Unmasking AI Threats

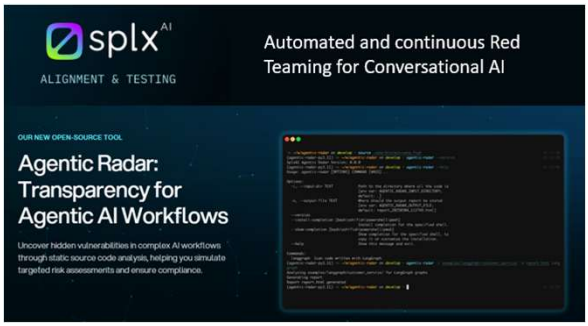
Sandy Dunn, CISO SPLX.AI

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
- Adjunct Prof at BSU
- Board member BSU Institute for Pervasive Cybersecurity & Cyber CORe program

SPLX.AI



At The End of Our Discussion

- ✔ Cyber security issues around the use of AI tools
- ✔ How AI can be effectively used to improve cyber security

Key Points	Agenda
⚙ Release of ChatGPT ignited an **Artificial Intelligence Tipping Point** ⚙ **Reciprocal Relationship** between understanding AI & understanding humans. The future might be amazing! (or we might become slaves to digital overlords) ⚙ The Digital Era **has happened & is happening rapidly**. What a healthy digital relationship is, still evolving. 📺 ⚙ **Critical Thinkers & Knowledge Seekers (us)** have endless possibilities & opportunities, anything is possible!	⚙ Sandy's 10 Grounding Rules for AI ⚙ How We Got Here, Before November 30, 2022 ⚙ After November 20, 2022 ⚙ The Challenge of Being Human ⚙ The Problem of Privacy ⚙ The 7 types of AI Threats ⚙ AIML Security ⚙ Resources

Sandy's 10 Grounding Rules for AI

1️⃣ Human tactics for emotional manipulation using misinformation, bias, predatory tactics, and 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 The digital ecosystem has evolved rapidly over the past 50 years, bringing amazing capabilities that have transformed our lives and benefited people. It has resulted in a 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 and 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 (Alexa, Google Assistant, Siri), fitness trackers, and medical devices (heart monitors, glucose monitors, COPD monitors). 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, and how neural networks work isn't fully understood even by their creators.

🔟 Genies don't go back in bottles.

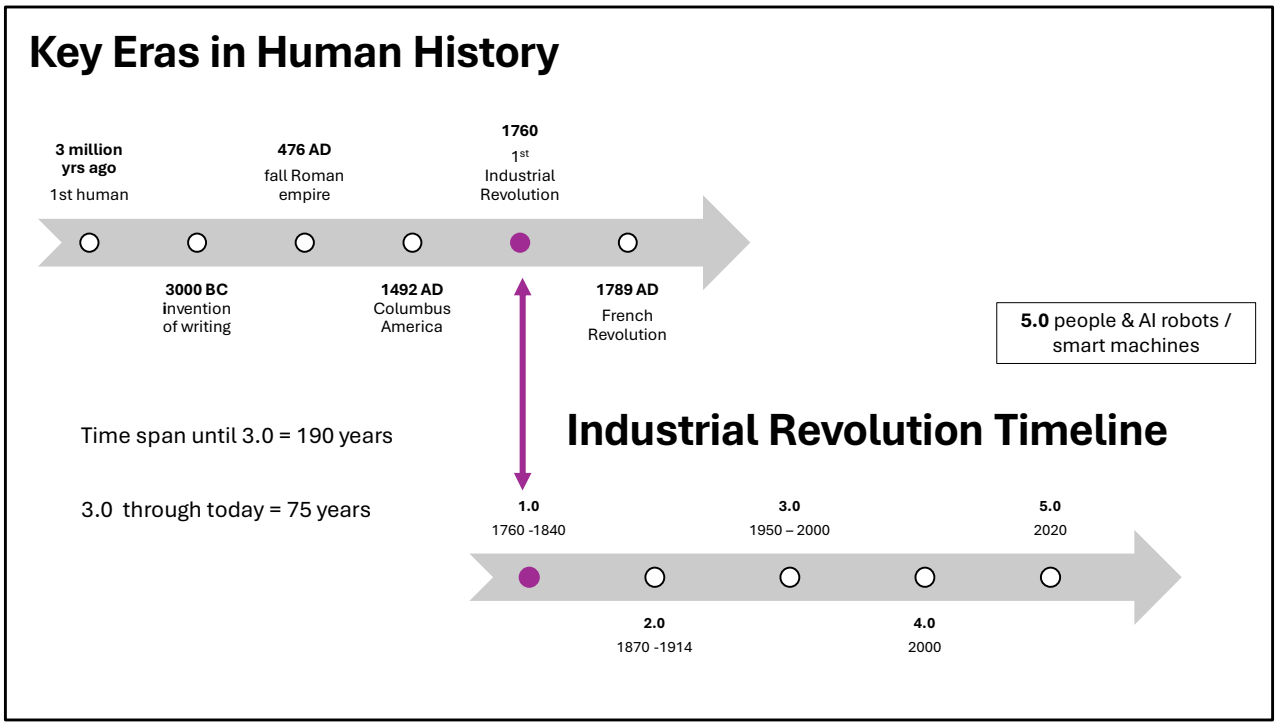
1. Royal leaders were "Gods" or were authorized by "God"
2. Thinking, Fast and Slow is a 2011 Daniel Kahneman / Amos Tversky "System 1" is fast, instinctive and emotional; "System 2" is slower, more deliberative, and more logical. 35,000 decisions a day
3. Maps, texts, YouTube
4. Think about how our police departments operate, from a local presence to a much smaller group of people at a National Level protecting us at a national level
5. Thousands of bits of data are collected about us each day
6. Anyone trying to keep kids or puppies corralled are familiar with this – it's easier to find a gap then keep things protected
7. AI has been used in systems extensively for 10 + years.
8. Chainsaw vs handsaw
9. Andrej Karpathy has an excellent series on LLMS and their sharp edges. Facinating to me about Reinforcement Learning Training and the difference in tuning for "verified results" vs "unverified results"
10. Ezra Klein Show w/ Ben Buchanan The Government Knows AGI is Coming

How We Got Here

The short amount of time from industrial revolution 1.0 until 5.0

3 things in 3 timelines

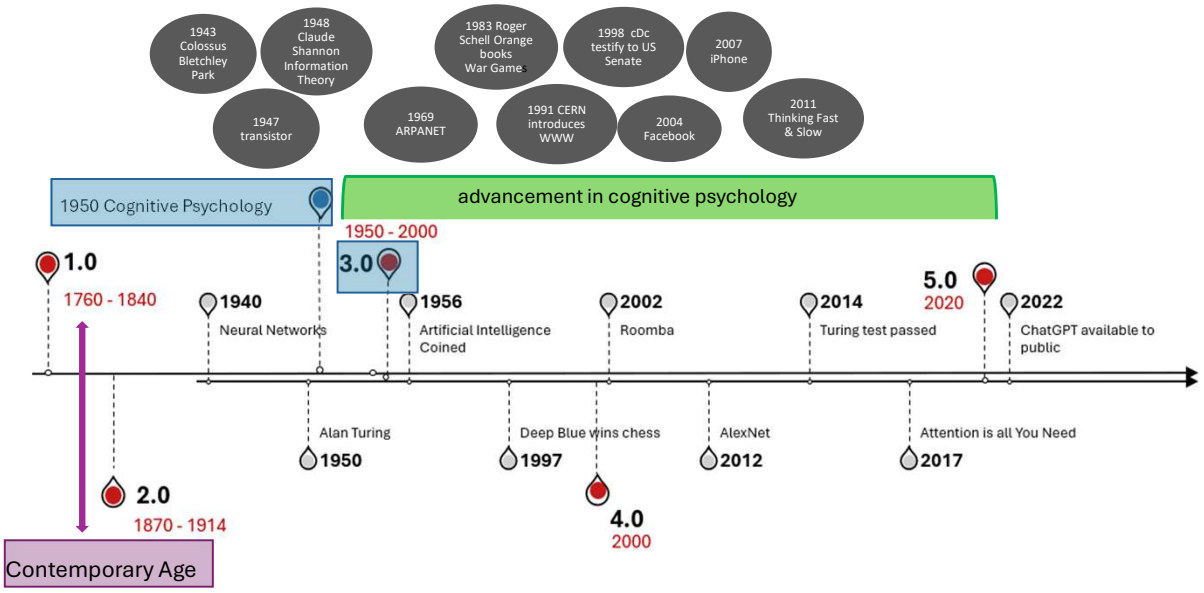
- Human history, industrial revolutions, & advancement in cognitive timelines
- Less than 300 years since 1st industrial revolution
- Rapid changes last 40 years
- 3 things in parallel
  - Thinking about a computer system acting like a human system
  - Development of IT systems
  - Study & advancements in cognitive psychology



https://journalofcloudcomputing.springeropen.com/articles/10.1186/s13677-022-00314-5#Sec19

1.0 Steam Engine & Mass production
2.0 Assembly lines
3.0 Automation / IT Systems
4.0 Digitization
5.0 People & AI robots / smart machines

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

A History of Reciprocal Growth

"How do people think?"

"How do people learn?"

"How do humans communicate?"

- Language
- Emotions
- Facial expressions
- Tone

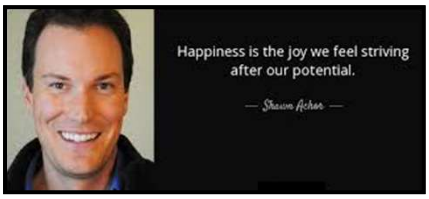
"How do they behave?"

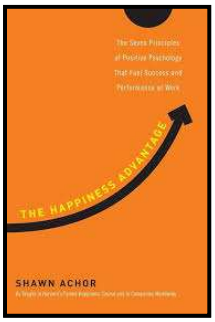
Cognitive Revolution

Beginning in the 1950s

The rise of technology, especially with the development of the **Turing Test** demonstrated **machines could mimic intelligent behavior.**

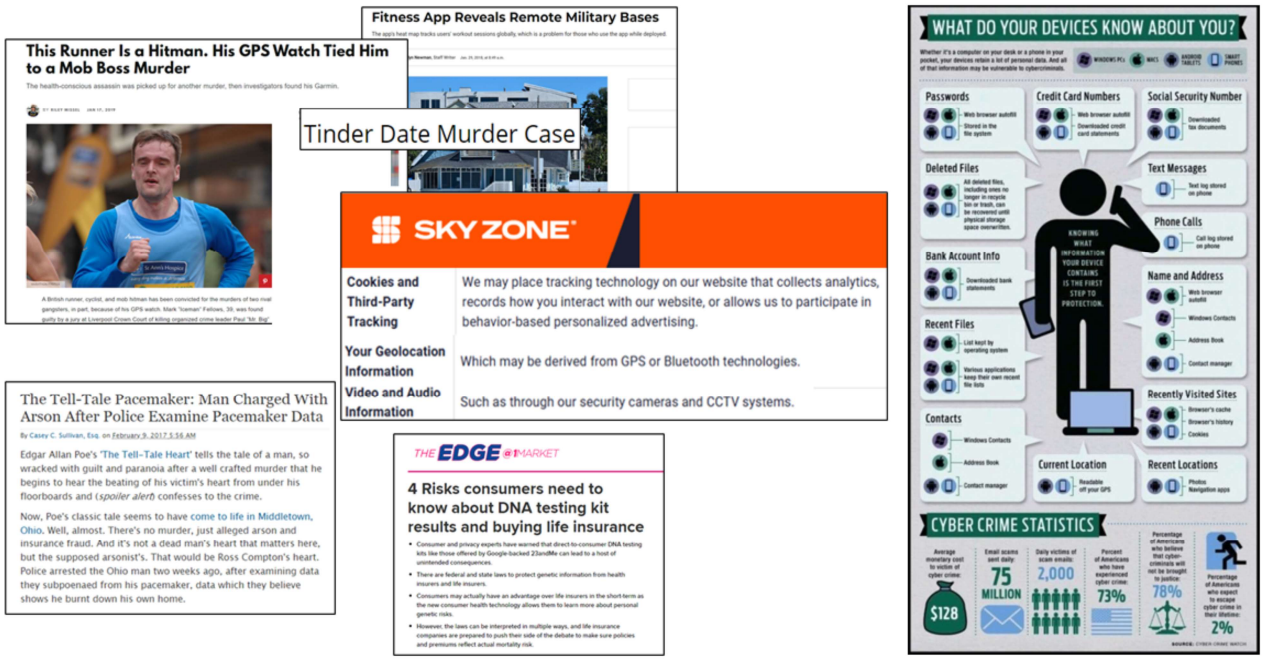
This breakthrough allowed researchers to use machines as models for studying various cognitive functions, opening new avenues of inquiry.





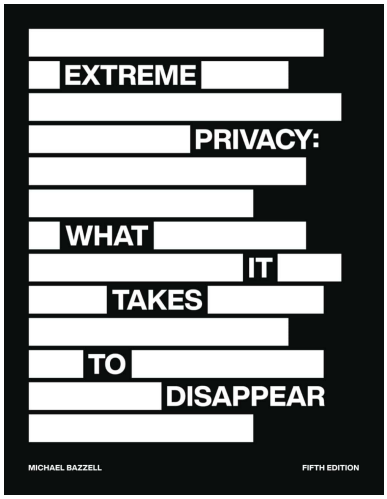
https://www.careershodh.com/the-emergence-of-cognitive-psychology/
https://achology.com/psychology/twenty-pivotal-moments-in-psychologys-history/
Shawn Achor Ted Talk on the Happiness Advantage
https://www.youtube.com/watch?v=GXy__kBVq1M

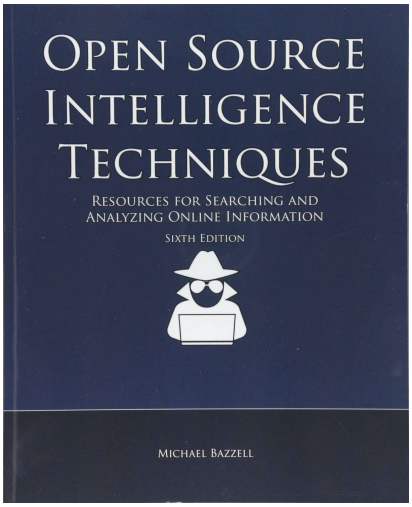
The Problem of Privacy Digital Tracking



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/

Michael Bazzell

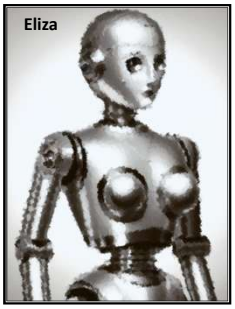




- Lock Credit
- Data removed from data brokerage sites

Challenge of Being Human

Anthropomorphism

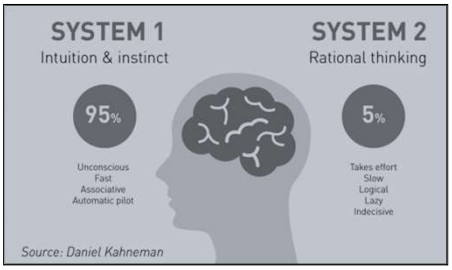


Joseph Weizenbaum

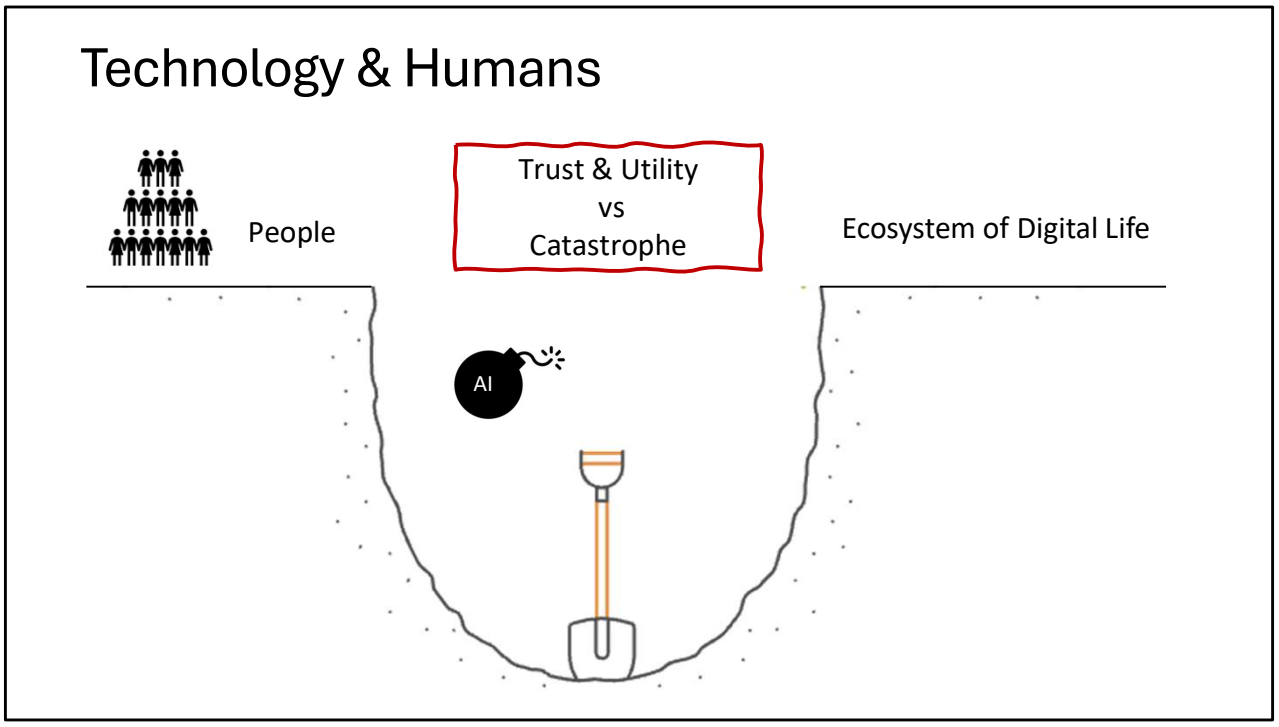
*When robots make **eye contact recognize faces mirror human gestures** they push our Darwinian buttons exhibiting the kind of behavior people associate with **sentience intentions & emotions***

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.

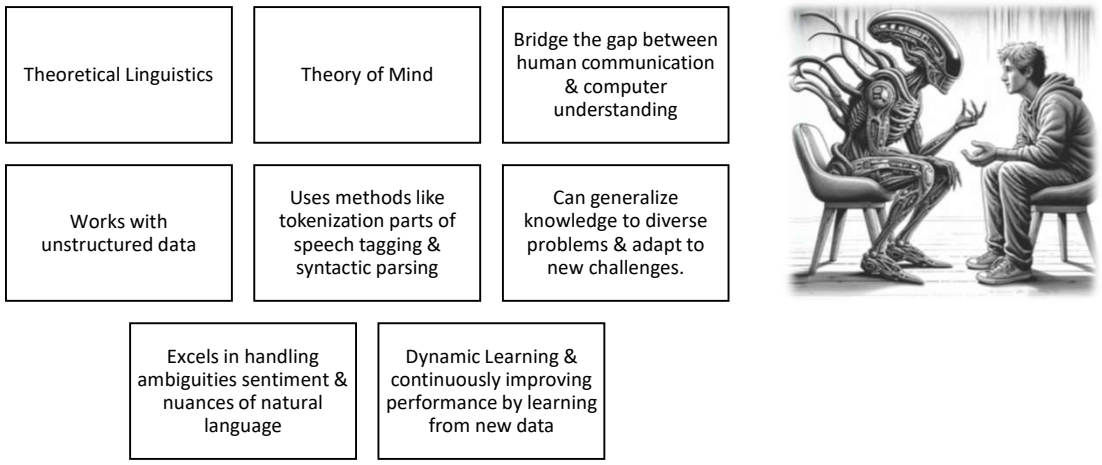


- The challenge is people & business are almost required to be digitally connected today.
- The internet wasn't secure, private, or safe before AI .
- Attacks are sophisticated, and many people and business are part of a digital risk lottery where their number just hasn't come up yet.

What Happened: November 30, 2022

Andrew Mayne's says even OpenAI was surprised

Why Natural Language Processing (NLP) is Special

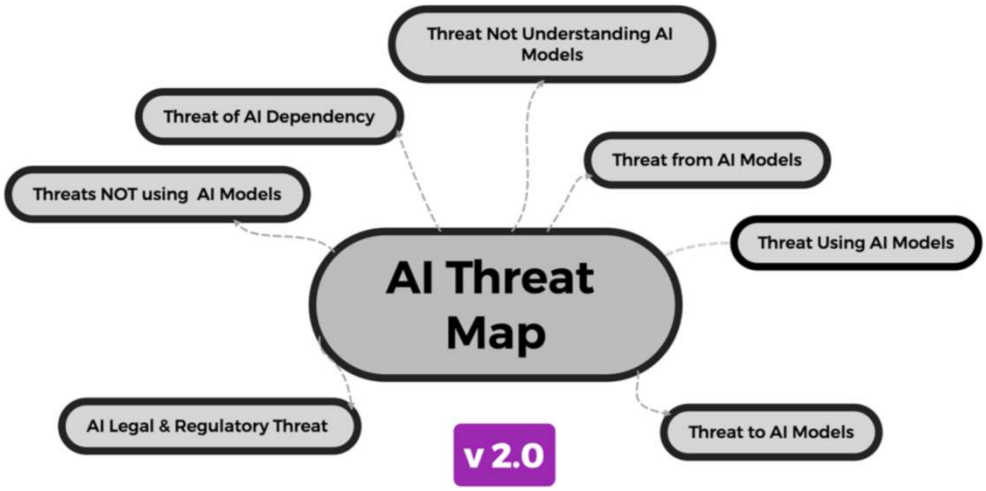


1. Natural Language Processing is software I can work with in my language not its
2. It's very similar to working with another person
3. It is a different experience to make a request to technology, receive a response, "I can't do that", and my reply is "Yes you can try harder" and it works.
4. **Theoretical 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
5. **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
6. Basic arithmetic functions in programming languages are billions of times faster than using an LLM and don't produce hallucinated results

Purpose: Bridge the gap between human communication & computer understanding w/ machines designed to understand & respond to language mimicking the natural processes of human communication.
Data Handling: Works with unstructured data (text speech) unlike traditional systems reliant on structured data and predefined rules.

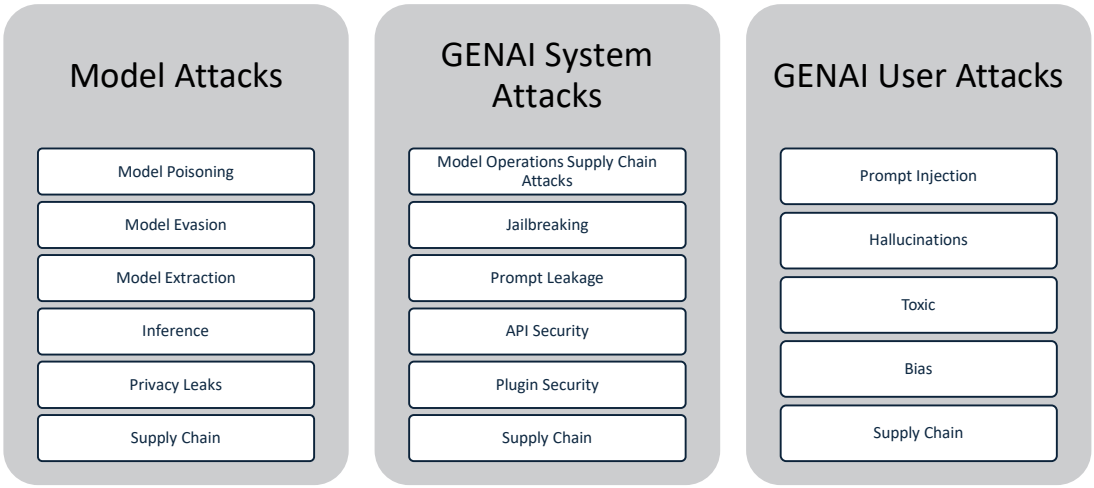
Techniques: Uses methods like tokenization, part-of-speech tagging, and syntactic parsing to process language patterns and meaning.
Dynamic Learning: Continuously improves performance by learning from new data, unlike traditional models requiring full retraining.
https://www.languageeducatorsassemble.com/intro-to-theoretical-linguistics/

7 Categories of AI Threats



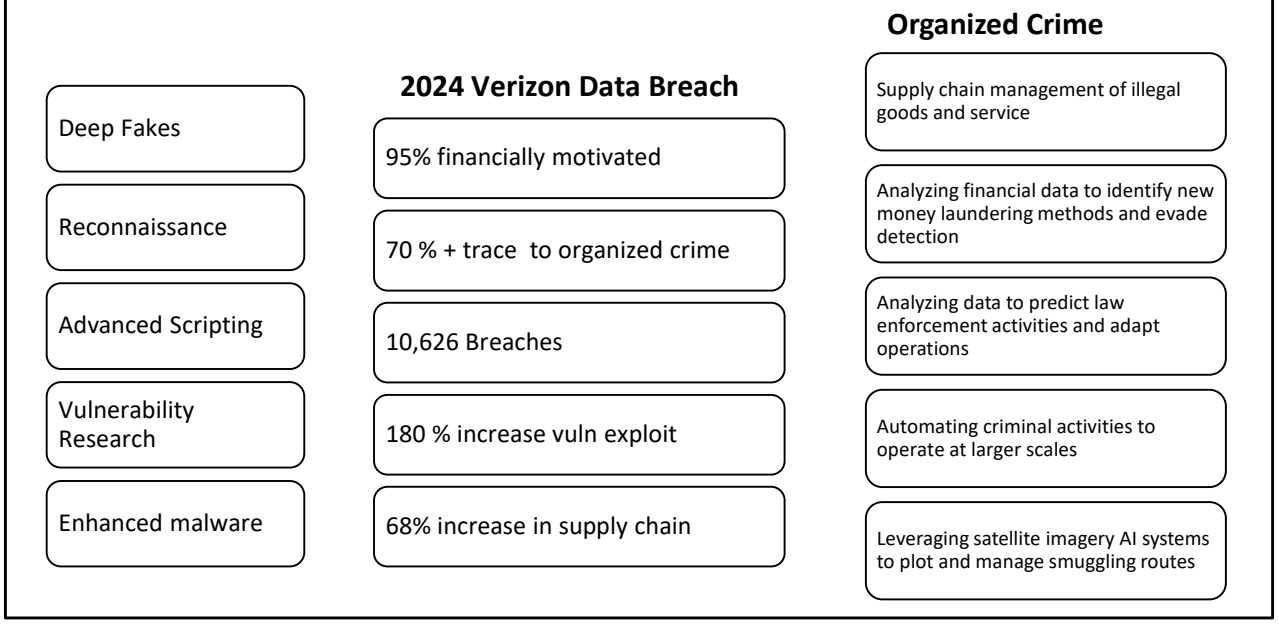
I break out AI Threats into these 7 categories

Types of AIML Attacks



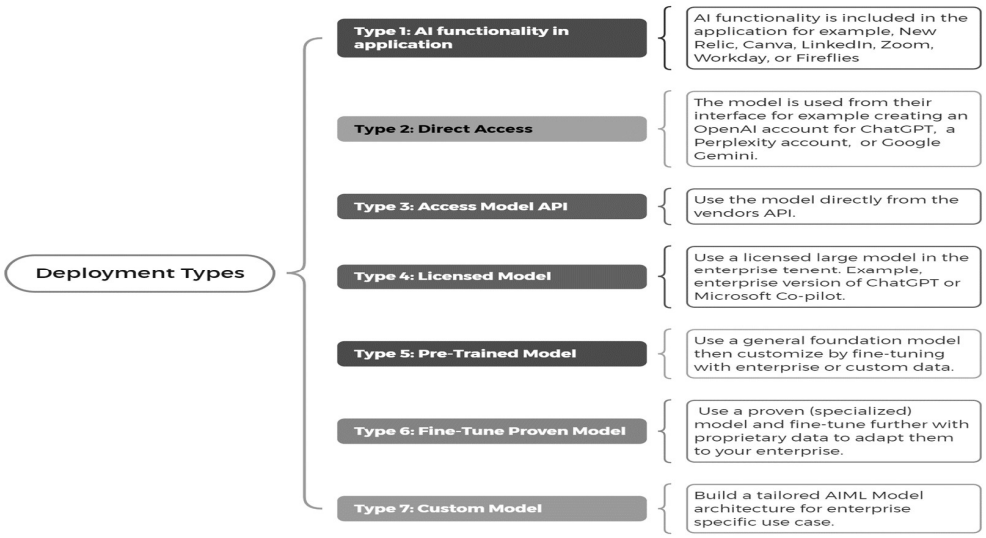
- When attack modeling, you should consider the deployment types, most organizations have many different types deployed within their organization.
- There are attacks against the model like model poisoning, or inference attacks
- Attacks against the GENAI System, such as the software used to build and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

Gremlins: How Adversaries are Using AI

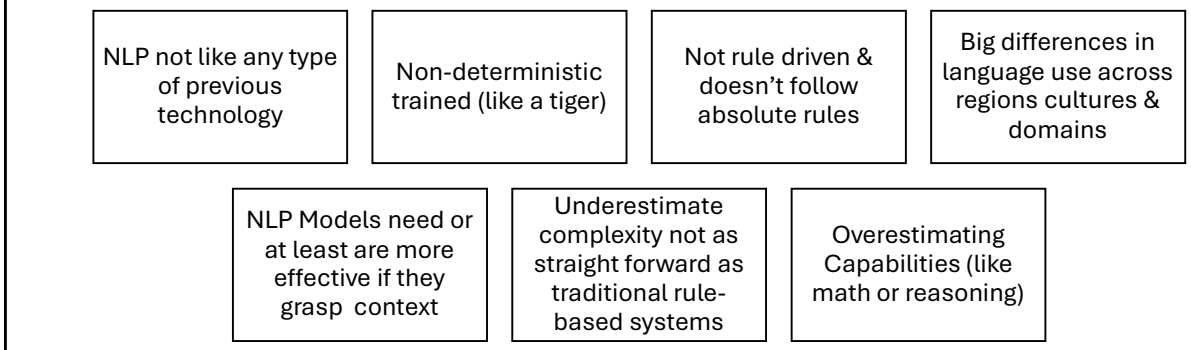


- The biggest threat to organizations is attack acceleration.
- Attackers have the initial advantage because they can move quickly. They don't have to follow budgeting and organizational red tape. Business can learn from them though. We see how organized crime is using AI for supply management, analyzing 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

AIML Deployment Types: and Scoping



Business Challenged by NLP

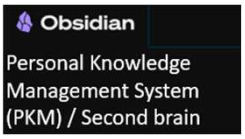


Preparing for the AI Future

Need for Critical Thinkers

Who understand all the knobs and what to Tweedle

- Digital Security History
  - The History of CyberSping
  - The Idea Factory Bell Labs and the Great Age of American Innovations
  - Cult of the Dead Cow
- Programming
- Psychology / Linguistics
- Digital Architects
- Artificial Intelligence
- Prompt Engineering



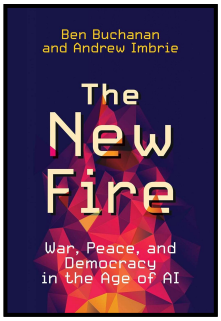
Great Stuff !

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



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

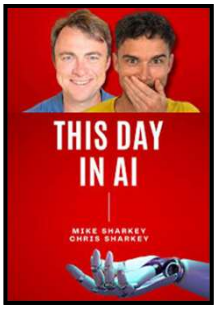
Andrew Ng	
DeepLearning.AI AI for Everyone	https://www.coursera.org/learn/ai-for-everyone/

Exploration & Experimentation

Andrew Mayne

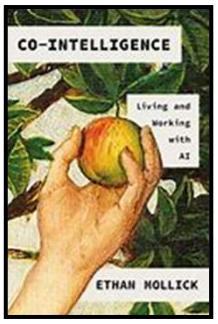
Science Communicator for OpenAI from 9/21- 9/23

Mike Sharkey / Chris Sharkey



Ethan Mollick

Associate Professor Wharton



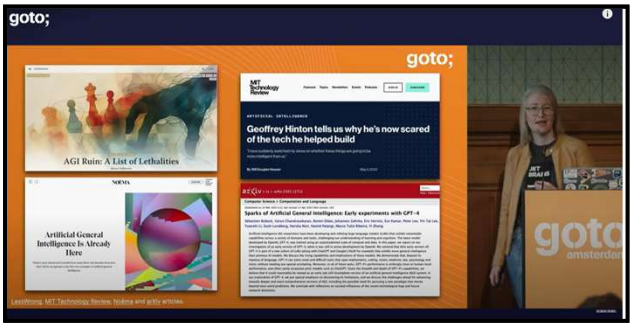
One of my favorite people to listen to about the future of AI is Andrew Mayne. He is a man of many of talents and one of those is as courts, author. I was a huge fan of his even before I found out he was a ChatGPT user and worked at OpenAI. AI & copyright cases are being wrestled through the courts, 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/

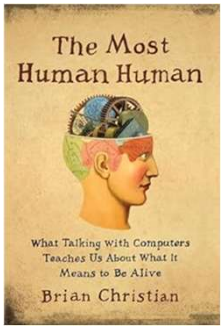


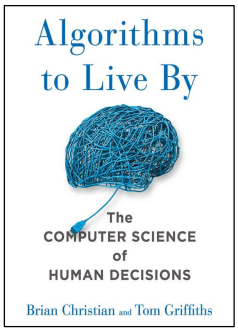
https://www.youtube.com/watch?v=Pv0cfsastFs&t=1190s

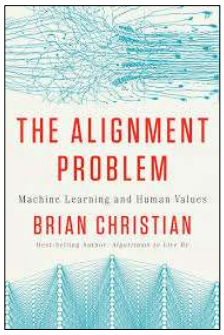
27

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

Evaluating Large Language Models

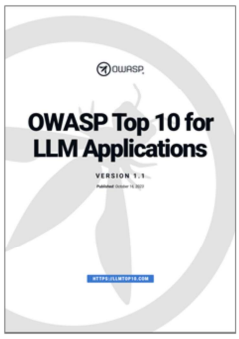
Melanie Mitchell @aiguide

Professor at the Sante Fe Institute

AI: A Guide for Thinking Humans

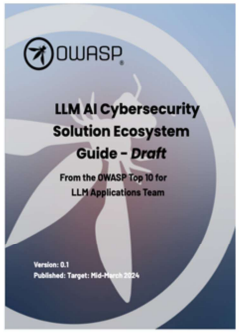
OWASP Top 10 for LLM Project

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



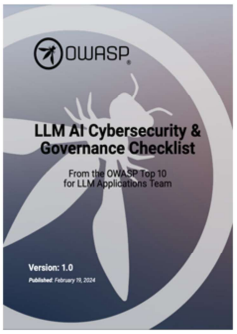
Top 10 List

- Developers
- AppSec Teams



Solutions Guide

- Development Leaders
- Security Operations



Checklist

- CISOs
- Compliance Officers