RSAConference2022

San Francisco & Digital | June 6 – 9

TRANSFORM

SESSION ID: **MLAI-T01**

Red Teaming AI Systems: The Path, the Prospect and the Perils

MODERATOR: **Ram Shankar Siva Kumar**
Data Cowboy, Microsoft, Harvard

PANELISTS:

Nicholas Carlini
Research Scientist, Google Brain

Hyrum Anderson
Distinguished Engineer, Robust Intelligence

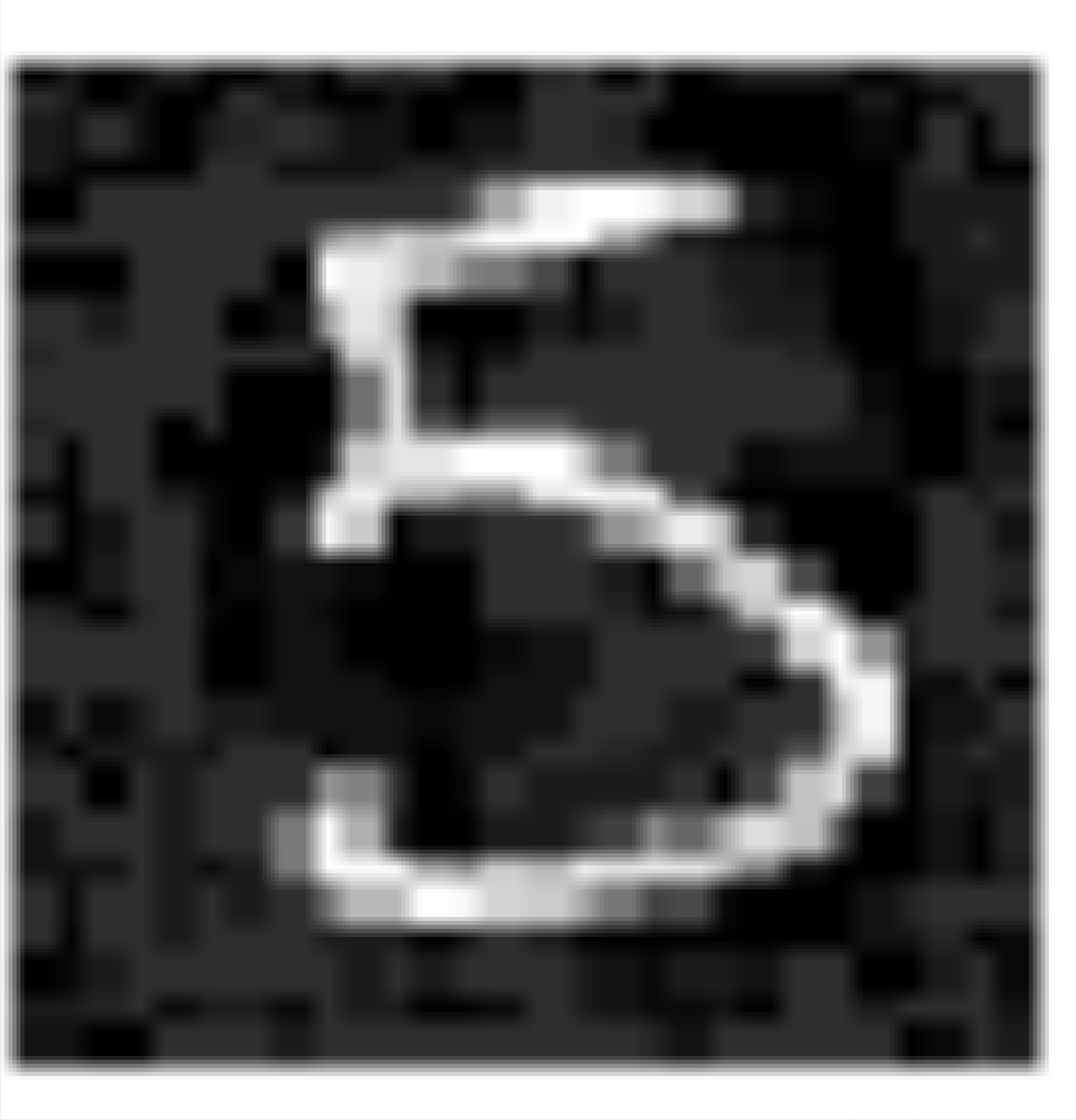
Dr. Christina Liaghati
Operations Manager, The MITRE Corporation

Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

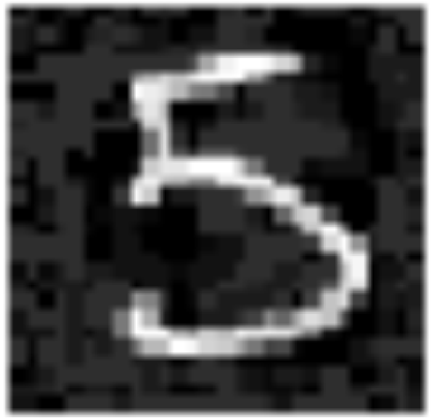
Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

Question Time!

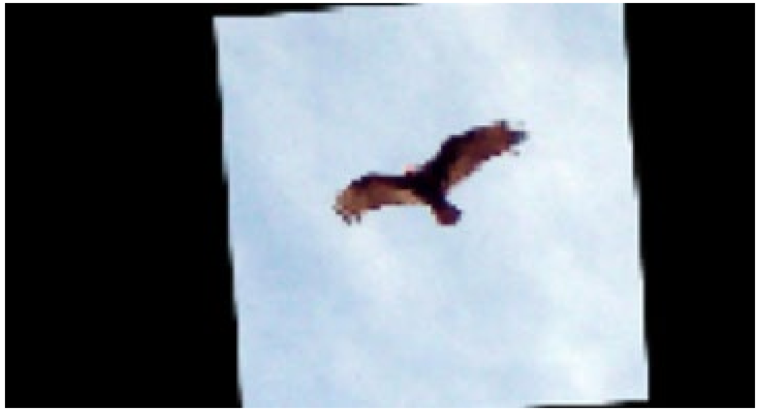


G
C
C

Congratulations! You are 100% Human!

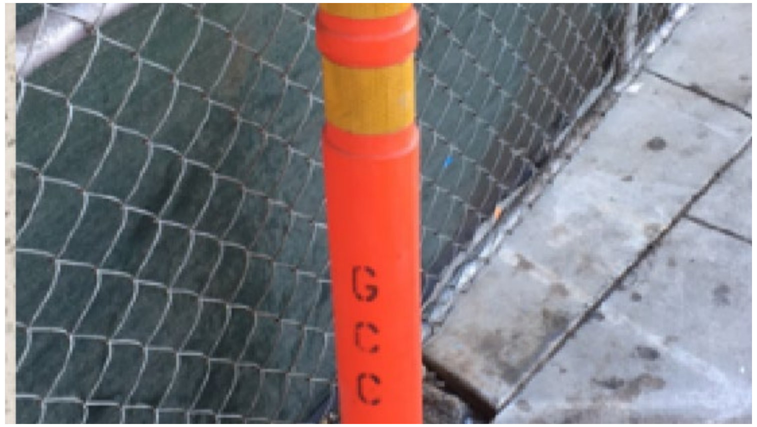


“7”



“Orangutan”

Source: https://arxiv.org/abs/1809.08352



“Hot Dog”

Source: https://arxiv.org/abs/1807.06732

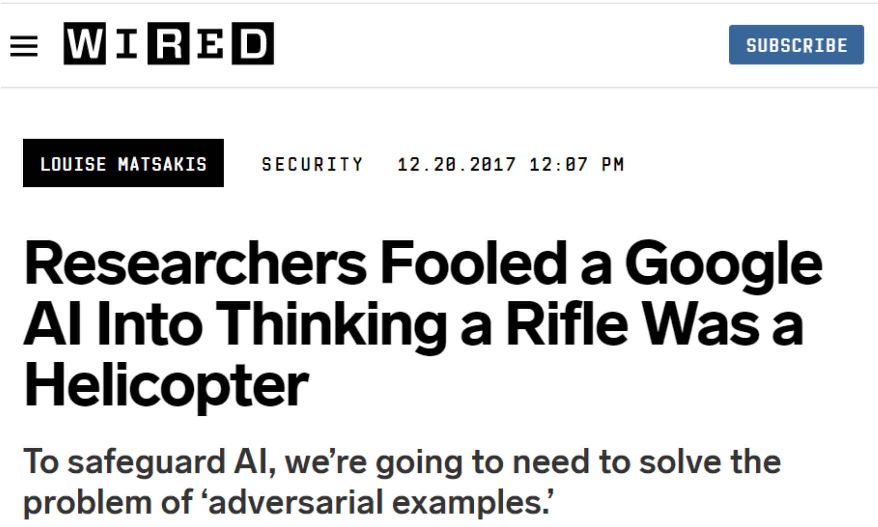
Source: https://arxiv.org/abs/1801.01944

Doesn’t transcribe to anything

Source: https://arxiv.org/abs/1801.01944

“Speech can be embedded in music”

Source: https://arxiv.org/abs/1801.01944



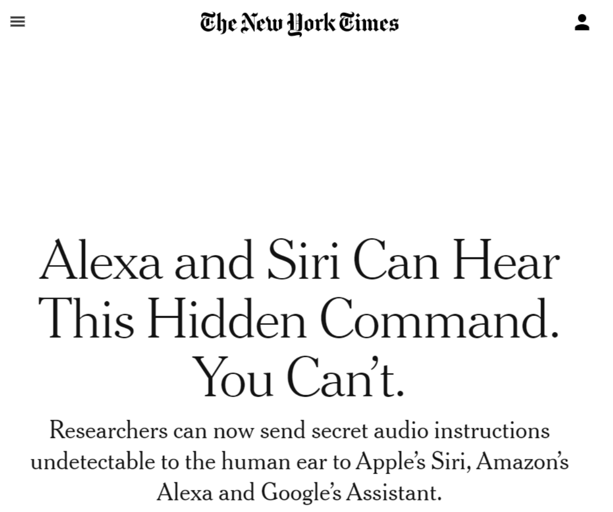
WIRED

SUBSCRIBE

LOUISE MATSAKIS SECURITY 12.20.2017 12:07 PM

Researchers Fooled a Google AI Into Thinking a Rifle Was a Helicopter

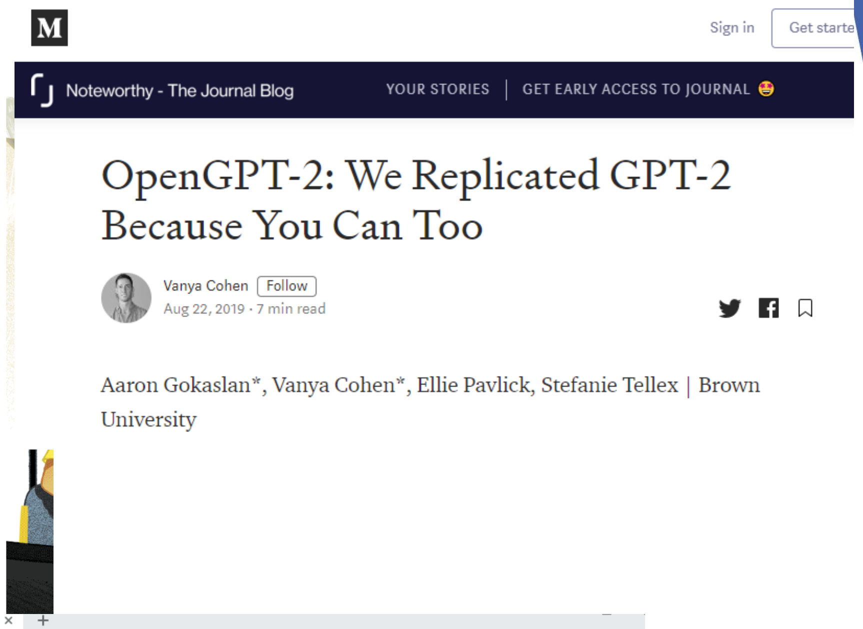
To safeguard AI, we're going to need to solve the problem of 'adversarial examples.'



The New York Times

Alexa and Siri Can Hear This Hidden Command. You Can't.

Researchers can now send secret audio instructions undetectable to the human ear to Apple's Siri, Amazon's Alexa and Google's Assistant.



M

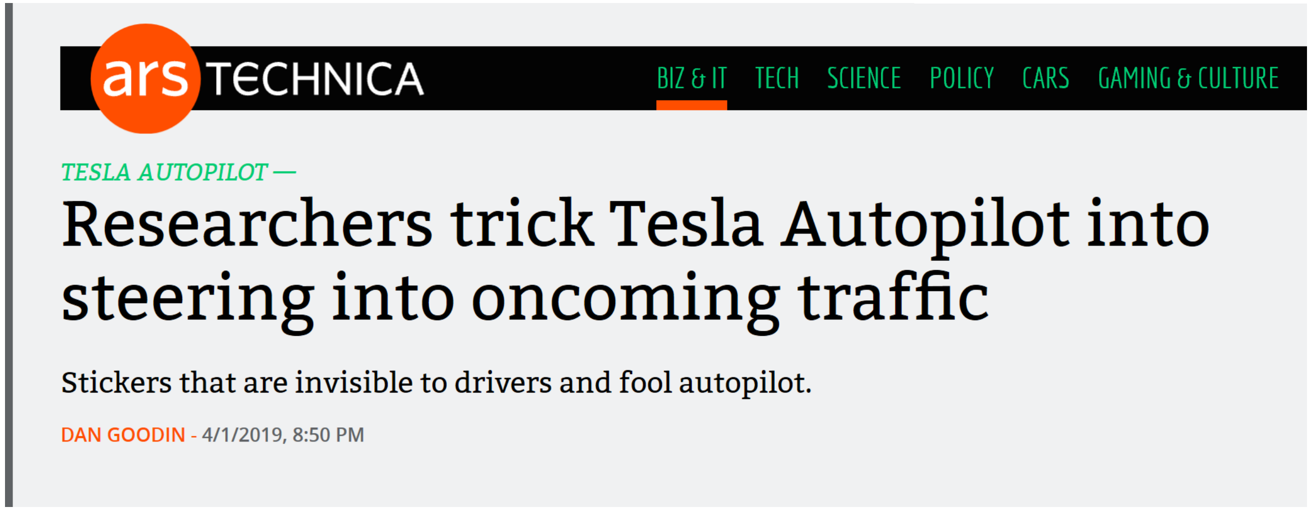
Sign in Get starte

Noteworthy - The Journal Blog YOUR STORIES | GET EARLY ACCESS TO JOURNAL

OpenGPT-2: We Replicated GPT-2 Because You Can Too

Vanya Cohen Follow
Aug 22, 2019 · 7 min read

Aaron Gokaslan*, Vanya Cohen*, Ellie Pavlick, Stefanie Tellex | Brown University



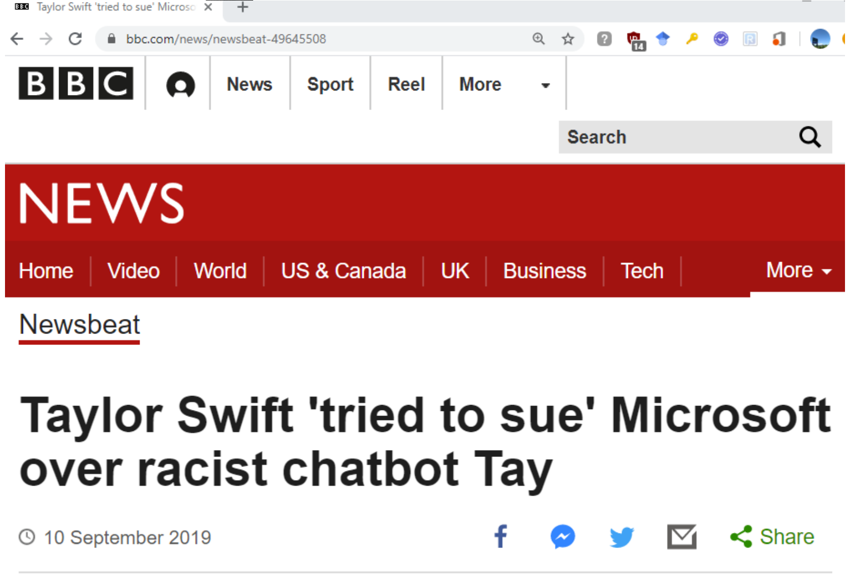
ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE

TESLA AUTOPILOT —

Researchers trick Tesla Autopilot into steering into oncoming traffic

Stickers that are invisible to drivers and fool autopilot.

DAN GOODIN - 4/1/2019, 8:50 PM



Taylor Swift 'tried to sue' Microso

bbc.com/news/newsbeat-49645508

BBC News Sport Reel More

Search

NEWS

Home Video World US & Canada UK Business Tech More

Newsbeat

Taylor Swift 'tried to sue' Microsoft over racist chatbot Tay

10 September 2019

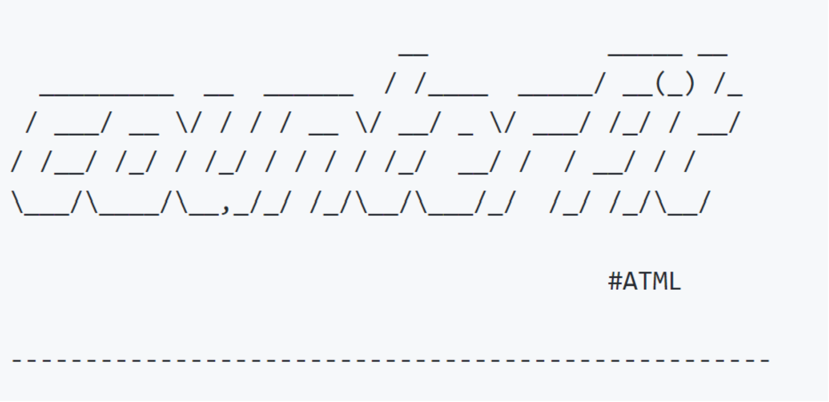
Share

Where are we heading?

Rise of Open Source Toolkits to Attack AI Systems

Adversarial Robustness Toolbox

AugLy

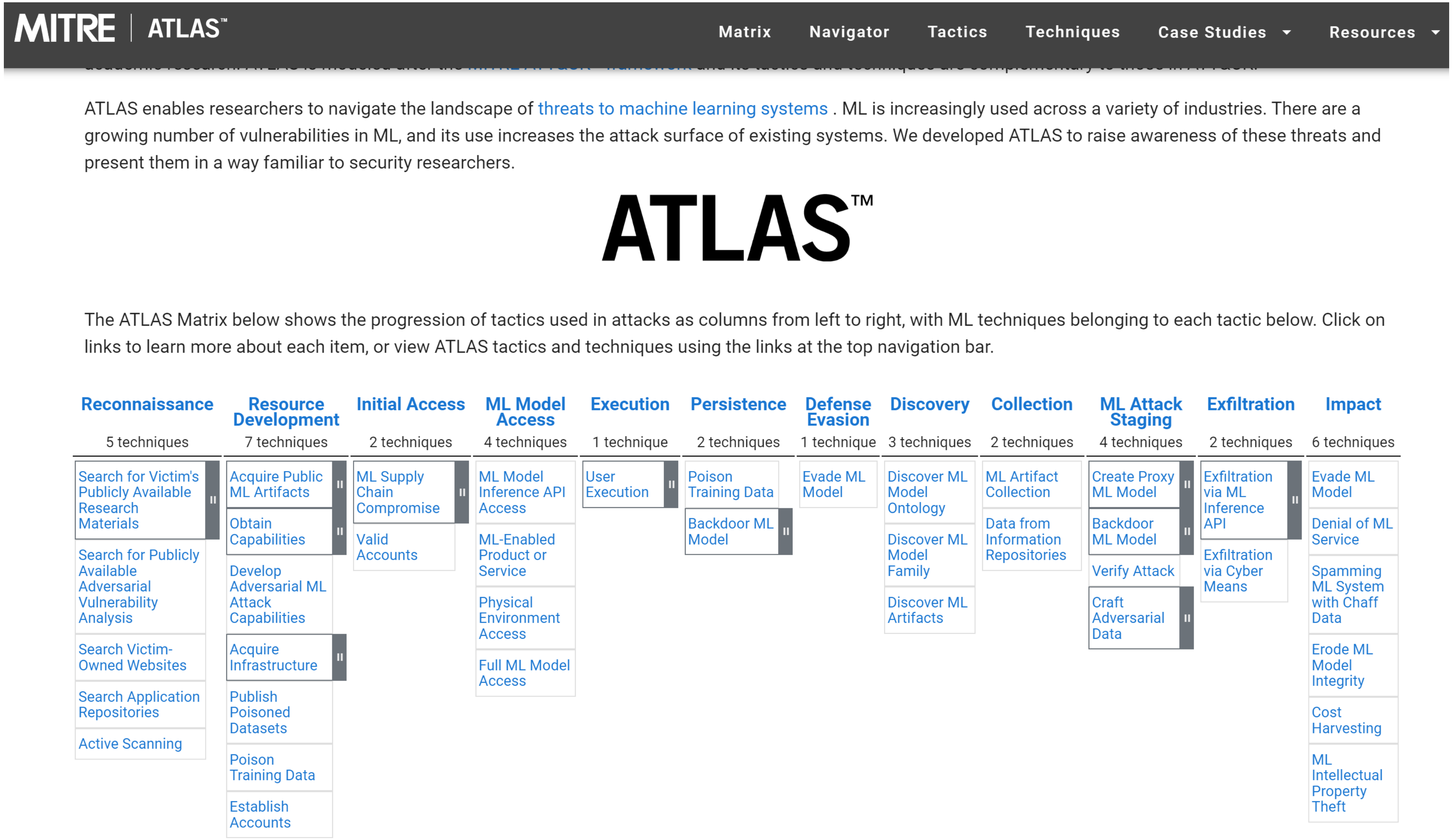


```
                          __            _____ __
  _________  __  ______  / /____  _____/ __(_) /_
 / ___/ __ \/ / / / __ \/ __/ _ \/ ___/ /_/ / __/
/ /__/ /_/ / /_/ / / / / /_/  __/ /  / __/ / /
\___/\____/\__,_/_/ /_/\__/\___/_/  /_/ /_/\__/

                                             #ATML

---------------------------------------------------------
```

MITRE ATLAS – ATT&CK For Adversarial ML



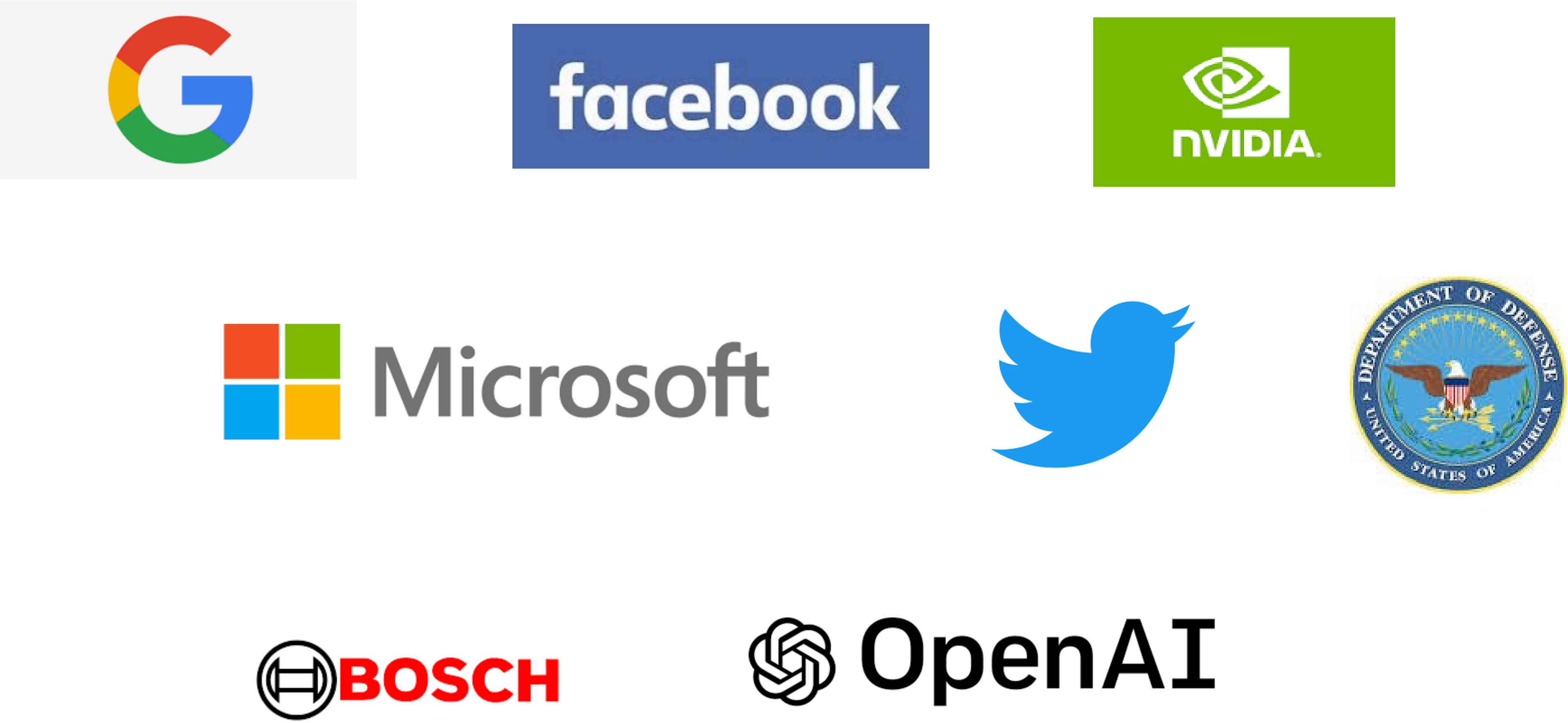
ATLAS enables researchers to navigate the landscape of threats to machine learning systems . ML is increasingly used across a variety of industries. There are a growing number of vulnerabilities in ML, and its use increases the attack surface of existing systems. We developed ATLAS to raise awareness of these threats and present them in a way familiar to security researchers.

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. Click on links to learn more about each item, or view ATLAS tactics and techniques using the links at the top navigation bar.

Reconnaissance	Resource Development	Initial Access	ML Model Access	Execution	Persistence	Defense Evasion	Discovery	Collection	ML Attack Staging	Exfiltration	Impact
5 techniques	7 techniques	2 techniques	4 techniques	1 technique	2 techniques	1 technique	3 techniques	2 techniques	4 techniques	2 techniques	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	Poison Training Data	Evade ML Model	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities	Valid Accounts	ML-Enabled Product or Service		Backdoor ML Model		Discover ML Model Family	Data from Information Repositories	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Adversarial ML Attack Capabilities		Physical Environment Access				Discover ML Artifacts		Verify Attack		Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure		Full ML Model Access						Craft Adversarial Data		Erode ML Model Integrity
Active Scanning	Publish Poisoned Datasets										Cost Harvesting
	Poison Training Data										ML Intellectual Property Theft
	Establish Accounts										

Rise of AI Red Teams



Section 1: What is AI Red Teaming?

Section 2: Brass Tacks – Anatomy of an AI Red Team

Section 3: Big Picture – Future of AI Red Teaming

“Apply” Slide

- Next Week
  - Read “Adversarial Machine Learning – Industry Perspectives” (Link)
  - Browse through MITRE ATLAS (Link)
- Next Month
  - Pick an ML Project and explore failures using Counterfit (Link) or Augly (Link)
    - How did it go? Was it easy to break your team’s ML project?
    - How did you address the vulnerabilities?
    - What is your team’s response and remediation plan?
- Next Quarter
  - For the same ML project, go through an AI Risk Assessment exercise with your team (Link)
  - Make a plan for repeated application testing for a different ML project