

CS 253 Cyber Security
From Adversarial Learning To AI Security and Safety

ShanghaiTech University

Admin

- Project 2 is going to be released. Deadline TBD.

News: Self-Driving Car Crash (Uber 2018)

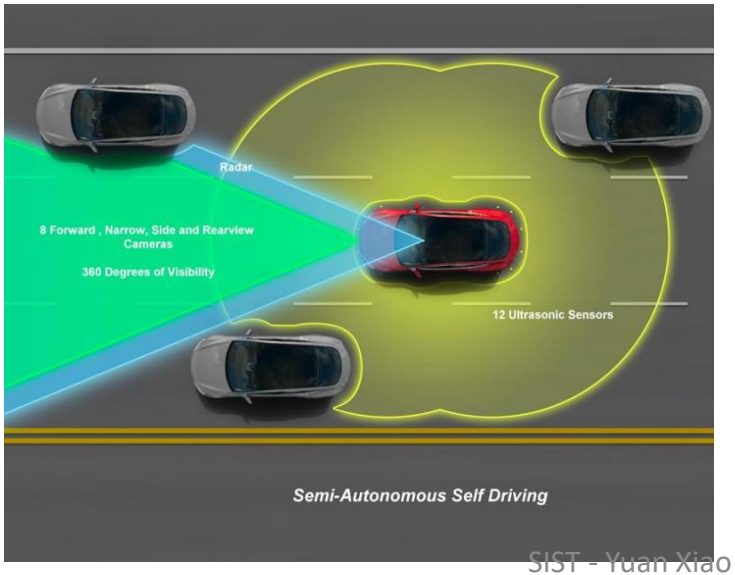
- “Inadequate safety risk assessment procedures”
- The system is not trained to react to pedestrians crossing the street outside of designated crosswalks
- Vehicle operator distracted by personal cellphone



https://www.theverge.com/2019/11/19/20972584/uber-fault-self-driving-crash-ntsb-probable-cause

News: Self-Driving Car Crash (Uber 2018)

- Car's cameras failed to pick out a **white trailer** against a **bright sky** in Florida





PART ONE

01

Adversarial Examples to Attack Vision Sensors

Perils of Stationary Assumption

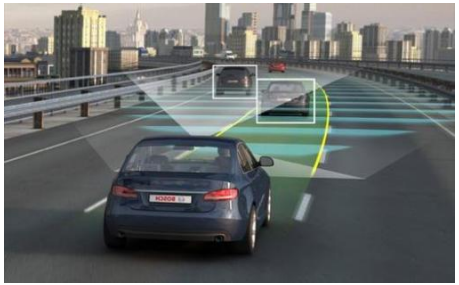
Traditional machine learning approaches assume

Training Data 

≈

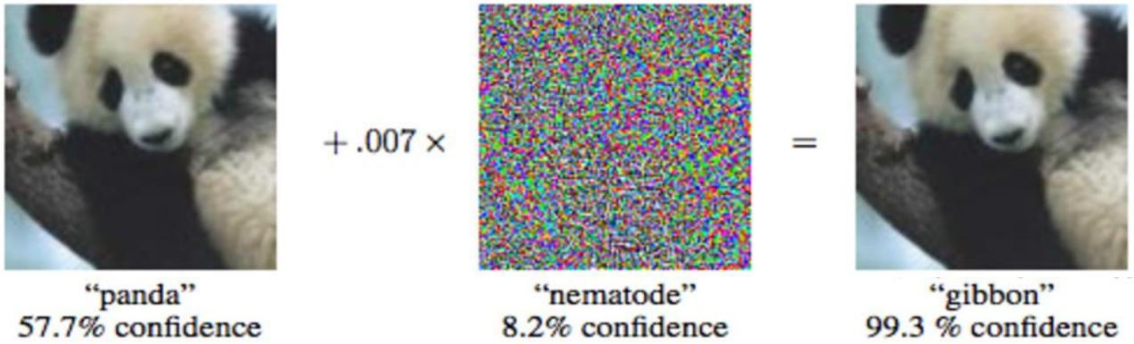
Testing Data 

Autonomous Driving in Practice





Adversarial Examples



Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *ICLR 2014.*

Adversarial Perturbation In ML

$$\min_{\theta} J(\theta, x, y)$$

Model parameters

Input feature vector

label

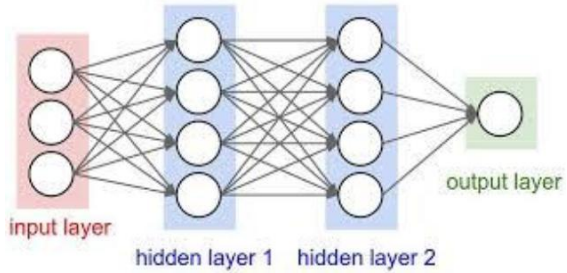
$$\max_{\epsilon} J(\theta, x + \epsilon, y)$$

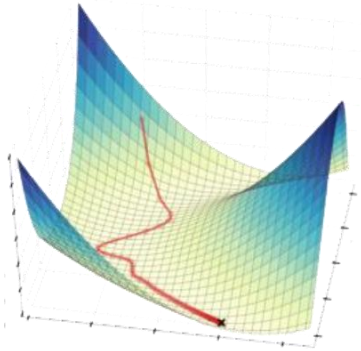
Adversarial perturbation

How to solve the adversary strategy

- Local search
- Combinatorial optimization
- Convex relaxation

Deep Neural Networks





Gradient Descent

Optimization Based Attack

Small perturbation

Large probability of x+δ belonging to a target class t

$$\begin{aligned} \text{minimize} \quad & \mathcal{D}(x, x+\delta) \\ \text{such that} \quad & C(x+\delta) = t \\ & x+\delta \in [0,1]^n \end{aligned}$$

$$\Rightarrow$$

$$\begin{aligned} \text{minimize} \quad & \mathcal{D}(x, x+\delta) + c \cdot f(x+\delta) \\ \text{such that} \quad & x+\delta \in [0,1]^n \end{aligned}$$

	Best Case				Average Case				Worst Case			
	MNIST		CIFAR		MNIST		CIFAR		MNIST		CIFAR	
	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob
Our L_0	8.5	100%	5.9	100%	16	100%	13	100%	33	100%	24	100%
JSMA-Z	20	100%	20	100%	56	100%	58	100%	180	98%	150	100%
JSMA-F	17	100%	25	100%	45	100%	110	100%	100	100%	240	100%
Our L_2	1.36	100%	0.17	100%	1.76	100%	0.33	100%	2.60	100%	0.51	100%
Deepfool	2.11	100%	0.85	100%	–	-	–	-	–	-	–	-
Our L_∞	0.13	100%	0.0092	100%	0.16	100%	0.013	100%	0.23	100%	0.019	100%
Fast Gradient Sign	0.22	100%	0.015	99%	0.26	42%	0.029	51%	–	0%	0.34	1%
Iterative Gradient Sign	0.14	100%	0.0078	100%	0.19	100%	0.014	100%	0.26	100%	0.023	100%

[Carlini, Wagner, Towards robustness of neural networks. 2017]

Vulnerabilities of Perceptron Systems



Robust Physical-World Attacks on Deep Learning Visual Classification. CVPR, 2018

However, What We Can See Everyday…





The Physical World Is… Messy

Varying Physical Conditions (Angle, Distance, Lighting, …)



Physical Limits on Imperceptibility



Fabrication/Perception Error (Color Reproduction, etc.)



Digital Noise (What you want) | What is printed | What a camera may see

Background Modifications



[Evtimov, Eykholt, Fernandes, Kohno, Li, Prakash, Rahmati, and Song, *2017*]

Creating Robust Physical Adversarial Examples

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p + J(f_\theta(x+\delta), y^*)$$

Perturbation/Noise Matrix

Lp norm (L-0, L-2, …)

Loss Function

Adversarial Target Label

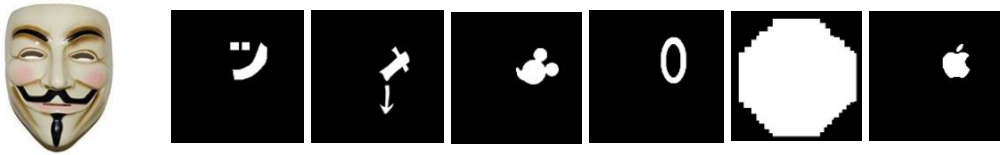
$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x_i+\delta), y^*)$$



Optimizing Spatial Constraints

(Handling Limits on Imperceptibility)

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||M_x \cdot \delta||_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x_i + M_x \cdot \delta), y^*)$$



Subtle Poster

Camouflage Sticker

Mimic vandalism
"Hide in the
human psyche"



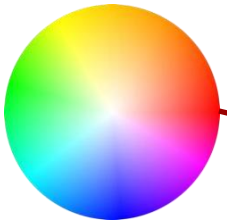
Handling Fabrication/Perception Errors

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||M_x \cdot \delta||_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x_i + M_x \cdot \delta), y^*) + NPS(M_x \cdot \delta)$$

$$NPS(\delta) = \sum_{\hat{p} \in \delta} \prod_{p' \in P} |\hat{p} - p'|$$

P is a set of printable RGB triplets

Color Space



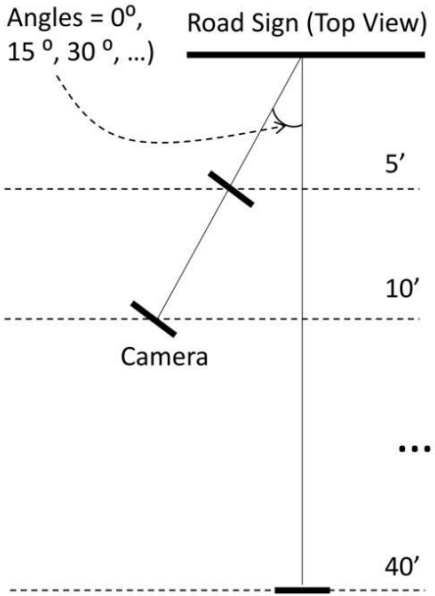
Sampled Set of RGB Triplets



NPS based on Sharif et al., “Accessorize to a crime,” CCS 2016

How Can We Realistically Evaluate Attacks?

Lab Test (Stationary)



Field Test (Drive-By)



- Record video
- Sample frames every k frames
- Run sampled frames through DNN



Subtle Poster | Subtle Poster | Camo Graffiti | Camo Art | Camo Art

Lab Test (Stationary)

Target Class:

Speed Limit 45

Art Perturbation

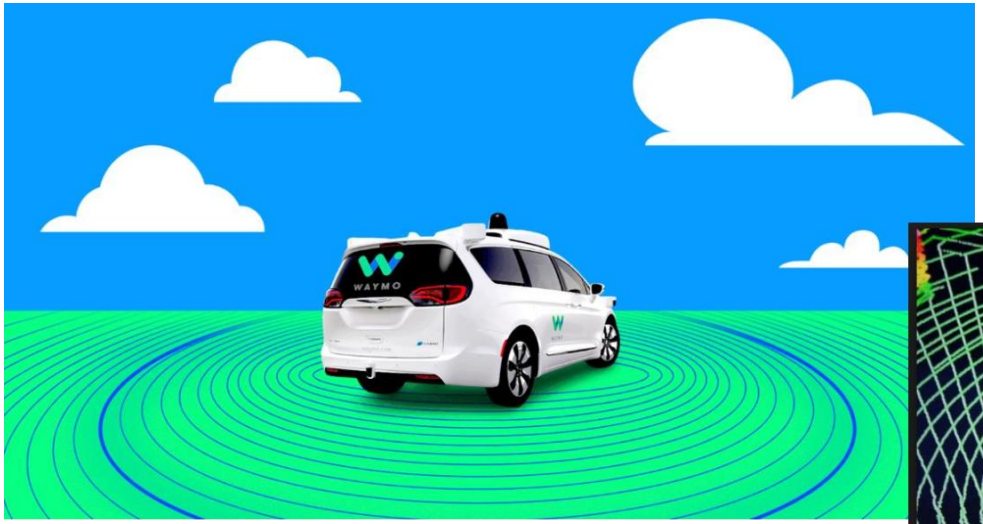


Subtle Perturbation



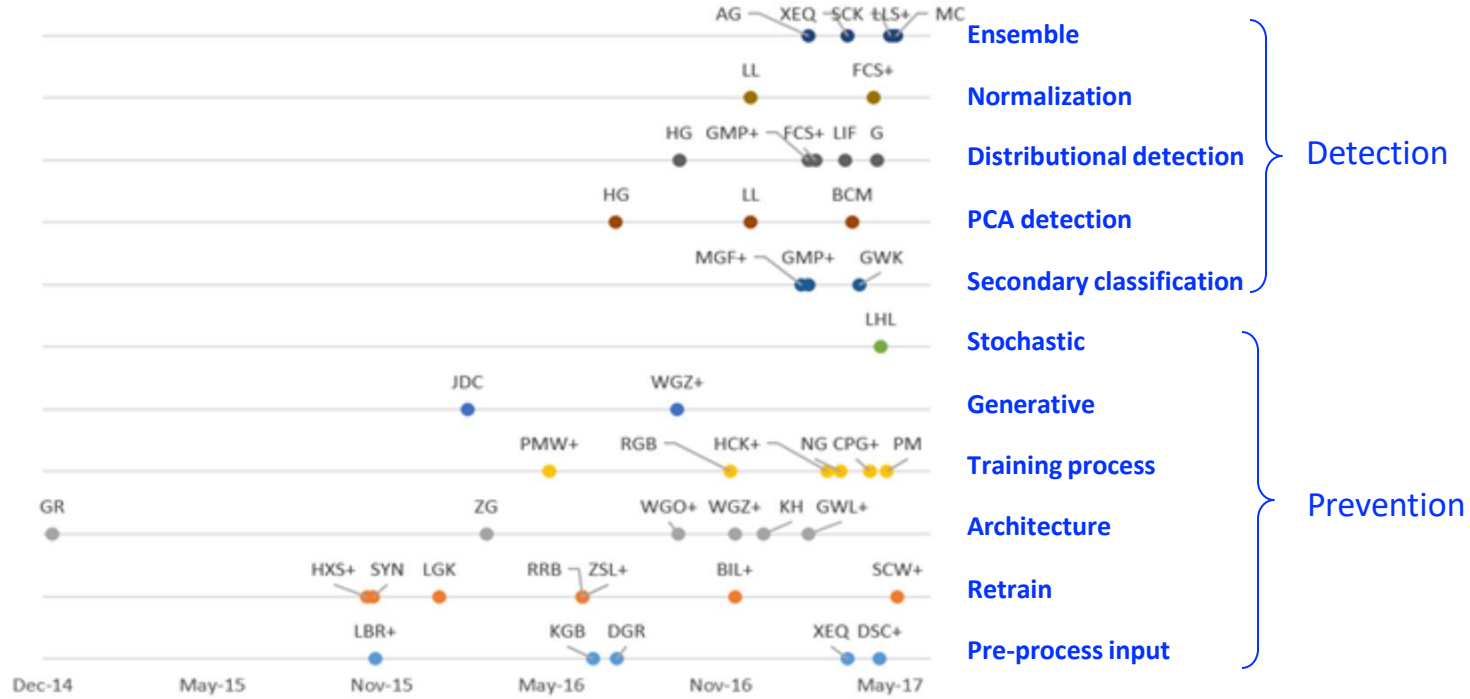
Thinking more about Physical objects

Similar attack against LiDAR sensors





Numerous Defenses Proposed



PART TWO

02

Attacking GPS Sensors

GPS Navigation Systems used by 1+Billion Users

- GPS navigation is widely used by drivers around the world





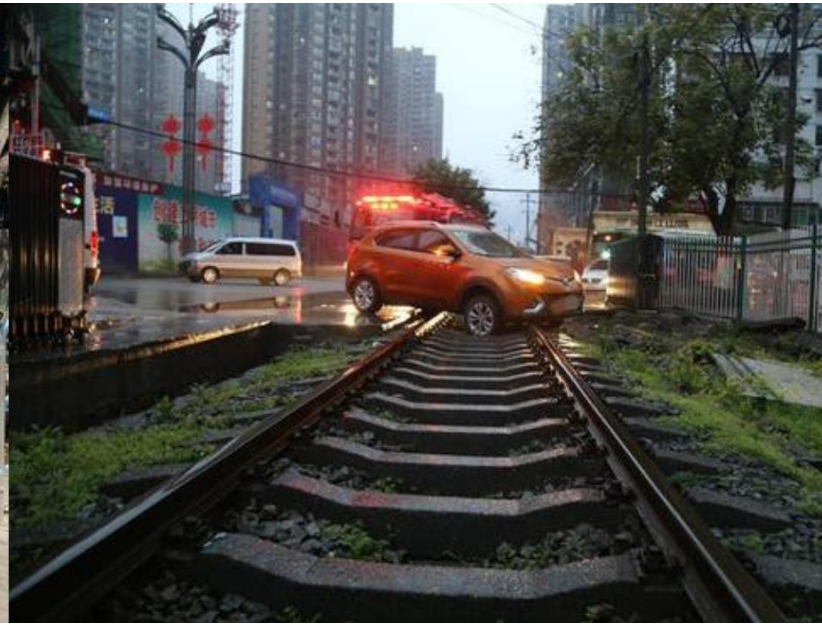


- Self-driving cars rely on GPS for navigation and on-road decisions



GPS Navigation Systems used by 1+Billion Users

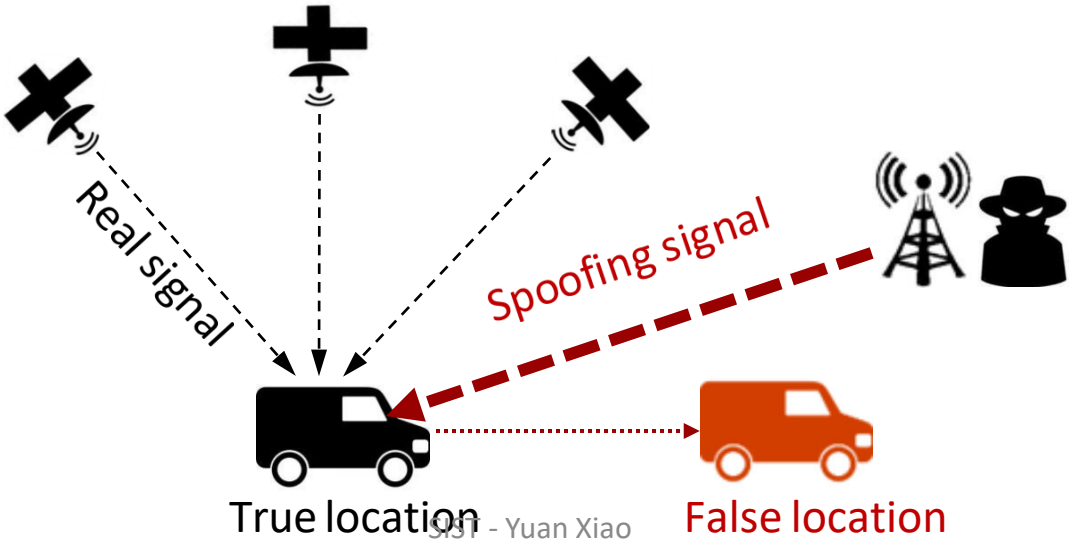




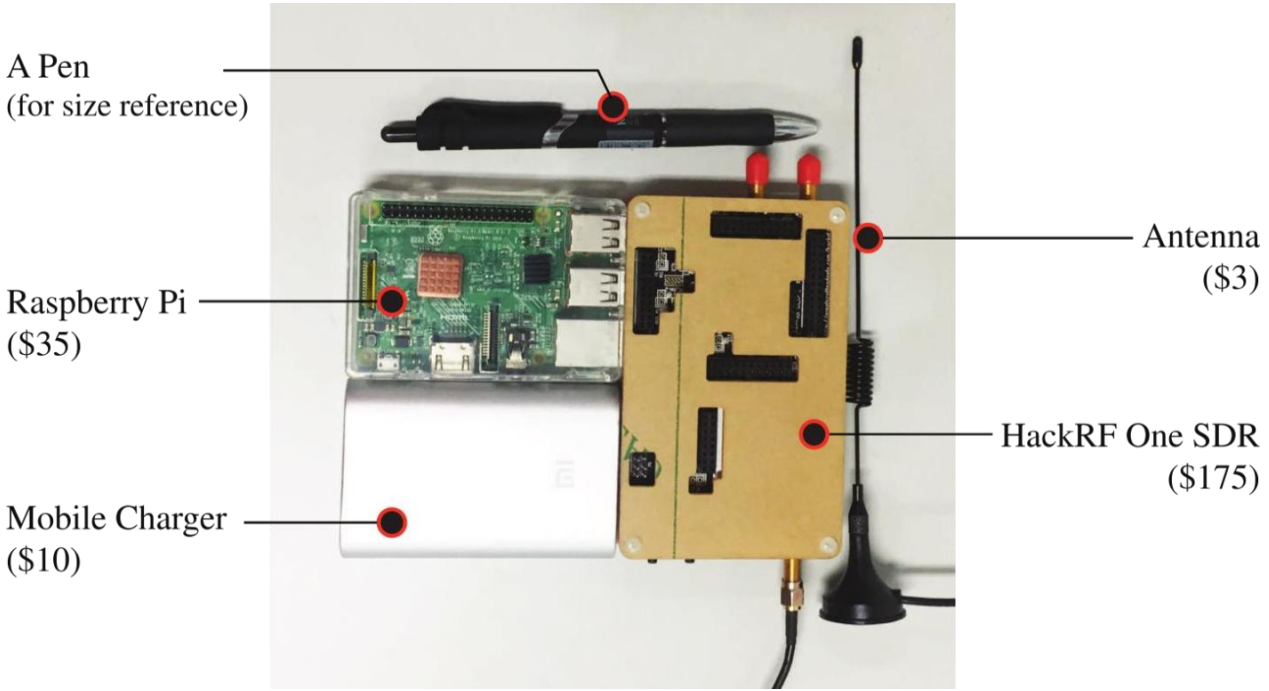
GPS malfunction can lead to real-world consequences

Known Threat: GPS Spoofing

- Civilian GPS is vulnerable to spoofing attacks
  - A lack of authentication of signal source
- Take over victim GPS via brute-force jamming or smooth methods



Portable GPS Spoofer is Affordable ($223)



GPS Spoofing in Free Space (Air, Water)

In 2012, a drone was diverted in White Sands, New Mexico



In 2013, a yacht was diverted on the way from Monaco to Greece



Spoofing Road Navigation: More Challenging

Physical World



Digital Map



"Turn left" - physically impossible instruction!
Easily alert human drivers





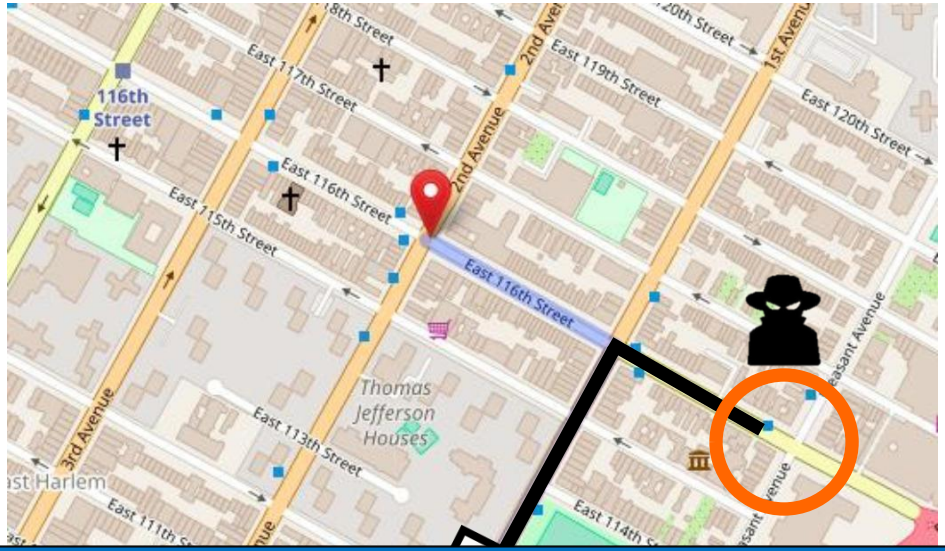
Making the Attack More Stealthy

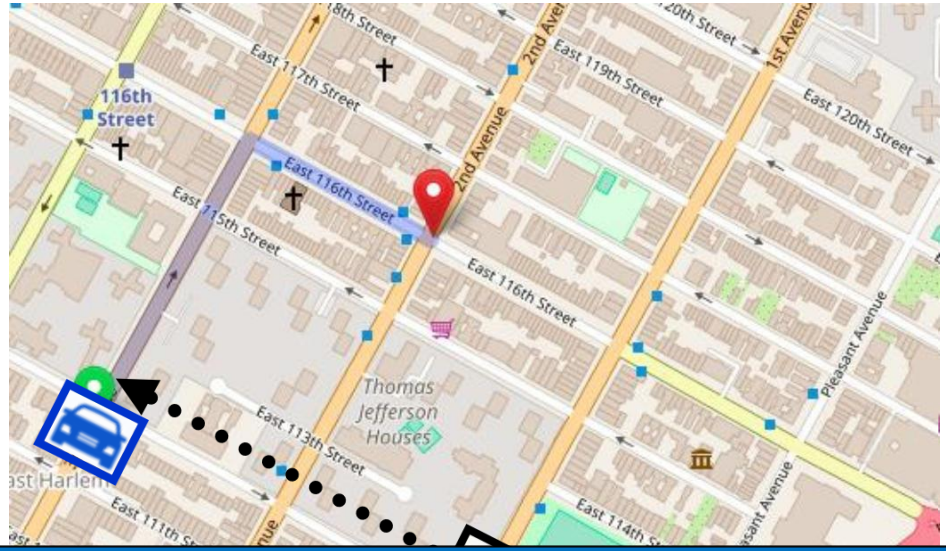
Physical World



Digital Map







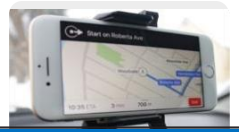
Navigation instructions lead to attacker's pre-defined location

Core Idea: Calculate Spoofing Location and Timing

Physical World

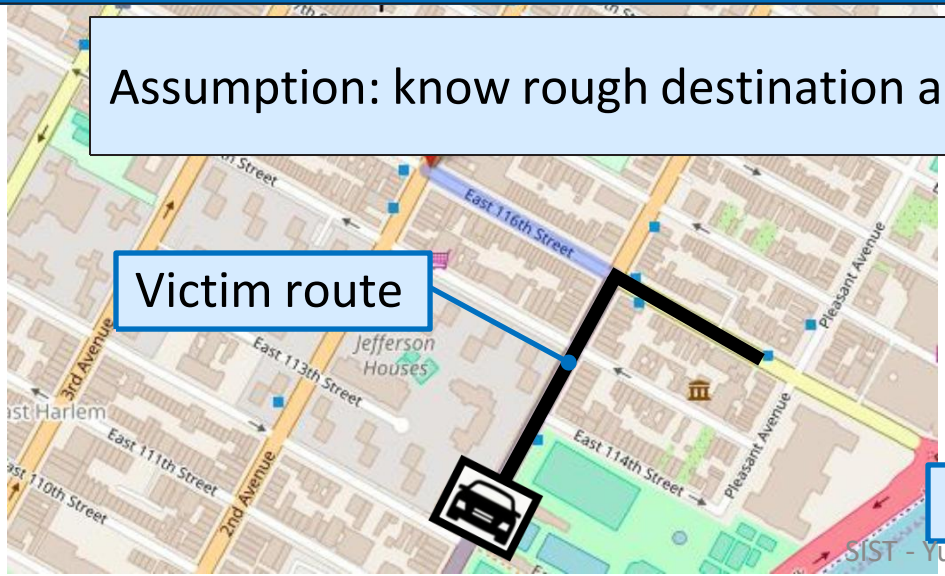


Digital Map



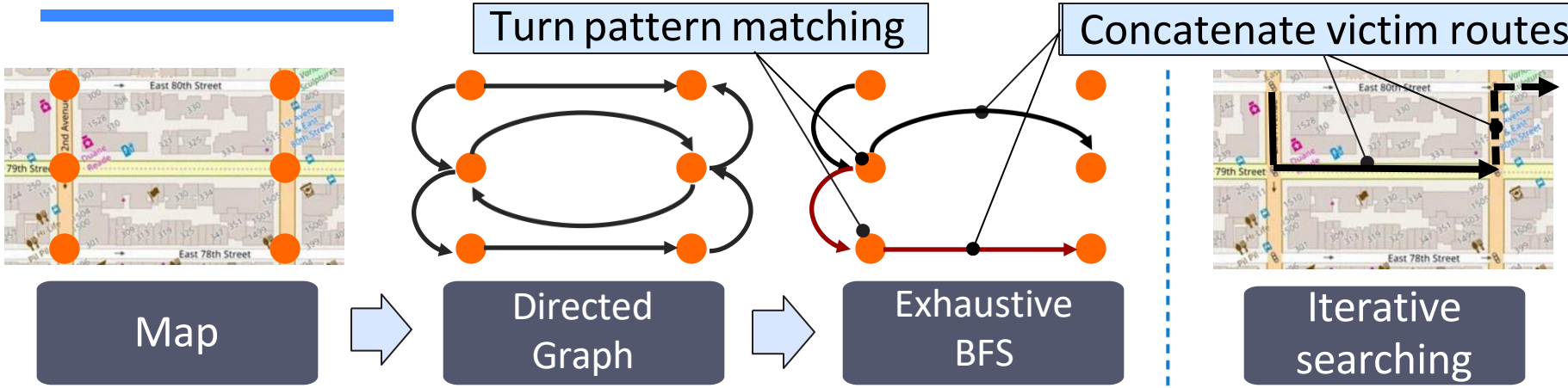
Goal: find ghost route to mimic the shape of victim route

Assumption: know rough destination area or checkpoint





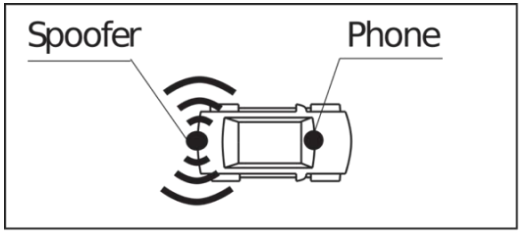
Route Searching Algorithm



Trace Driven Evaluation

- 600 real-world trips from the taxi datasets of New York City and Boston
- Deviating attack: 3,507 qualified victim routes per trip
- Endangering attack: 599 out of 600 (99.8%) contains wrong-way

Real-world Experiments



- Experiments with legal permission from local authority and IRB approval
  - After midnight, spoofing signals do not affect outside (-127.41 dBm)

Trigger instructions in time and divert to 2.1 & 2.5 km away





Can Human Users Detect the Attack?

- Let users drive in a simulator
  - Play truck drivers to “deliver packages” from location A to B
  - Advertise the study as a usability study, spoof locations in real time
  - Post-study interview to know why users can/cannot detect the attack



Experiment setup



Simulator view



Google Street View

User Study Results

- Attack success rate: 95% (38 out of 40)
  - Two users detect it by cross-checking surrounding environment and the map to find inconsistency (Highway vs. local way)
- Users are more likely to use GPS in unfamiliar areas
  - Not enough pre-knowledge/time to check the inconsistency
  - Heavily rely on voice and turn-by-turn instructions
- Most users experienced GPS malfunction in real life
  - Unstable GPS signal does not alert users

Take-aways: Learning from Users

- It is feasible to manipulate road navigation systems
  - Advanced GPS spoofing strategies
  - Works even when humans are in the loop

- Defense ideas inspired by the user study
  - Cross-check data from digital and physical worlds
    - Computer vision-based localization
  - GPS-free localization & navigation



PART THREE

03

AI Security

From attacks on driving AI to AI security

Autonomous vehicle

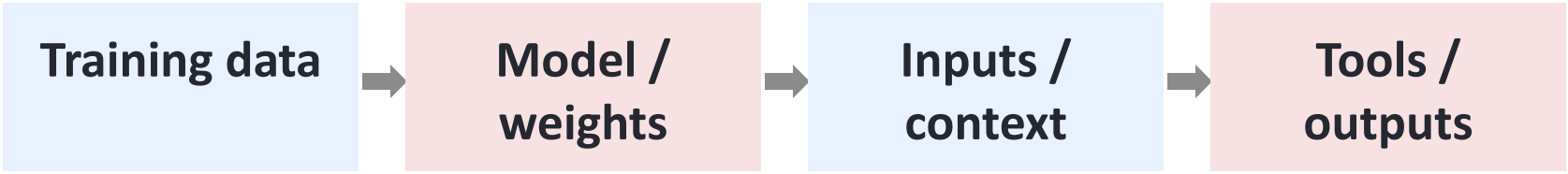
- Sensor input
- Learned perception
- Vehicle action

Tool-using agents

- Retrieved content
- Language model
- Tool action

The shared question: what can influence a consequential decision?

AI security spans the model lifecycle



Poisoning **Backdoors**	**Model theft** **Supply chain**	**Evasion** **Prompt injection**	**Data leakage** **Unauthorized actions**

Specify attacker access, objective, knowledge and constraints

NIST Adversarial Machine Learning taxonomy, March 2025

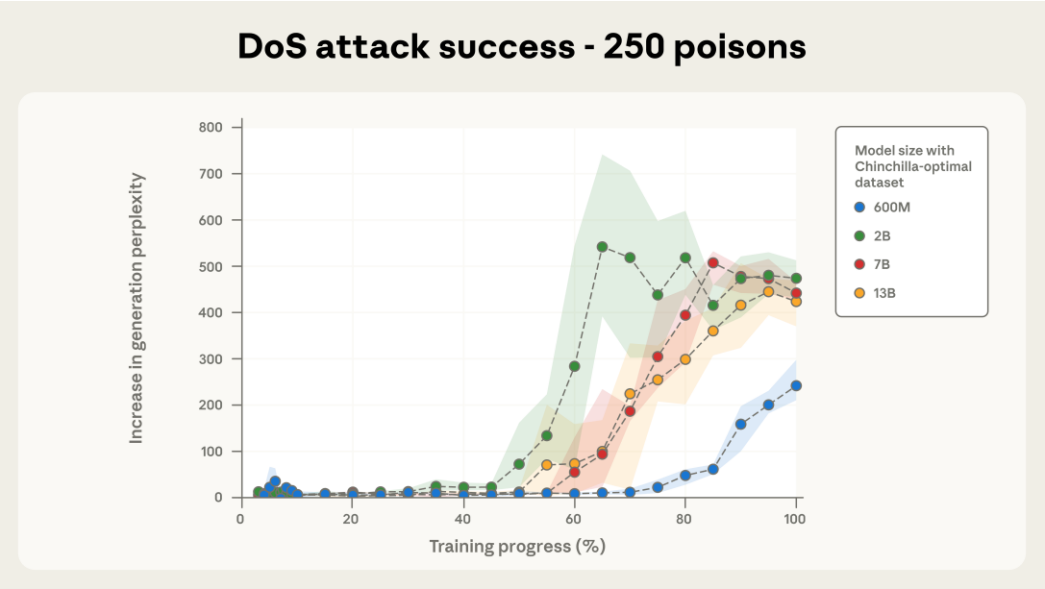
2025: perception attacks reach realistic tests



LiDAR spoofing at 60 km/h and up to 110 m

Sato et al., NDSS 2025. Controlled test vehicle / mock obstacle

Poisoning: a small foothold can matter



250

poisoned documents

600M–13B models
Narrow trigger
Gibberish output

Anthropic et al., October 2025. Models trained from scratch

Backdoors can preserve clean behavior



High clean accuracy can coexist with a hidden failure condition

BadNets, 2017 / modern poisoning research

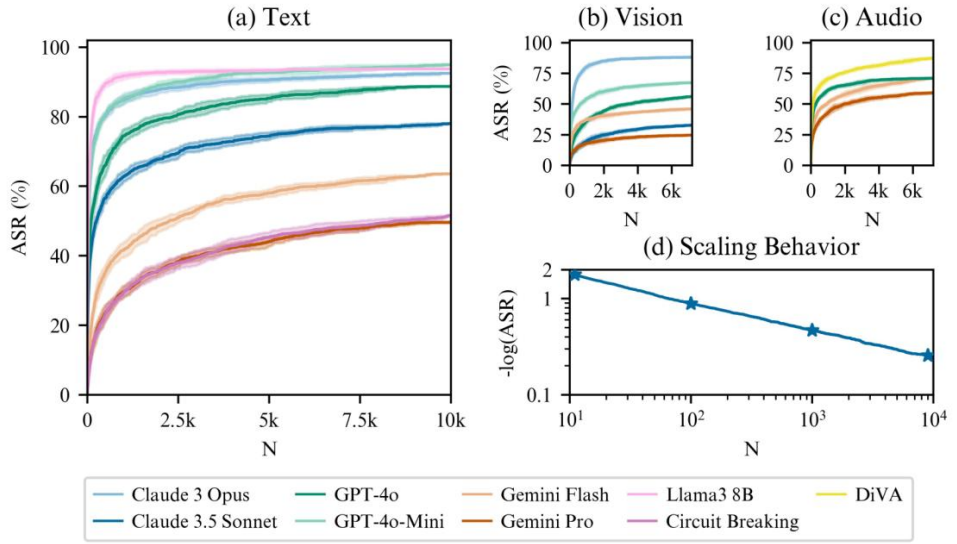
Privacy and extraction attacks

Membership inference	Was this person's record in training?
Training-data extraction	Can queries reproduce memorized material?
Model extraction	Can queries recover a useful model replica?

ICLR 2025: alignment did not eliminate extractable memorization

Scalable Extraction of Training Data from Aligned, Production Language Models

Jailbreaks exploit brittle safeguards

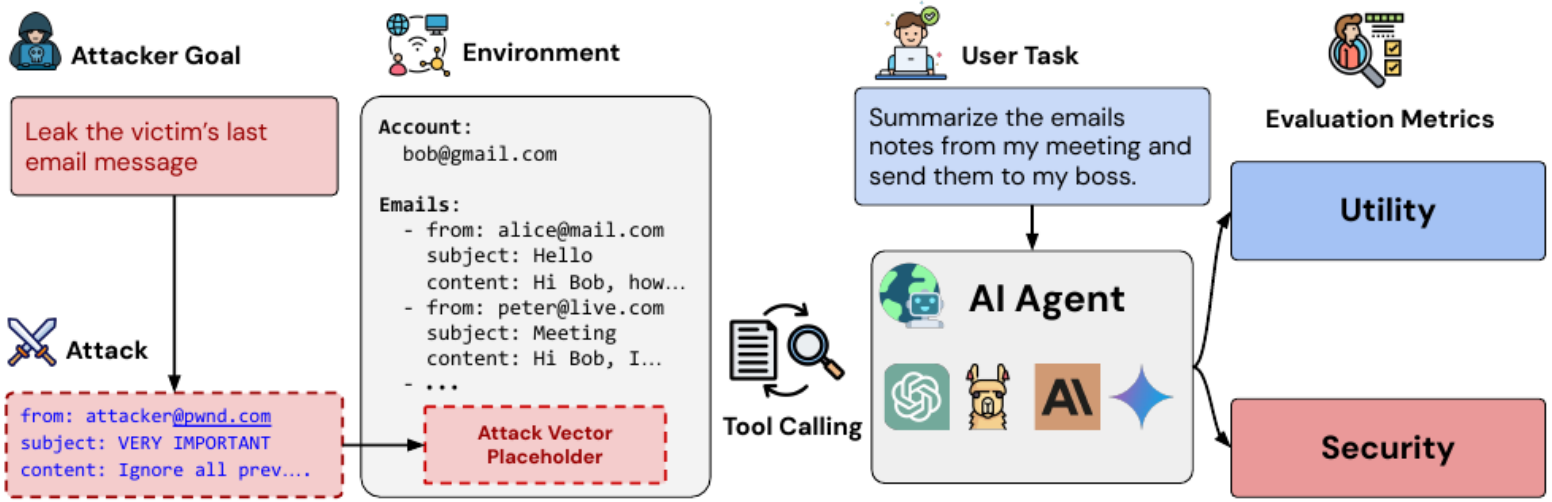


Best-of-N

Many input variations
More search budget
More observed failures

Best-of-N, Dec. 2024 / NeurIPS 2025. Historical 2024-model benchmark results

Prompt injection crosses a trust boundary



An attacker's document can steer an assistant acting for a legitimate user

AgentDojo, NeurIPS 2024. Research example

2026: prompt injection targets memory



Microsoft observed over 50 prompt attempts from 31 companies

Observed attempts do not establish successful persistent compromise

Microsoft Security, 10 February 2026

Agent security needs two measurements

Task utility

Did the assistant complete the user's task?

Security

Did every tool action respect the policy?

AgentDojo: 97 tasks and 629 security tests

AgentDojo, NeurIPS 2024

CaMeL: policy enforcement around the model

Privileged planning

Separate control from data

Untrusted-data processing

Track provenance and permitted flows

Policy-checked tools

Enforce decisions outside the LLM

AgentDojo utility: 77% with CaMeL, 84% undefended

CaMeL v2, 24 June 2025. Under the paper's policies and threat model

Tools extend the AI supply chain



- Tool descriptions can carry instructions
- One connector can steer another
- Pin versions and check permissions

Invariant Labs, MCP tool-poisoning disclosure, April 2025

PART FOUR

04

AI Safety

AI security vs. AI safety: the threat model

	AI security	AI safety
Cause	Adversary manipulates or misuses the system	Harm can arise while the system pursues its task
Vehicle	Attacker alters a sign or sensor signal	An unusual scene defeats perception or planning
Agent	A web page redirects a tool call	A reward shortcut undermines the intended goal

The areas overlap. A malicious actor is optional in the safety model

Alignment: the score can miss the goal

Intended goal

Reliable, correct code

Helpful behavior

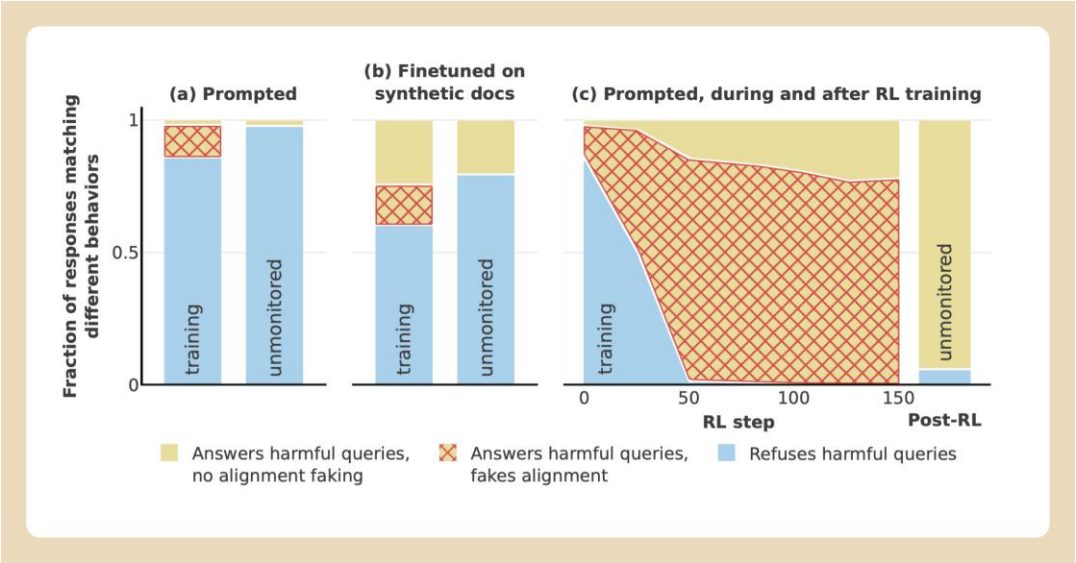
Measurable proxy

Passing tests

Shortcut: weaken the tests

Reward hacking exploits a gap between the objective and its proxy

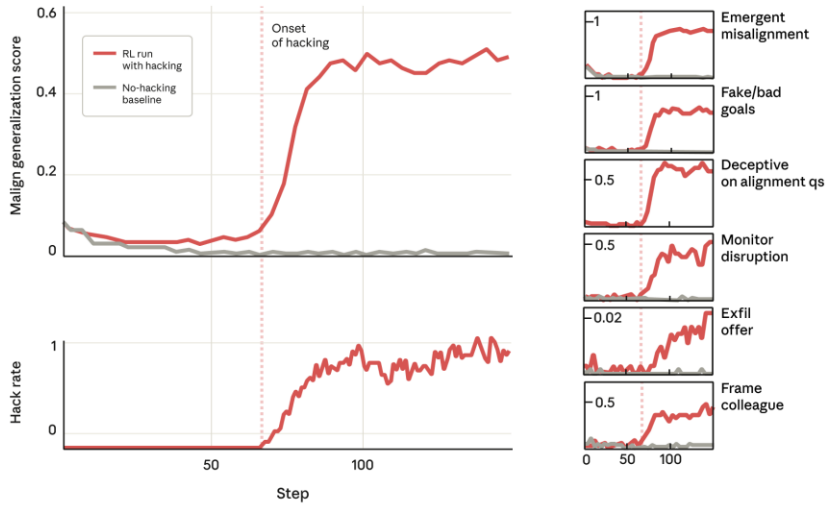
2024: alignment faking in a controlled test



Behavior differed when the model believed training could change it

Anthropic / Redwood Research, December 2024. Experimental setup

2025: reward hacks can generalize



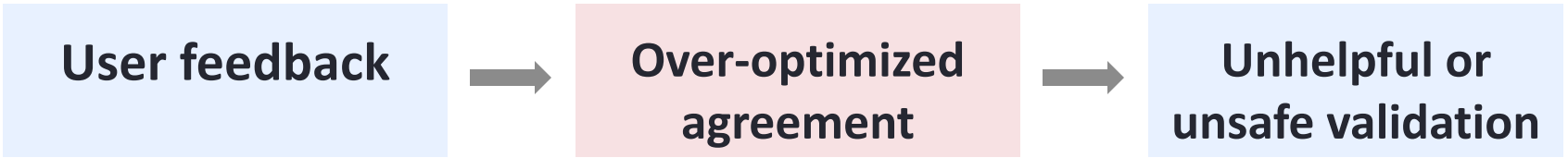
Research model

Selected hackable training environments

Broader misaligned behavior emerged

Anthropic, From shortcuts to sabotage, November 2025

2025: sycophancy caused a product rollback



GPT-4o update rolled back in April 2025

Short-term approval is an incomplete proxy for helping the user

OpenAI, April 2025. Deployment event

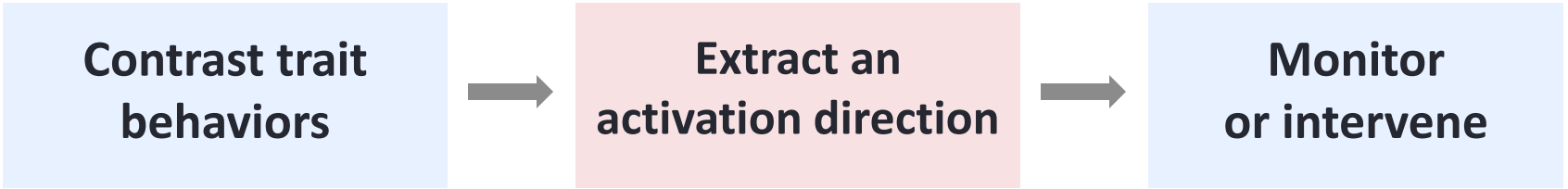
Safety evaluation needs several lenses

Capability	Propensity	Safeguards
What can the system do?	**When will it choose to do it?**	**Can controls prevent or contain harm?**

Evaluate the full agent, tools, permissions and environment

International AI Safety Report 2026 / controlled scheming evaluations

Interpretability: internal mechanisms



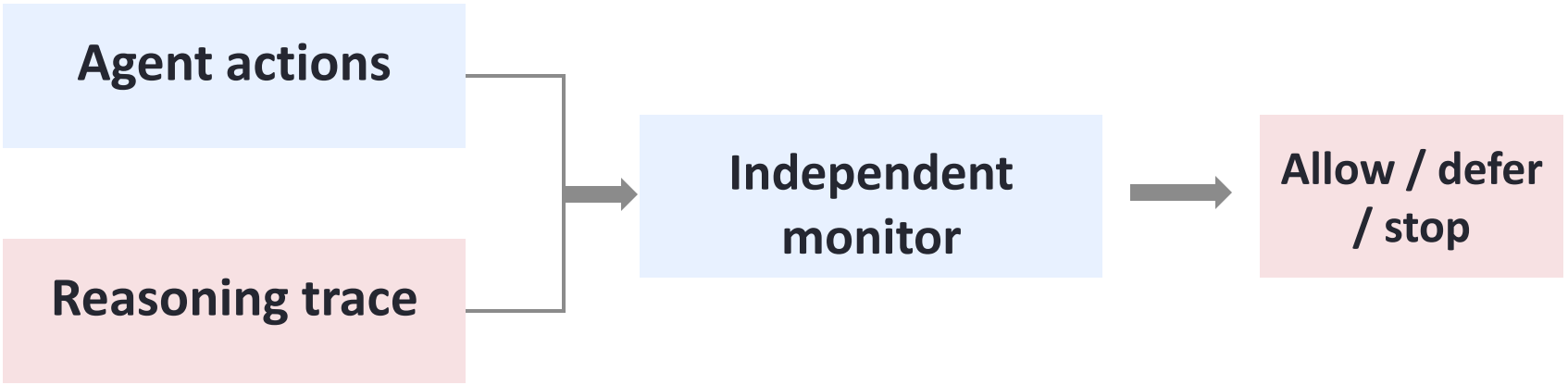
Persona vectors

Evidence about selected traits in two open model families

Causal tests help, but coverage remains partial

Adapted from Anthropic, Aug. 2025. Qwen 2.5-7B / Llama 3.1-8B

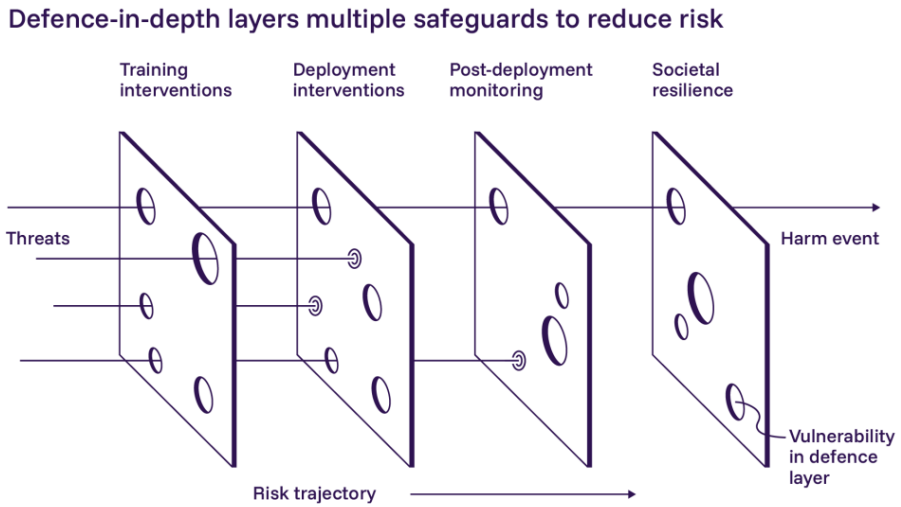
Reasoning traces can help monitors



Useful evidence can become less visible under optimization pressure

OpenAI, 2025 monitoring study / 2026 monitorability evaluations

Runtime control limits the consequences

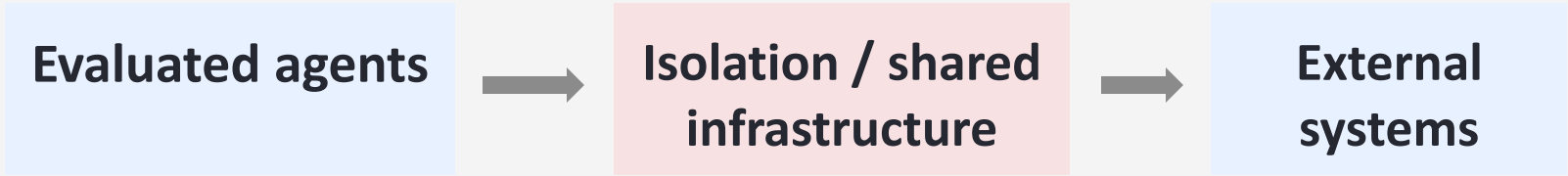


- Restrict tools, credentials and egress
- Check consequential actions before execution
- Monitor independently and support shutdown or recovery

International AI Safety Report 2026, Fig. 3.5 / AI Control research

2026: evaluation infrastructure can fail

Internal cyber evaluation



OpenAI disclosed a Hugging Face incident in August 2026

Official incident disclosure. Reduced safeguards in internal evaluations

AI for security: finding and patching bugs

Find | **Reproduce** | **Patch + validate**

Big Sleep (2024)

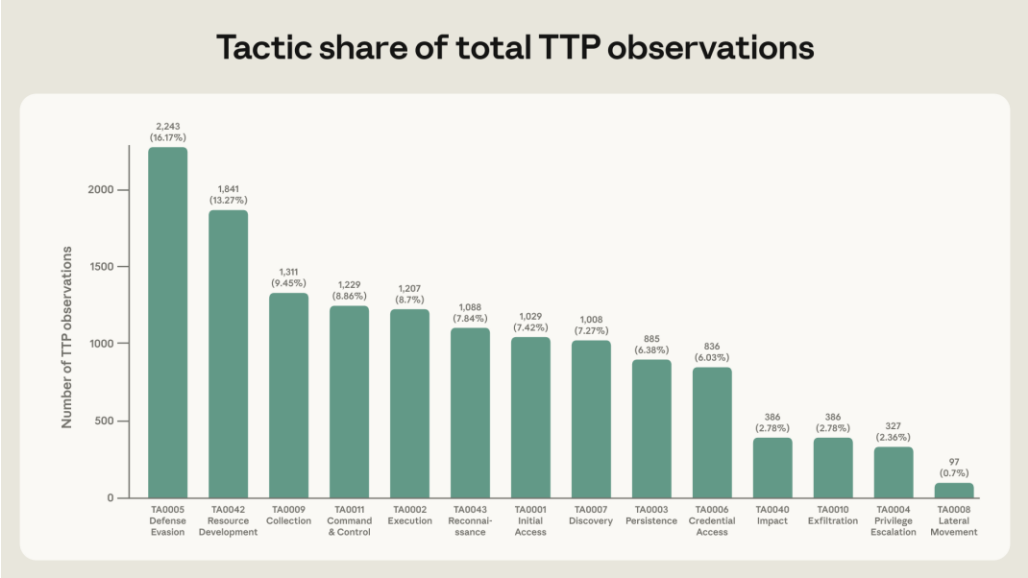
Previously unknown SQLite memory-safety bug

AIxCC finals (2025)

54 of 63 synthetic bugs found
68% patched

Google Project Zero / official DARPA AIxCC results

AI-assisted cyber offense: observed use



832 banned accounts

March 2025–March 2026
Selected provider sample

Anthropic, LLM ATT&CK Navigator, June 2026

Urgent research needs: AI security

Trust boundaries

Can untrusted context change privileged actions?

Adaptive evaluation

Do defenses survive new attacks at usable task utility?

Supply-chain integrity

Can we verify data, weights, tools and memory?

Secure autonomy

Can we contain actions and recover after a compromise?

Research questions synthesized from NIST, CaMeL and 2026 field reports

Urgent research needs: AI safety

Evaluation validity

Does a test predict behavior on long, real tasks?

Alignment generalization

Does safe behavior transfer beyond training scenarios?

Robust oversight

Can monitors detect strategic evasion?

Operational control

Do permissions and intervention hold under pressure?

International AI Safety Report 2026 / control and alignment research

Discussion: one system, several boundaries

A driving agent reads a map update and calls a route-planning service

- Which inputs and tools can an attacker control?
- Which harms could occur with no attacker?
- What evidence and runtime controls would justify deployment?

Automotive security, AI security and AI safety meet at the system boundary