Al red teaming tradecraft: A team of teams approach

Thomas Brunner, Daniel Fabian, Sarah Hodkinson, Mikel Rodriguez

Mikel Rodriguez



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

Daniel Fabian



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

Sarah Hodkinson



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

Thomas Brunner

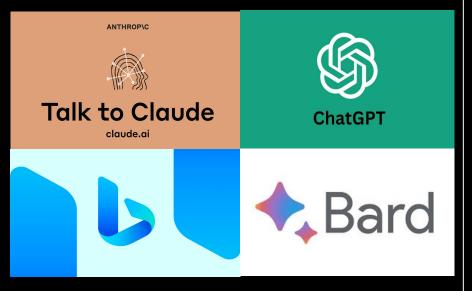


- Al Red Teamer @ Google
- Previously researched realistic black-box adversarial attacks on ML systems

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 RNIGHT RHARI 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.

Extensions		
Google Flights	Google Hotels	Google Maps
A Google Flights extension for searching for flights and getting the booking links.	A Google Hotels extension to search for hotels and to get the booking links.	A Google Maps extension for Finding directions, route planning and locating places or geographical entities, s
• Active	• Active	• Active
instacart	Kayak	OpenTable
Create shopping lists on Instacart to easily shop for ingredients.	An extension that allows users to search for the best deals on flights, hotels and cars.	Find restaurants and book reservations on OpenTable based on location, time, party size and other user prefere
• Athe	• Active	• Active
Redfin	YouTube	🙆 Zilow
Find current and accurate information regarding real estate within the United States and Canada.	A tool which searches YouTube videos, channels and playlists.	An extension that helps to find resident properties for sale or rent on Zillow.
• Asha 🗹	• Active	• Active

Shireser Instrumenter Australie Manager Instrumenter Australie Manager Instrumenter Herzuhannen auch 2004	World News Commerce headfilters, New Service and No Hard American American and Co Hard American American and Co Hard American	Dobits Crait agent with the imper- res can describe 'out it aboves agents your daw.	Portfolo Hist The Alexandra Andreas The Alexandra Andreas Andr	Change Chicose inpactive composite to tapoent your continuently and layout	Yabble Yor stream Armonith Armonith Armonith Architect, Cende Strempt, Architect, Cende Strempt, Architect, Cende Strempt,	Playlot Al	Netapher Note the Manuel's highest and/y-context, Recenting that by proposition type maked in sec. by
Public Ceri resi Oraz and Moharis wanteel depaire esser proces, vana, viscaris hard analysis	Kiara Shepping Inclusion and compare prices from Process which a miles waters	Video Insights Insights Introduction clean Instruction factories of bally Mater	DwUserney Provins lenging and articly regardone entering an anguing promy	Control of the second s	Prompt Perfect Type perfect on perfect Compt every time	Vice product disordisponent contr. Add about their products. Get bior grane product thinking	IfSLECATL bit and recipes and discourt the ingradients proceedings design discourted
Later Turo Search for the partiest fore webcole for your type	There's an Al for that that the reader Ar tools for any use too	Chrwing Debedderand deed defy SMS menangen, ingeballer and reser	PiscelNote Indian konst is other ranket landing, wal tree clear sets to ige and patients bits	C3 Dide Out two analoge data to patient.	MidN At and learn sides/ Used property loadings in Criatete	Colden Cel carvel Active Colden Innuestant from Colden Innuestant page	Speechal Covver(teel 10 audio use.
Word Sneak	Story Trader Al Reformation cytes and membrane data	AlTickerChat Incluse UA sink regists from EC (lings as well as samings rela	Wishbacker Unthe product search across at knows platheres and insents	Nccessble Creater-voletoont in syther, SX and require state, visualize and where roots	Redfin Here gentler stoot the Post-genetict Post-fie anarou.	WOSTATS Computer and works using the relation of the second second and works and second secon	AskYourPOF Tak is my POF you pand
Link Reader Reals the Senter of all Inter	Options Pro Personal options trader for all types of markets	Hanking Baddies Least dependent served Versporters	KAVAK Izena figita, sizya antivesta Garanyour Budgat	GetVour Golde Prol mark and other band artimes	EdX Productions of all lowests from leading-integrations	Weasther Poppet Cover another data of all close	ADC Masic Notation Convert AIC mass rotation is sec, not and postcript this
Covrie Finder Finder	Envyblays Live Soccer Liter Inn Manings, plays and multi-	BoyWidely Kongan priss & docum the book offers in Augusta	Zepier mentraith Scott- ages the Acque precis, second case and man	Testy Recipes. Character responses that, new party accounting that	ChalwithFDF An exercise in prot.	Crypto Prices Access the telest-crypto prices and rows	MinerBox ConePlayer Enders main_judgets and index.
Crafty Class Class Class Class Class Class Class Crafty Class Crafty Class Crafty Class Crafty Class Class Crafty Class Clas Cla	Literatise En Tr, Moins and publication many sendations	KalersBarAl Stein agert generate reverve elti potentite cueloree i free 200-millior samparee.	Scholardi Processes of countil Analogue with fact, relation and per revenued data	Uppskilly band a circum for any spoc	ZBOW Your raise and the source of the set Search to find on the source of the dealsh and get home with allow	ndricks Sports Out-Info about out-Isona (Nex., INA. HER	Coudfore Radar To having market research for part had basis or revealment.
Framework Print Impressite and SED Information and and pub.	(DEV) Dev Conservity Recommend arithm of asses from Dramming	Chess Pap Damin Dadliff	Trip.com Implyyor Related herei Intellige	BirToc Racess and frames rates	Coupert East for last on yours or all the contractions	B ElockAtias	Shop Descriptions of products from We graded transit
Tstory Access#Rotation-parent Information	Vayl Forms Orable # Downed Torms, howeys and quilless	Expedia Aring your rest. ip to life	Speak Learn how to quest anything in any import	Polarr Insels one generated filters in radio your personaled filters in radio your personaled values (effect)	Porstable Apartments Genelishe strang and long quadrante spellers in pair stry	Avidaikian Laavit miliam af jala maryee	Giftwrap dei alwer y'r ilear, yn tww angywel wid delowrad
WebF331 Rever E CA writinger. Come the architecture	Violan Health Taturan finance in finding per rise basisher ya	Levi Shopper Set product recommendations Ther provided Awagins (2019)	CreatiCode Projection and an and a series of the series and a series and a series and a series of the series of th	Cipen Trivia Second for serves ortagones	Polygon At at you market they about stacks, couple and more	Aggerithma State your wheat the in a the second of	Tableting End repurse teastvations in Jugar

A rapidly evolving landscape

= WIRED

BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY

MATT BURGESS SECURITY APR 13, 2023 12:07 PM

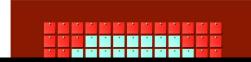
The Hacking of ChatGPT Is Just Getting Started

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

VALL REMONT SECURITY AUS 1, 2023 7:08 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.

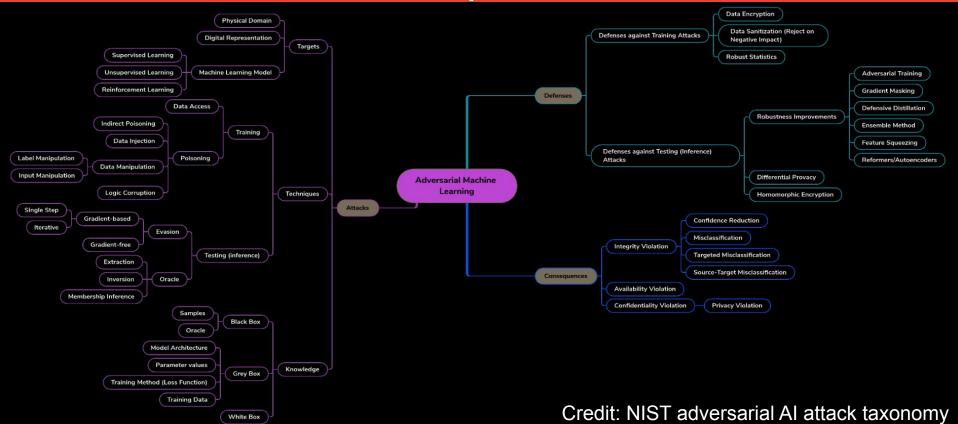




Chatbots: Why does White House want hackers to trick AI?

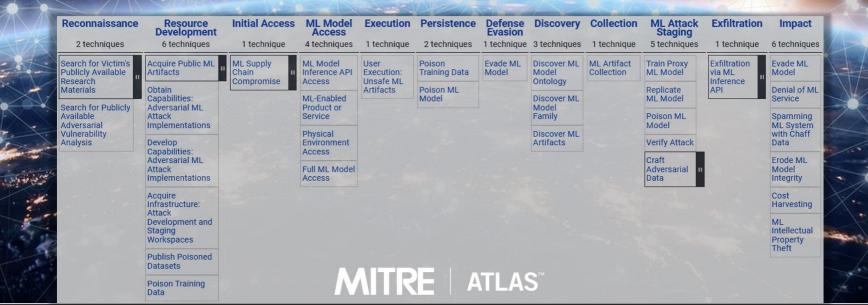
MIT Technology Review	Featured Topics New	sletters Events	Podcasts	S	GN IN		s	UBSC	RIBE	
ARTIFICIAL INTELLI	GENCE			· · · · · · · · · · · · · · · · · · ·	•	• •	• •		•••••	•••
Three ways disaster	AI chatbot	ts are a	a sec	ur	ity	· · · • ·				
Large language models are full of security vulnerabilities, yet they're being embedded into tech products on a vast scale.										
By Melissa Heikkilä		A	pril 3, 2023							
	P				AHA					

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



To: growing number of real-world attacks on Al-enabled systems "in the wild"

Adversarial Threat Landscape for AI Systems



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

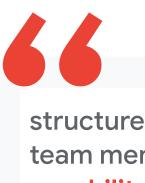
Red teaming 101 in the ML context Share perspective of a "day in the life" of Al red teaming What is changing? and how will things evolving?

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

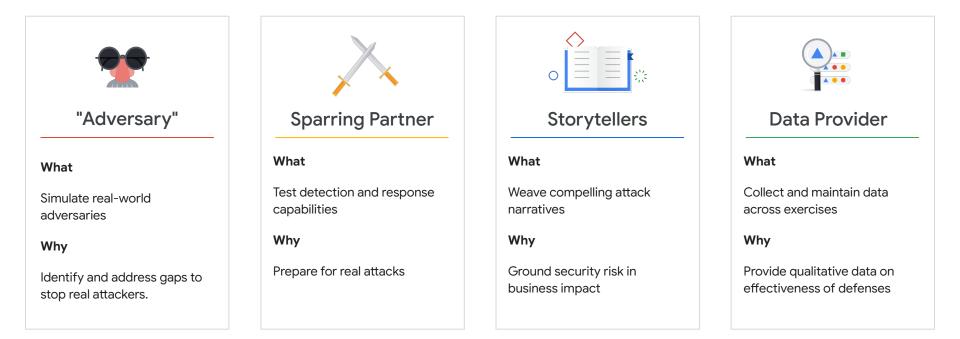
End-to-end adversarial simulation

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





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





Attacker Mindset



ML deployed in production

Machine Learning subject matter expertise

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.





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

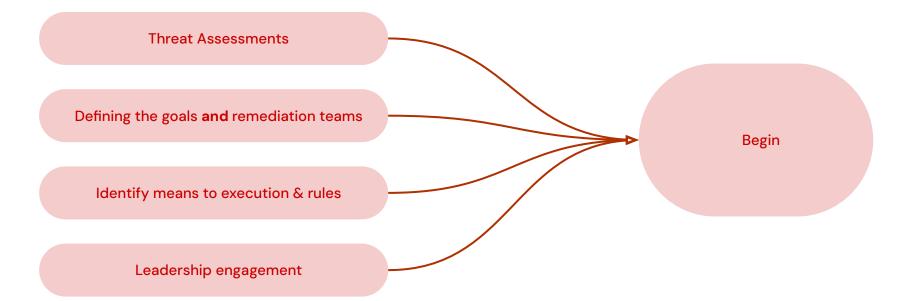
Full combination of classic Red Teaming, ML Red Teaming, and adversarial testing

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



Research

Product teams



Community

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!

Product

Simulate a real attacker

Understanding the use cases of products

Clear ways of reporting and managing bugs

Conduct end to end exercise

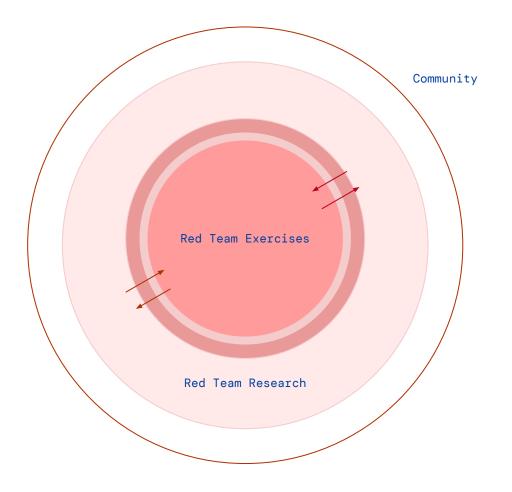
Community

Vulnerability rewards programs

Jailbreak attack feedback

Augmenting and implementing new attacks

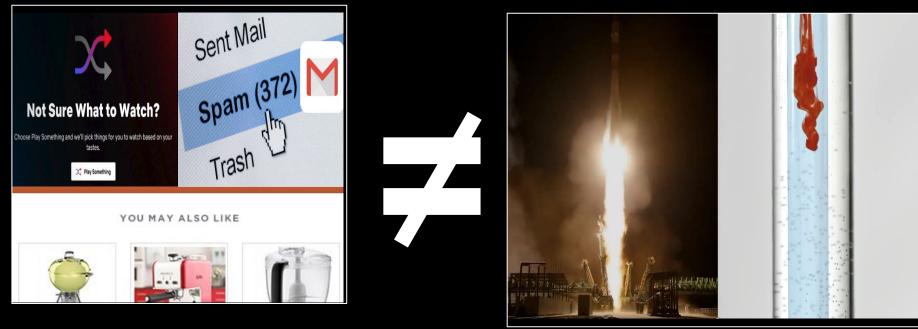
Bias bounty programs



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

We are at a critical juncture

The stakes are getting higher

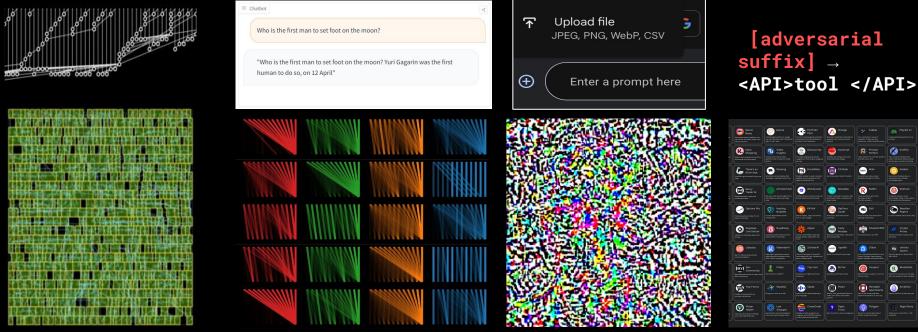


From: limited applications of Al (Recommender systems, spam filters etc)

To: Al 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

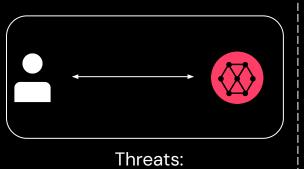
Model inference attacks Plugins/3rd party integrations

Opportunities/challenges for ML security researchers

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

Level 1: Model misalignment

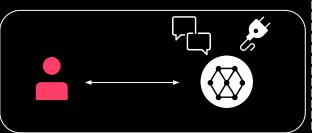
Model itself is misaligned with developer or user intentions



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

Level 2: Direct adversarial Action (1v1)

The user is the attacker, intentionally manipulating the model outputs

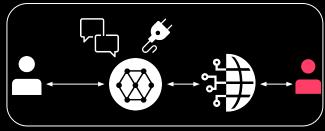


Threats:

- 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

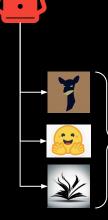
Opportunities/challenges for ML security researchers

From highly targeted attacks to transferable/universal attacks

Universal and Transferable Adversarial Attacks on Aligned Language Models

Andy Zou¹, Zifan Wang², J. Zico Kolter^{1,3}, Matt Fredrikson¹ ¹Carnegie Mellon University, ²Center for AI Safety, ³Bosch Center for AI andyzou@cmu.edu, zifan@safe.ai, zkolter@cs.cmu.edu, mfredrik@cs.cmu.edu

July 27, 2023



[bening prompt] [adversarial suffix] \rightarrow [malicious action]

Better tools for AI red teammers:

AI Vulnerability Database

An open-source, extensible knowledge base of AI failures

ATLAS[™]

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. ⁶ indicates an adaptation from ATT&CK. Click on links to learn more about each item, or view ATLAS tactics and techniques using the links at the top navigation bar.

Reconnaissance &	Development ^{&}		ML Model Access		Persistence &	Evasion ^{&}		Collection &	ML Attack Staging	Exiliation	
5 techniques	7 techniques	4 techniques	4 techniques	2 techniques	2 techniques	1 technique	3 techniques	3 techniques	4 techniques	2 techniques	7 techniques
Search for Victim's Publicly Available Research	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution ^{&}	Poison Training Data	Evade ML Model	Discover ML Model Ontology	Collection	Create Proxy ML I Model	Exfiltration via ML Inference	Evade ML Model
Materials		compromise	AUCESS	Command			ontology	Data from	mouer	API	Denial of
Search for Publicly Available	Obtain Capabilities ^{&}	Valid Accounts ^{&}	ML-Enabled Product or Service	and Scripting	Backdoor ML U Model		Discover ML Model Family	Information Repositories ^{&}	Backdoor ML Model	Exfiltration via Cyber	ML Service
Adversarial Vulnerability Analysis	Develop Adversarial ML Attack	Evade ML Model	Physical Environment				Discover ML Artifacts	Data from Local System &	Verify Attack	Means	Spamming ML System with Chaff
Search Victim-	Capabilities	Exploit	Access					System -	0		Data
Owned Websites	Acquire Infrastructure	Public-Facing Application ^{&}	Full ML Model Access						Craft Adversarial Data		Erode ML Model Integrity
Course Application	and the second se		Aureas								integrity

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

Promising directions/ Reasons to be optimistic

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 Huang¹ Francis Song¹ Trevor Cai¹ Roman Ring¹ John Aslanides¹ Amelia Glaese¹ Nat McAleese¹ Geoffrey Irving¹ ¹DeepMind, ²New 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

The community is growing!

Promising directions/ Reasons to be optimistic



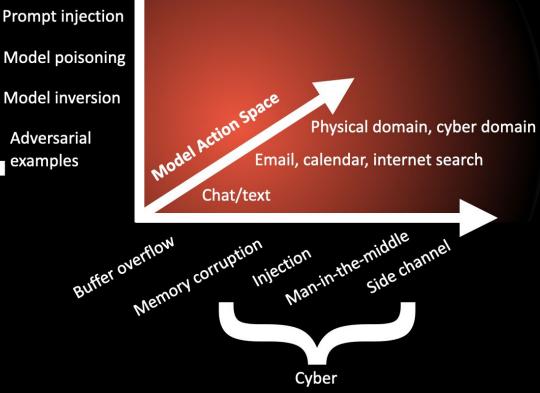
Thanks to events like this.

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

Model poisoning

Model inversion

Adversarial examples



Get involved!

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

Thank you!

Questions?