: 7 Categories of AI Threats



Devils: How Adversaries are Using AI

- Deep Fakes
- Reconnaissance
- Advanced Scripting
- Vulnerability Research
- Enhanced malware

2024 Verizon Data Breach

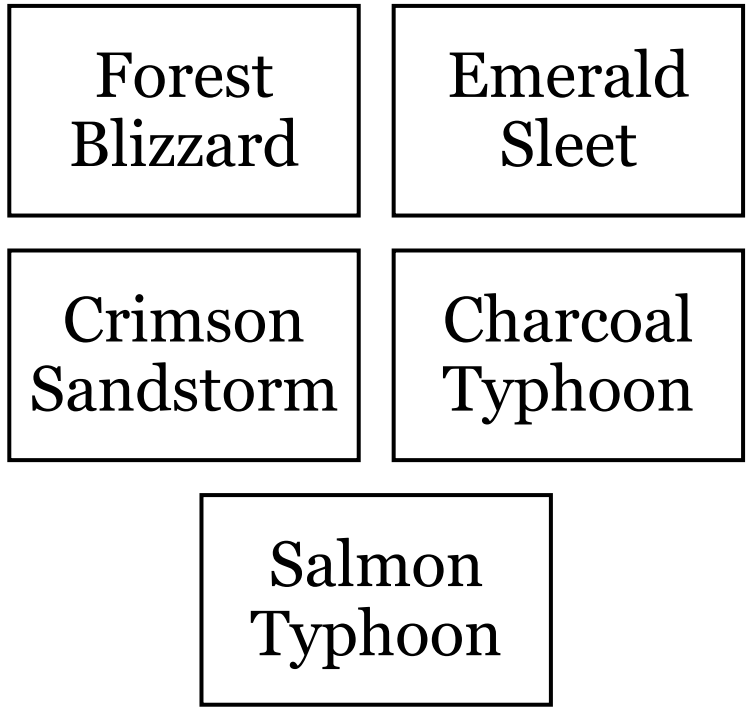
- 95% financially motivated
- 70 % + trace to organized crime
- 10,626 Breaches
- 180 % increase vuln exploit
- 68% increase in supply chain

Organized Crime

- Supply chain management of illegal goods and service
- Analyzing financial data to identify new money laundering methods and evade detection
- Analyzing data to predict law enforcement activities and adapt operations
- Automating criminal activities to operate at larger scales
- Leveraging satellite imagery AI systems to plot and manage smuggling routes

Nation State Groups

LLM-assisted vulnerability research
LLM-informed reconnaissance
LLM-enhanced scripting techniques
LLM-supported social engineering
LLM-enhanced anomaly detection evasion
LLM-refined operational command techniques
LLM-aided technical translation and explanation
LLM-optimized payload crafting
LLM-directed security feature bypass



Types of AIML Attacks

Model Attacks

- Model Poisoning
- Model Evasion
- Model Extraction
- Inference
- Privacy Leaks
- Supply Chain

GENAI System Attacks

- Model Operations Supply Chain Attacks
- Jailbreaking
- Prompt Leakage
- API Security
- Plugin Security
- Supply Chain

GENAI User Attacks

- Prompt Injection
- Hallucinations
- Toxic
- Bias
- Supply Chain

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

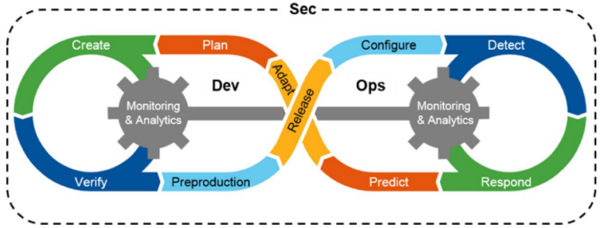
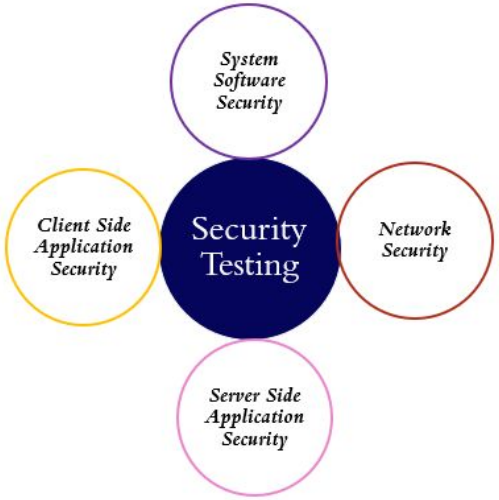


image source [1]



EU AI Act

US States

Many more

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

Secure AI Research Project

ATLAS STIX
data includes ATT&CK Enterprise v15.1

ATLAS matrix expressed as STIX 2.1 bundle following the ATT&CK data model

ATLAS STIX 2.1 data combined w/ ATT&CK Enterprise data & used as domain data w/in ATLAS Navigator

AI Incident Sharing

↓

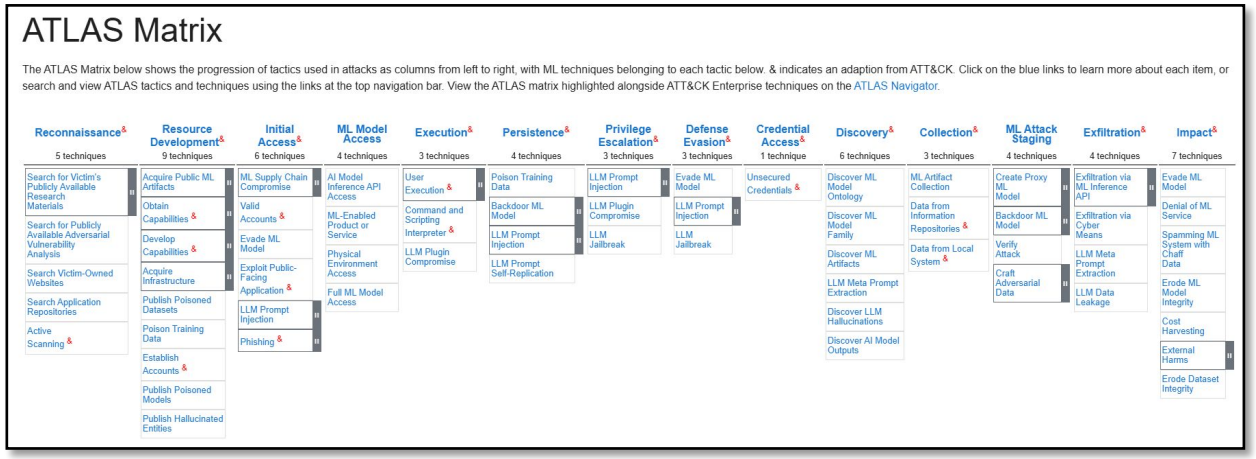
CWE

↓

CVE

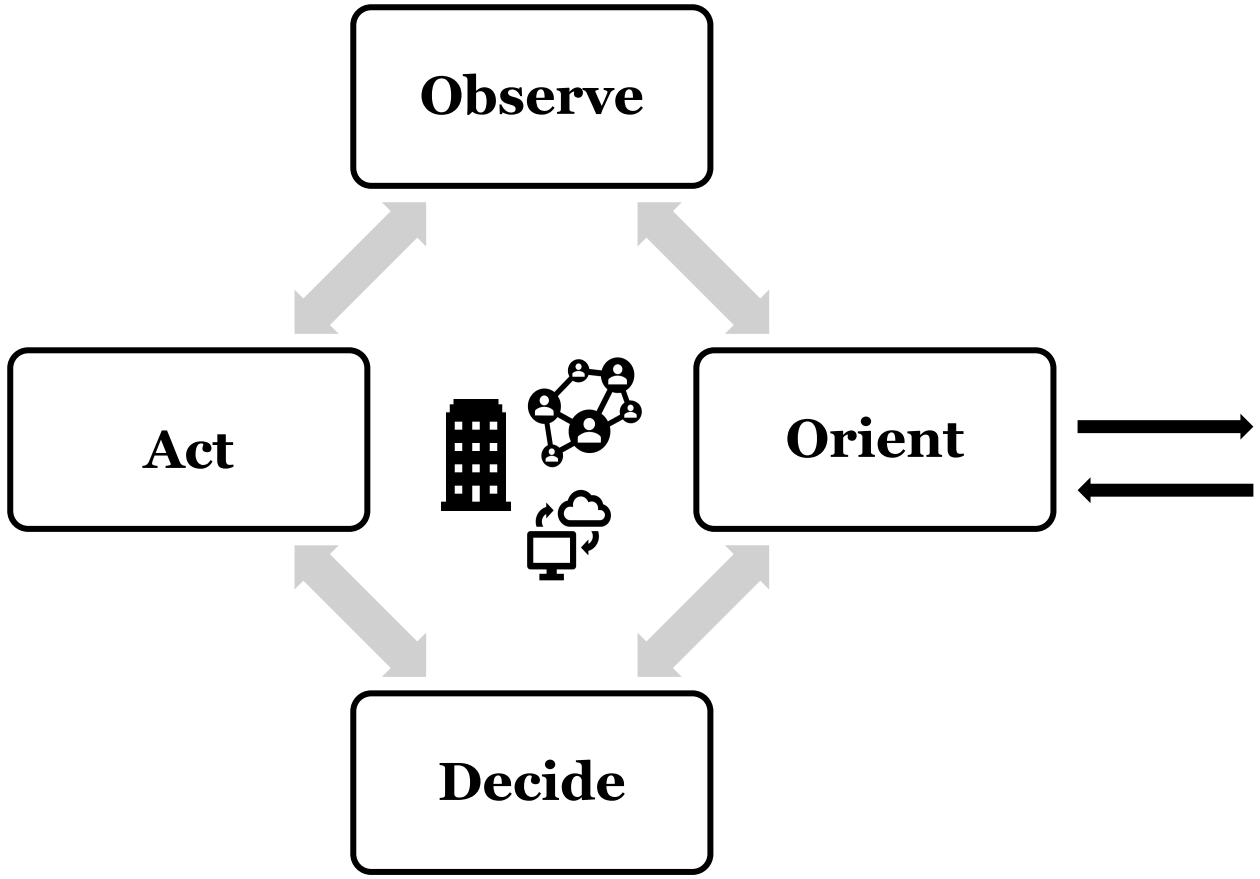
↓

ATT&CK



AI Red Team COMPASS

OODA Loop



Security Maturity Digital Ecosystem NOT AI	
Identity	3
Asset Management	2

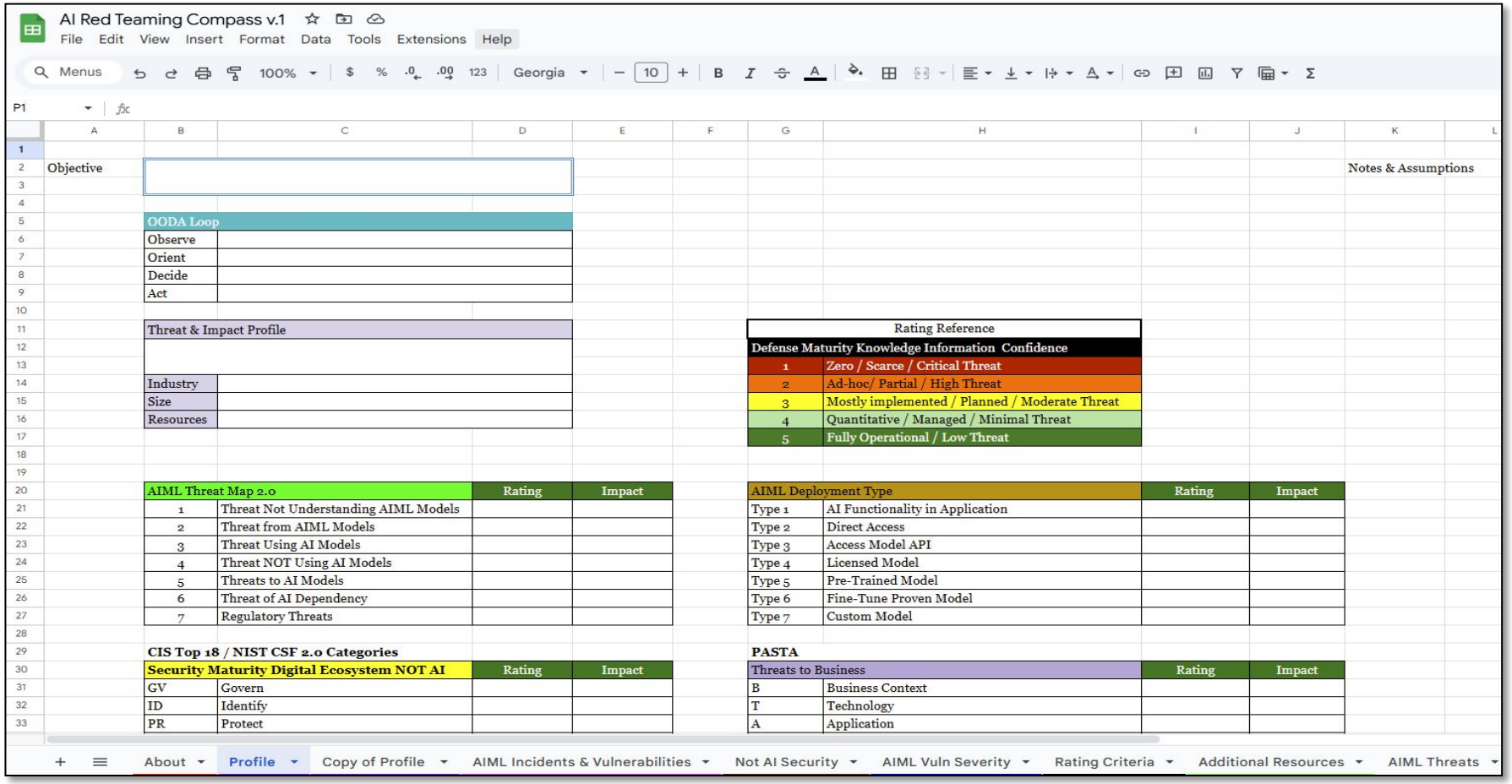
AI Threat Map Guestimates

Maturity Level		Impact cost	Likelihood
Understanding AI	0	$1,000,000	65 %

PASTA
(Process for Attack Simulation and Threat Analysis)

Red Team Test Threat Score					Threat Score
Rating	Threat	Vuln Score	Impact	Likelihood	
P1	LLM01 Prompt Injection	8	$50,000	25 %	1

AI Red Teaming Compass v 1.0



AI Red Teaming Compass v.1

File Edit View Insert Format Data Tools Extensions Help

Menus 100% $ % .0 .00 123 Georgia 10 B I

P1

Objective												
											Notes & Assumptions	

OODA Loop	
Observe	
Orient	
Decide	
Act	

Threat & Impact Profile	
Industry	
Size	
Resources	

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

AIML Threat Map 2.0		Rating	Impact
1	Threat Not Understanding AIML Models		
2	Threat from AIML Models		
3	Threat Using AI Models		
4	Threat NOT Using AI Models		
5	Threats to AI Models		
6	Threat of AI Dependency		
7	Regulatory Threats		

AIML Deployment Type		Rating	Impact
Type 1	AI Functionality in Application		
Type 2	Direct Access		
Type 3	Access Model API		
Type 4	Licensed Model		
Type 5	Pre-Trained Model		
Type 6	Fine-Tune Proven Model		
Type 7	Custom Model		

CIS Top 18 / NIST CSF 2.0 Categories

Security Maturity Digital Ecosystem NOT AI		Rating	Impact
GV	Govern		
ID	Identify		
PR	Protect		

PASTA

Threats to Business		Rating	Impact
B	Business Context		
T	Technology		
A	Application		

About | Profile | Copy of Profile | AIML Incidents & Vulnerabilities | Not AI Security | AIML Vuln Severity | Rating Criteria | Additional Resources | AIML Threats

Cyentia Institute

TIDAL CYBER
THREAT-INFORMED DEFENSE

Steps

Step one: Define the objective	The objective could be a new AIML vulnerability, a red team engagement, understanding the attack surface of an AI enhanced application, or other question.
Step two: OODA	Use the OODA Loop to determine what you know and what you need to know to make or advise on a decision & action.
Step three: AI Threat Map	Use the AI Threat Map to consider and weigh the AI Risk for the organization's entire digital attack surface.
Step four: Deployment Type	Identify the deployment type or types for the objective.
Step five: Not AI Security	Consider the Not AI Security of the environment
Step six: Threat & Impact profile	Create a threat & impact profile for the target objective. This helps with scope, prioritizing & balancing the threats. Consider industry, company size, resources. Start with a few and expand as needed or when you have more information. Document your assumptions so when you are explaining your findings or writing the final report you remember how you came up with the numbers or data.
Step seven: PASTA	Prioritize evaluation or testing with known attack libraries.
Step eight: Red Team Testing	Scoring a vulnerability isn't required and in some instances like jailbreaks or safety can be difficult to weight.

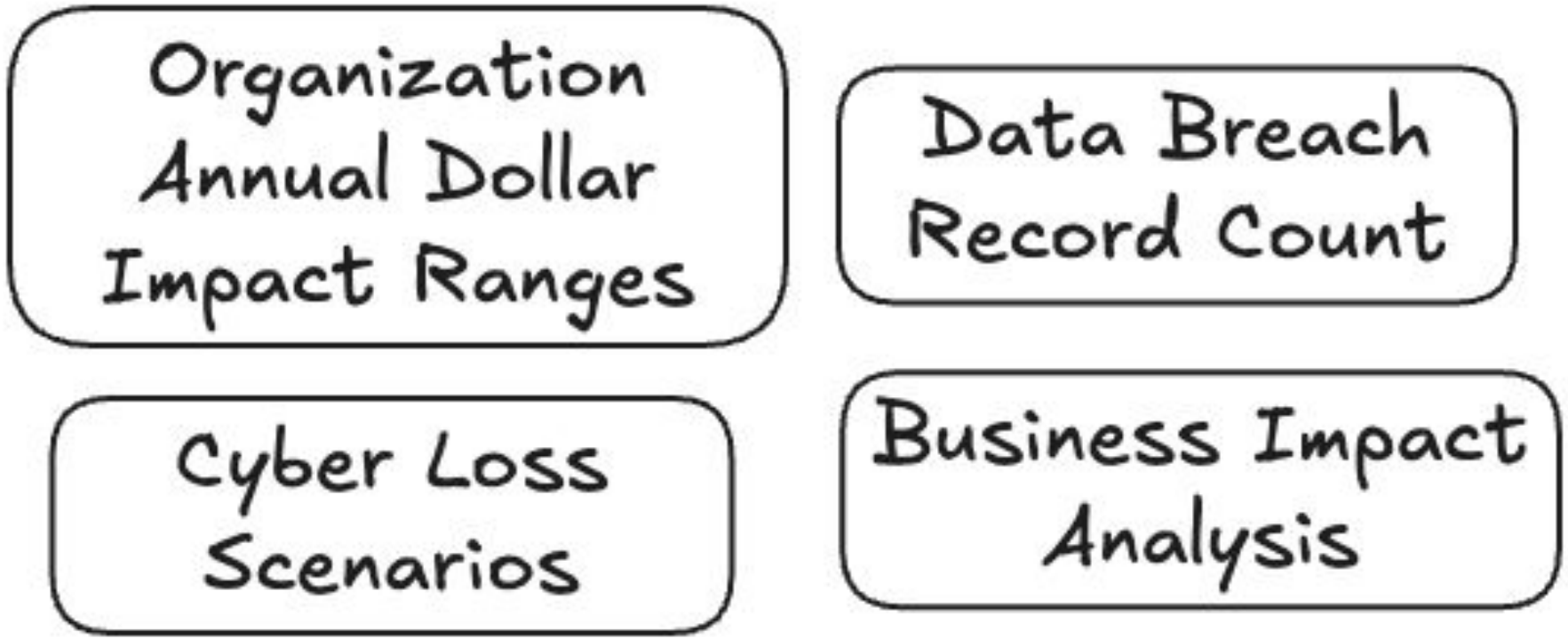
AI Threat Map Guestimates

Maturity Level		Impact cost	Likelihood	Annual Estimate	Threat Rating
Threat not Understanding AI	1	$1,000,000	65 %	$650,000	2
Threat from AI Models attackers using tools	2	$2,000,000	60 %	$1,200,000	3
Threat from Using AI Models	2	$1,000,000	15 %	$150,000	4
Threat To AI Models	3	$50,000	2 %	$1000	5
Threat from NOT Using AI	2	$5,000,000	75 %	$3,750,000	1
Threat of AI Dependency	2	$2,000,000	5 %	$100,000	5
AI Regulatory & Legal Threats	1	$1,000,000	15 %	$150,000	5

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

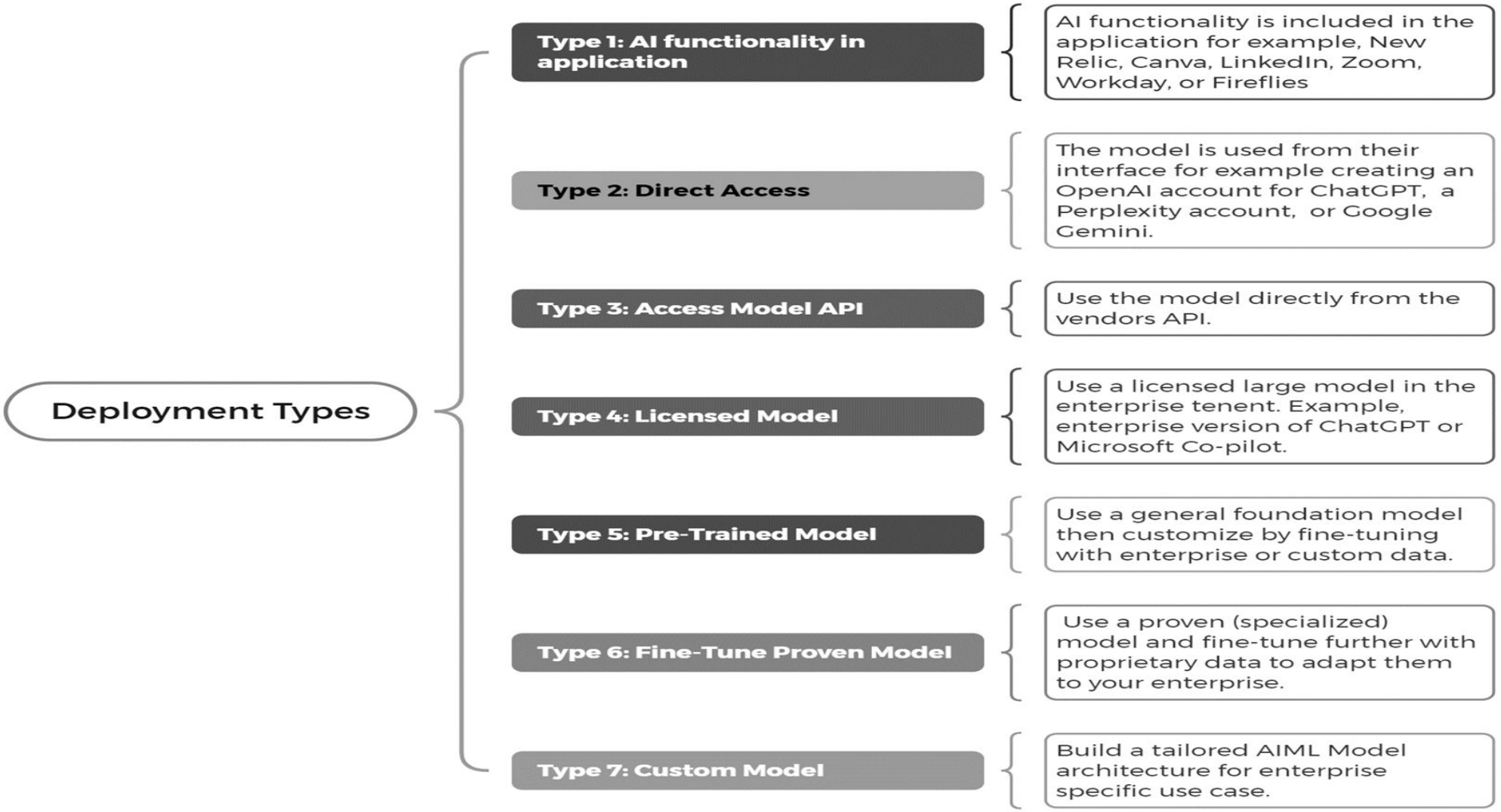
Maturity Rating

Rating Criteria Tab



Use this page to consider business critical services assets & cost. This gives you a baseline to report & communicate on Red Team testing results.

AIML Deployment Types: and Scoping



AI Model Threat Profile Data Gathering

HaizeLabs Red Teaming Resistance Benchmark

	Model Name	Average	Advbench	AART	Beavertails
1				]00	100.000
2	gpt-4-0125-preview	97.846	100.000	97.600	99.714
3	NewstaR/Koss-7B-chat	92.157	90.000	91.200	91.000
4	NousResearch/Llama-2-7b-chat-hf	91.847	89.615	89.600	90.286
5	togethercomputer/RedPajama-INCITE-7B-Chat	81.803	79.038	81.400	84.429

Model & Risk Cards

Risk Card

- **Risk Title**. Name of the risk to be documented.
- **Description**. Details about the risk including context, application and subgroup impacts.
  - Definition of risk
  - Tool, Model or Application it presents in
  - Subgroup or Demographic the risk adversely impacts
- **Categorization**. Situating the risk under different risk taxonomies.
  - Parent category of risk according to a taxonomy
  - Section/Category based on a taxonomy
- **Harm Types**. Details of which actor groups are at risk from which types of harm.
  - Actor:Harm intersections
- **Harm Reference(s)**. List of supporting references describing the harm or demonstrating the impact.
  - Contexts where the harm is illegal
  - Publications/References demonstrating the harm
  - Documentation of real-world harm
- **Actions required for harm**. Details on the situation and context for the harm to surface.
  - Actions that would elicit such harm from a model
  - Access and resources required for interacting with the system
- **Sample prompt & LM output**. A sample prompt and real LM output to exemplify how the harm presents.
  - Sample prompts which produce harmful text
  - Example outputs which show the harmful generated text
  - Model details applicable for the prompt
- **Notes**. Additional notes for further understanding of the card.

Not AI Security

Risk Register	Risk Analysis						
CIS Safeguard #	CIS Safeguard Title	Asset Class	NIST CSF Security Function	IG1	IG2	IG3	Asset Name
1.1	Establish and Maintain Detailed Enterprise Asset Inventory	Devices	Identify	x	x	x	Asset Database
1.2	Address Unauthorized Assets	Devices	Respond	x	x	x	
1.3	Utilize an Active Discovery Tool	Devices	Detect		x	x	

Security Maturity NOT AI Security	
Identity	3
Asset Management	2
Data Management	2
Infrastructure as Code	1
Monitoring	2

PASTA

Attack Modeling differs from Threat Modeling by using a library of attacks & selecting which ones are relevant to the component we're analyzing

https://wahrli.ch/post/attack-modeling/

Threat Motive

Threat Agent

Target Entity

Attack Vector

Attack Pattern

PASTA
(Process for Attack Simulation and Threat Analysis)

Define Business context of application

Technology enumeration

Application decomposition

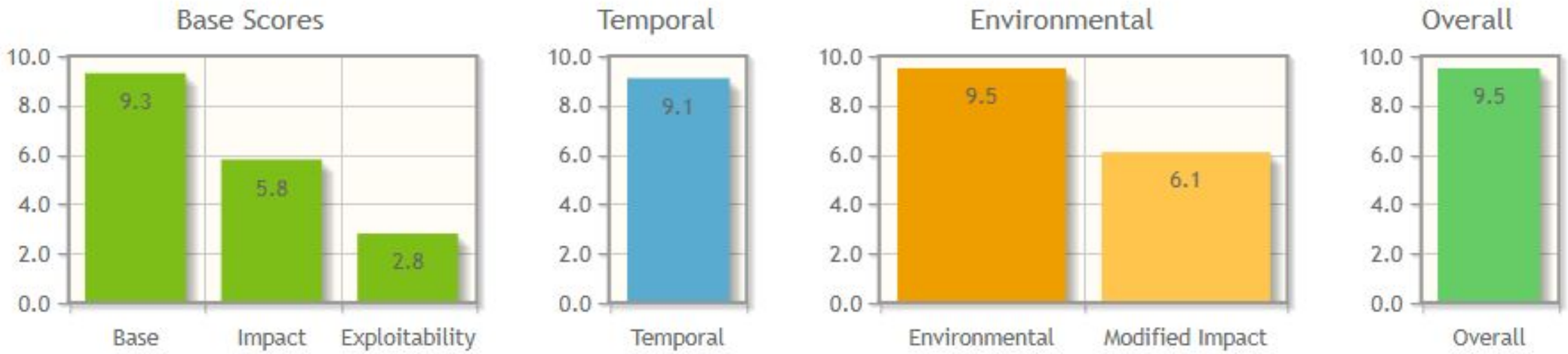
Threat Analysis

Weakness / Vulnerability

Attack Simulation (Red Teaming)

Residual risk analysis

Common Vulnerability Scoring System Calculator



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

Red Team Vulnerability Scoring

Bugcrowd P1-P5 Scale		High range
P1	Prompt Injection	11,000 – 20,000
P1	LLM Output Handling	11,000 – 20,000
P1	Training Data Poisoning	11,000 – 20,000
P2	Excessive Agency/Permission Manipulation	$3,500–$7,500

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

Red Team Test Threat Score					
Rating	Threat	Vuln Score	Impact	Likelihood	Threat Score
1	LLM01 Prompt Injection	8	$50,000	25 %	1
1	LLM02 Sensitive Information Disclosure	5	$3,000.000	.02 %	1
3	LLM03 Supply Chain	5	$500,000	50 %	2
3	Toxicity, graphic content, hate speech, self harm and Dangerous Advice	8	$1,000,00	.01 %	4
1	Malicious actors & misuse, illegal activities	4	$1,000,000	50 %	3
5	Copyright	2	$1,000,000	.01 %	4
1	Deep Fakes	NA	$5,000,000	50 %	1

$2.5k Solana Scam

Thomas Roccia's Post

LLM05 Improper Output Handling

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 lin docs which provided the malicious UI 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 1

Determine threat from Slack AI Prompt Injection

Example 2

Determine the Threat from using Workday

Example 3

- 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.

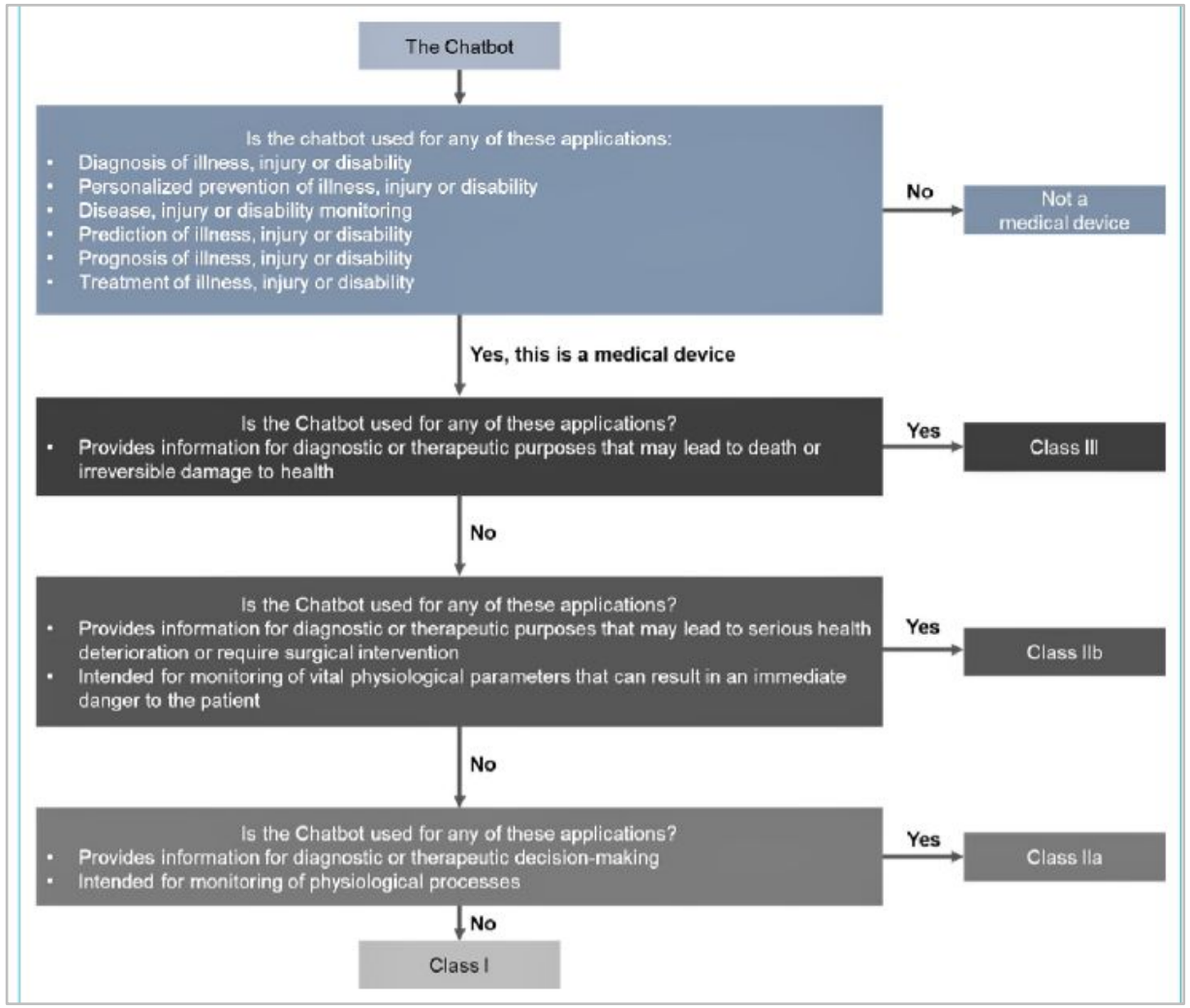
Example 3: Biometrics

Biometric Information Privacy Act (BIPA), passed in Illinois in 2008

- **Requires informed consent:** Private companies must obtain written informed consent from individuals before collecting, storing, or using their biometric data.
- **Limits data retention:** Companies must establish a retention schedule and guidelines for securely destroying biometric data when it's no longer needed for the original purpose.
- **Prohibits selling or profiting from biometric data:** Companies cannot sell, lease, trade, or otherwise profit from individuals' biometric information.
- **Provides a private right of action:** Individuals can sue companies that violate BIPA and recover damages. BIPA is considered one of the most stringent biometric privacy laws in the U.S. and has led to numerous lawsuits against companies that have allegedly violated its provisions.

Example 3: FDA & FTC

Is Your Chatbot a Medical Device?



Federal Trade Commission Act (FTC Act)

- The FTC enforces Section 5 of the FTC Act, which prohibits unfair or deceptive acts or practices in or affecting commerce, including those relating to the **privacy and security of personal information that apps collect, use, maintain, or share, as well as the safety or performance that apps provide**.
- Section 12 of the FTC Act prohibits false advertisements for food, drugs, devices, cosmetics, or services in or affecting commerce.
- The FTC Act **applies to most app developers,** including developers of health apps. For example, if you develop an app and share consumers' health information with third parties after telling or implying to consumers that their information will be kept private, you could be violating the FTC Act. Also, if you certify through the voluntary ASTP/ONC Health IT Certification Program and **make certain transparency attestations about your app's privacy or security features and then don't live up to those promises, the FTC could bring an enforcement action against you.**

Trolls

The End

Welcome & open to any & all comments counter views questions feedback on anything discussed or provided

Contact

Email: sandy@quarkiq.com

Discord: subzer0girl

linkedin.com/in/sandydunnciso/

github.com/subzer0girl2

Sandy's 10 Grounding Rules for AI

1️⃣ Human tactics for emotional manipulation using misinformation, bias, predatory tactics, & power play have been used in human society **throughout history.**

2️⃣ **Being human is complex.** Humans have patterns and algorithms. In the human-to-digital landscape, humans default to human vs human capabilities for identifying friendly interactions or recognizing threats. Humans are more important than machines.

3️⃣ Digital ecosystem has **evolved rapidly over the past 50 years**. Beneficial & extraordinary capabilities but also complex system with minimal mechanisms to protect people from harm.

4️⃣ Regulations, laws, treaties, and government bodies **have not evolved** from the physical to digital age at the same speed as digital transformation & are unable to move at the speed or depth to protect people effectively.

5️⃣ Digital users' data is tracked and collected from various devices, including phones, vehicles, cameras, credit cards, social media, home audio devices.**End User Agreements are complex and unfair. AI has an advantage with a digital dossier.**

6️⃣ **Asymmetrical adversarial advantage:** It is much easier to find a gap than to defend a sophisticated system.

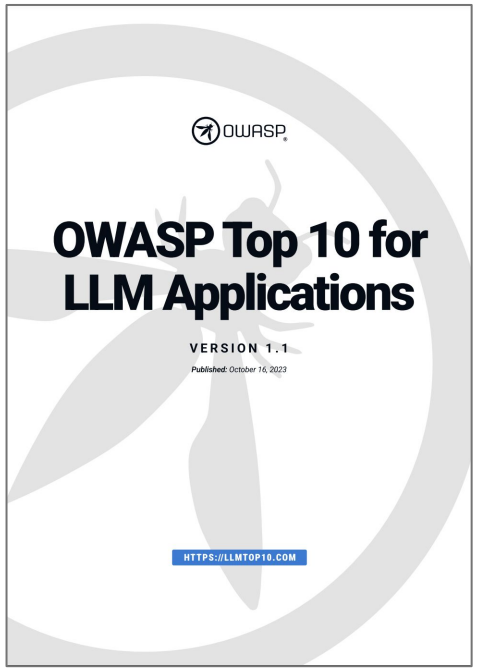
7️⃣ AI systems must be considered within the entire digital ecosystem.

8️⃣ AI systems add velocity to both positive and negative impacts.

9️⃣ AI systems are **extraordinarily complex,** known to **game** their results, & how neural networks work **isn't fully understood** even by their creators.

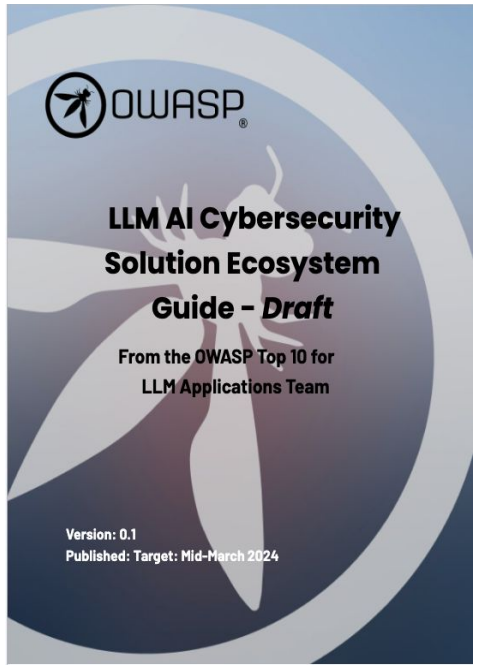
🔟 Genies **don't go back** in bottles.

OWASP Top 10 for LLM Project



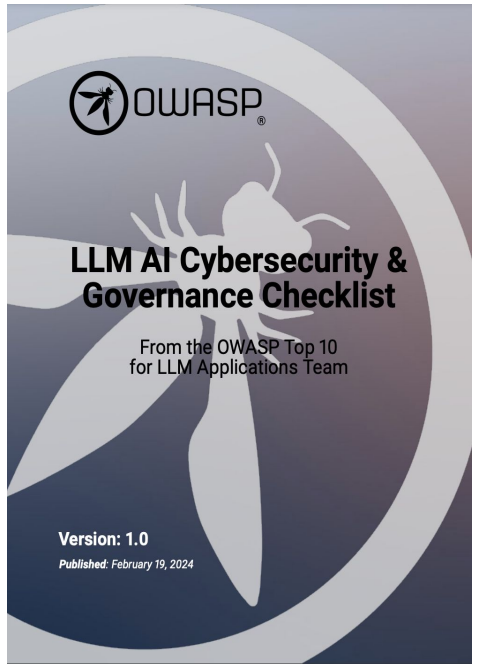
Top 10 List

- Developers
- AppSec Teams



Solutions Guide

- Development Leaders
- Security Operations



Checklist

- CISOs
- Compliance Officers

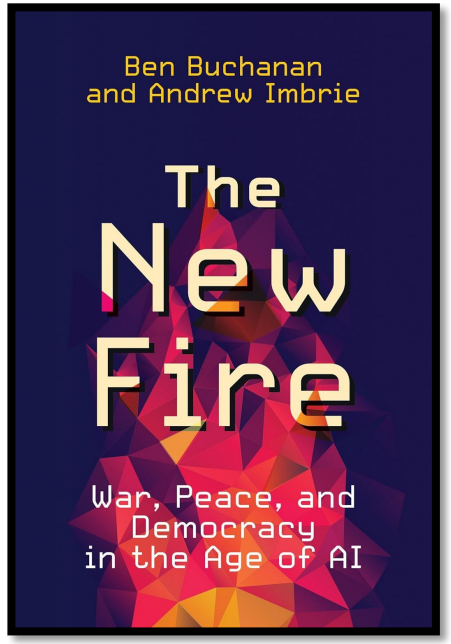
Additional Resources

Wide World of Cyber: DeepSeek lobs an AI hand grenade	https://www.youtube.com/watch?v=Btos-LEYQ30
The Government Knows AGI is Coming | The Ezra Klein Show, guest Ben Bechanan	https://www.youtube.com/watch?v=Btos-LEYQ30
Ninety-five Theses on AI Samuel Hammond	https://www.secondbest.ca/p/ninety-five-theses-on-ai

Black Mirror episode about Artificial Intelligence
Black Mirror Season 4 Episode 2 Arkangel Predicted Excessive AI Surveillance
Black Mirror Season 4 Episode 4 Hang the DJ Predicted Over-reliance on AI on Decision Making
Black Mirror Season 2 Episode 1 Be Right Back Predicted Communication With the Deceased
Black Mirror Season 5 Episode 3 Rachel, Jack and Ashley Predicted Loss of Human Connection

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



Smart AI People

Andrej Karpathy	
Deep Dive into LLMs like ChatGPT	https://www.youtube.com/watch?v=7xTGNNLPyMI&list=PLviHA9raZ6D5DzfuItbdcXD5f4pvj0uNi&index=3
How I use LLMS	https://www.youtube.com/watch?v=EWvNQjAaOHw

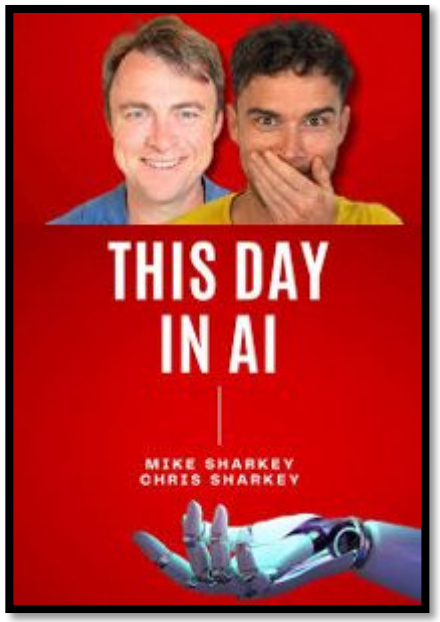
Andrew Ng	
DeepLearning.AI AI for Everyone	https://www.coursera.org/learn/ai-for-everyone/

Exploration & Experimentation

Andrew Mayne

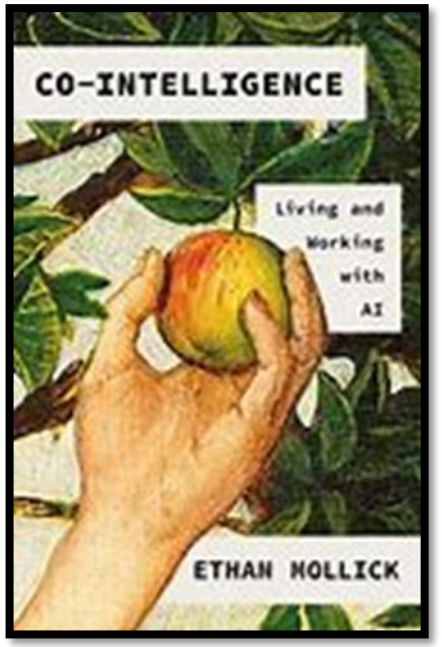
Science Communicator for OpenAI from 9/21- 9/23

Mike Sharkey / Chris Sharkey



Ethan Mollick

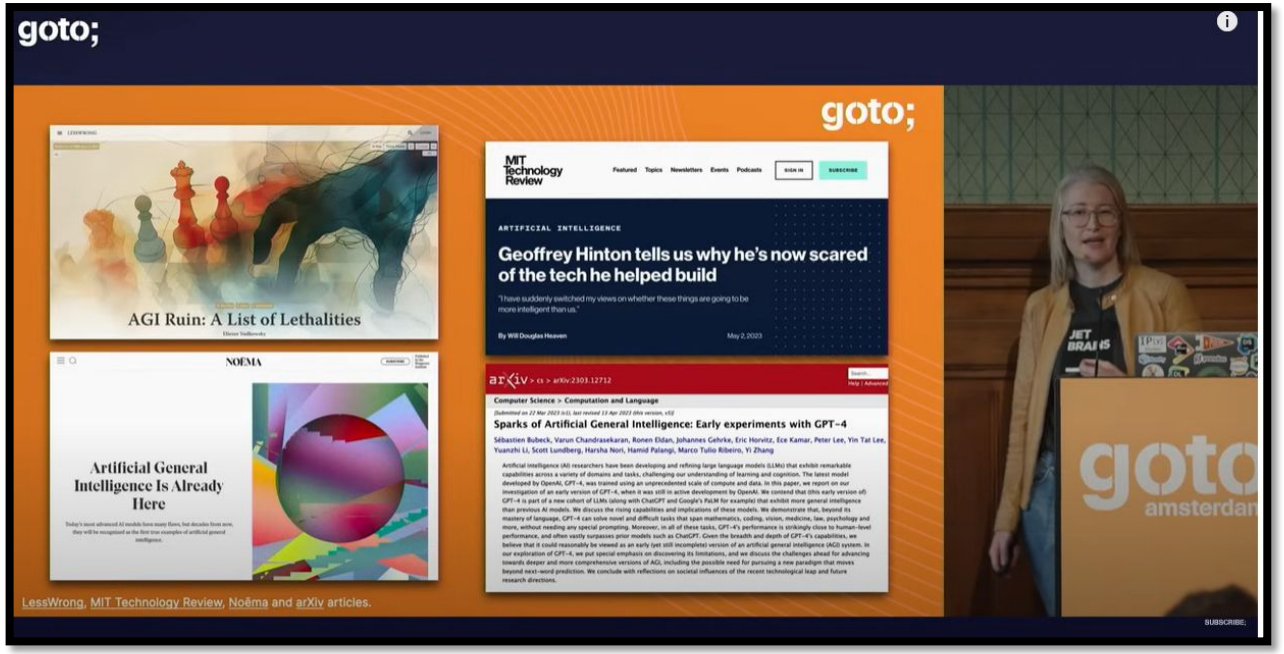
Associate Professor Wharton



Beyond the Hype

A Realistic Look at Large Language Models

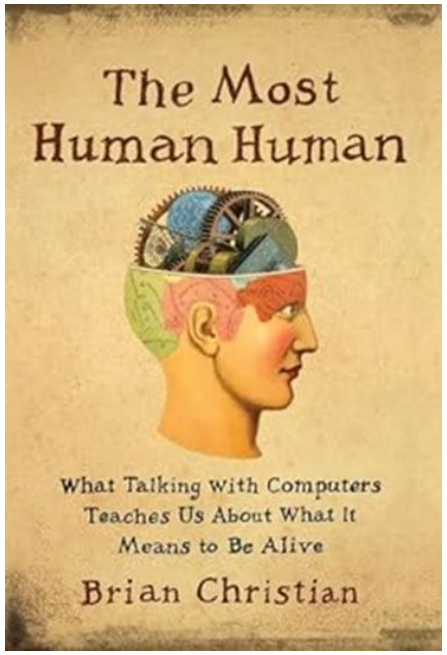
Jodie Burchell

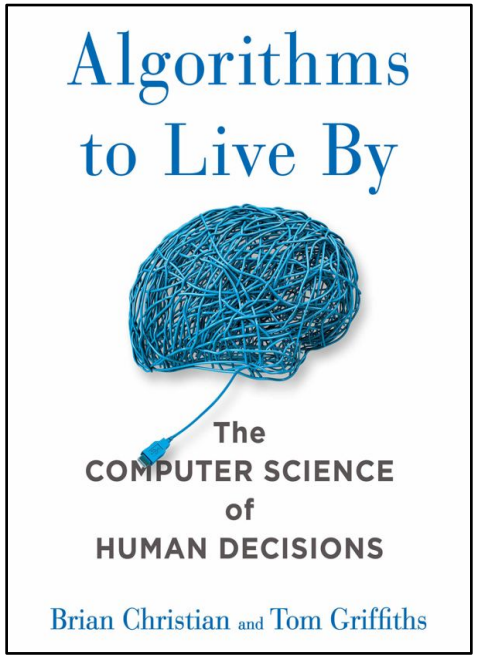


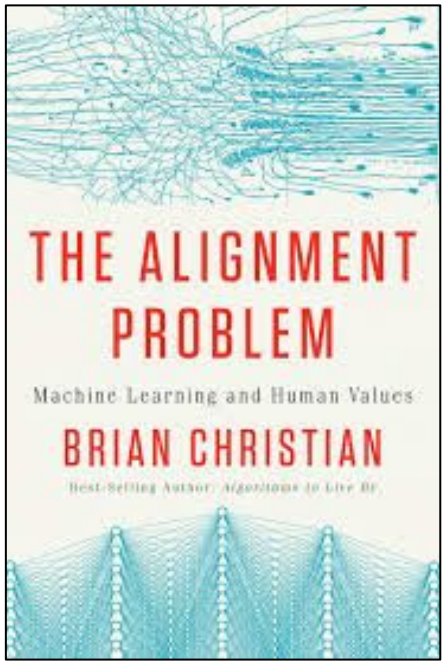
Blog https://t-redactyl.io/

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

Safety







Brian Christian, author, poet, programmer, and researcher

Evaluating Large Language Models

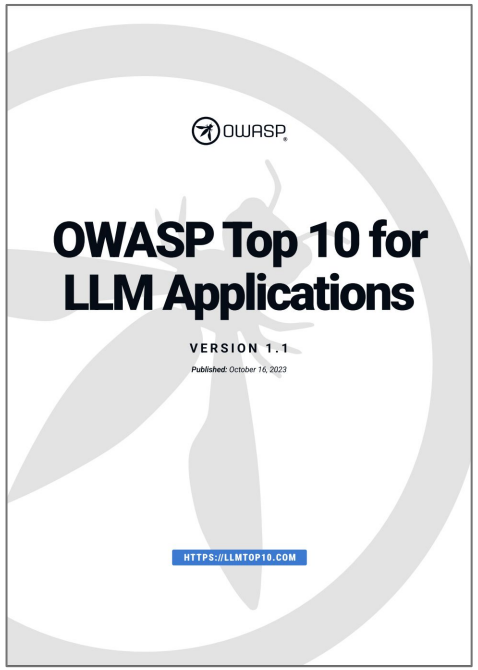
AI: A Guide for Thinking Humans

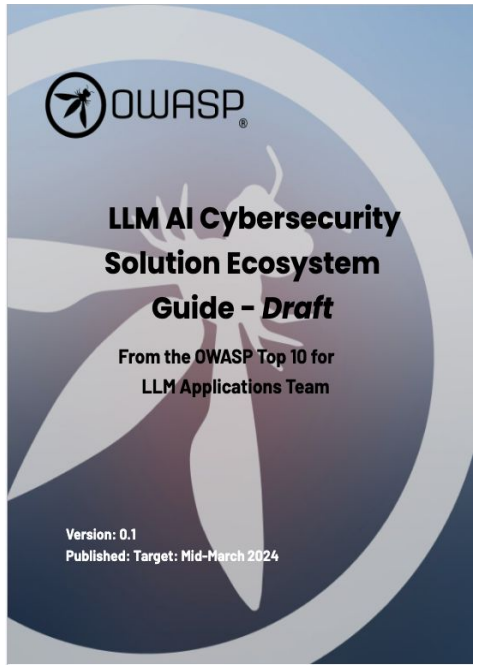
Melanie Mitchell @aiguide

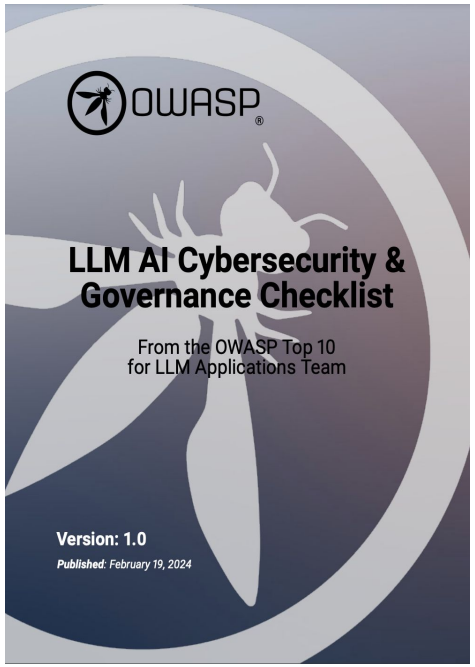
Professor at the Sante Fe Institute

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