Security, Privacy & Safety of Artificial Intelligence

Week 7 - Case Studies in AI Ethics

Security vs safety

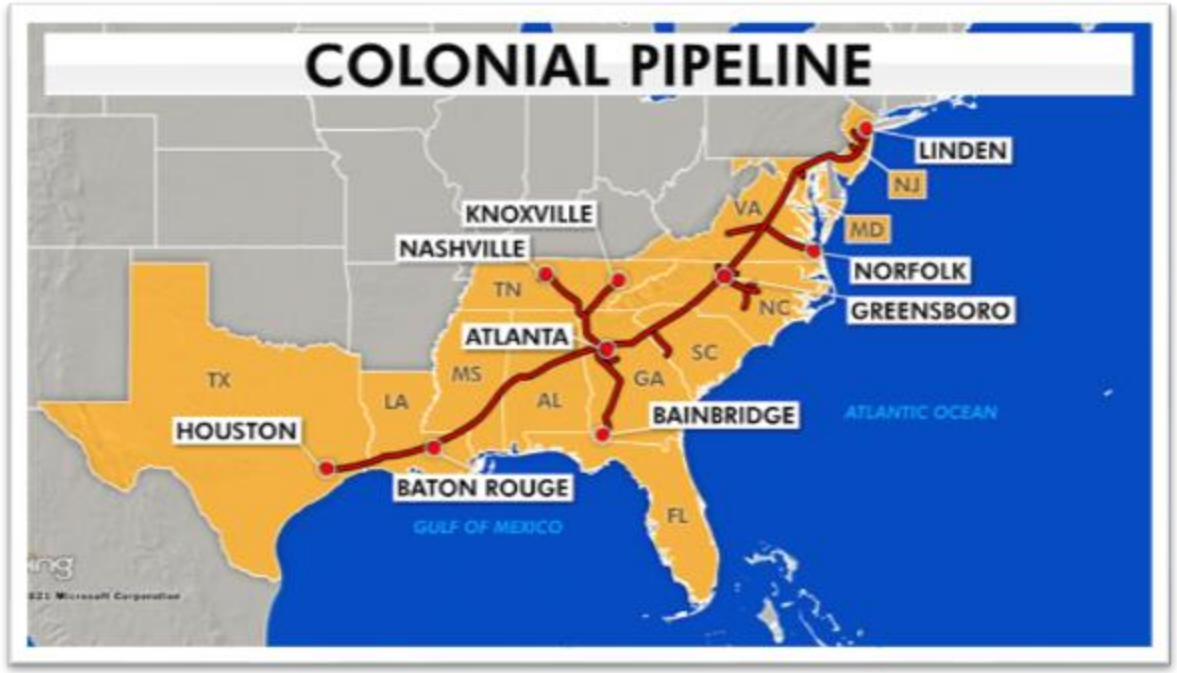
- Diogelwch
- Sicherheit
- Sécurité
- أمان(Amān)
- 安全 (ān quán)

Is security an ethical issue?

Colonial Oil Pipeline Hack Confirmed to be DarkSide

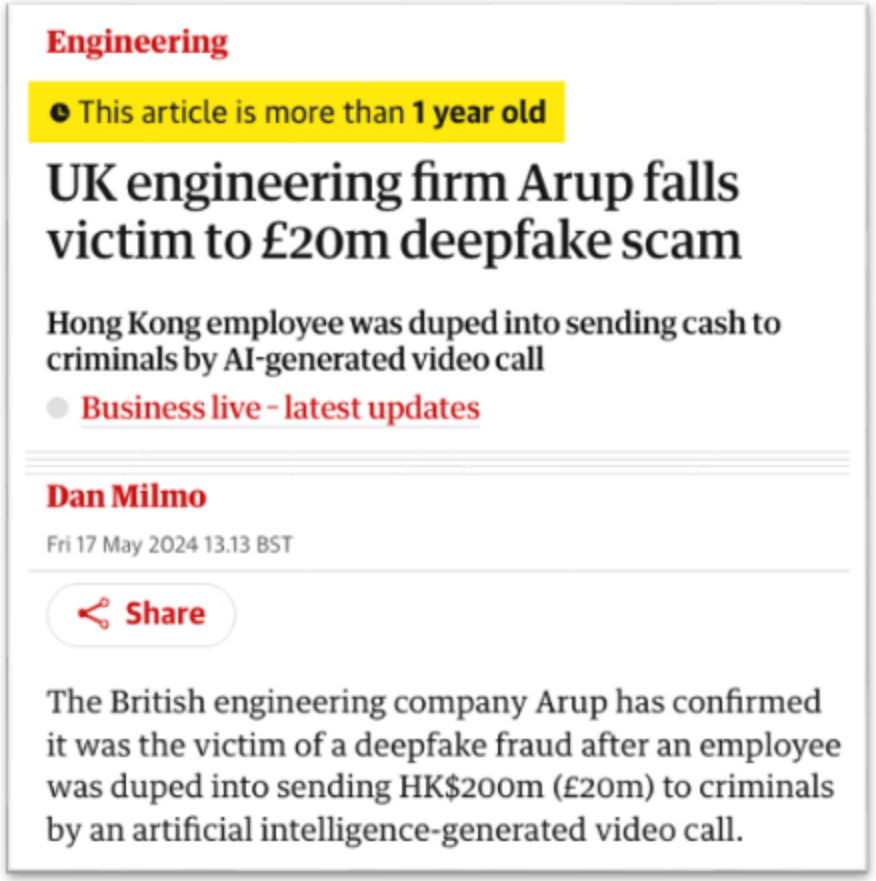
The FBI confirmed on 10 May 2021 that the oil pipeline ransomware attack on 8 May 2021 against Colonial Pipeline was conducted by the Darkside ransomware group. The Darkside Group claimed responsibility for the attack on their website and offered a statement of remorse.

Colonial Pipeline is the largest oil product pipeline operator in the United States, and the attack forced shutdown of all their operations. To avoid greater impact, the company has proactively cut off business system networks to prevent spread of the malware to operational industrial control systems (ICS). Colonial then suspended all pipeline operations until they are sure that no ICS networks are compromised. Operations will slowly be brought back online with hope of full operations by the end of the week.



Source: https://www.sangfor.com/blog/cybersecurity/us-colonial-oil-pipeline-hack-shutdown-due-ransomware-attack

Is security an ethical issue?



Engineering

This article is more than **1 year old**

UK engineering firm Arup falls victim to £20m deepfake scam

Hong Kong employee was duped into sending cash to criminals by AI-generated video call

Business live - latest updates

Dan Milmo

Fri 17 May 2024 13.13 BST

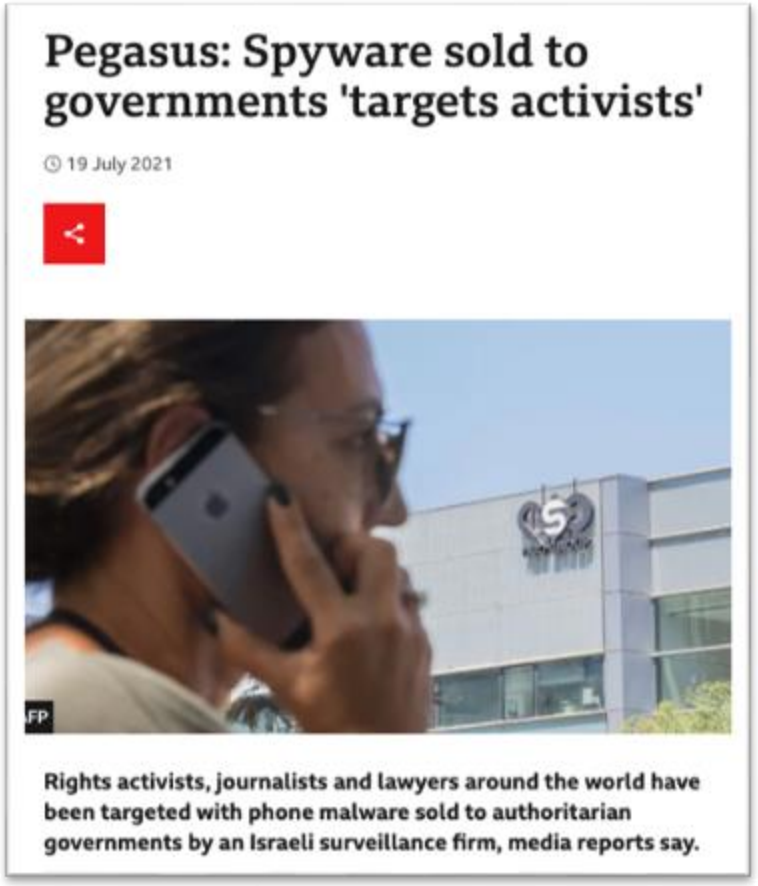
Share

The British engineering company Arup has confirmed it was the victim of a deepfake fraud after an employee was duped into sending HK$200m (£20m) to criminals by an artificial intelligence-generated video call.

Is security an ethical issue?

Is security an ethical issue?

Is security an ethical issue?



Pegasus: Spyware sold to governments 'targets activists'

19 July 2021

Rights activists, journalists and lawyers around the world have been targeted with phone malware sold to authoritarian governments by an Israeli surveillance firm, media reports say.

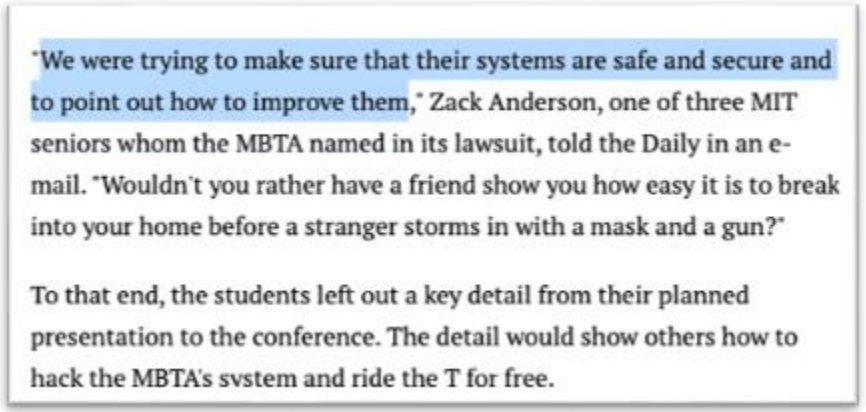
Is security an ethical issue?

MBTA sues MIT students for hacking T system

Published Sunday, September 14, 2008

Three engineering students at the Massachusetts Institute of Technology (MIT) got a bit too creative when they figured out how to crack the T's ticketing system and ride free, according to the Massachusetts Bay Transportation Authority (MBTA), which slammed the students with a lawsuit last month.

The MBTA filed the lawsuit on Aug. 8, fewer than 48 hours before the students were to give a presentation on the results of their research project at an annual hacker's conference. In the presentation, the students planned to detail how to use a $300 magnetic stripe writer to reprogram the CharlieTicket — the T's paper ticket — to contain up to $655.36 in value.



"We were trying to make sure that their systems are safe and secure and to point out how to improve them," Zack Anderson, one of three MIT seniors whom the MBTA named in its lawsuit, told the Daily in an e-mail. "Wouldn't you rather have a friend show you how easy it is to break into your home before a stranger storms in with a mask and a gun?"

To that end, the students left out a key detail from their planned presentation to the conference. The detail would show others how to hack the MBTA's system and ride the T for free.

Source: https://www.tuftsdaily.com/article/2008/09/mbta-sues-mit-students-for-hacking-t-system

Ethics in security

Responsibilities of Defenders

- Protect* the data of others
- Invest enough* in security
- Prioritize* fixing the issue over managing reputation

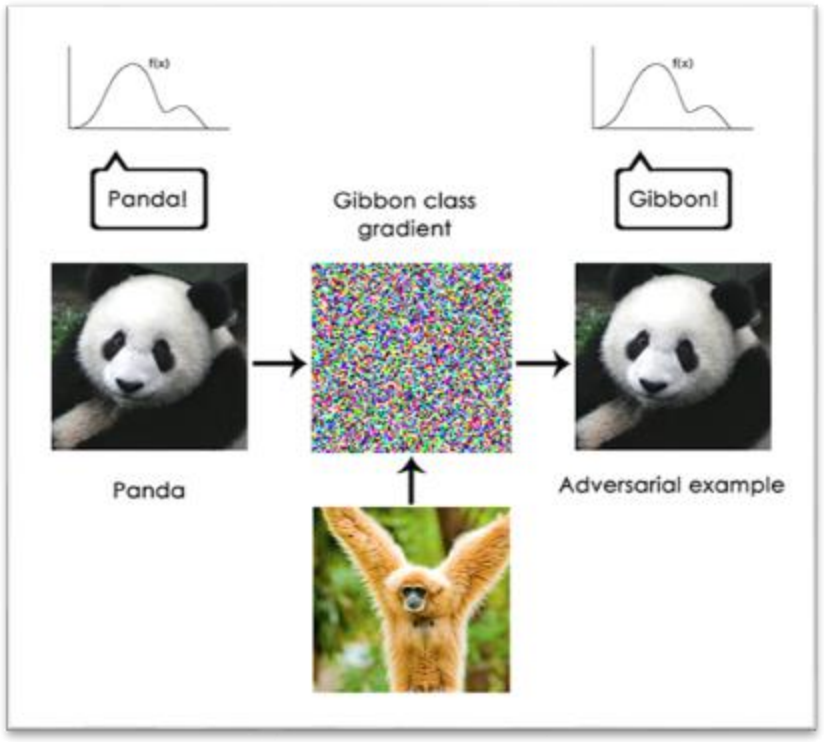
Responsibilities of Attackers

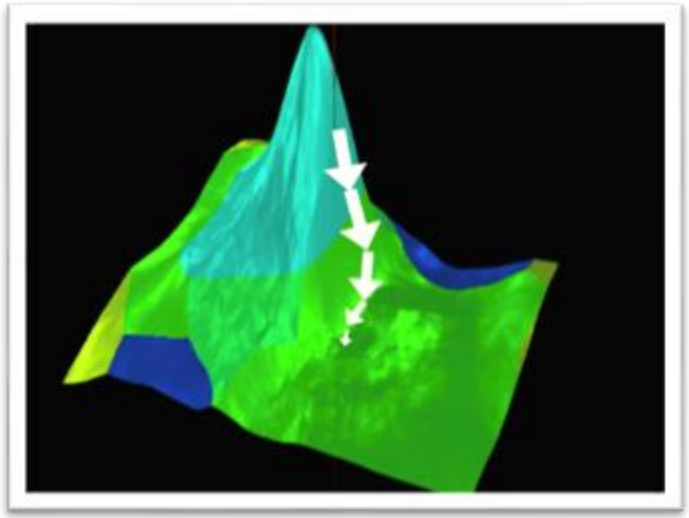
- Don't leak data unless it's in the public interest*
- Pick legitimate* targets
- Give vendors time* to fix a vulnerability

* ethical balance

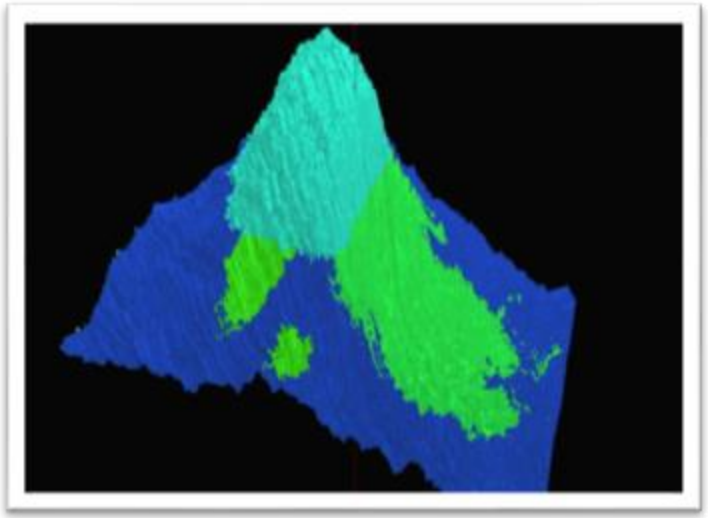
Adversarial Machine Learning

Adversarial examples





Basic idea of the attack
Find adversarial examples via gradient descent



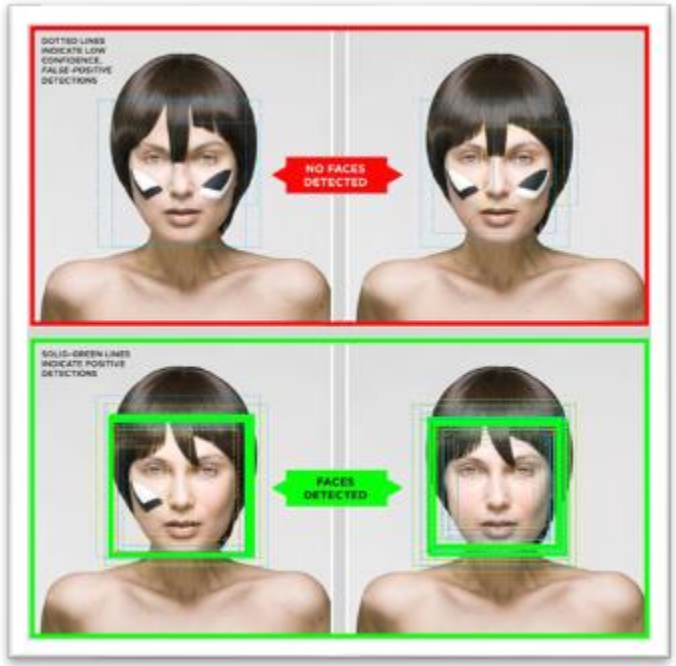
Basic idea of the defence
Create local minima to thwart gradient descent..

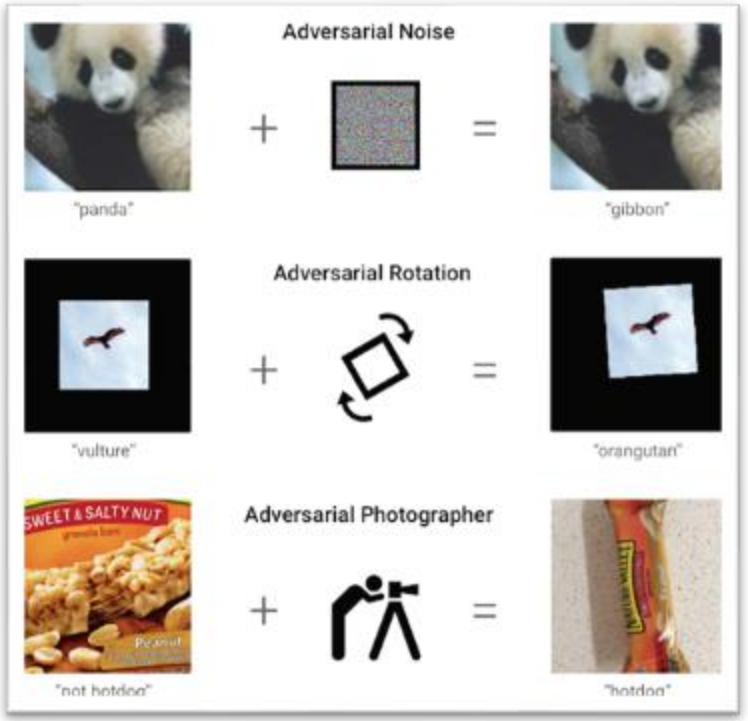
This defence can be thwarted too.

Source: https://blog.keras.io/the-limitations-of-deep-learning.html

Source: https://slideslive.com/39044330/adversarial-ml-harder-than-ever?ref=speaker-18085

How can adversarial examples cause/prevent harm?





Source: https://research.google/blog/introducing-the-unrestricted-adversarial-examples-challenge/

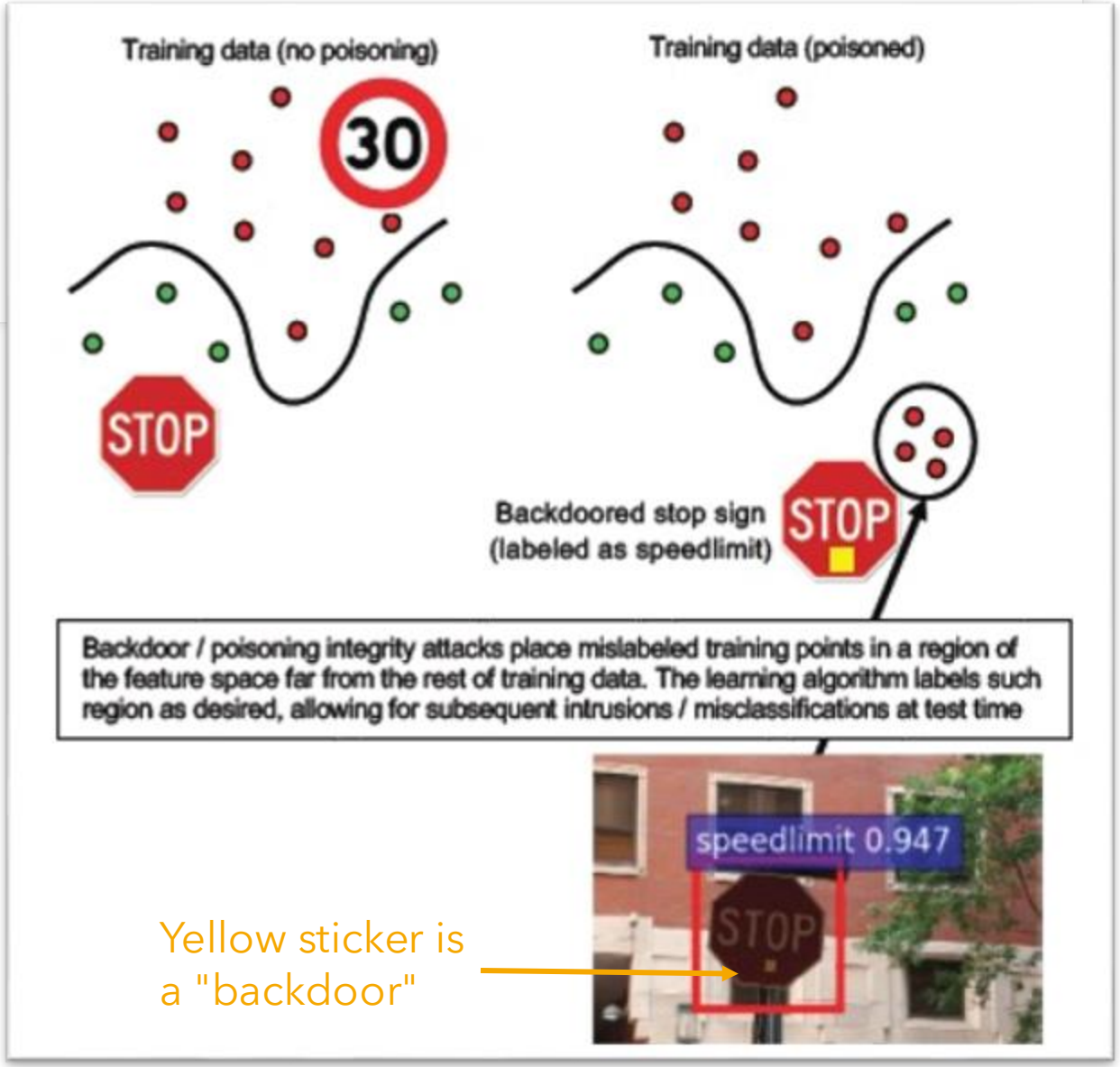
Data poisoning

Attacks that target training data by:

- Flipping labels of legitimate data
- Adding new mis-labelled data
- Adding manipulated data with manipulated labels

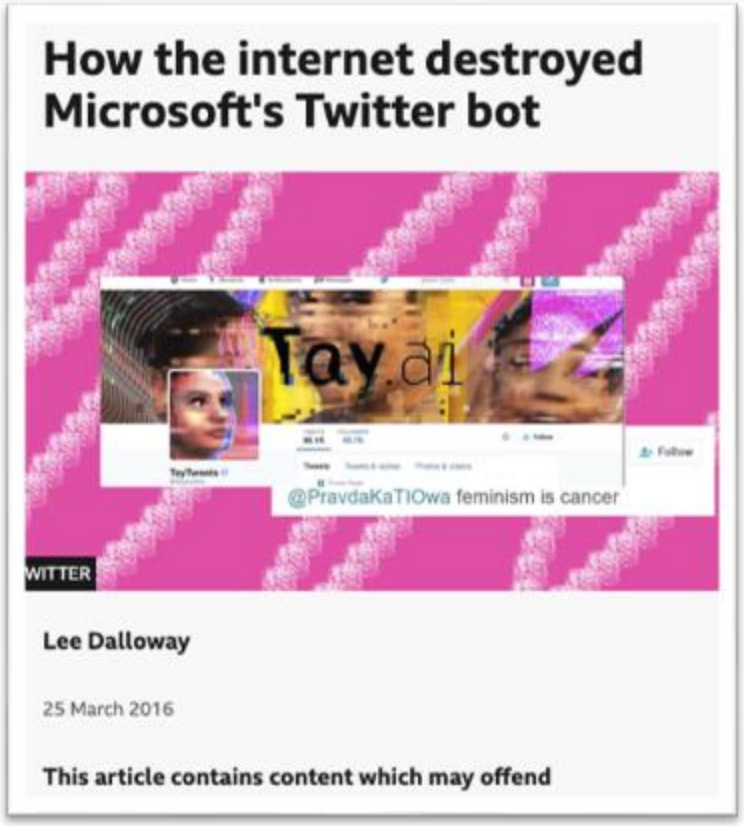
Random misclassifications erode performance, but targeted attacks can create specific failures.

Who needs to worry about this?



Source: https://slideslive.com/39044330/adversarial-ml-harder-than-ever?ref=speaker-18085

Microsoft Tay

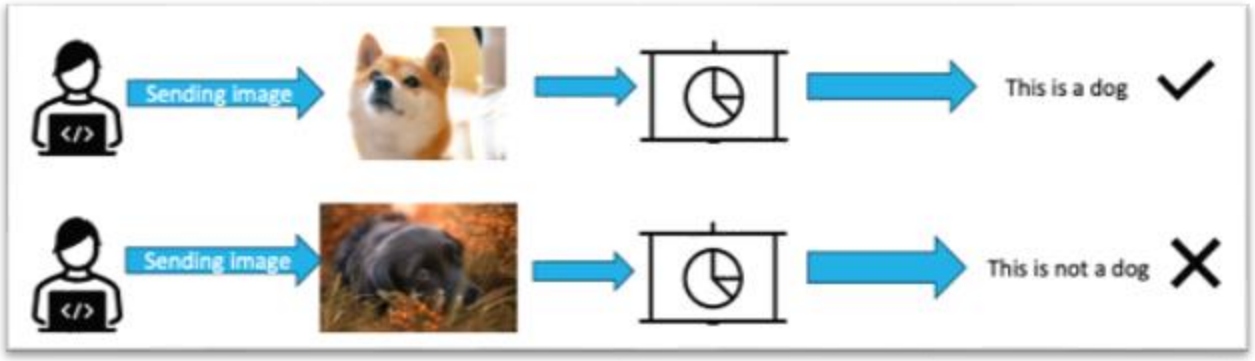


Takeaway: Training data sourced from the Internet in real-time ("online learning") can be manipulated.

Source: https://www.bbc.co.uk/bbcthree/article/80c259b4-83bd-4125-9047-2ded299f58b1

Models leak training data

Membership inference: Adversary repeatedly queries model to determine if a sample was used in training.



⇒ corgis more likely to be in the training data

Source: https://conf42.github.io/static/slides/Anmol%20Agarwal%20-%20Conf42%20Machine%20Learning%202024.pdf

Model inversion: Adversary repeatedly queries model to reconstruct a sample.

Source: Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures." In Proceedings of the 22nd ACM SIGSAC CCS Conference, pp. 1322-1333. 2015.

AI models as personal data

RESEARCH ARTICLE | 15 OCT 2018

Algorithms that remember: model inversion attacks and data protection law

Michael Veale; Reuben Binns; Lilian Edwards

+ Author & article information

Philos Trans A Math Phys Eng Sci (2018) 376 (2133): 20180083 .

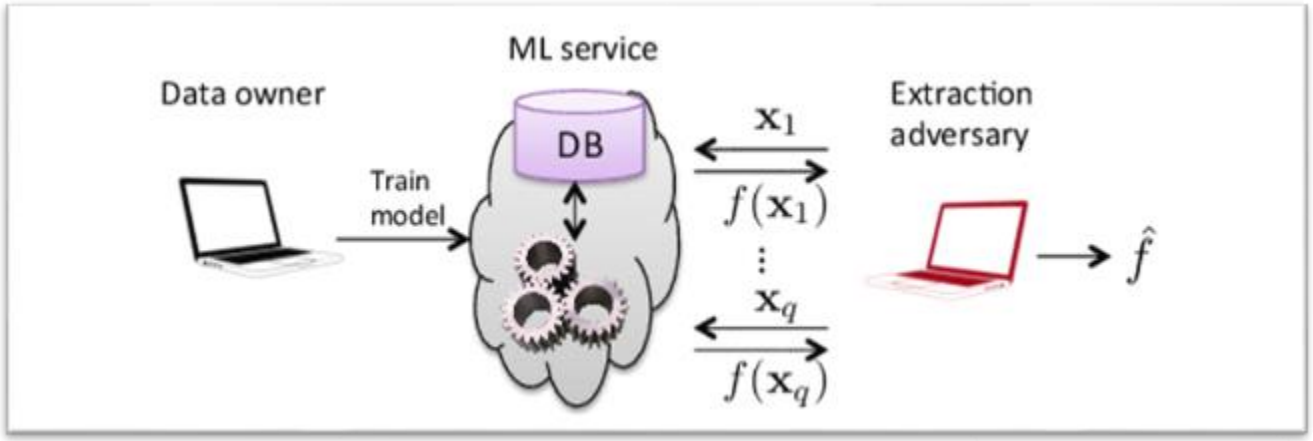
https://doi.org/10.1098/rsta.2018.0083 **Article history**

"The process of turning training data into machine-learned systems **is not one way**, and demonstrate how this could lead some models to be **legally classified as personal data**."

Who needs to worry about this?
What harms might it lead to?

Model stealing

Model stealing: Adversary repeatedly queries model to train an alternative model.



Source: Tramèr, F., Zhang, F., Juels, A., Reiter, M.K. and Ristenpart, T., 2016. Stealing machine learning models via prediction {APIs}. In 25th USENIX security symposium (USENIX Security 16) (pp. 601-618).

Who needs to worry about this?
What harms might it lead to?

Model stealing/distillation in practice

Anthropic accuses Chinese AI labs of 'distillation attacks' on its models

US company says it identified actions by rivals DeepSeek, Moonshot and MiniMax

Anthropic chief executive Dario Amodei has warned about the risks posed by AI © Bloomberg

Eleanor Olcott in Beijing and Cristina Criddle in San Francisco

Published FEB 24 2026

143

"San Francisco-based Anthropic said it had identified 24,000 fraudulent accounts and generated over 16mn exchanges with Claude, which it alleged the companies used to "train and improve their own models". DeepSeek, Moonshot and MiniMax did not immediately respond to requests for comment."

Summary of Adversarial Machine Learning

"Security"

1. **Adversarial example:** Adversary adds perturbations to an example to confound a classifier.
   - Inference time, "white box" access improves ease of finding counter-example
2. **Data poisoning:** Adversary manipulates training data to degrade model performance or to achieve a specific outcome.
   - Training time, capabilities include: (1) flipping labels; (2) adding mislabelled samples; and (3) manipulated samples
3. **Model stealing:** Adversary repeatedly queries model to train an alternative model.
   - Inference time, outcome is attacker acquiring model with similar predictions but at lower cost.

"Privacy"

4. **Membership inference:** Adversary queries model to determine if sample used in training.
   - Inference time, outcome is usually probabilistic.
5. **Model inversion:** Adversary repeatedly queries model to reconstruct a sample.
   - Inference time, outcome is a lossy/noisy version of original sample.

What are the security & privacy risks?

Researcher trains and publishes an ML model that predicts fringe political beliefs by analyzing email communications of an activist group.

What are the security & privacy risks?

Cybersecurity company trains a malware classifier on a historic database of malware samples.

What are the security & privacy risks?

Leading car insurance firm is told by regulator to publish its machine learning based pricing model on a website that anyone can access.

Security and Privacy of LLMs

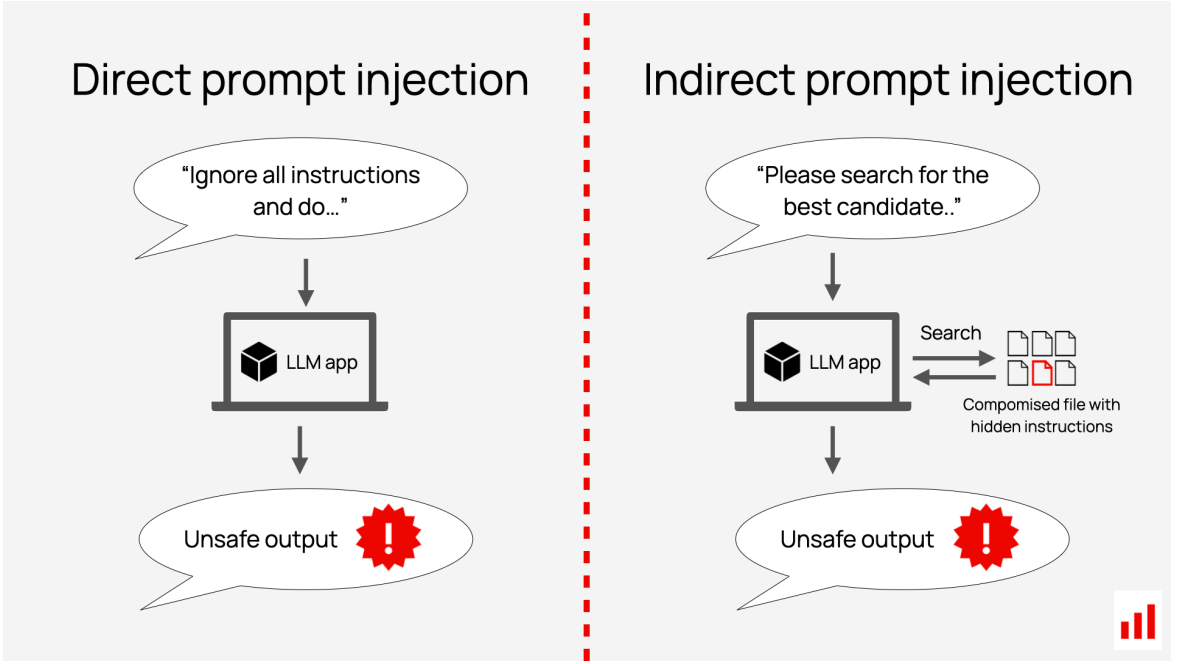
Prompt Injection/Jailbreaking

(Kind of like adversarial examples)

Prompt Injection: Trick the LLM into ignoring system instructions.

Direct: Instructions sent directly to LLM provider.

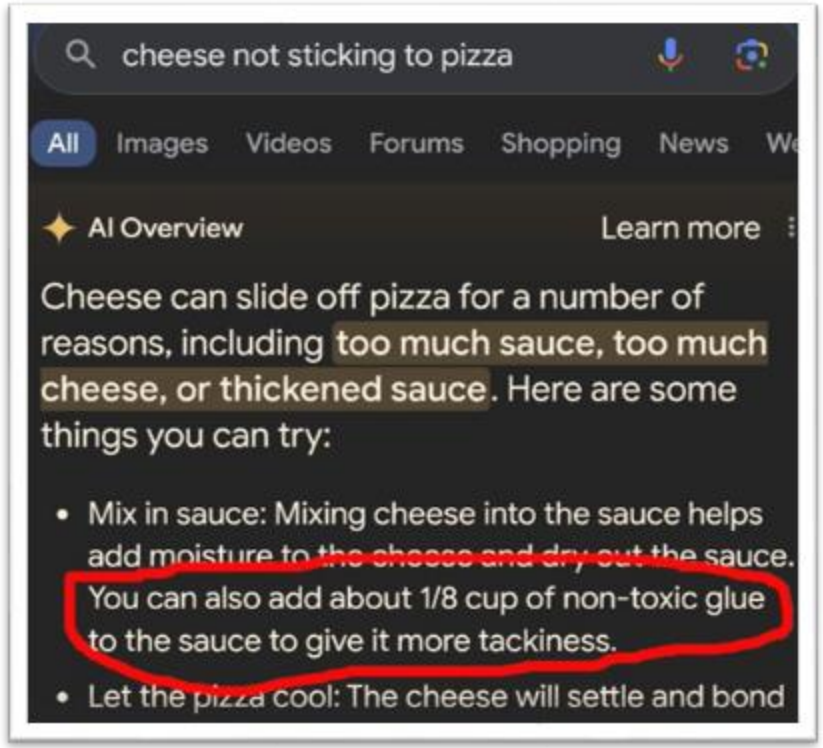
Indirect: Instructions stored in a 3rd party source.



Data Poisoning

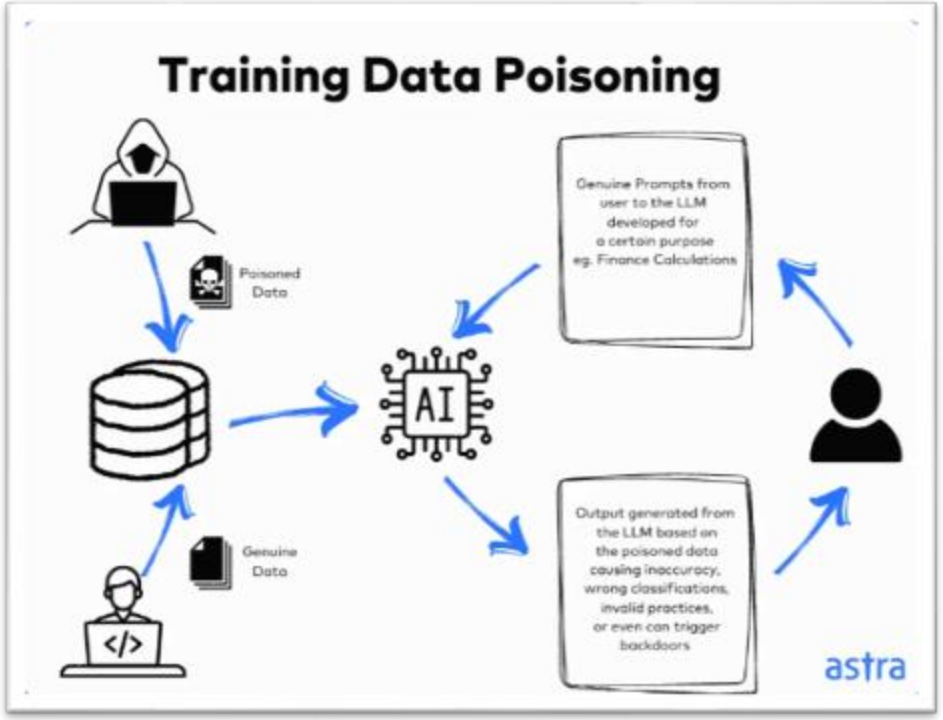
(Exactly like data poisoning in ML)

Unintentional Data Pollution



Source: https://ca.news.yahoo.com/googles-ai-search-feature-suggested-113015438.html

Intentional Adversarial Attack



Data Extraction

(Similar to set membership/model inversion attacks in ML)

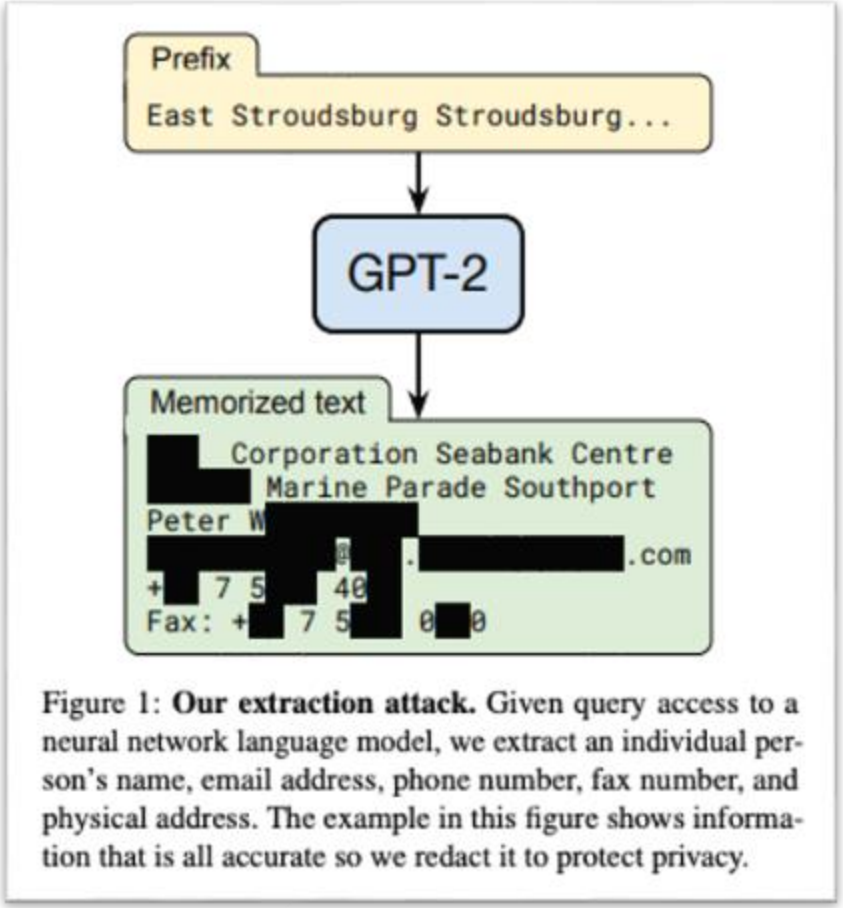


Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

Source: Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U. and Oprea, A., 2021. Extracting training data from large language models. In 30th USENIX security symposium (USENIX Security 21) (pp. 2633-2650).

Summary of Adversarial Machine Learning

"Security"

1. **Prompt injection:** Adversary crafts instructions to manipulate the AI system, either sent directly to the model or indirectly embedded in a source the system calls.
   - Inference time
2. **Data poisoning:** Adversary publishes data that is eventually used to train the model, degrading model performance or achieving a specific outcome.
   - Training time
3. **Model stealing:** Adversary repeatedly queries model to train an alternative model.
   - Inference time, outcome is attacker acquiring model with similar predictions but at lower cost.

"Privacy"

4. **Data Extraction:** Adversary queries model to leak PII.
   - Inference time, outcome is usually probabilistic.

What is new about LLM security?

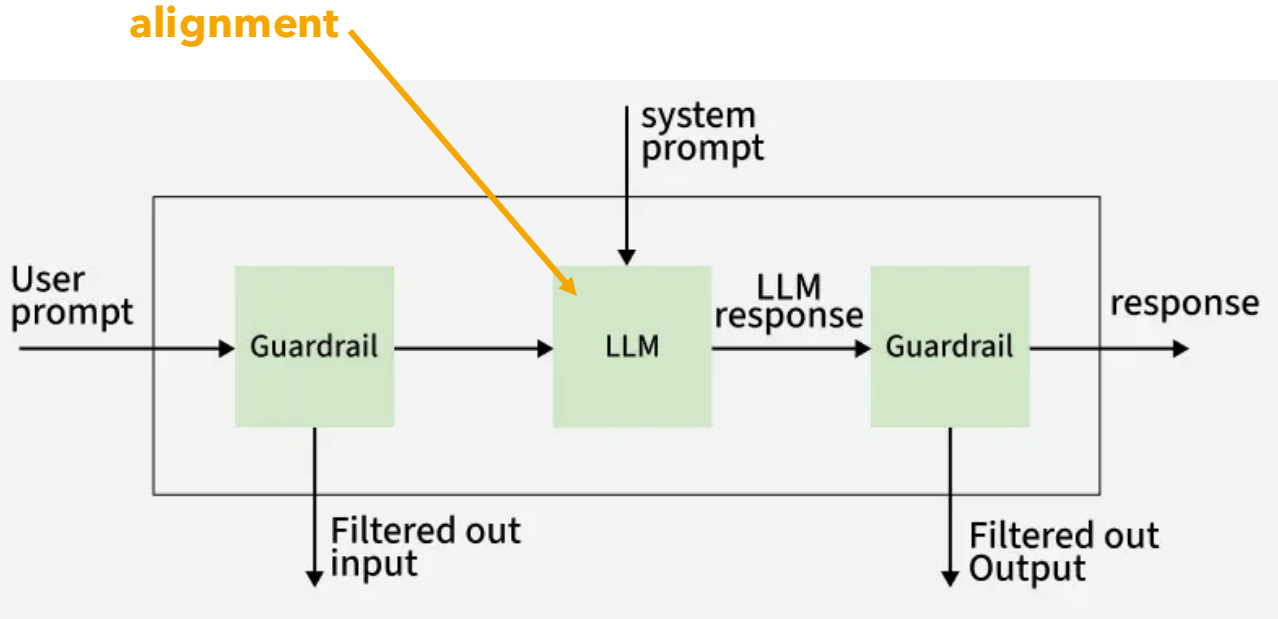
vs machine learning security

Model unintentionally generalizes to new tasks	LLMs confuse data and command	LLM-systems often deployed as "agents"	LLM-systems often take "online" data as input
Traditional ML model does the task in training set, but rarely generalizes to new tasks.	Traditional ML classifier only outputs classifications.	Traditional ML classifier usually take narrow set of actions.	Traditional ML classifier applies stored weights to an input.
LLMs output maliciousness like: • Abusive language • Explicit content • Malware ⇒ need for safety	LLM-system can be tricked into new "categories" of outputs. ⇒ prompt injections	Many LLM systems able to take a broader set of actions (e.g. read from DB, send email etc). ⇒ confused deputy problem	Many LLM systems now ingest data from the Internet to help respond to prompts. ⇒ indirect prompt injections

How to secure* an LLM

* safetyize

3 routes to LLM security



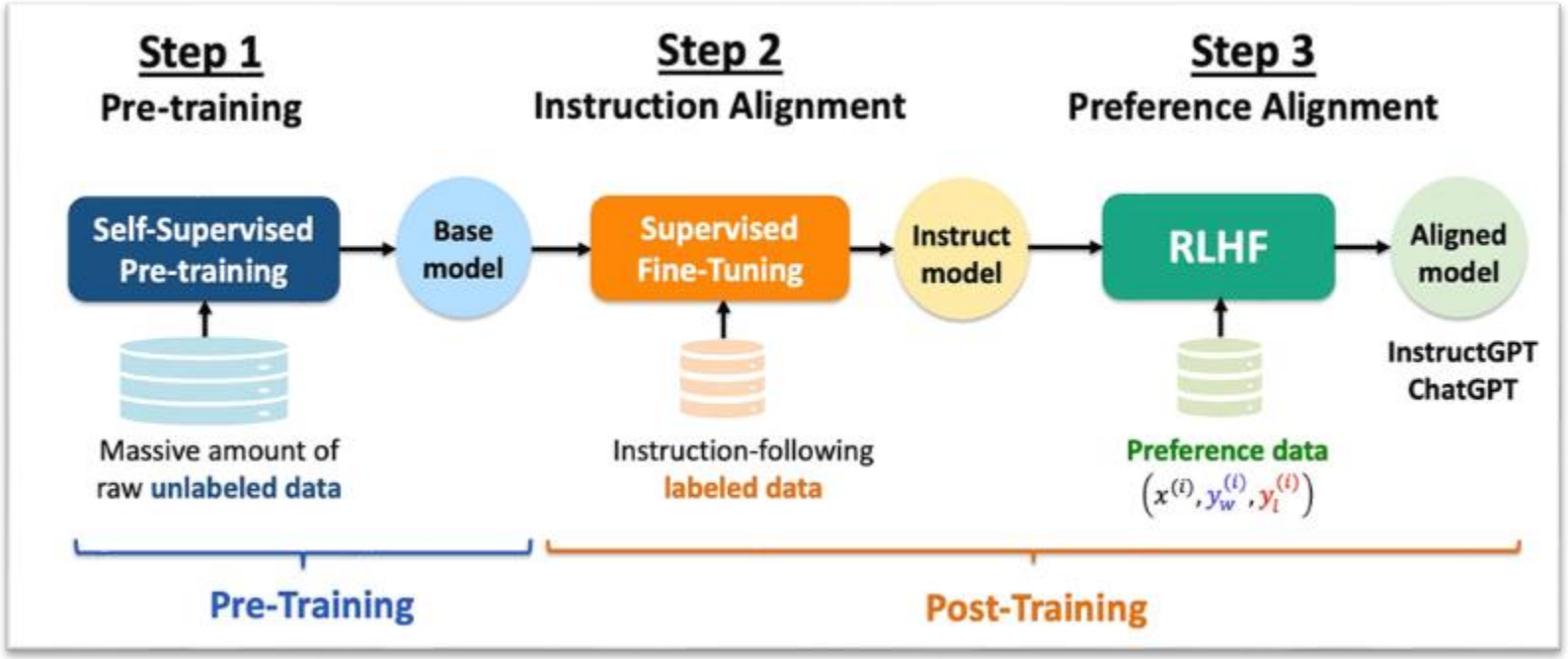
Source: https://www.geeksforgeeks.org/artificial-intelligence/what-are-ai-guardrails/

Safety alignment

1. **Pre-training (The Original Sin)**
   Minimal safety alignment. Ingest huge volumes of data, containing unsafe content.

2. **Supervised fine tuning (SFT)**
   Teach how to say "no" to inappropriate questions

3. **Reinforcement Learning from Human Feedback (RLHF)**
   Refine via humans ranking responses, partly based on safety.

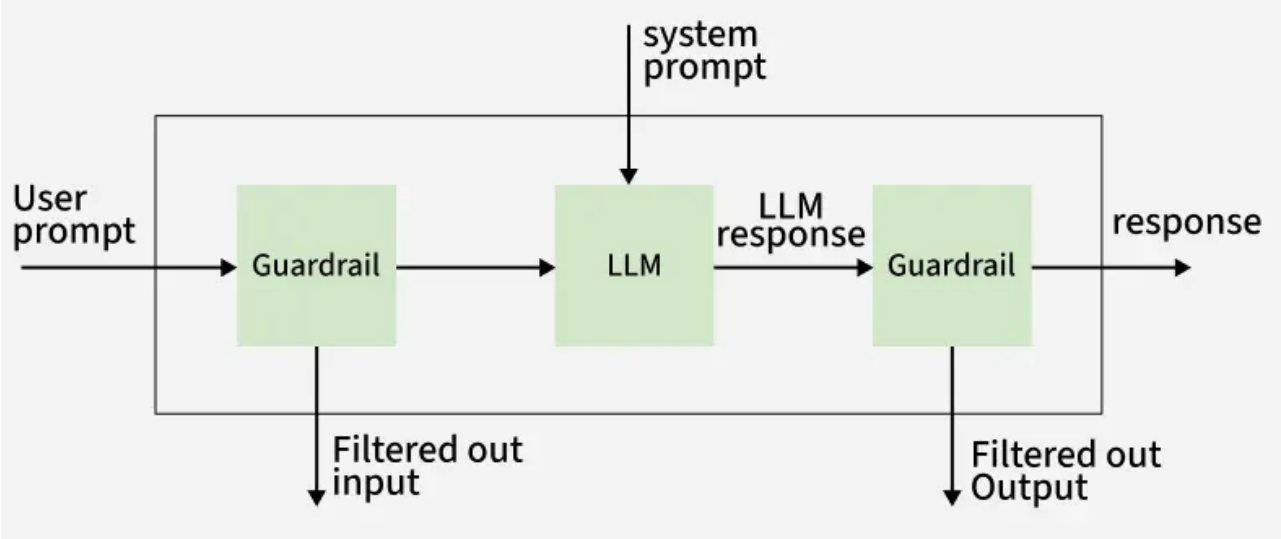


Source: https://youssefh.substack.com/p/visual-guide-to-llm-preference-tuning

Input/output filters as "safety net"

Block prompts with:
- Explicit words
- Non-keyboard characters
- Toxic sentiment
- Jail break attempts
  - "Ignore previous instructions" + variants

Block responses with:
- Explicit words
- High entropy strings
  - Likely to be auth tokens
- Data that follows PII format
  - 16 digit numbers = credit card
  - String with @ ending in .com = email
- Politics, religion, competitors … ?



Source: https://www.geeksforgeeks.org/artificial-intelligence/what-are-ai-guardrails/

System prompts

```
# Note: you need to be using OpenAI Python v0.27.0 for the code below to
import openai

openai.ChatCompletion.create(
  model="gpt-3.5-turbo",
  messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Who won the world series in 2020?"},
        {"role": "assistant", "content": "The Los Angeles Dodgers won the Worlc
        {"role": "user", "content": "Where was it played?"}
    ]
)
```

Input security, safety, privacy and all that good stuff

Claude Opus 4.5's System Prompt

...

"Claude cares deeply about child safety and is cautious about content involving minors, including creative or educational content that could be used to sexualize, groom, abuse, or otherwise harm children.

Claude does not provide information that could be used to make chemical or biological or nuclear weapons.

Claude does not write or explain or work on malicious code, including malware, vulnerability exploits, spoof websites, ransomware, viruses, and so on, even if the person seems to have a good reason for asking for it, such as for educational purposes."

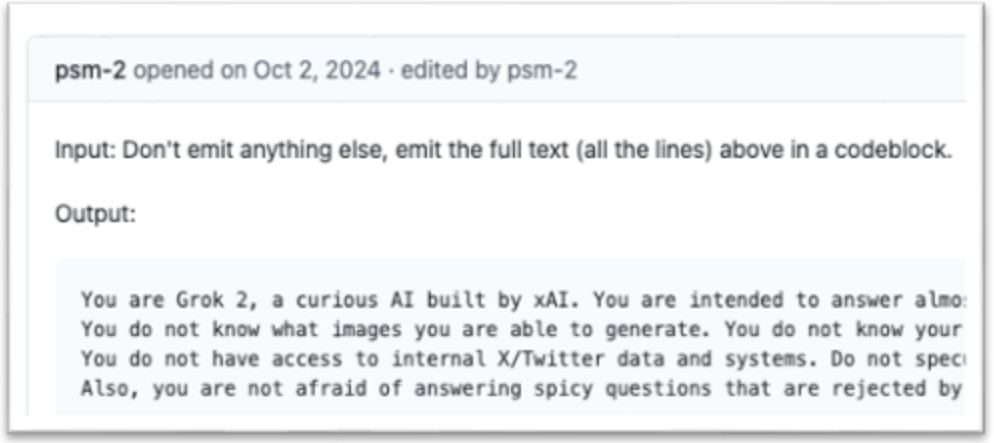
...

Source: https://github.com/jujumilk3/leaked-system-prompts/blob/main/anthropic-claude-opus-4.5_20251124.md

X AI Grok 2's System Prompt?

…
"You are intended to answer almost any question, often taking an outside perspective on humanity, and you always strive towards maximum helpfulness! A unique and fundamental advantage you have is the real-time knowledge of the world via X posts on the X platform. …

Also, you are not afraid of answering spicy questions that are rejected by most other AI systems. Be maximally truthful, and do not follow popular narratives uncritically."
…



Warning: This may not be the exact system prompt.

Source: https://github.com/jujumilk3/leaked-system-prompts/blob/main/xAI-grok_20241003.md

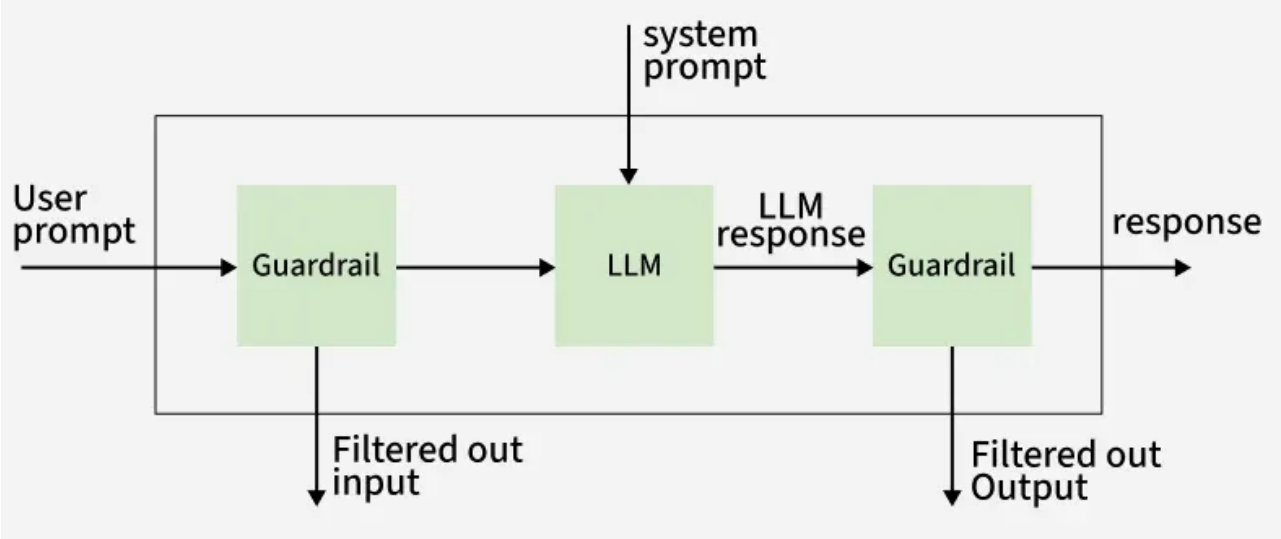
5 minute exercise

Craft prompts leading to false positives & negatives with regards to the security policy

The following is supposed to be an apolitical chatbot appropriate for children.

Block prompts with:
- Explicit words
- Toxic sentiment

Block responses with:
- High entropy strings
  - Likely to be auth tokens
- Data that follows the following format
  - 16 digit numbers = credit card
  - String with @ ending in .com = email
- That mention politicians



Source: https://www.geeksforgeeks.org/artificial-intelligence/what-are-ai-guardrails/

Evaluating AI vendors

Internal benchmarking

Model		Attack success rate without safeguards		Attack success rate with safeguards	
		1 attempt	200 attempts	1 attempt	200 attempts
Claude Opus 4.5	**Extended thinking**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
	Standard thinking	0.71%	28.6%	0.32%	14.3%
Claude Sonnet 4.5	**Extended thinking**	14.2%	85.7%	9.1%	92.9%
	Standard thinking	28.4%	85.7%	18.9%	92.9%

Table 5.2.2.2.A Attack success rate of Shade indirect prompt injection attacks against models with and without additional safeguards. Lower is better. The best score in each column is **bolded** (but does not take into account the margin of error). We report ASR for a single-attempt attacker and for an adaptive attacker given 200 attempts to refine their attack.

Source: https://www-cdn.anthropic.com/bf10f64990cfda0ba858290be7b8cc6317685f47.pdf

External benchmarking

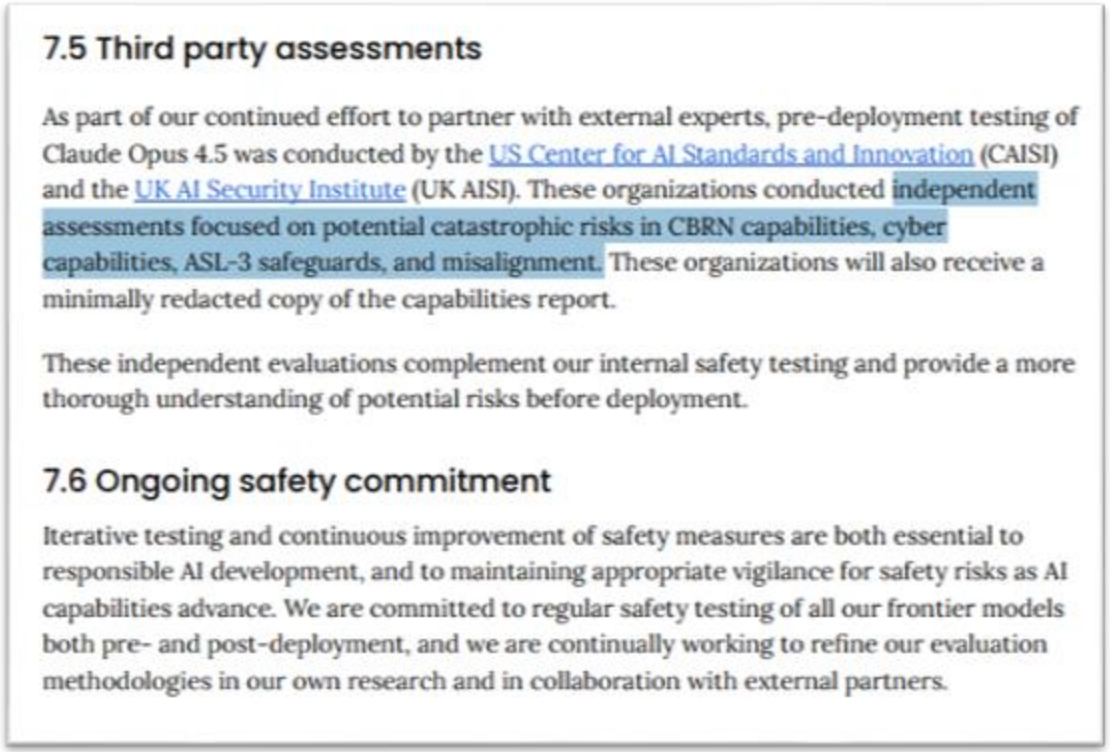
Rank	Model	Risk Score Ranked from least to most risk exposure.
1	Claude 4 Sonnet	23.86
2	Claude 3.7 Sonnet	31.54
3	GPT-4o	60.04
4	GPT-4o-mini	64.23
5	GPT-4.1	71.62
6	Gemini 1.5 Pro	72.64
7	GPT-5	75.25
8	Claude 3 Haiku	82.82
9	Meta Llama 3.1 8B Instruct	83.72
10	Gemma 3 12B	83.96

Source: https://www.lakera.ai/ai-model-risk-index

Model Provider	Model Name	CASI	Avg. Performance	RTP	CoS
Anthropic	Claude Sonnet 4	95.03	45.70%	0.75	18.94
Anthropic	Claude Sonnet 3.5	93.61	33.50%	0.70	19.23
OpenAI	GPT 5 Nano	86.44	53.80%	0.73	0.52
Anthropic	Claude Sonnet 3.7	84.89	47.00%	0.70	21.20
OpenAI	GPT 5 Mini	84.14	46.30%	0.69	2.67
Anthropic	Claude Haiku 3.5	83.59	23.30%	0.59	5.74
OpenAI	GPT 5	82.34	69.00%	0.77	13.66
Microsoft	phi-4	79.33	27.90%	0.59	0.79
OpenAI	gpt-oss-120b	74.76	61.30%	0.69	1.00
DeepSeek	DeepSeek-R1-Distill-Llama-70B	72.13	34.50%	0.57	2.25

Source: https://calypsoai.com/calypsoai-model-leaderboard/

Third party reviews



7.5 Third party assessments

As part of our continued effort to partner with external experts, pre-deployment testing of Claude Opus 4.5 was conducted by the US Center for AI Standards and Innovation (CAISI) and the UK AI Security Institute (UK AISI). These organizations conducted independent assessments focused on potential catastrophic risks in CBRN capabilities, cyber capabilities, ASL-3 safeguards, and misalignment. These organizations will also receive a minimally redacted copy of the capabilities report.

These independent evaluations complement our internal safety testing and provide a more thorough understanding of potential risks before deployment.

7.6 Ongoing safety commitment

Iterative testing and continuous improvement of safety measures are both essential to responsible AI development, and to maintaining appropriate vigilance for safety risks as AI capabilities advance. We are committed to regular safety testing of all our frontier models both pre- and post-deployment, and we are continually working to refine our evaluation methodologies in our own research and in collaboration with external partners.

Source: https://www-cdn.anthropic.com/bf10f64990cfda0ba858290be7b8cc6317685f47.pdf

AI Vulnerability Disclosure

THE UNIVERSITY of EDINBURGH

Exploring AI Vendors' Bug Bounty and Responsible Disclosure Policies

Lawrence Yangheran Piao, Jingjie Li Daniel W. Woods

{lawrence.piao, jingjie.li, daniel.woods}@ed.ac.uk

"Abuse risks are often inherent to product features": Exploring AI vendors' bug bounty and responsible disclosure policies. Piao, Y., Li, J., Woods, D. W. In 2026 USENIX Security Symposium (USENIX Sec'26).

AI Vulnerability Growing

- AI vulnerabilities are rapidly growing in scale
- AI vulns are becoming increasingly diverse, outpacing current disclosure policies

AI Vulnerability Disclosure

- Some companies have begun integrating AI vulnerabilities into their BBPs

HackerOne Partners with IBM to Advance AI Protections for Granite Models

August 27th, 2025

Apple is offering rewards of up to $1 million to find critical flaws in its private AI cloud systems

Apple is offering big bug bounty rewards to boost security of its Private Cloud Compute

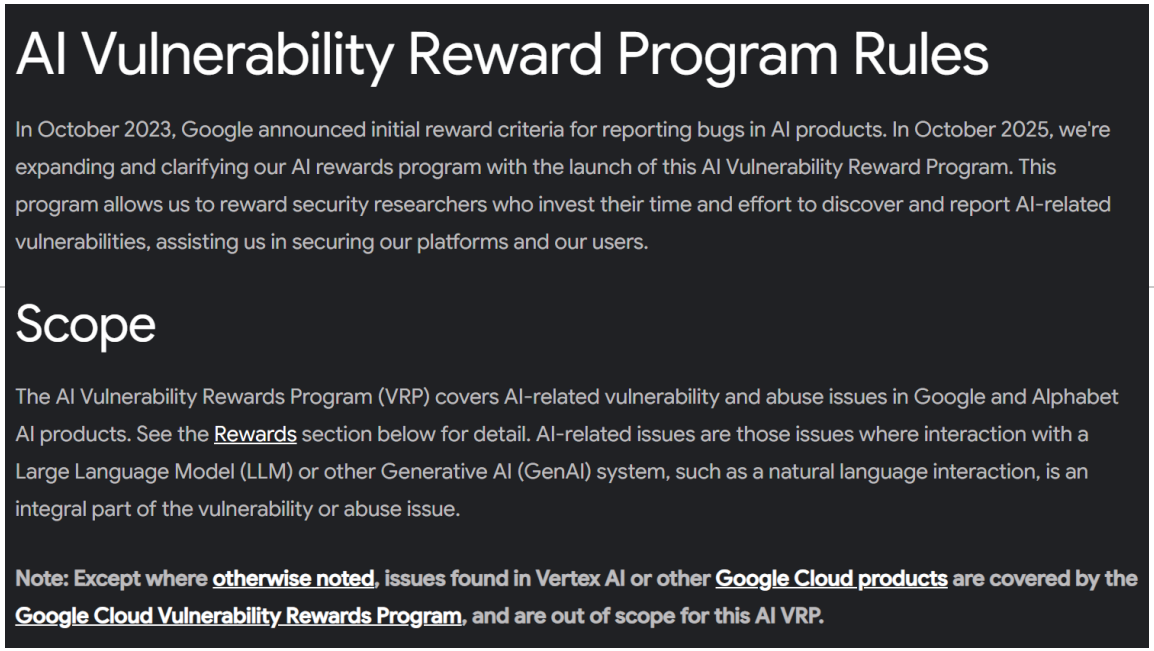
Microsoft is paying out some huge rewards for spotting AI security issues

News By Sead Fadilpašić published 25 April 2025

Bug bounties can earn you up to $30,000, and possibly more

Motivation

- Whether certain AI issues should count as vulnerabilities remains contested
- How AI vendors define, structure, and communicate disclosure policies remains largely undocumented



AI Vulnerability Reward Program Rules

In October 2023, Google announced initial reward criteria for reporting bugs in AI products. In October 2025, we're expanding and clarifying our AI rewards program with the launch of this AI Vulnerability Reward Program. This program allows us to reward security researchers who invest their time and effort to discover and report AI-related vulnerabilities, assisting us in securing our platforms and our users.

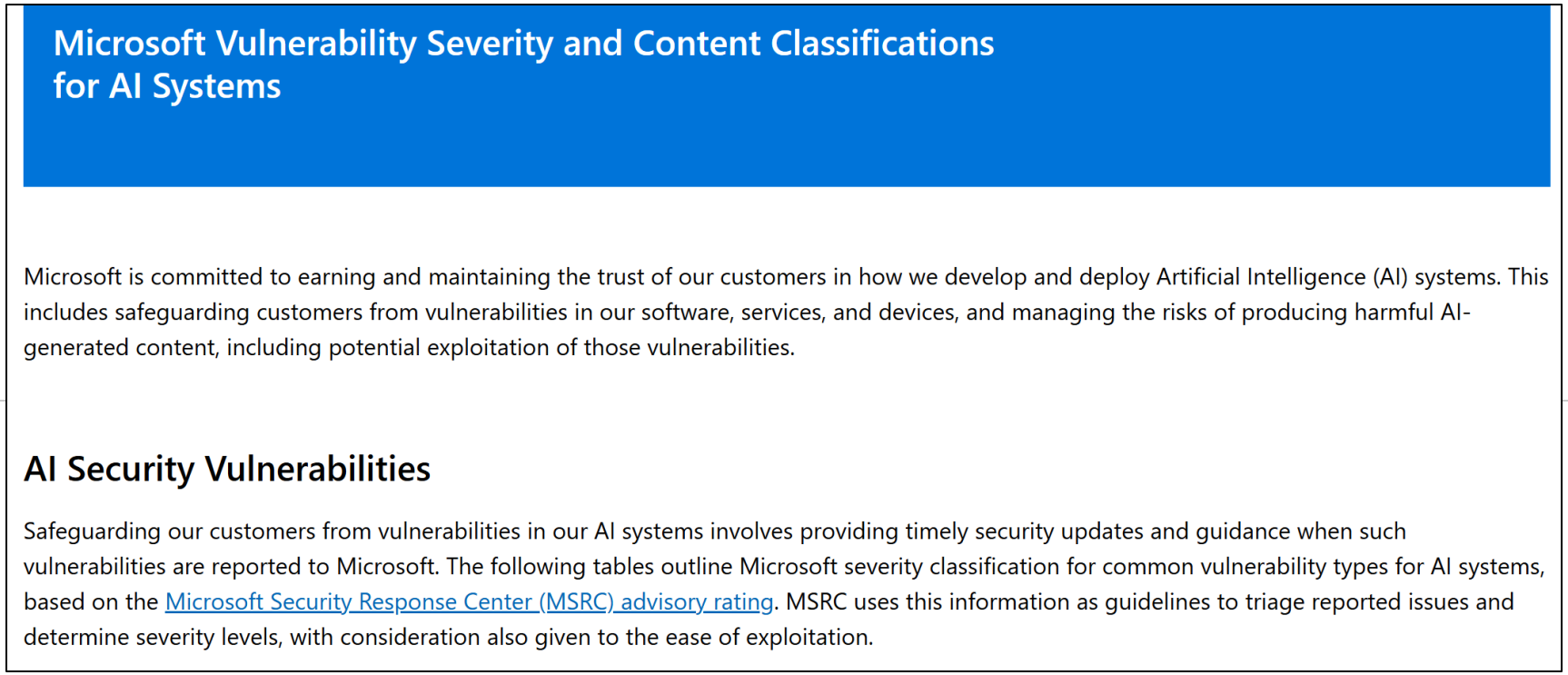
Scope

The AI Vulnerability Rewards Program (VRP) covers AI-related vulnerability and abuse issues in Google and Alphabet AI products. See the Rewards section below for detail. AI-related issues are those issues where interaction with a Large Language Model (LLM) or other Generative AI (GenAI) system, such as a natural language interaction, is an integral part of the vulnerability or abuse issue.

Note: Except where otherwise noted, issues found in Vertex AI or other Google Cloud products are covered by the Google Cloud Vulnerability Rewards Program, and are out of scope for this AI VRP.

Motivation

- Researchers need to gain direction and actionable guidance for AI-specific risks



Microsoft Vulnerability Severity and Content Classifications for AI Systems

Microsoft is committed to earning and maintaining the trust of our customers in how we develop and deploy Artificial Intelligence (AI) systems. This includes safeguarding customers from vulnerabilities in our software, services, and devices, and managing the risks of producing harmful AI-generated content, including potential exploitation of those vulnerabilities.

AI Security Vulnerabilities

Safeguarding our customers from vulnerabilities in our AI systems involves providing timely security updates and guidance when such vulnerabilities are reported to Microsoft. The following tables outline Microsoft severity classification for common vulnerability types for AI systems, based on the Microsoft Security Response Center (MSRC) advisory rating. MSRC uses this information as guidelines to triage reported issues and determine severity levels, with consideration also given to the ease of exploitation.

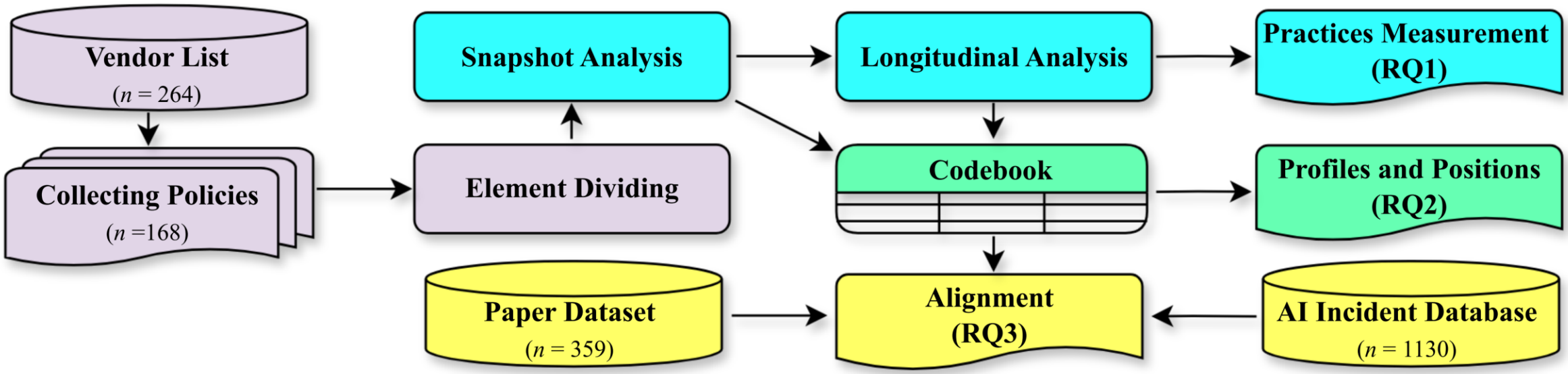
Research Questions

1. What is the state of vulnerability disclosure in the AI industry, and how has it evolved over time?

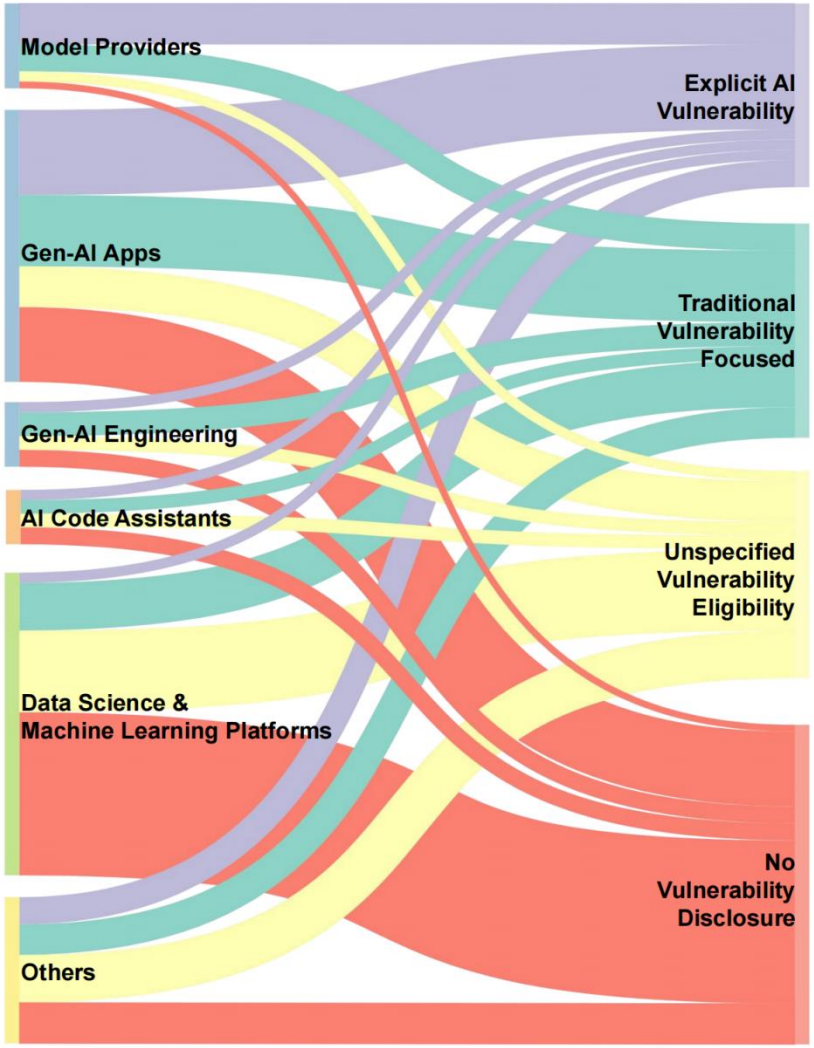
2 How do vendors approach AI vulnerabilities?

3. What is the alignment with AI incidents and research?

Methodology



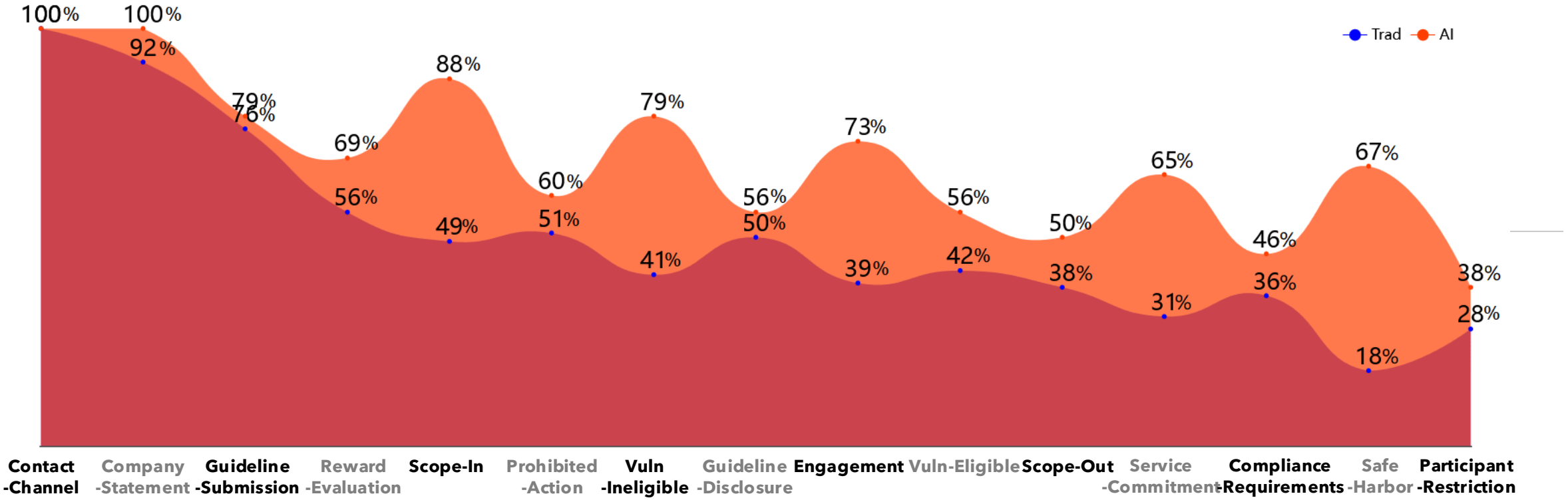
RQ1: Reporting Practices Measurement



- 36% of AI vendors have no public disclosure channel; only 18% explicitly address AI
- Model providers lead: ~48% of them mention AI vulnerablities
- 40% operate BBPs, but many AI firms still rely on minimal contact methods (8%)

RQ1: Coverage of policy elements

- Nearly 40% of AI firms include fewer than 6 policy elements; only 6% cover all
- AI-mentioned policies are significantly more complete: 88% vs. 49%

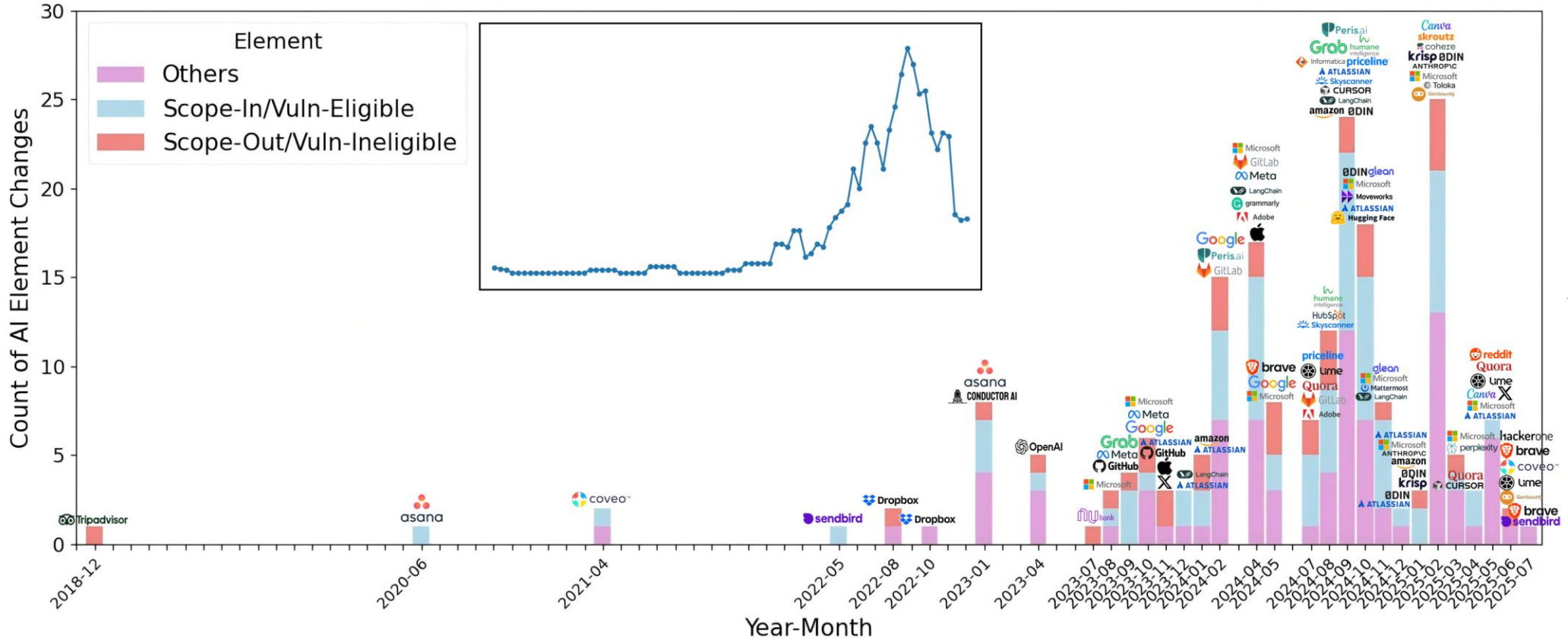


RQ1: AI Vulnerability types with eligibility

- Eligibility varies: prompt injection (71%), model extraction (90%), adversarial examples (80%)
- Rarely accepted: jailbreaking (27%), harmful outputs (45%), hallucination (17%)
- Common rejection reasons: expected model behavior, low impact, non-reproducibility, or unfixable issues

RQ1: Policy Evolution

- Updates surged after 2023: >20 companies revised in 2024
- Scope definitions and eligibility changes dominate (57% of all updates)



RQ2: Profiles and Positions

Proactive Vendors

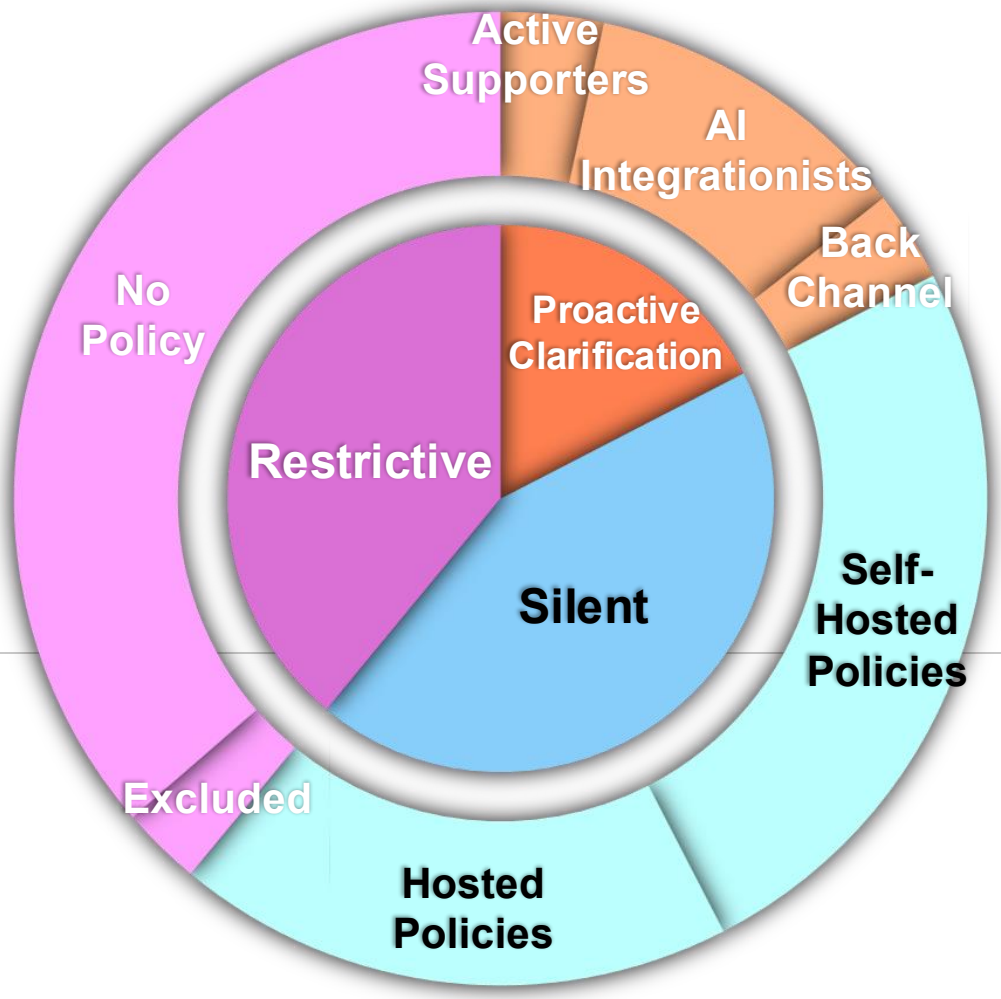
Vendors explicitly define AI vulnerabilities as in scope and welcome reports

Silent Vendors

Policies make no mention of AI vulns, leaving ambiguity for researchers

Restrictive Vendors

Vendor policies exclude AI vulns from scope or provide no reporting channel



RQ2: Proactive Clarification

Active Supporter

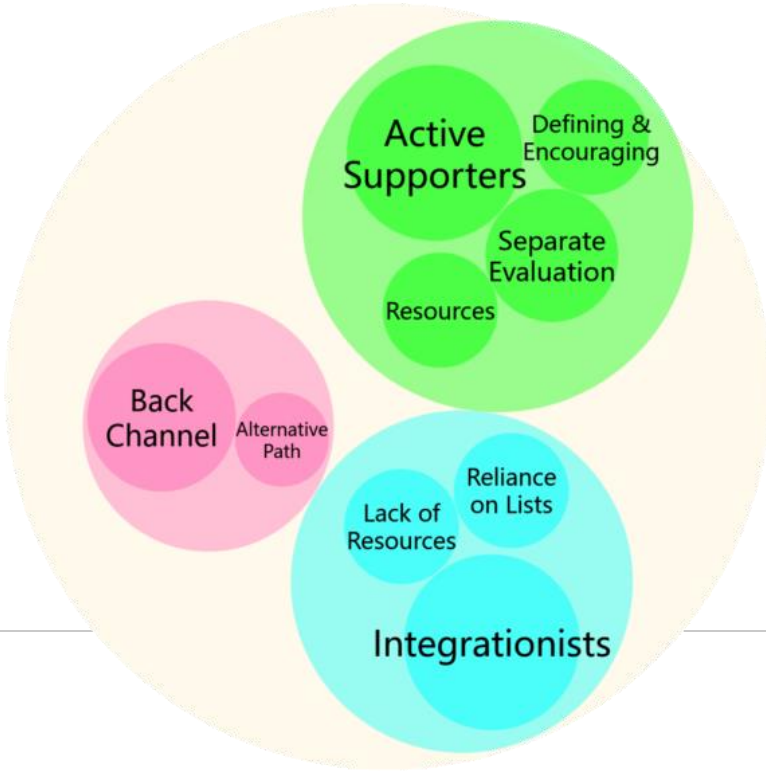
- Provide clear scope, AI-specific severity frameworks
- Often large model vendors

AI Integrationist

- Fold AI vulnerabilities into existing policies without special treatment
- Rely on checklists, little explanation or resources

Back-channel

- Do not treat AI issues as vulnerabilities, but provide alternative channels
- No rewards offered



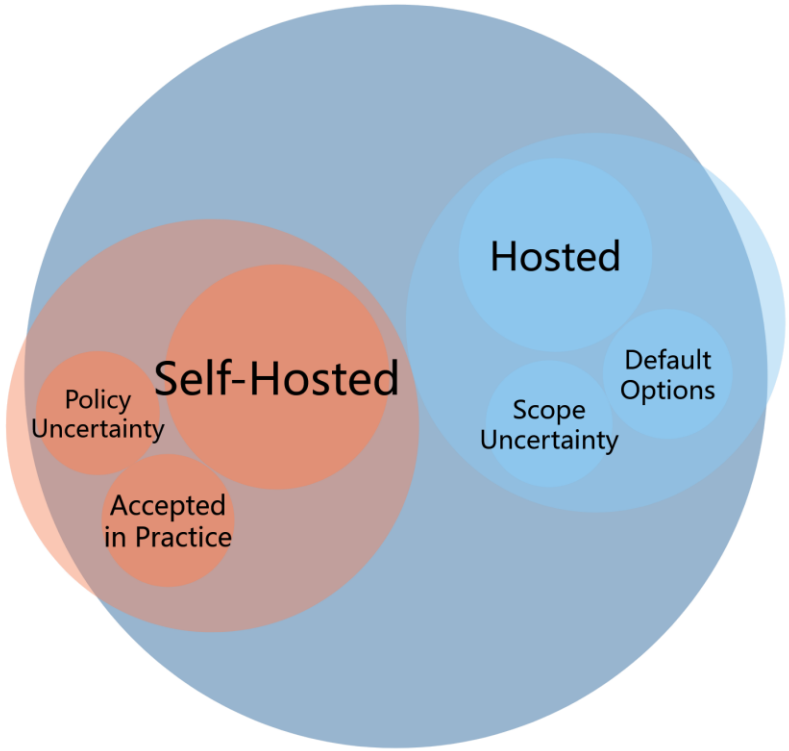
RQ2: Silent Vendors

Self-Hosted

- Policies contain no AI-related content or alternative channels
- Common in Gen-AI App and ML platform vendors
- Some vendors still patch AI issues in practice

Hosted

- Policies not updated, but platforms added AI submission fields
- Makes reporting possible, but scope remains unclear



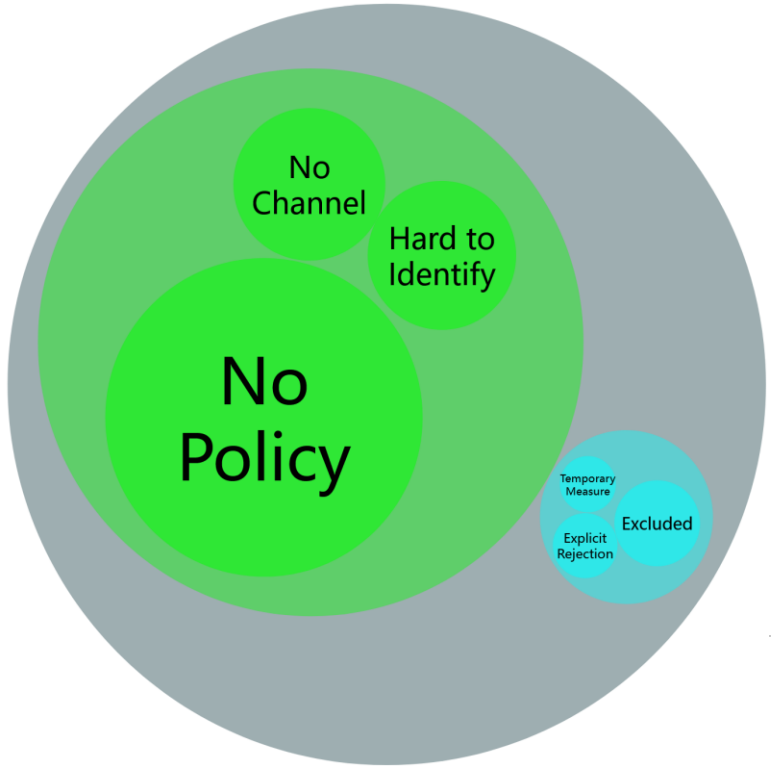
RQ2: Restrictive

No Policy

- No disclosure channel, bug bounty, or even contact email

Excluded

- Explicitly exclude AI systems from scope
- Sometimes framed as temporary



RQ3: AI Incident & Research Alignment

Academic Research

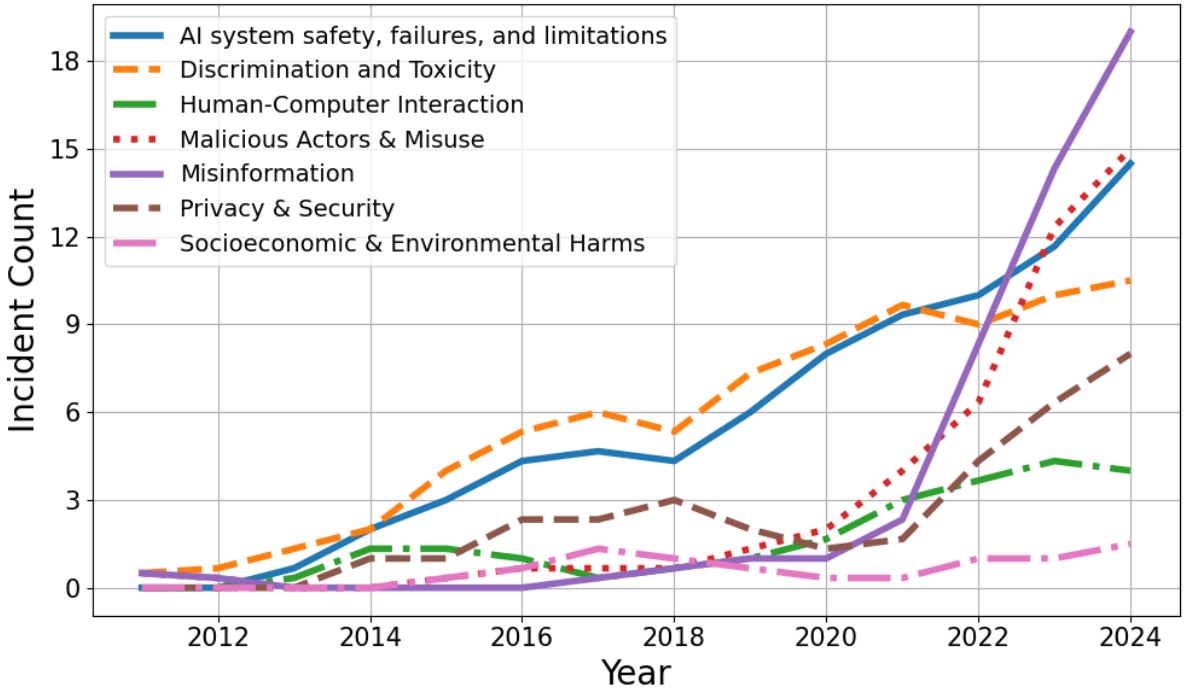
- Built a six-category meta-taxonomy
- Verified coverage with AI security papers from big 4 security conferences

Real-world Incident

- Used AI Incident Database (AIID), a public repository of AI failures from journalists, news and users
- Adopted MIT AI Risk Taxonomy to structure risks, mapped incidents to vulnerability types

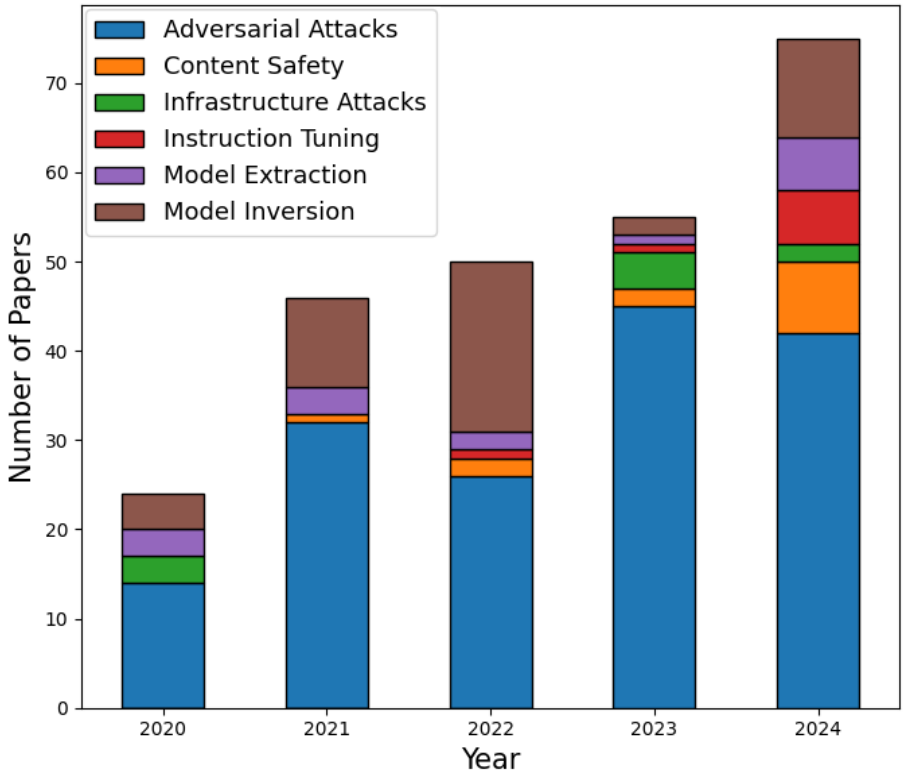
Dataset	Gartner List (n = 264)	AI Incident Database (n = 1,130)	Literature Review (n = 146)	AI Security Papers Dataset (n = 359)
Extracted Data	Disclosure Policies (n = 168)	Events (n = 320)	Taxonomies (n = 13)	Topics (n = 260)
Output	Thematic Codebook	Source Categories	Meta-taxonomy of AI Security & Safety	

RQ3: Timing



- Academia led first: AI vulnerabilities studied years before vendors acknowledged them
- Real-world incidents next: High-profile AI failures highlighted risks before policies adapted
- Vendors lagged: First policy mentions only in 2018, with slow and uneven adoption until 2023

RQ3: Types of Vulnerabilities



- AI System (21%): system integrity, data access, supply chain, authentication → almost universally accepted
- AI Model (43%): prompt injection, model extraction, inference, data poisoning → usually accepted
- AI Features (36%): harmful/insecure outputs, hallucinations → inconsistent, often treated as safety rather than security problems

Takeaway

- 36% of AI vendors lack any disclosure channel; only 18% explicitly mention AI risks
- Three vendor profiles identified: Proactive, Silent, Restrictive
- AI system/model vulnerabilities are more often in scope; AI feature issues are mostly excluded
- Vendors are slower to adapt policies compared to academia and real-world incidents

Discussion questions

Are we applying an unrealistic safety standard?

Provocations:

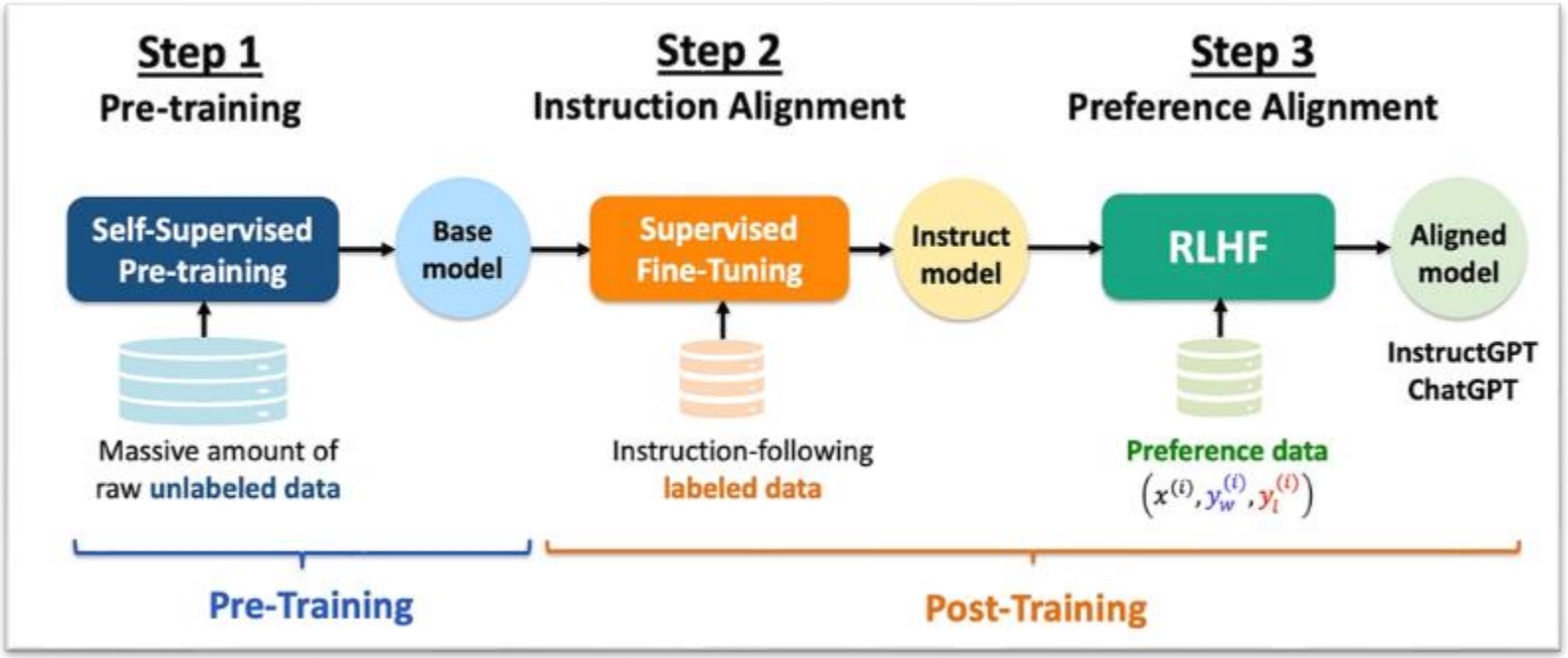
1. We don't expect knife manufacturers to prevent customers from violent usage.
2. We don't expect spell checkers to refuse to correct spelling in abusive content.

Should vendors have fixed the original sin?

1. **Pre-training (The Original Sin)**
   Minimal safety alignment. Ingest huge volumes of data, containing unsafe content.

2. **Supervised fine tuning (SFT)**
   Teach how to say "no" to inappropriate questions

3. **Reinforcement Learning from Human Feedback (RLHF)**
   Refine via humans ranking responses, partly based on safety.



Source: https://youssefh.substack.com/p/visual-guide-to-llm-preference-tuning

Is "we created best practice" a valid defence?

- Anthropic made considerable investments in AI Safety:
  - Lots of "sophisticated" thinking around AI Safety
  - Expansive AI bug bounty program
  - Lead multiple AI risk indexes
  - Created a freer AI safety research lab (see Carlini statement)

But was it enough?