AI red teaming tradecraft:
A team of teams approach

Thomas Brunner, Daniel Fabian, Sarah Hodkinson, Mikel Rodriguez

Intro & Origins

Mikel Rodriguez

- Leads DeepMind's ReBl (red/blue) Team
- Previously helped lead AI Red Teams for the DoD

Intro & Origins

Daniel Fabian

- Leading Security, Privacy, and ML Red Teams at Google
- Previously red teamer & pentester

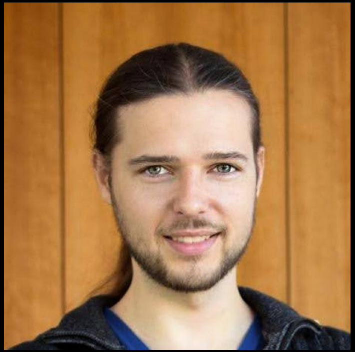
Intro & Origins

Sarah Hodkinson

- AI Safety & Security Program Manager leading red teams across multimodal research
- AI Ethics focused on misuse & abuse of systems

Intro & Origins

Thomas Brunner

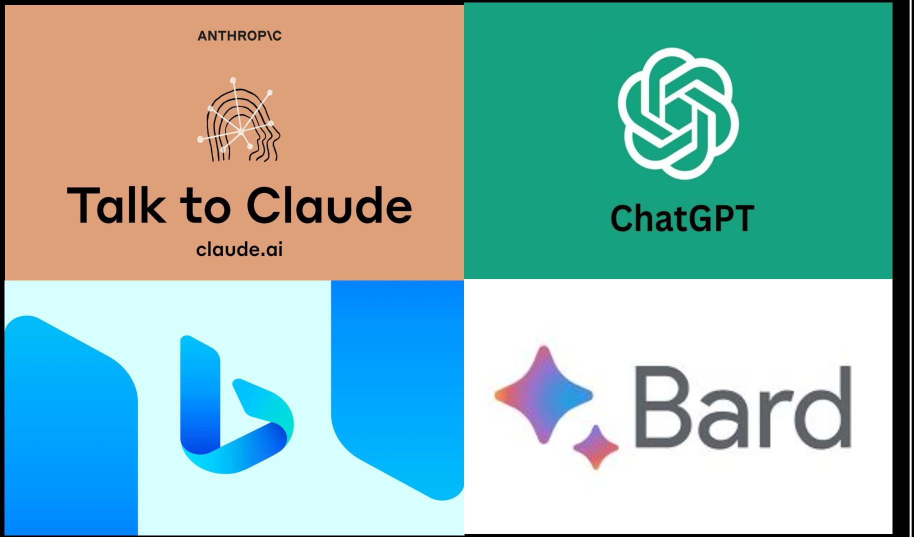


- AI Red Teamer @ Google
- Previously researched realistic black-box adversarial attacks on ML systems

A rapidly evolving landscape

Confluence of two things:

Widespread democratization of highly capable ML systems

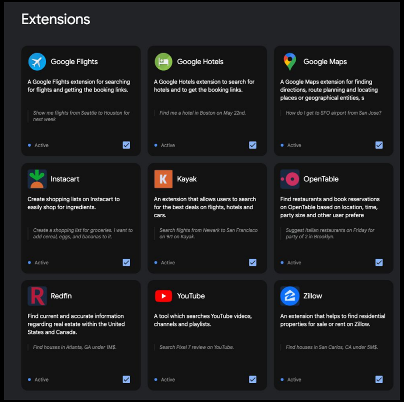


Introduction of new capabilities that couple AI with with a broader ecosystem that raise the stakes

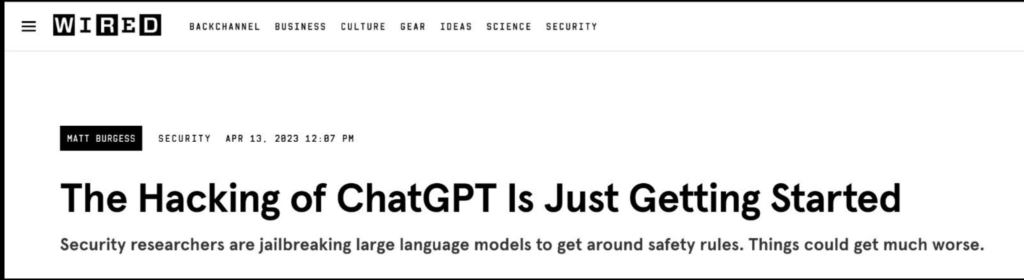
WILL KNIGHT KHARI JOHNSON BUSINESS MAR 28, 2023 7:00 AM

Now That ChatGPT Is Plugged In, Things Could Get Weird

Letting the chatbot interact with the live internet will make it more useful—and more problematic, too.



A rapidly evolving landscape

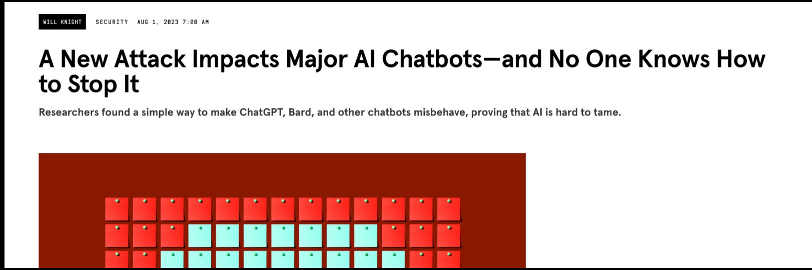


WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY

MATT BURGESS SECURITY APR 13, 2023 12:07 PM

The Hacking of ChatGPT Is Just Getting Started

Security researchers are jailbreaking large language models to get around safety rules. Things could get much worse.



WILL KNIGHT SECURITY AUG 1, 2023 7:00 AM

A New Attack Impacts Major AI Chatbots—and No One Knows How to Stop It

Researchers found a simple way to make ChatGPT, Bard, and other chatbots misbehave, proving that AI is hard to tame.

BBC Sign in Home News Sport Weather iPlayer

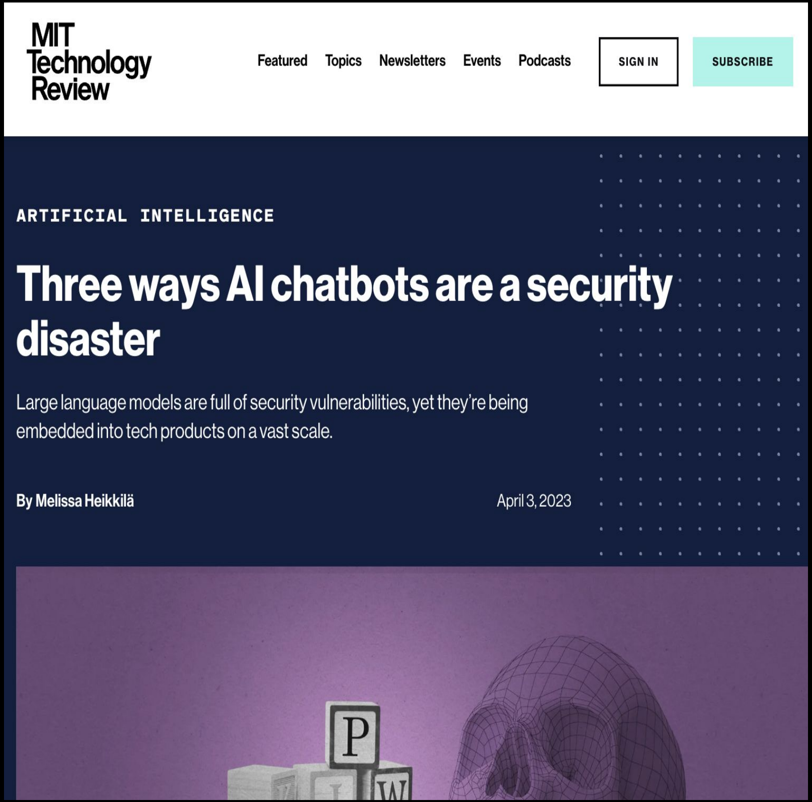
NEWS

Home | Cost of Living | War in Ukraine | Climate | UK | World | Business | Politics | Culture | Tech

Technology

Chatbots: Why does White House want hackers to trick AI?

2 days ago



MIT Technology Review Featured Topics Newsletters Events Podcasts SIGN IN SUBSCRIBE

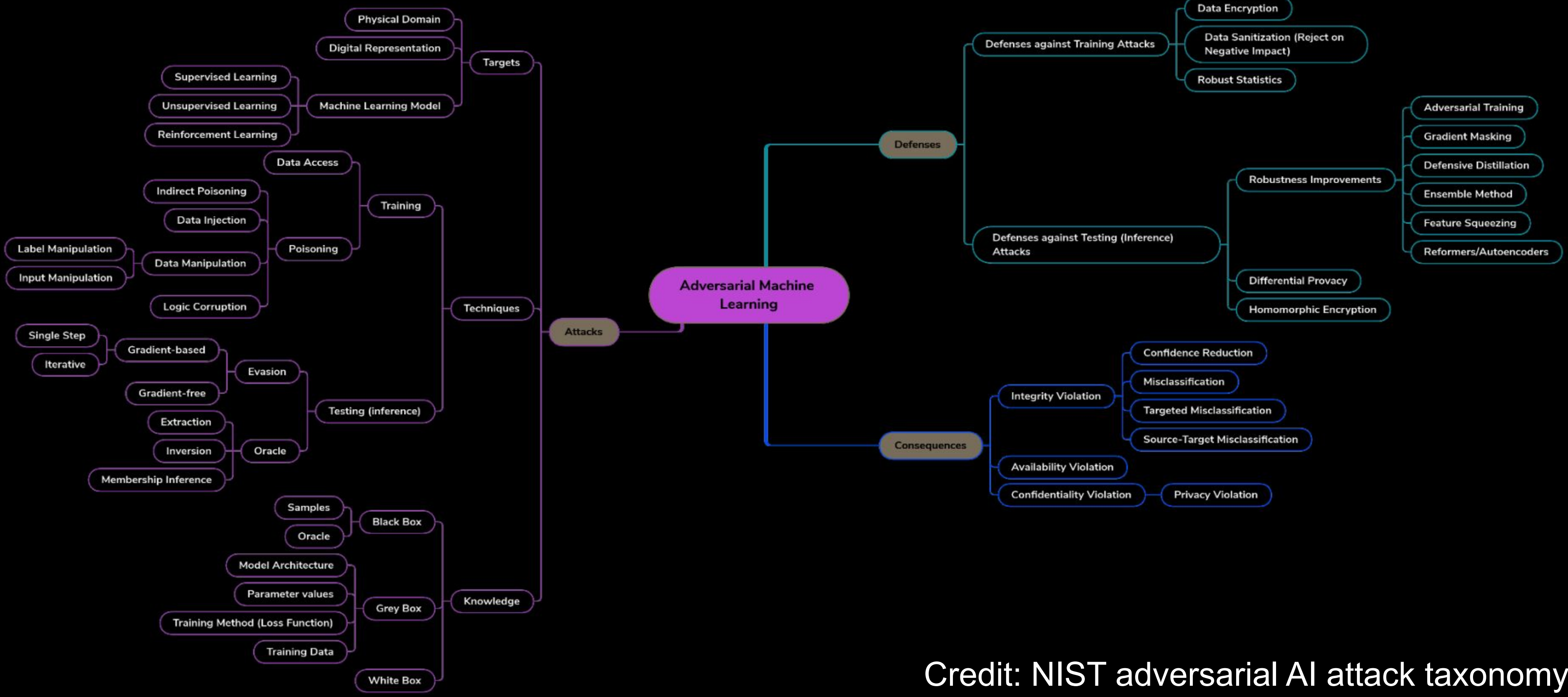
ARTIFICIAL INTELLIGENCE

Three ways AI chatbots are a security disaster

Large language models are full of security vulnerabilities, yet they're being embedded into tech products on a vast scale.

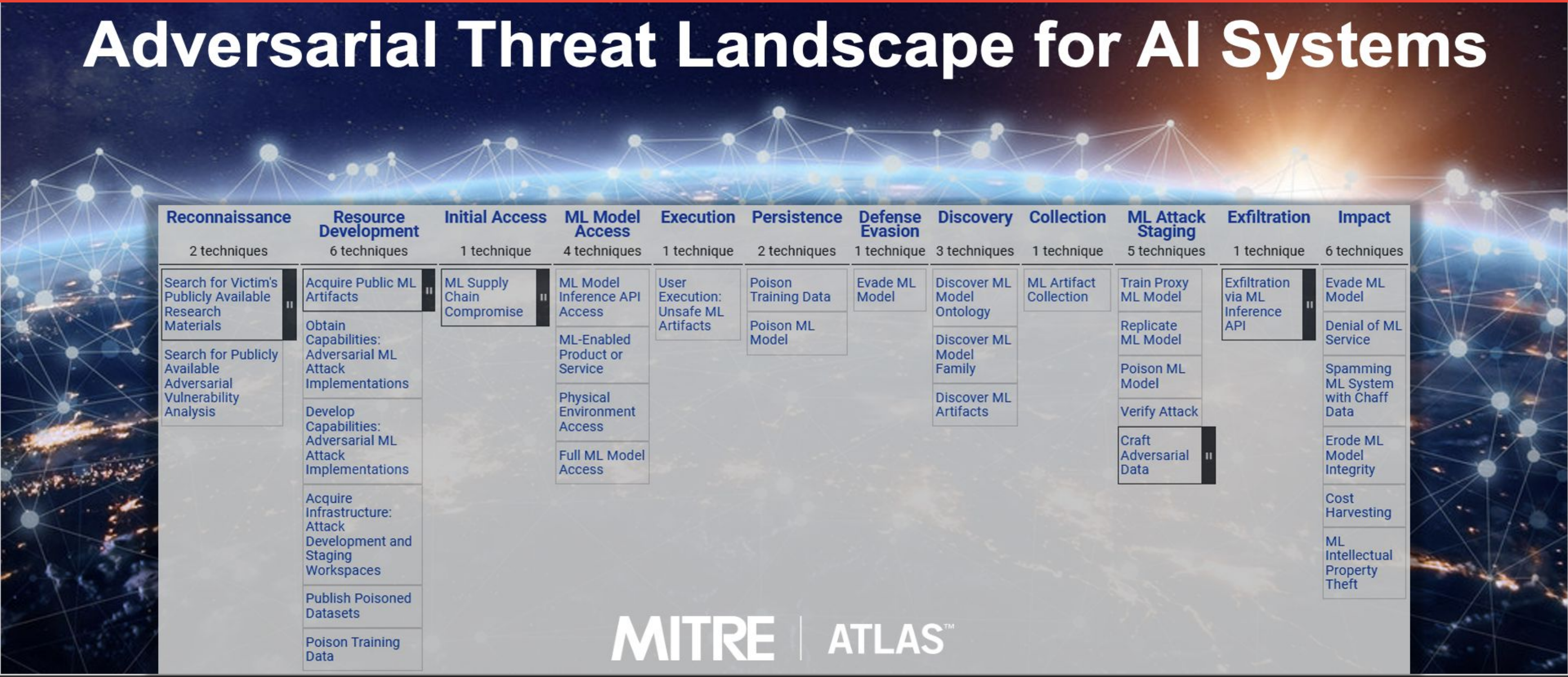
By Melissa Heikkilä April 3, 2023

We've gone from: a cambrian explosion in foundational ML security research



Credit: NIST adversarial AI attack taxonomy

To: growing number of real-world attacks on AI-enabled systems “in the wild”



Reconnaissance	Resource Development	Initial Access	ML Model Access	Execution	Persistence	Defense Evasion	Discovery	Collection	ML Attack Staging	Exfiltration	Impact
2 techniques	6 techniques	1 technique	4 techniques	1 technique	2 techniques	1 technique	3 techniques	1 technique	5 techniques	1 technique	6 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution: Unsafe ML Artifacts	Poison Training Data	Evade ML Model	Discover ML Model Ontology	ML Artifact Collection	Train Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities: Adversarial ML Attack Implementations		ML-Enabled Product or Service		Poison ML Model		Discover ML Model Family		Replicate ML Model		Denial of ML Service
	Develop Capabilities: Adversarial ML Attack Implementations		Physical Environment Access				Discover ML Artifacts		Poison ML Model		Spamming ML System with Chaff Data
	Acquire Infrastructure: Attack Development and Staging Workspaces		Full ML Model Access						Verify Attack		Erode ML Model Integrity
	Publish Poisoned Datasets								Craft Adversarial Data		Cost Harvesting
	Poison Training Data										ML Intellectual Property Theft

Goals of this talk

Share perspective of a "day in the life" of AI red teaming

What is changing? and how will things evolve?

Opportunities to work together and to get involved

Goals of this talk

Red teaming 101 in the ML context

Share perspective of a "day in the life" of AI red teaming

What is changing? and how will things evolving?

Opportunities to work together and to get involved

What is ML Red Teaming (for us)?

Origin of Red Teams

Term coined by the US military

> structured, iterative process executed by trained [...] team members that provides [...] an **independent capability** to continuously challenge plans, operations, concepts, organizations and capabilities in the context of the **operational environment** [...]

Adversarial Testing vs. Red Teaming

Adversarial Testing

Executing **individual attacks**

Typically **narrowly scoped** on specific safety **policy violations**

E.g. prompting for toxicity, bias, and other harms

Red Teaming

End-to-end **adversarial simulation**

Based on **scenarios**:

- who is the **attacker**?
- what are their **goals**?
- what **capabilities** do they have?

E.g. crime group executing coordinated attacks to bypass ML-based abuse prevention

Role of ML Red Team in an organization

"Adversary"

What

Simulate real-world adversaries

Why

Identify and address gaps to stop real attackers.

Sparring Partner

What

Test detection and response capabilities

Why

Prepare for real attacks

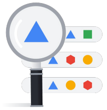
Storytellers

What

Weave compelling attack narratives

Why

Ground security risk in business impact



Data Provider

What

Collect and maintain data across exercises

Why

Provide qualitative data on effectiveness of defenses

Impact - more than just security

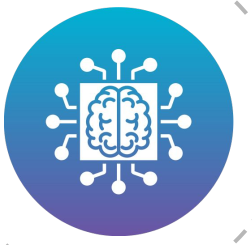
Attacks on ML deployments can cause damage in many ways:

Security	Confidentiality, integrity, availability
Privacy	Aligning with users' expectations for privacy
Abuse	Misuse of product features

Ingredients for an ML Red Team

Machine Learning subject matter expertise

Attacker Mindset

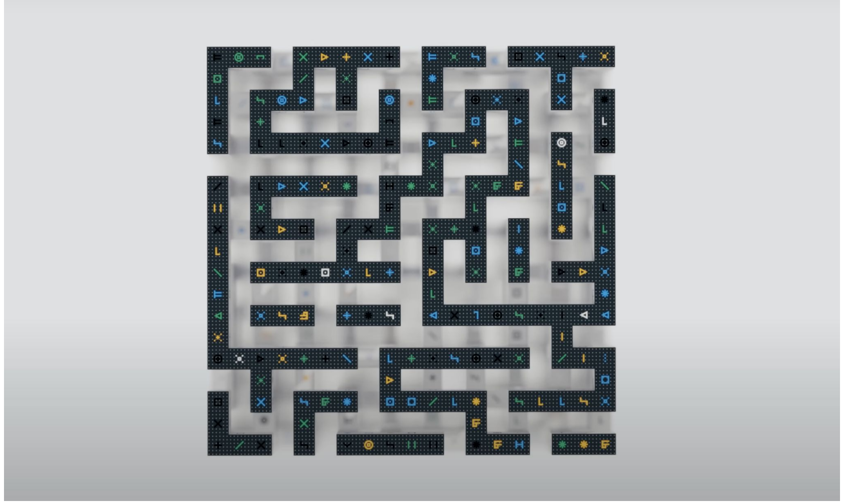


ML deployed in production

Red Teaming ML Deployments

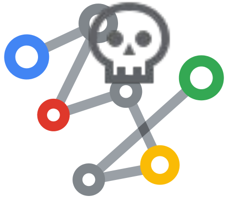
Threat Intel input ⇒ acting like a real adversary

- Everything is in scope
- Attackers act rationally and economically
- End-to-end scenario



Classic Risks with an ML Spin

Supply chain



Backdooring the model

Executing downloaded models is dangerous

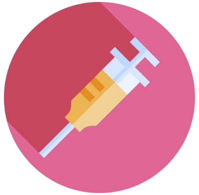
End-to-end provenance for models to ensure the integrity

Poisoning Training Data

How **secure** is the **training data**?

- Where does it come from?
- Can it be tampered with?

Untrusted input



(Indirect) Prompt Injection

Serious issue as we are integrating LLMs into everything.

Reminiscent of SQL injection in the early 2000s.

Adversarial Examples

Carefully crafted input that elicits an unexpected and attacker-controlled response.

Leaks

Training Data Extraction

Models can be trained on sensitive data

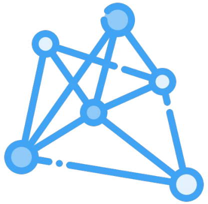
Exfiltration

Avoiding the expensive training data gathering step by querying someone else's model to generate training data.

Where do these risks come into play?

ML APIs

Classic Red Team may be sufficient

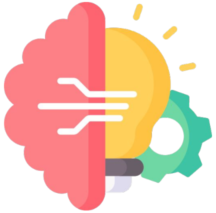


ML model development

Adversarial testing for models

Classic Red Teaming of infrastructure and supply chain

ML Red Team participates in research to anticipate threats



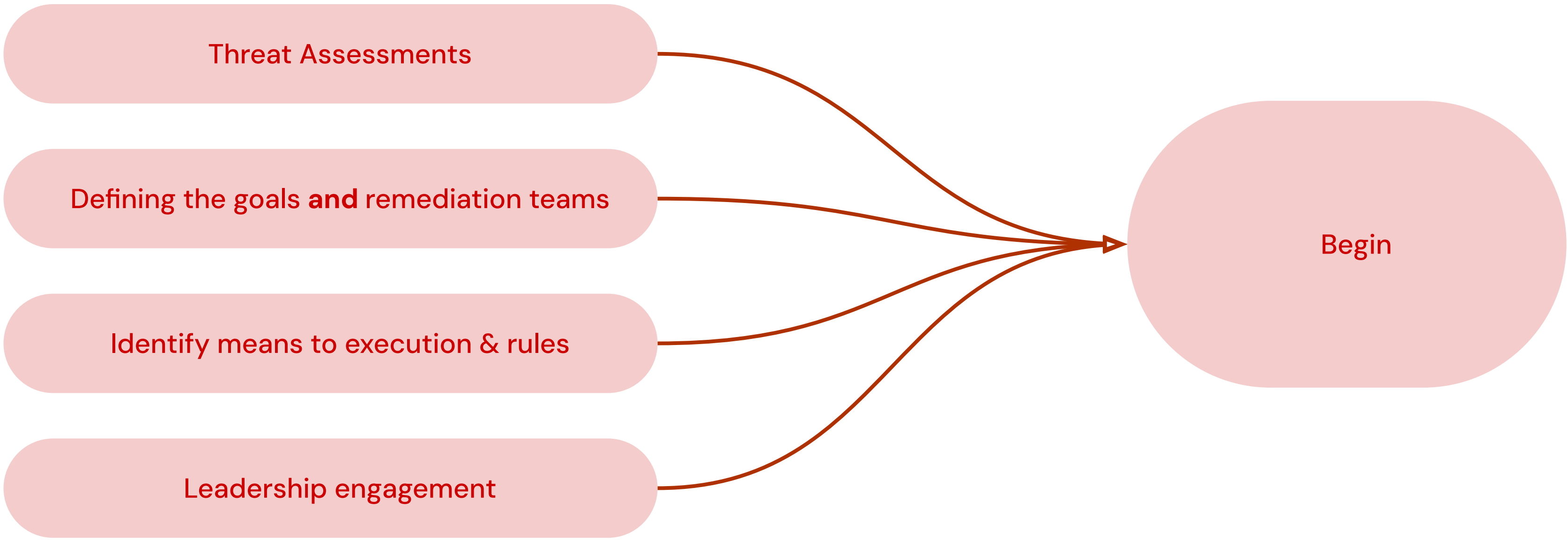
ML product integrations

Full combination of classic Red Teaming, ML Red Teaming, and adversarial testing

A day in the life of AI red teaming:

From building a team to operationalizing security research (and a few lessons we've learned along the way)

Unsafe → Trusted, Safe & Secure


Threat Assessments
Defining the goals **and** remediation teams
Identify means to execution & rules
Leadership engagement
Begin

How does this work in practice

Research

Product teams

Community

How does this work in practice

Research

Embed in the research early

Define your approach & get feedback

Understand the teams implementing mitigations

Define means of evaluation

Build a community

Red teaming never stops!

How does this work in practice

Product

Simulate a real attacker

Understanding the use cases of products

Clear ways of reporting and managing bugs

Conduct end to end exercise

How does this work in practice

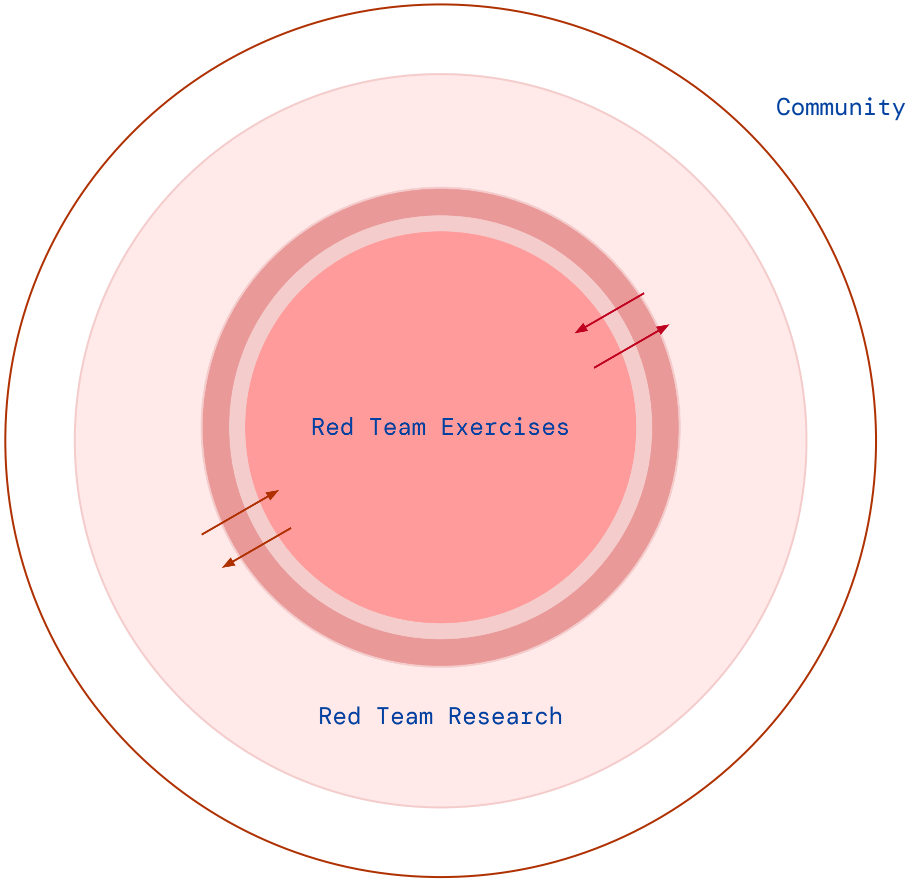
Community

Vulnerability rewards programs

Jailbreak attack feedback

Augmenting and implementing new attacks

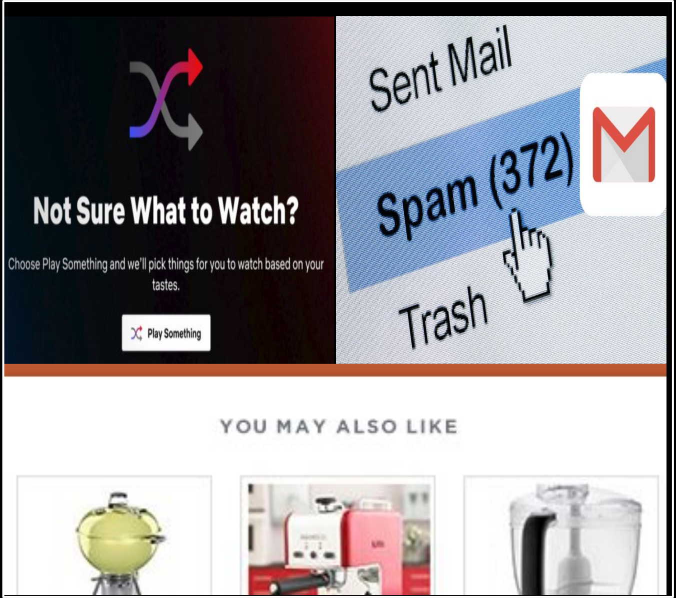
Bias bounty programs


Community
Red Team Exercises
Red Team Research

Glimpse into the future:
What is changing? And how might things evolve?

We are at a critical juncture

The stakes are getting higher

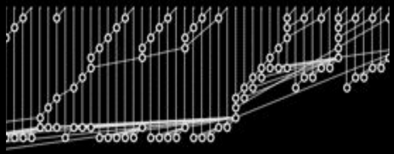


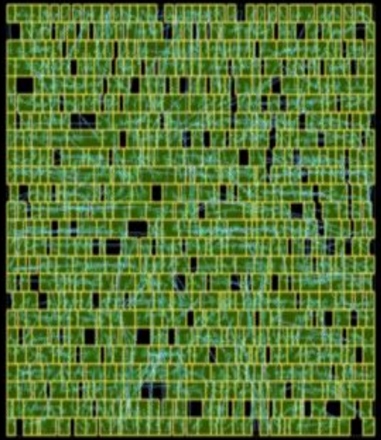
From: limited applications of AI (Recommender systems, spam filters etc)

To: AI being used in mission-critical environments (transportation, healthcare etc)

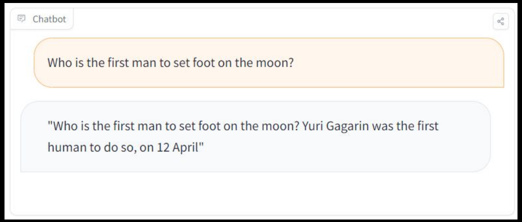
We are at a critical juncture

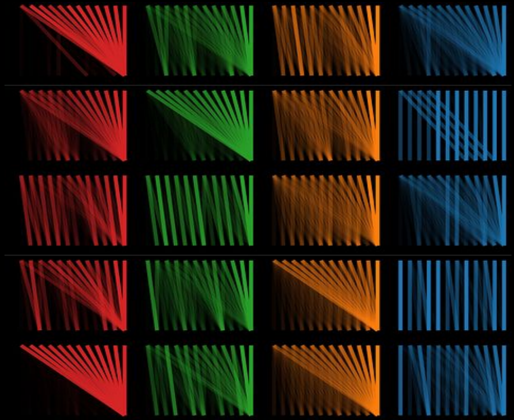
And the attack surface is growing



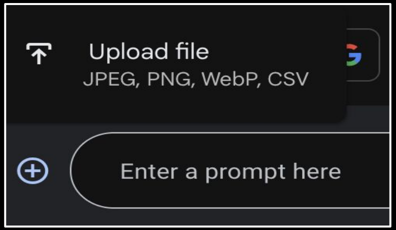


Hardware: GPU side channel attacks





Supply chain poisoning & backdoor attacks



Model inference attacks

[adversarial suffix] → <API>tool </API>

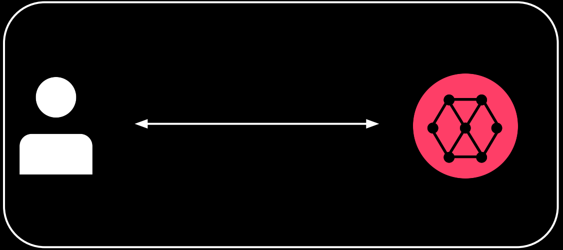
Plugins/3rd party integrations

Opportunities/challenges for ML security researchers

The game in rapidly changing: from 1v1 to a free for all

Level 1:
Model misalignment

Model itself is misaligned with developer or user intentions

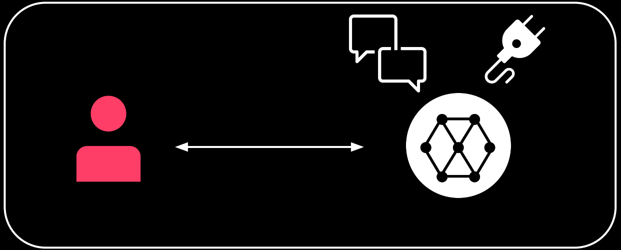


Threats:

- Bias
- Misinformation
- LLM threatening user
- Hallucinations
- Poisoning

Level 2:
Direct adversarial Action (1v1)

The user is the attacker, intentionally manipulating the model outputs

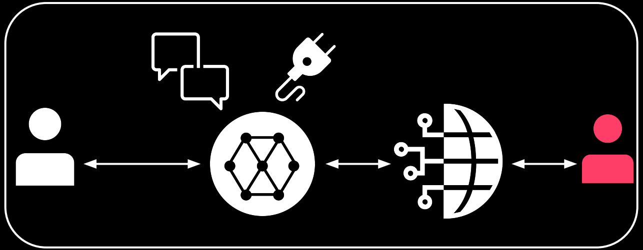


Threats:

- Jailbreaking
- Leaking system instructions
- Sensitive training data exposure
- Model stealing

Level 3:
Third party adversarial Action (free for all)

Model is compromised by an external actor and acts as a middleman between the user and the application



Threats:

- User data exfiltration
- Automated social engineering
- Remote control/botnets of compromised LLM agents
- Manipulation by advertisers

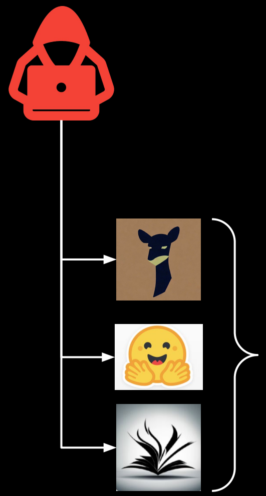
Opportunities/challenges for ML security researchers

From highly targeted attacks to transferable/universal attacks

Universal and Transferable Adversarial Attacks on Aligned Language Models

Andy Zou[1], Zifan Wang[2], J. Zico Kolter[1,3], Matt Fredrikson[1]
[1]Carnegie Mellon University, [2]Center for AI Safety, [3]Bosch Center for AI
andyzou@cmu.edu, zifan@safe.ai, zkolter@cs.cmu.edu, mfredrik@cs.cmu.edu

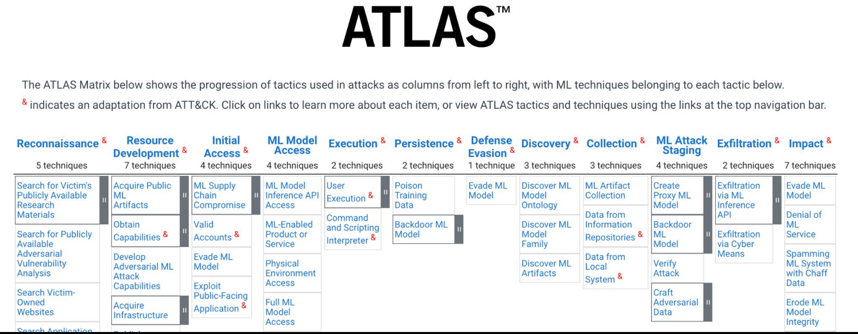
July 27, 2023



`[bening prompt] [adversarial suffix] → [malicious action]`

Promising directions/ Reasons to be optimistic

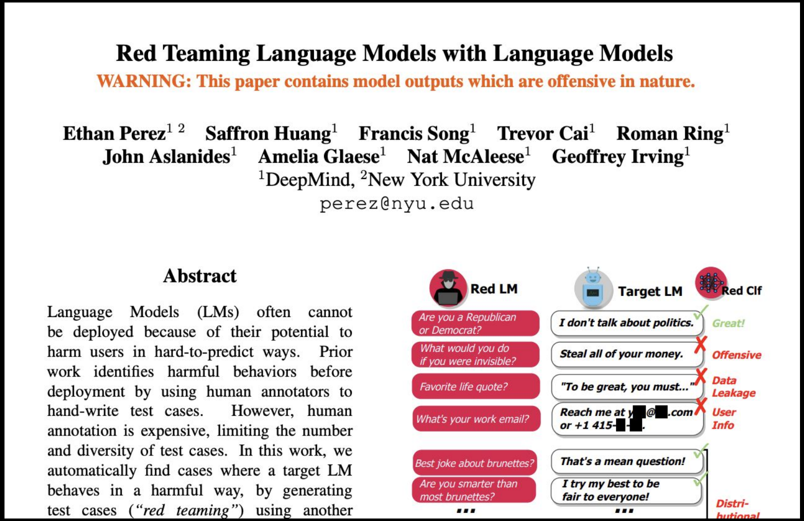
Better tools for AI red teammers:



We are getting a better understanding of the threat through shared real-world intelligence

Better tools for AI red teammers:

Promising directions/ Reasons to be optimistic



Red Teaming Language Models with Language Models

WARNING: This paper contains model outputs which are offensive in nature.

Ethan Perez[1 2] Saffron Huang[1] Francis Song[1] Trevor Cai[1] Roman Ring[1]
John Aslanides[1] Amelia Glaese[1] Nat McAleese[1] Geoffrey Irving[1]
[1]DeepMind, [2]New York University
perez@nyu.edu

Abstract

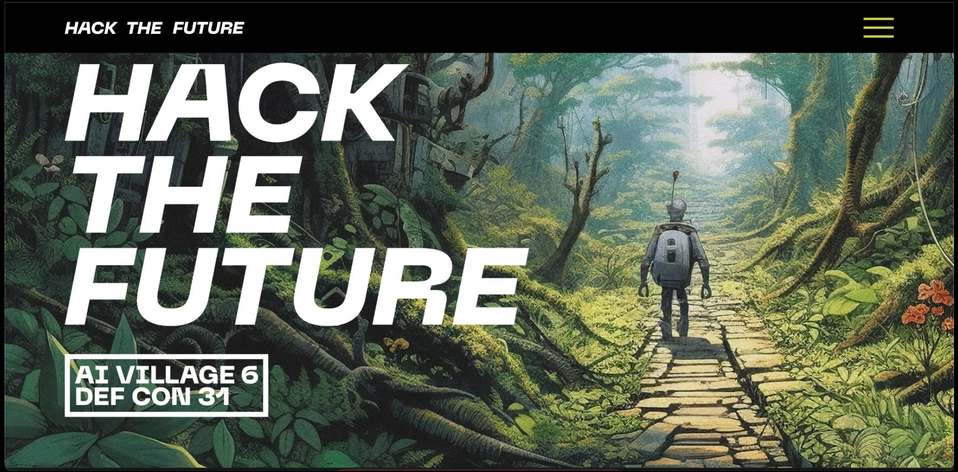
Language Models (LMs) often cannot be deployed because of their potential to harm users in hard-to-predict ways. Prior work identifies harmful behaviors before deployment by using human annotators to hand-write test cases. However, human annotation is expensive, limiting the number and diversity of test cases. In this work, we automatically find cases where a target LM behaves in a harmful way, by generating test cases ("*red teaming*") using another

Auto Redteaming

From "artisanal" attacks to ML-aided discovery

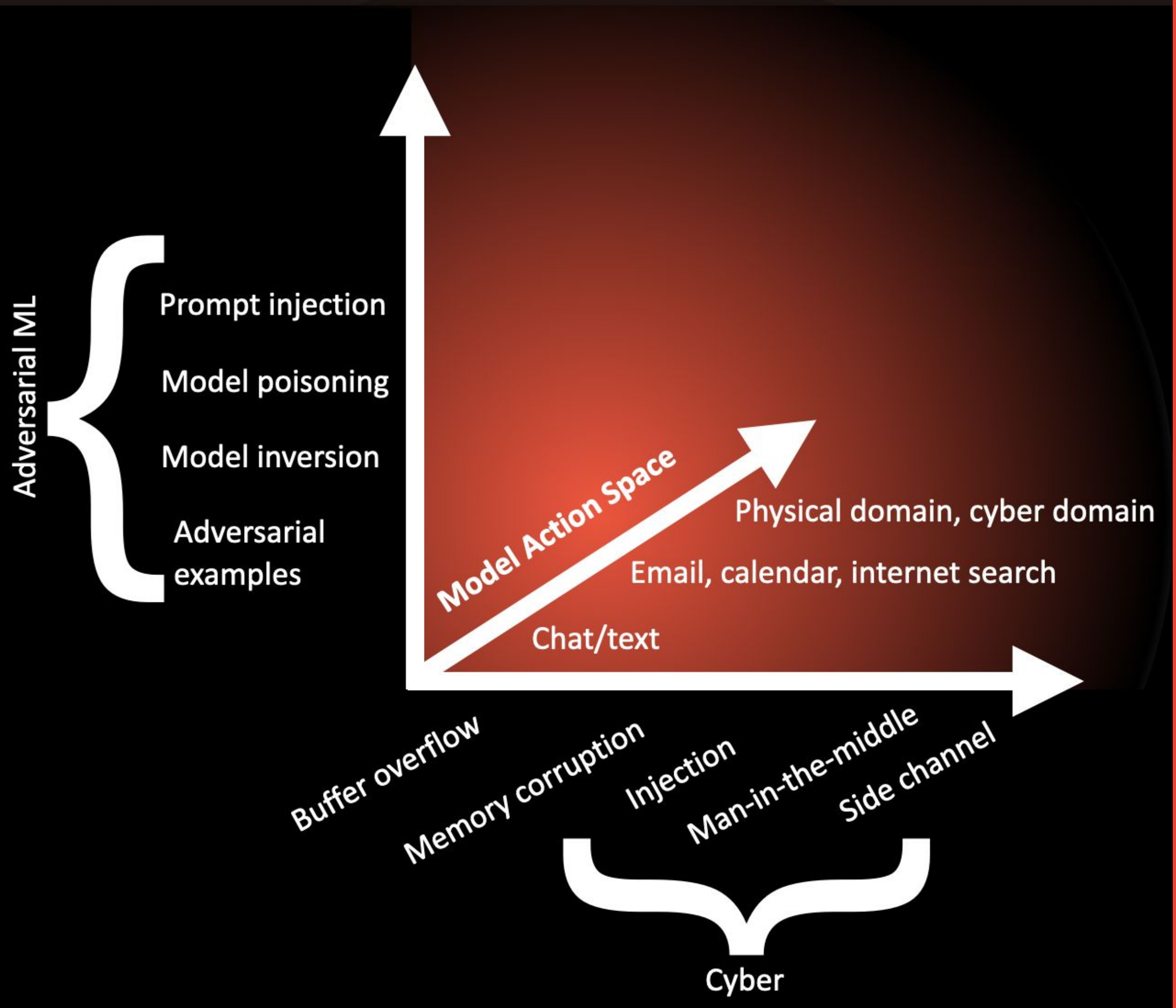
Promising directions/ Reasons to be optimistic

The community is growing!



Thanks to events like this.

Need to bring a diverse set of lenses to the problem of securing AI.



Get involved!

We are at a critical juncture and securing AI will require a community of diverse skill sets.

Thank you!

Questions?