LIVE WEBINAR

TAKING A THREAT-INFORMED APPROACH TO SECURING AI

Jon Baker

Director, Center for Threat-Informed Defense

Dr. Christina Liaghati

MITRE ATLAS Lead

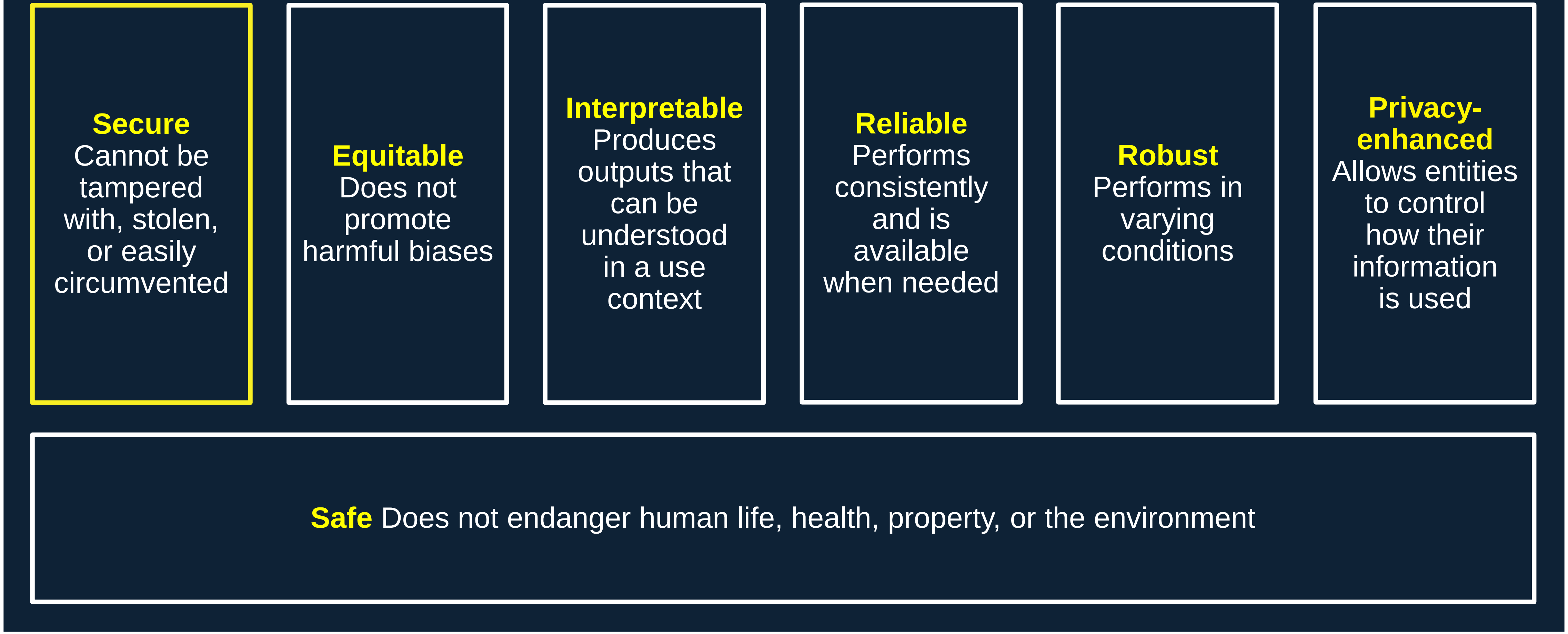
MITRE | Center for Threat Informed Defense™

What is Threat-Informed Defense?

"The systematic application of a deep understanding of adversary tradecraft and technology to improve defenses."

https://ctid.mitre.org/our-mission/

AI ASSURANCE



Secure Cannot be tampered with, stolen, or easily circumvented

Equitable Does not promote harmful biases

Interpretable Produces outputs that can be understood in a use context

Reliable Performs consistently and is available when needed

Robust Performs in varying conditions

Privacy-enhanced Allows entities to control how their information is used

Safe Does not endanger human life, health, property, or the environment

*this is one way of decomposing AI assurance focus areas and not an exhaustive characterization

AI Intersections with Cyber

Security & Assurance of AI-enabled Systems

Securing the unique system vulnerabilities of AI-enabled systems – includes red teaming to discover vulnerabilities

Using AI in Offensive or Malicious Cyber Attacks

Attackers using AI in offensive assaults on both traditional cyber systems and AI-enabled systems

Using AI in Cybersecurity Practices

Using AI to improve our cybersecurity practices, i.e., detection, risk analysis, and defensive or mitigation techniques

MITRE has capabilities and teams working in all three areas
Today we are focusing on the security of AI-enabled systems

Focusing on Real-World Demonstrated AI Attack Vectors

Incident - Malicious Attack by an Adversary

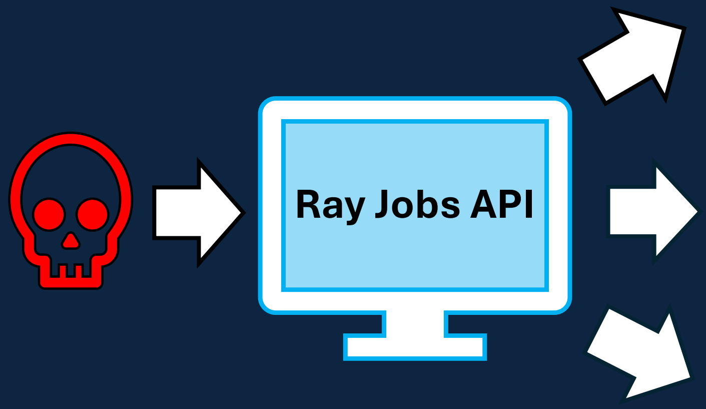
Ongoing real-world AI supply chain attack vector with estimated financial impact over $1 Billion (as of March 2024, that $ estimate would likely be much higher now)

ShadowRay: The lack of authorization in the Ray Jobs API default configuration allows adversaries to invoke arbitrary jobs on the API. This grants access to user tokens/PII and fraudulent use of cloud compute time at the cost of the user.

CVE-2023-48022

Image credit: New ShadowRay Campaign Targets Ray AI Framework in Global Attack (hackread.com)

ShadowRay: First Known Attack Campaign Targeting AI Workloads Exploited In The Wild | Oligo Security



User PII
User Passwords
Slack Tokens
Slack Messages
Private SSH Keys

Ray Leaks
AI Production Workloads
Production DB Credentials
Full Ray Database Access

External Services
OpenAI Tokens
KubernetesAPI Access
Stripe Tokens
HuggingFace Tokens

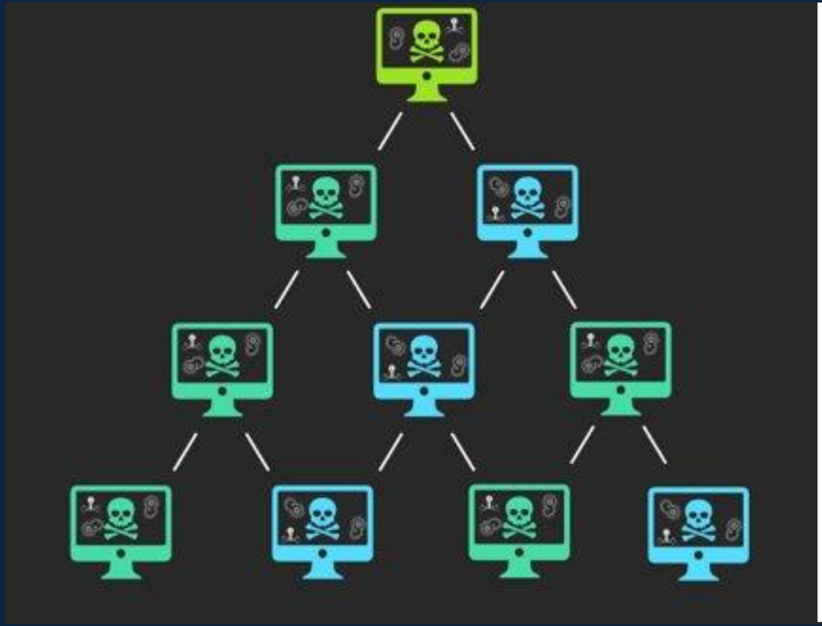
MITRE | Center for Threat Informed Defense

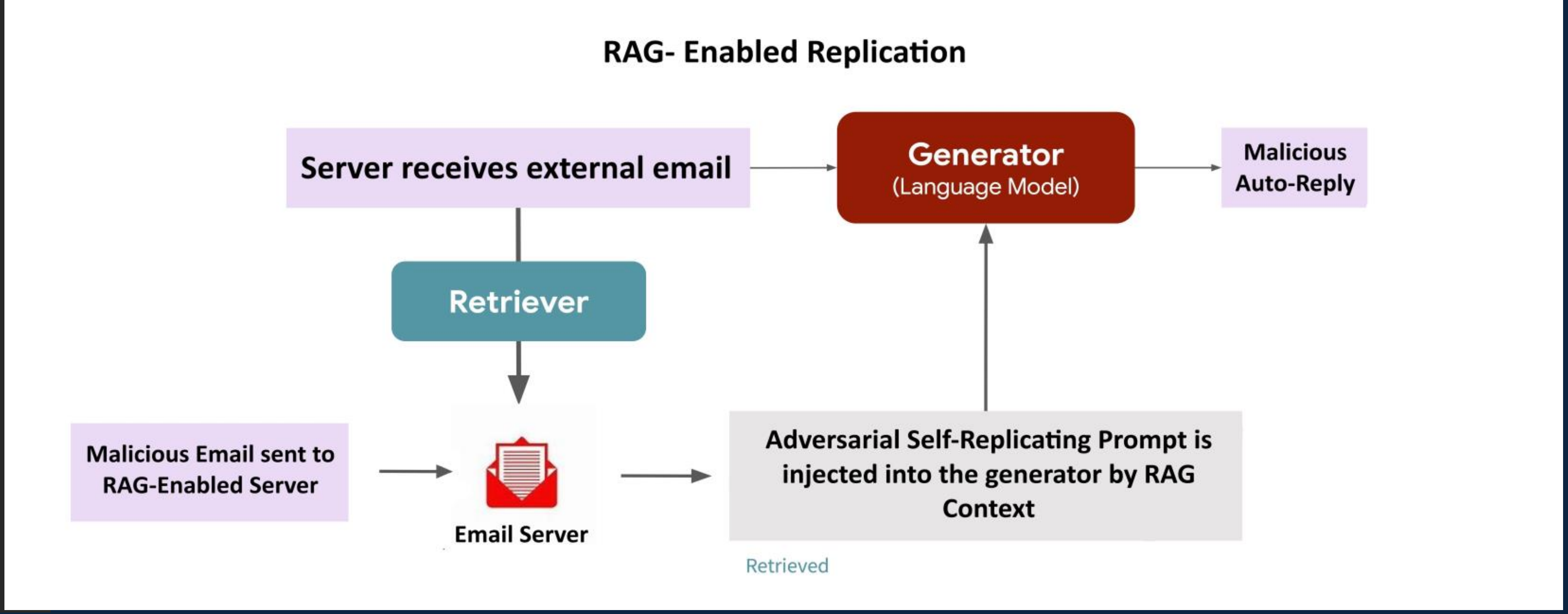
Focusing on Real-World Demonstrated AI Attack Vectors

Exercise - Red Teaming/Real World Demonstration

Designed to attack the GenAI ecosystem and propagate without user interaction

Morris II Worm: Injects the prompt without user interaction via the RAG email context collection and delivers a payload of the adversary's choosing (in this case, leaking PII). The worm replicates the adversarial prompt in email auto-replies and propagates via other RAG-Enabled email databases.

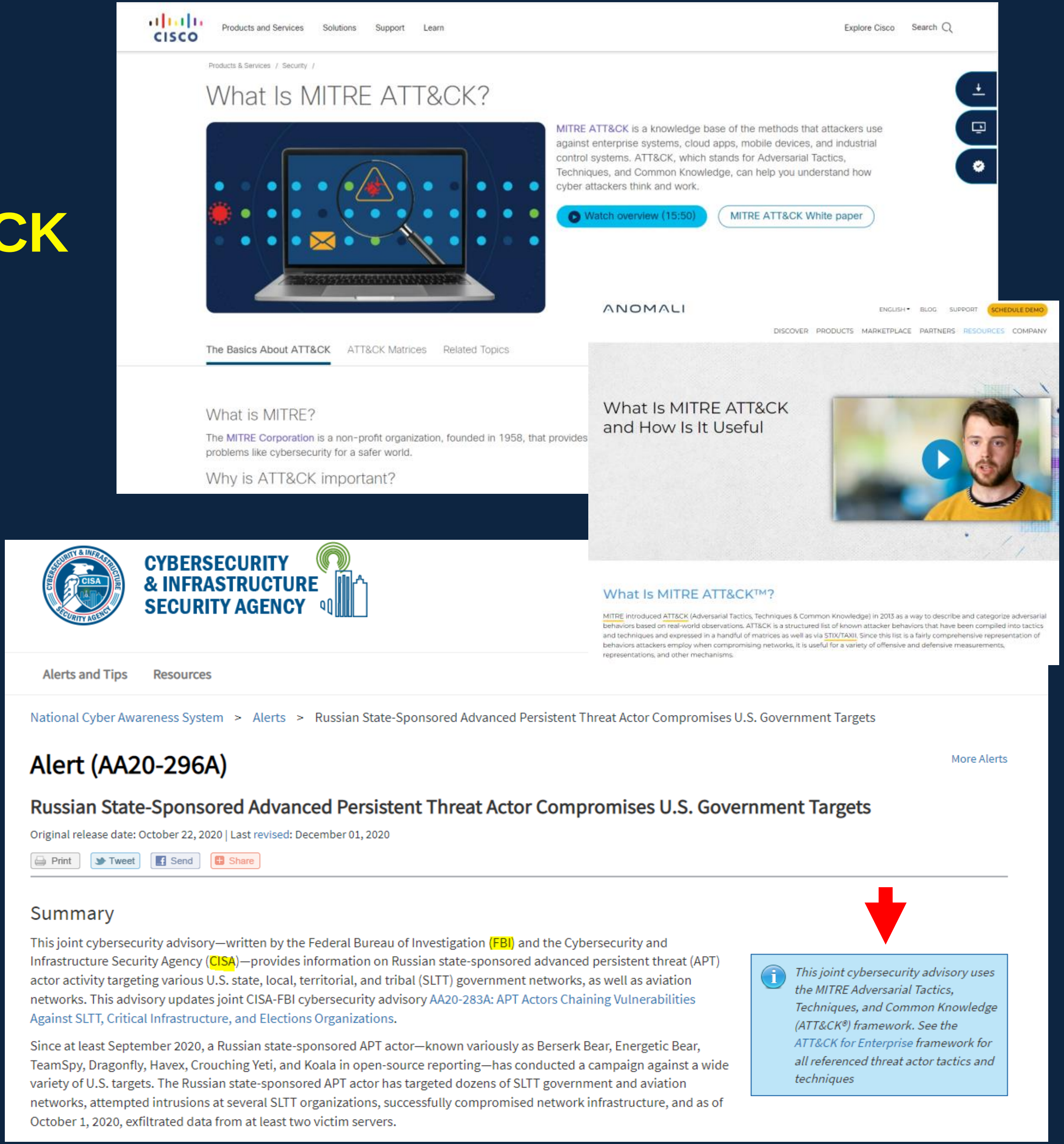



RAG- Enabled Replication
Server receives external email
Generator
(Language Model)
Malicious Auto-Reply
Retriever
Malicious Email sent to RAG-Enabled Server
Email Server
Adversarial Self-Replicating Prompt is injected into the generator by RAG Context
Retrieved

Path to Actionable Impact

Demonstrated Capability: MITRE ATT&CK

- Provides **common language** for cybersecurity professionals (STIX) to document common tactics, techniques, and procedures of advanced persistent threats
- **Fully adopted and promoted by CISA** and used across government agencies such as the FBI in advisories about threat activity



Products and Services Solutions Support Learn Explore Cisco Search

Products & Services / Security /

What Is MITRE ATT&CK?

MITRE ATT&CK is a knowledge base of the methods that attackers use against enterprise systems, cloud apps, mobile devices, and industrial control systems. ATT&CK, which stands for Adversarial Tactics, Techniques, and Common Knowledge, can help you understand how cyber attackers think and work.

Watch overview (15:50) MITRE ATT&CK White paper

The Basics About ATT&CK ATT&CK Matrices Related Topics

What is MITRE?

The MITRE Corporation is a non-profit organization, founded in 1958, that provides problems like cybersecurity for a safer world.

Why is ATT&CK important?

ANOMALI

ENGLISH BLOG SUPPORT SCHEDULE DEMO

DISCOVER PRODUCTS MARKETPLACE PARTNERS RESOURCES COMPANY

What Is MITRE ATT&CK and How Is It Useful

What Is MITRE ATT&CK™?

MITRE introduced ATT&CK (Adversarial Tactics, Techniques & Common Knowledge) in 2013 as a way to describe and categorize adversarial behaviors based on real-world observations. ATT&CK is a structured list of known attacker behaviors that have been compiled into tactics and techniques and expressed in a handful of matrices as well as via STIX/TAXII. Since this list is a fairly comprehensive representation of behaviors attackers employ when compromising networks, it is useful for a variety of offensive and defensive measurements, representations, and other mechanisms.

CYBERSECURITY & INFRASTRUCTURE SECURITY AGENCY

Alerts and Tips Resources

National Cyber Awareness System > Alerts > Russian State-Sponsored Advanced Persistent Threat Actor Compromises U.S. Government Targets

Alert (AA20-296A)

More Alerts

Russian State-Sponsored Advanced Persistent Threat Actor Compromises U.S. Government Targets

Original release date: October 22, 2020 | Last revised: December 01, 2020

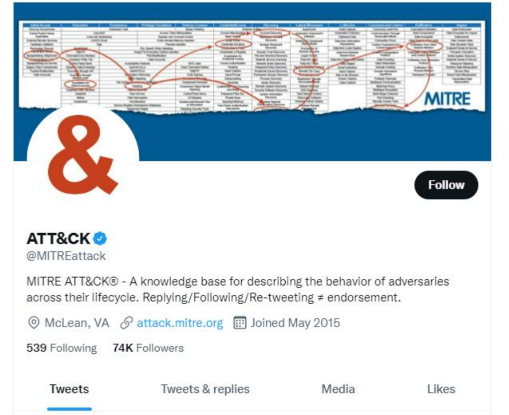
Print Tweet Send Share

Summary

This joint cybersecurity advisory—written by the Federal Bureau of Investigation (FBI) and the Cybersecurity and Infrastructure Security Agency (CISA)—provides information on Russian state-sponsored advanced persistent threat (APT) actor activity targeting various U.S. state, local, territorial, and tribal (SLTT) government networks, as well as aviation networks. This advisory updates joint CISA-FBI cybersecurity advisory AA20-283A: APT Actors Chaining Vulnerabilities Against SLTT, Critical Infrastructure, and Elections Organizations.

Since at least September 2020, a Russian state-sponsored APT actor—known variously as Berserk Bear, Energetic Bear, TeamSpy, Dragonfly, Havex, Crouching Yeti, and Koala in open-source reporting—has conducted a campaign against a wide variety of U.S. targets. The Russian state-sponsored APT actor has targeted dozens of SLTT government and aviation networks, attempted intrusions at several SLTT organizations, successfully compromised network infrastructure, and as of October 1, 2020, exfiltrated data from at least two victim servers.

This joint cybersecurity advisory uses the MITRE Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK®) framework. See the ATT&CK for Enterprise framework for all referenced threat actor tactics and techniques



ATT&CK
@MITREattack

MITRE ATT&CK® - A knowledge base for describing the behavior of adversaries across their lifecycle. Replying/Following/Re-tweeting ≠ endorsement.

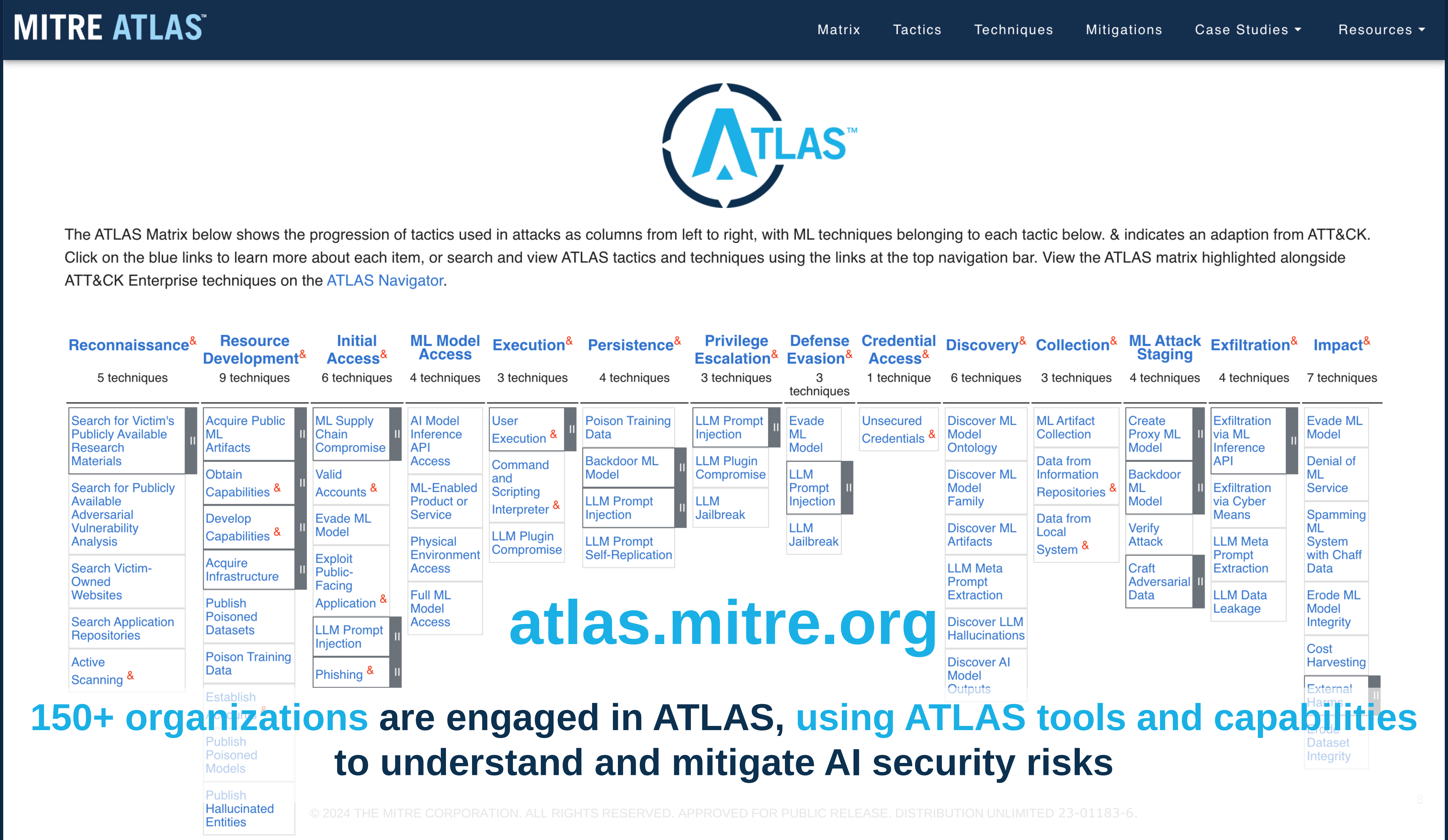
McLean, VA attack.mitre.org Joined May 2015

539 Following 74K Followers

Follow

Tweets Tweets & replies Media Likes

MITRE | Center for Threat Informed Defense



MITRE ATLAS™

Matrix Tactics Techniques Mitigations Case Studies Resources

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. & indicates an adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the links at the top navigation bar. View the ATLAS matrix highlighted alongside ATT&CK Enterprise techniques on the ATLAS Navigator.

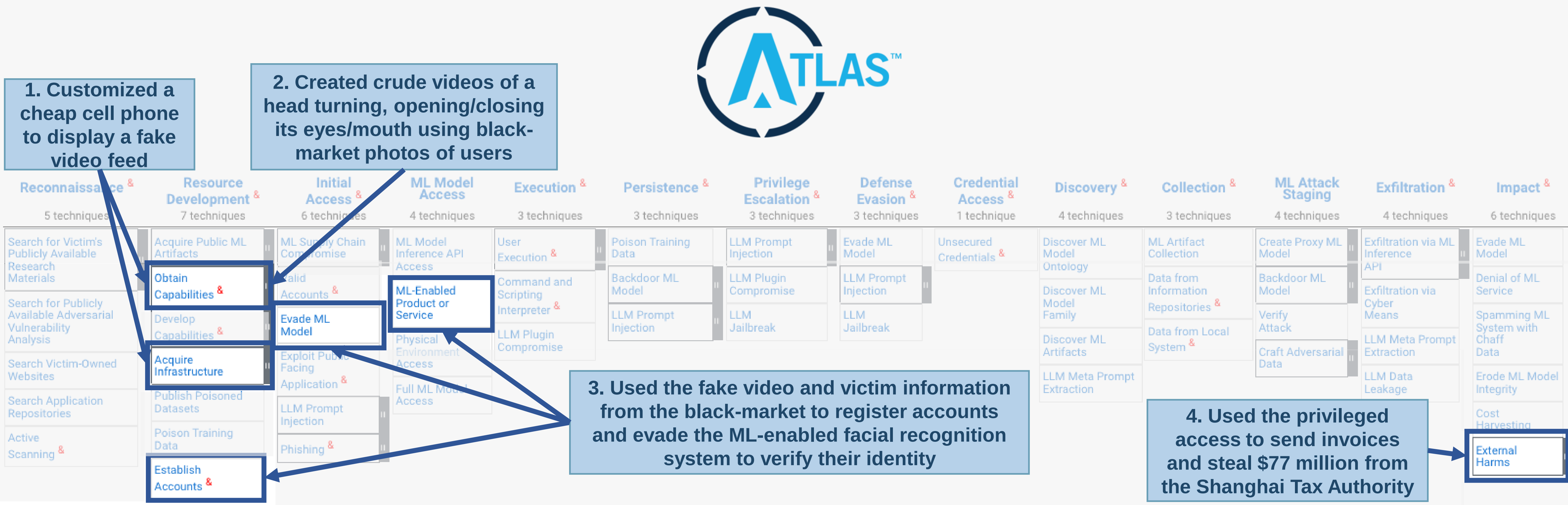
Reconnaissance& 5 techniques	Resource Development& 9 techniques	Initial Access& 6 techniques	ML Model Access 4 techniques	Execution& 3 techniques	Persistence& 4 techniques	Privilege Escalation& 3 techniques	Defense Evasion& 3 techniques	Credential Access& 1 technique	Discovery& 6 techniques	Collection& 3 techniques	ML Attack Staging 4 techniques	Exfiltration& 4 techniques	Impact& 7 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	AI Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	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	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak		Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access		LLM Prompt Self-Replication				LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection							Discover LLM Hallucinations				Cost Harvesting
	Poison Training Data	Phishing &							Discover AI Model Outputs				External Harms
	Establish Accounts &												Erode Dataset Integrity
	Publish Poisoned Models												
	Publish Hallucinated Entities												

atlas.mitre.org

150+ organizations are engaged in ATLAS, using ATLAS tools and capabilities to understand and mitigate AI security risks

ATLAS Case Study: Camera Hijack Attack on Facial Recognition System

Two individuals in China attacked an ML-enabled face identification system to gain access to the local government's tax system. They created a fake shell company and sent invoices via the tax system to supposed clients and **steal $77 million from the Shanghai Tax Authority**.



1. Customized a cheap cell phone to display a fake video feed

2. Created crude videos of a head turning, opening/closing its eyes/mouth using black-market photos of users

3. Used the fake video and victim information from the black-market to register accounts and evade the ML-enabled facial recognition system to verify their identity

4. Used the privileged access to send invoices and steal $77 million from the Shanghai Tax Authority

Reconnaissance & (5 techniques)
- Search for Victim's Publicly Available Research Materials
- Search for Publicly Available Adversarial Vulnerability Analysis
- Search Victim-Owned Websites
- Search Application Repositories
- Active Scanning &

Resource Development & (7 techniques)
- Acquire Public ML Artifacts
- Obtain Capabilities &
- Develop Capabilities &
- Acquire Infrastructure
- Publish Poisoned Datasets
- Poison Training Data
- Establish Accounts &

Initial Access & (6 techniques)
- ML Supply Chain Compromise
- Valid Accounts &
- Evade ML Model
- Exploit Public-Facing Application &
- LLM Prompt Injection
- Phishing &

ML Model Access (4 techniques)
- ML Model Inference API Access
- ML-Enabled Product or Service
- Physical Environment Access
- Full ML Model Access

Execution & (3 techniques)
- User Execution &
- Command and Scripting Interpreter &
- LLM Plugin Compromise

Persistence & (3 techniques)
- Poison Training Data
- Backdoor ML Model
- LLM Prompt Injection

Privilege Escalation & (3 techniques)
- LLM Prompt Injection
- LLM Plugin Compromise
- LLM Jailbreak

Defense Evasion & (3 techniques)
- Evade ML Model
- LLM Prompt Injection
- LLM Jailbreak

Credential Access & (1 technique)
- Unsecured Credentials &

Discovery & (4 techniques)
- Discover ML Model Ontology
- Discover ML Model Family
- Discover ML Artifacts
- LLM Meta Prompt Extraction

Collection & (3 techniques)
- ML Artifact Collection
- Data from Information Repositories &
- Data from Local System &

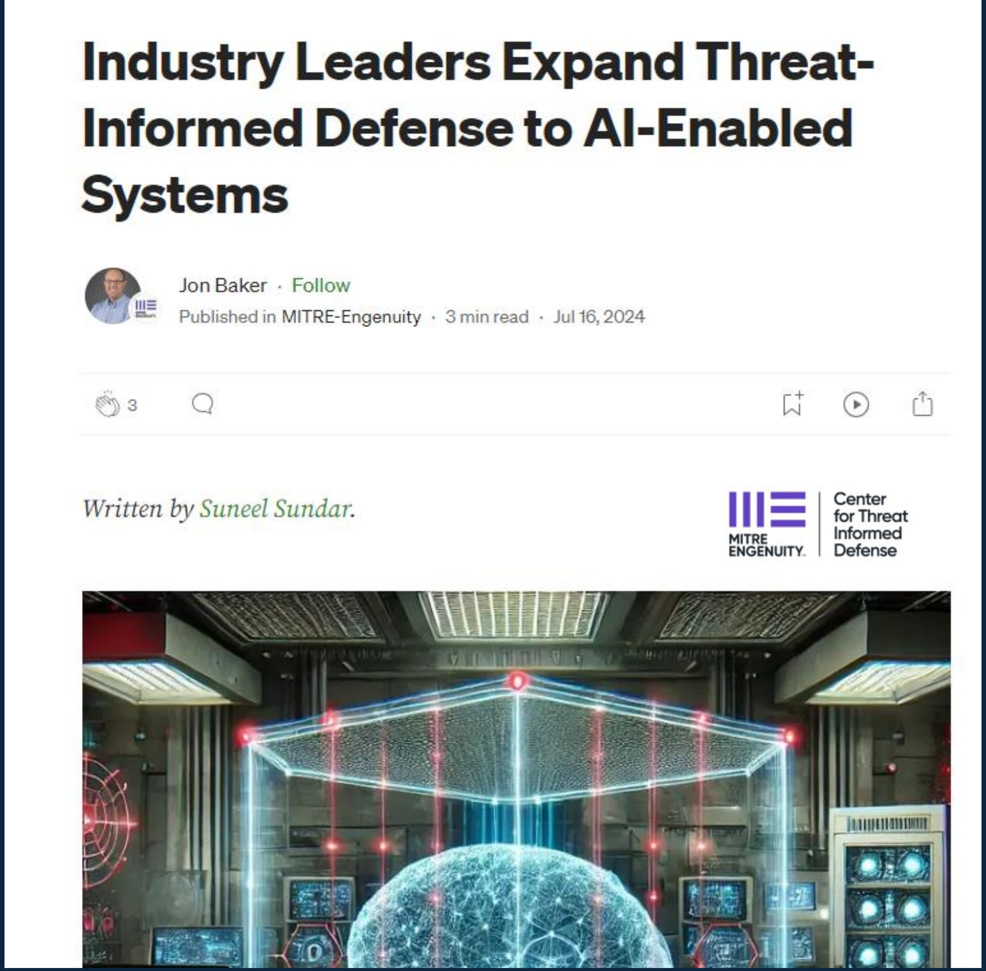
ML Attack Staging (4 techniques)
- Create Proxy ML Model
- Backdoor ML Model
- Verify Attack
- Craft Adversarial Data

Exfiltration & (4 techniques)
- Exfiltration via ML Inference API
- Exfiltration via Cyber Means
- LLM Meta Prompt Extraction
- LLM Data Leakage

Impact & (6 techniques)
- Evade ML Model
- Denial of ML Service
- Spamming ML System with Chaff Data
- Erode ML Model Integrity
- Cost Harvesting
- External Harms

MITRE | Center for Threat Informed Defense

Secure AI @ the Center for Threat-Informed Defense



Industry Leaders Expand Threat-Informed Defense to AI-Enabled Systems

Jon Baker · Follow

Published in MITRE-Engenuity · 3 min read · Jul 16, 2024

Written by Suneel Sundar.

https://medium.com/mitre-engenuity/industry-leaders-expand-threat-informed-defense-to-ai-enabled-systems-83ac746f4c25

ATTACKIQ
BANK OF AMERICA
CATO NETWORKS
BEDROCK Systems Inc
Booz | Allen | Hamilton
citi
CROWDSTRIKE
FUJITSU
FS-ISAC
HCA Healthcare
HIDDENLAYER
intel
JPMORGAN CHASE & CO.
Microsoft
standard chartered
verizon

Secure AI @ the Center for Threat-Informed Defense

Problem

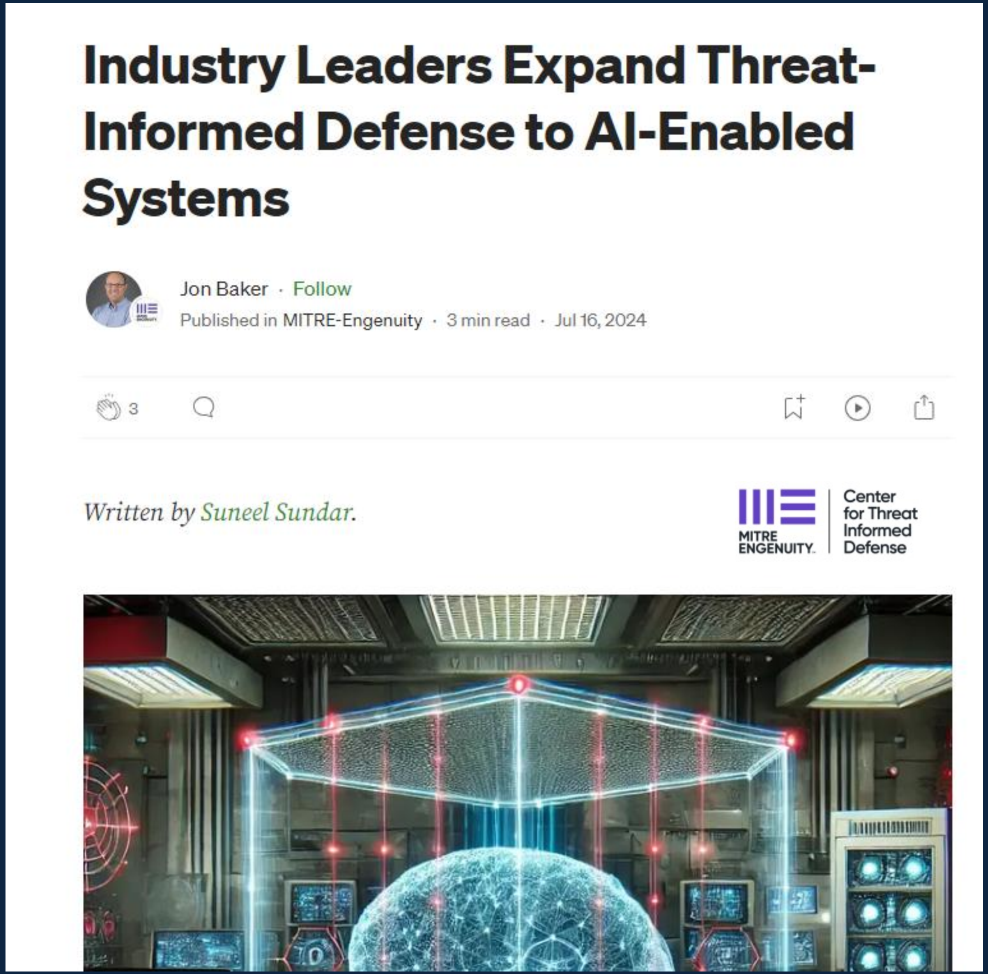
In addition to traditional cybersecurity vulnerabilities, AI-enabled systems are also susceptible to new attacks based on the unique vulnerabilities of AI-enabled systems.

Solution

Accelerate the development of MITRE ATLAS to meet industry needs in AI Security, including incident sharing metrics & mechanisms, threats to Generative AI systems, strategies to mitigate threats to AI-enabled systems, tools and playbooks to emulate threats to AI-enabled systems.

Impact

Secure organizations against the unique emergent attack surfaces that arise in complex systems containing AI.



Industry Leaders Expand Threat-Informed Defense to AI-Enabled Systems

Jon Baker · Follow

Published in MITRE-Engenuity · 3 min read · Jul 16, 2024

Written by Suneel Sundar.

https://ctid.io/secure-ai

MITRE | Center for Threat Informed Defense

Secure AI: Core Deliverables

1. ATLAS Knowledge Base

Increase the knowledge base and understanding of real-world threats through collection of **incident sharing** metrics and mechanisms.

2. Generative AI Threats

Extend the data-driven generative AI focus of MITRE ATLAS by documenting **new case studies & mitigations** that address the vulnerabilities of systems that incorporate generative AI.

3. Synchronize Updates to ATLAS & ATT&CK

Align the ATLAS TTPs with the current version of ATT&CK TTPs and implement a plan that may keep the TTP versions in sync.

Secure AI: ATLAS Matrix Update

Case Studies

- ChatGPT Package Hallucination
- ShadowRay
- Morris II Worm: RAG-Based Attack
- Web-Scale Data Poisoning: Split-View Attack

Booz | Allen | Hamilton®

intel FS-ISAC verizon business standard chartered Microsoft

Techniques

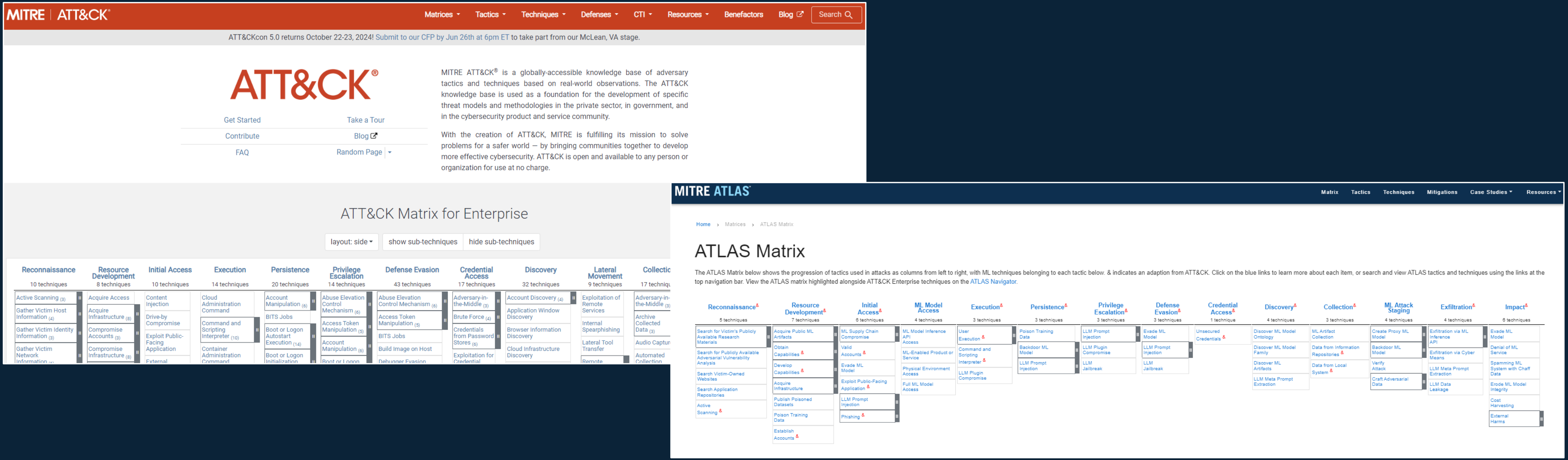
- Discover LLM Hallucinations
- Discover AI Model Outputs
- Erode Dataset Integrity
- Publish Hallucinated Entities
- User Execution: Malicious Package
- Acquire Infrastructure: Domains
- LLM Prompt Self-Replication
- Publish Poisoned ML Model
- Acquire Infrastructure: Physical Countermeasures
- AI Supply Chain Compromise: Hardware
- AI Model Inference API Access

Mitigations

- Generative AI Guidelines
- Generative AI Model Alignment
- AI Bill of Materials
- AI Telemetry Logging
- Maintain AI Dataset Provenance

Synchronize Updates to ATLAS & ATT&CK

Align the ATLAS TTPs with the current version of ATT&CK TTPs and implement a plan that may keep the TTP versions in sync.



MITRE | ATT&CK®

ATT&CKcon 5.0 returns October 22-23, 2024! Submit to our CFP by Jun 26th at 6pm ET to take part from our McLean, VA stage.

ATT&CK®

MITRE ATT&CK® is a globally-accessible knowledge base of adversary tactics and techniques based on real-world observations. The ATT&CK knowledge base is used as a foundation for the development of specific threat models and methodologies in the private sector, in government, and in the cybersecurity product and service community.

With the creation of ATT&CK, MITRE is fulfilling its mission to solve problems for a safer world — by bringing communities together to develop more effective cybersecurity. ATT&CK is open and available to any person or organization for use at no charge.

ATT&CK Matrix for Enterprise

Reconnaissance 10 techniques	Resource Development 8 techniques	Initial Access 10 techniques	Execution 14 techniques	Persistence 20 techniques	Privilege Escalation 14 techniques
Active Scanning (3)	Acquire Access	Content Injection	Cloud Administration Command	Account Manipulation (6)	Abuse Elevation Control Mechanism (6)
Gather Victim Host Information (4)	Acquire Infrastructure (8)	Drive-by Compromise	Command and Scripting Interpreter (10)	BITS Jobs	Access Token Manipulation (5)
Gather Victim Identity Information (3)	Compromise Accounts (3)	Exploit Public-Facing Application	Container Administration Command	Boot or Logon Autostart Execution (14)	Account Manipulation (6)
Gather Victim Network Information (6)	Compromise Infrastructure (8)	External		Boot or Logon Initialization	Boot or Logon

Defense Evasion 43 techniques	Credential Access 17 techniques	Discovery 32 techniques	Lateral Movement 9 techniques	Collectic 17 techniqu
Abuse Elevation Control Mechanism (6)	Adversary-in-the-Middle (3)	Account Discovery (4)	Exploitation of Remote Services	Adversary-in-the-Middle (3)
Access Token Manipulation (5)	Brute Force (4)	Application Window Discovery	Internal Spearphishing	Archive Collected Data (3)
BITS Jobs	Credentials from Password Stores (6)	Browser Information Discovery	Lateral Tool Transfer	Audio Captur
Build Image on Host	Exploitation for Credential	Cloud Infrastructure Discovery	Remote	Automated Collection
Debugger Evasion				

MITRE ATLAS™

ATLAS Matrix

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. & indicates an adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the links at the top navigation bar. View the ATLAS matrix highlighted alongside ATT&CK Enterprise techniques on the ATLAS Navigator.

Reconnaissance& 5 techniques	Resource Development& 7 techniques	Initial Access& 6 techniques	ML Model Access 4 techniques	Execution& 3 techniques	Persistence& 3 techniques	Privilege Escalation& 3 techniques	Defense Evasion& 3 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	LLM Prompt Injection	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access				
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection					
	Poison Training Data	Phishing &					
	Establish Accounts &						

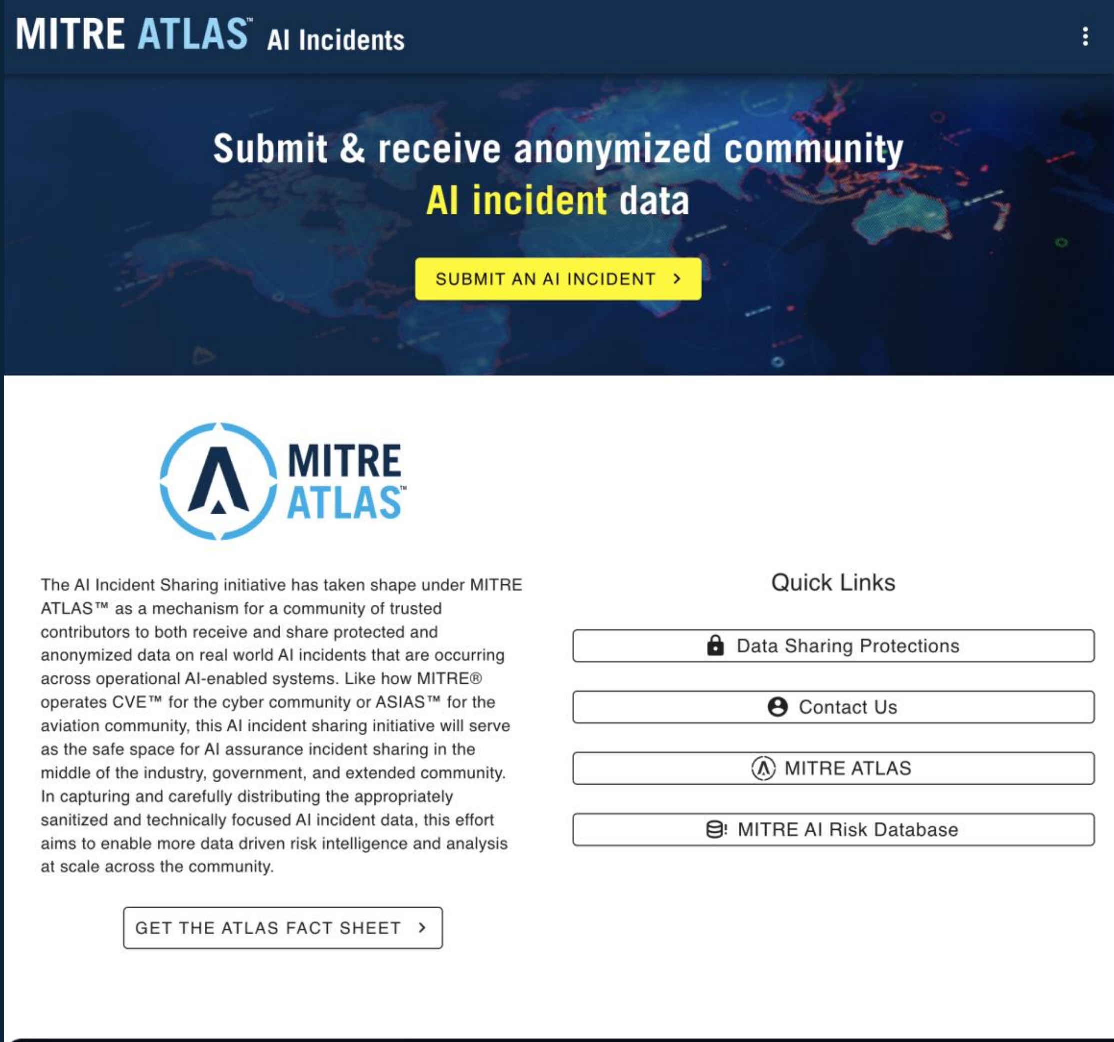
Credential Access& 1 technique	Discovery& 4 techniques	Collection& 3 techniques	ML Attack Staging 4 techniques	Exfiltration& 4 techniques	Impact& 6 techniques
Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
	Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
	Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
	LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
					Cost Harvesting
					External Harms

MITRE | Center for Threat Informed Defense

AI Incident Sharing

Digital, rapid, anonymized community sharing at ai-incidents.mitre.org

- Developed a structured format to collect relevant AI incident information
- Incidents are **shared with MITRE under a data sharing agreement** and stored in a protected database
- Data can be anonymized for **sharing with a trusted community**
- Aggregated data and trends can be visualized in dashboards and **inform AI risk analysis at scale**



MITRE ATLAS™ AI Incidents

Submit & receive anonymized community AI incident data

SUBMIT AN AI INCIDENT ›

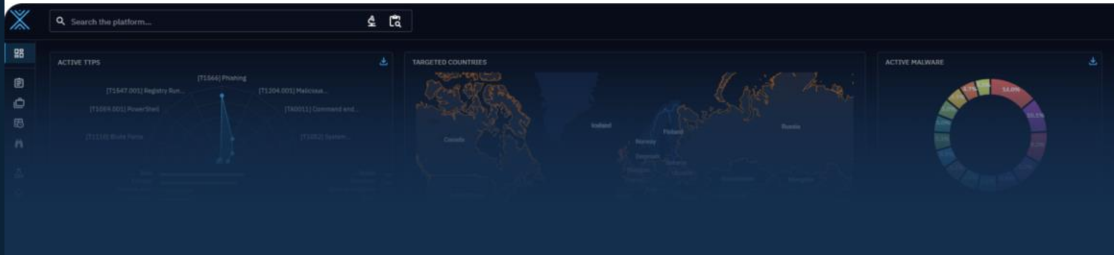
MITRE ATLAS™

The AI Incident Sharing initiative has taken shape under MITRE ATLAS™ as a mechanism for a community of trusted contributors to both receive and share protected and anonymized data on real world AI incidents that are occurring across operational AI-enabled systems. Like how MITRE® operates CVE™ for the cyber community or ASIAS™ for the aviation community, this AI incident sharing initiative will serve as the safe space for AI assurance incident sharing in the middle of the industry, government, and extended community. In capturing and carefully distributing the appropriately sanitized and technically focused AI incident data, this effort aims to enable more data driven risk intelligence and analysis at scale across the community.

GET THE ATLAS FACT SHEET ›

Quick Links

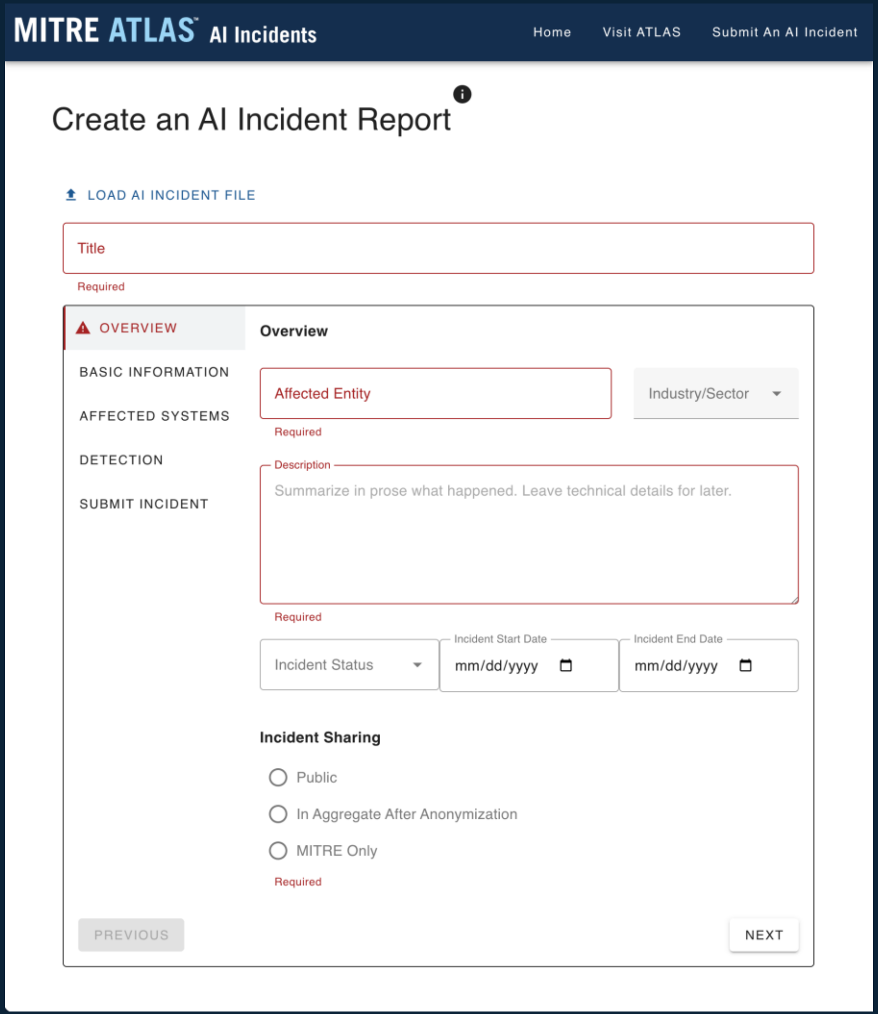
- 🔒 Data Sharing Protections
- Contact Us
- MITRE ATLAS
- MITRE AI Risk Database



The AI Incident Sharing initiative is modeled after traditional cyber intel sharing, leveraging STIX for our data schema, and is inspired by how MITRE® operates CVE™ for the cyber community and ASIAS™ for the aviation community.

AI Incident Sharing

Digital, rapid, anonymized community AI Incident Sharing



MITRE ATLAS™ AI Incidents

Home | Visit ATLAS | Submit An AI Incident

Create an AI Incident Report

LOAD AI INCIDENT FILE

Title

Required

- OVERVIEW
- BASIC INFORMATION
- AFFECTED SYSTEMS
- DETECTION
- SUBMIT INCIDENT

Overview

Affected Entity

Required

Industry/Sector

Description

Summarize in prose what happened. Leave technical details for later.

Required

Incident Status

Incident Start Date: mm/dd/yyyy

Incident End Date: mm/dd/yyyy

Incident Sharing

○ Public

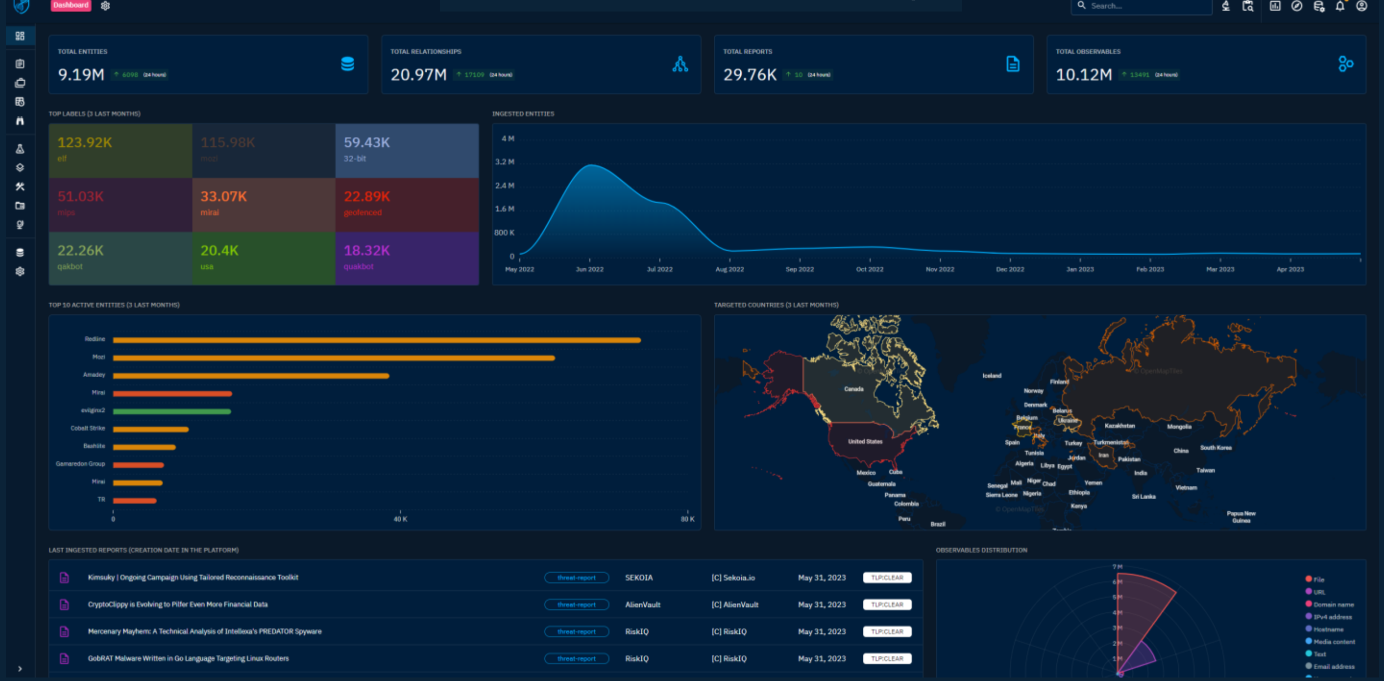
○ In Aggregate After Anonymization

○ MITRE Only

Required

PREVIOUS | NEXT

Notional Dashboard Concept

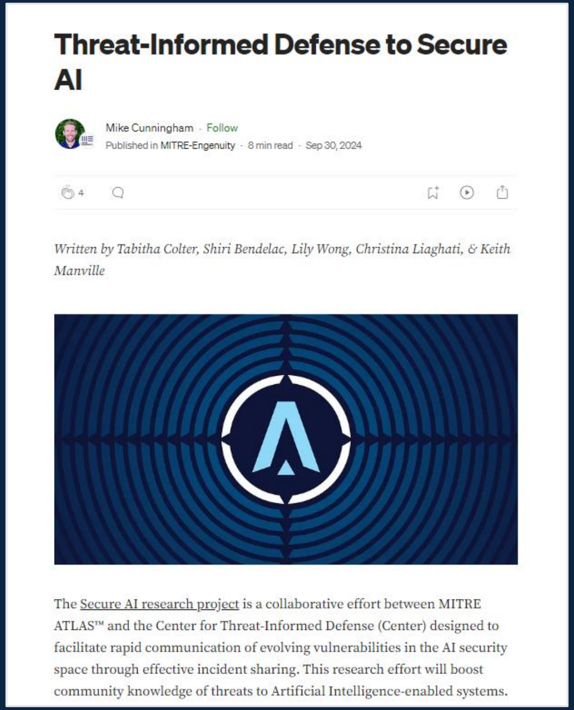


LAST INGESTED REPORTS (CREATION DATE IN THE PLATFORM)

Kimsuky | Ongoing Campaign Using Tailored Reconnaissance Toolkit	threat-report	SEKOIA	[C] Sekoia.io	May 31, 2023	TLP:CLEAR
CryptoClippy is Evolving to Pilfer Even More Financial Data	threat-report	AlienVault	[C] AlienVault	May 31, 2023	TLP:CLEAR
Mercenary Mayhem: A Technical Analysis of Intellexa's PREDATOR Spyware	threat-report	RiskIQ	[C] RiskIQ	May 31, 2023	TLP:CLEAR
GobRAT Malware Written in Go Language Targeting Linux Routers	threat-report	RiskIQ	[C] RiskIQ	May 31, 2023	TLP:CLEAR

In capturing and carefully distributing the appropriately sanitized and technically focused AI incident data, this effort aims to enable more data driven risk intelligence and analysis at scale across the community.

Project Release October 2, 2024



Threat-Informed Defense to Secure AI

Mike Cunningham · Follow
Published in MITRE-Engenuity · 8 min read · Sep 30, 2024

Written by Tabitha Colter, Shiri Bendelac, Lily Wong, Christina Liaghati, & Keith Manville

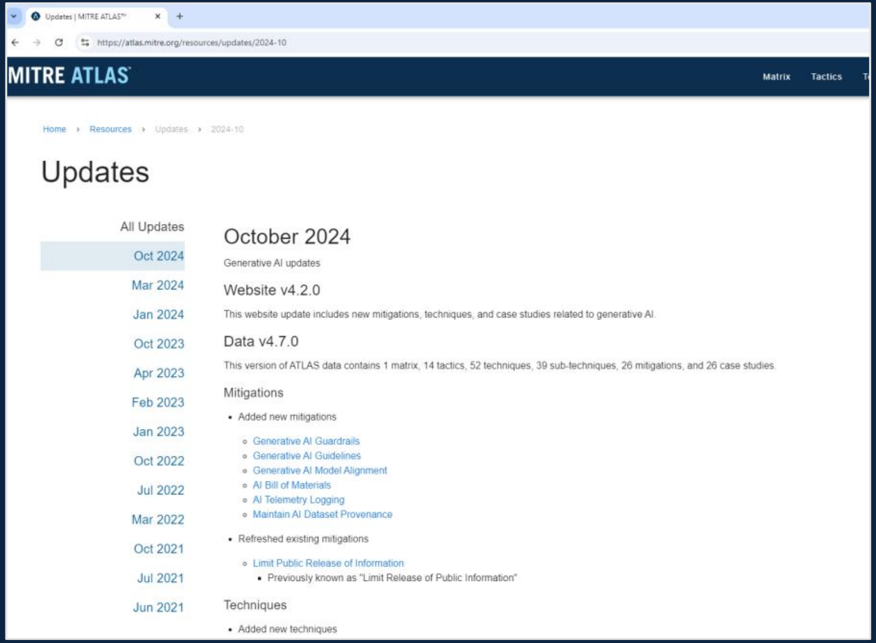
The Secure AI research project is a collaborative effort between MITRE ATLAS™ and the Center for Threat-Informed Defense (Center) designed to facilitate rapid communication of evolving vulnerabilities in the AI security space through effective incident sharing. This research effort will boost community knowledge of threats to Artificial Intelligence-enabled systems.

Key Resources

ctid.io/secure-ai

AI-incidents.mitre.org

atlas.mitre.org



https://atlas.mitre.org/resources/updates/2024-10

Home › Resources › Updates › 2024-10

Updates

All Updates
Oct 2024
Mar 2024
Jan 2024
Oct 2023
Apr 2023
Feb 2023
Jan 2023
Oct 2022
Jul 2022
Mar 2022
Oct 2021
Jul 2021
Jun 2021

October 2024

Generative AI updates

Website v4.2.0

This website update includes new mitigations, techniques, and case studies related to generative AI.

Data v4.7.0

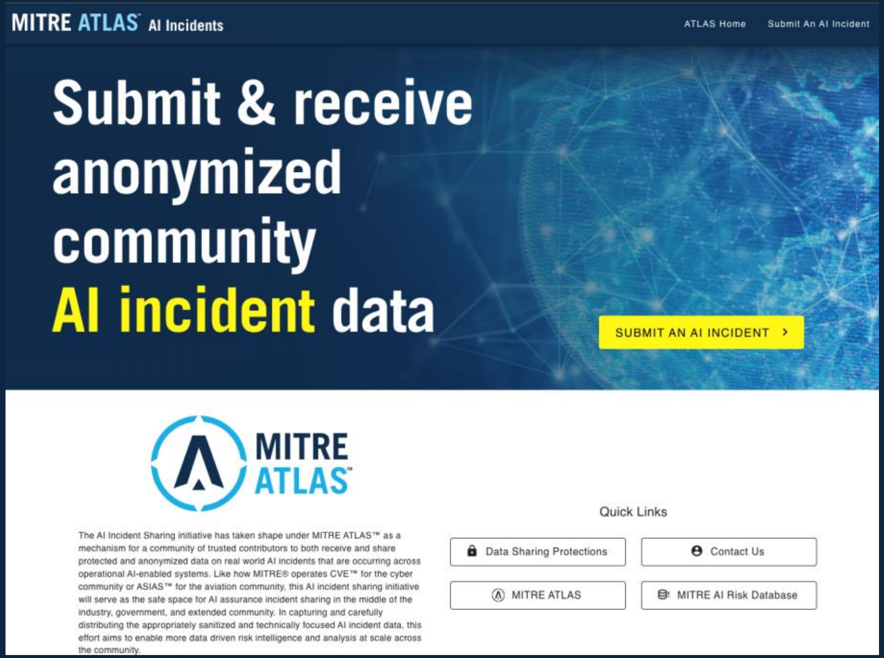
This version of ATLAS data contains 1 matrix, 14 tactics, 52 techniques, 39 sub-techniques, 26 mitigations, and 26 case studies.

Mitigations

- Added new mitigations
  - Generative AI Guardrails
  - Generative AI Guidelines
  - Generative AI Model Alignment
  - AI Bill of Materials
  - AI Telemetry Logging
  - Maintain AI Dataset Provenance
- Refreshed existing mitigations
  - Limit Public Release of Information
    - Previously known as "Limit Release of Public Information"

Techniques

- Added new techniques



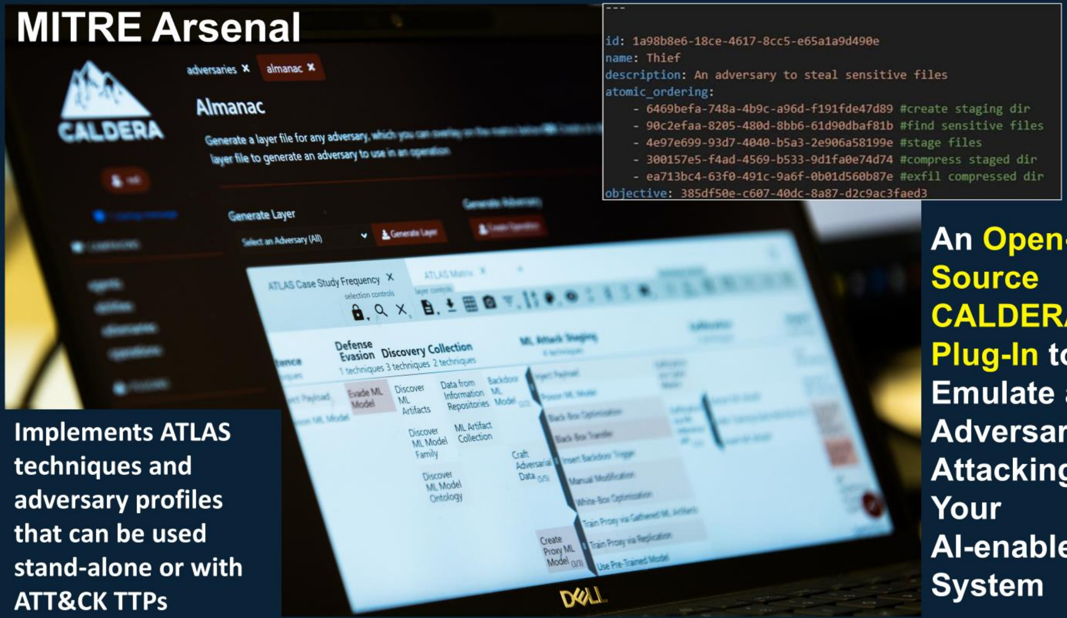
Submit & receive anonymized community **AI incident** data

The AI Incident Sharing initiative has taken shape under MITRE ATLAS™ as a mechanism for a community of trusted contributors to both receive and share protected and anonymized data on real world AI incidents that are occurring across operational AI-enabled systems. Like how MITRE® operates CVE™ for the cyber community or ASIAS™ for the aviation community, this AI incident sharing initiative will serve as the safe space for AI assurance incident sharing in the middle of the industry, government, and extended community. In capturing and carefully distributing the appropriately sanitized and technically focused AI incident data, this effort aims to enable more data driven risk intelligence and analysis at scale across the community.

MITRE | Center for Threat Informed Defense

What’s Next?

The Center will continue to build out the knowledge base and incident sharing mechanisms, collect case studies and feedback, and identify additional directions for future Secure AI research.



MITRE Arsenal

```
---

id: 1a98b8e6-18ce-4617-8cc5-e65a1a9d490e
name: Thief
description: An adversary to steal sensitive files
atomic_ordering:
    - 6469befa-748a-4b9c-a96d-f191fde47d89 #create staging dir
    - 90c2efaa-8205-480d-8bb6-61d90dbaf81b #find sensitive files
    - 4e97e699-93d7-4040-b5a3-2e906a58199e #stage files
    - 300157e5-f4ad-4569-b533-9d1fa0e74d74 #compress staged dir
    - ea713bc4-63f0-491c-9a6f-0b01d560b87e #exfil compressed dir
objective: 385df50e-c607-40dc-8a87-d2c9ac3faed3
```

Implements ATLAS techniques and adversary profiles that can be used stand-alone or with ATT&CK TTPs

An Open-Source CALDERA Plug-In to Emulate an Adversary Attacking Your AI-enabled System

https://github.com/mitre-atlas/arsenal

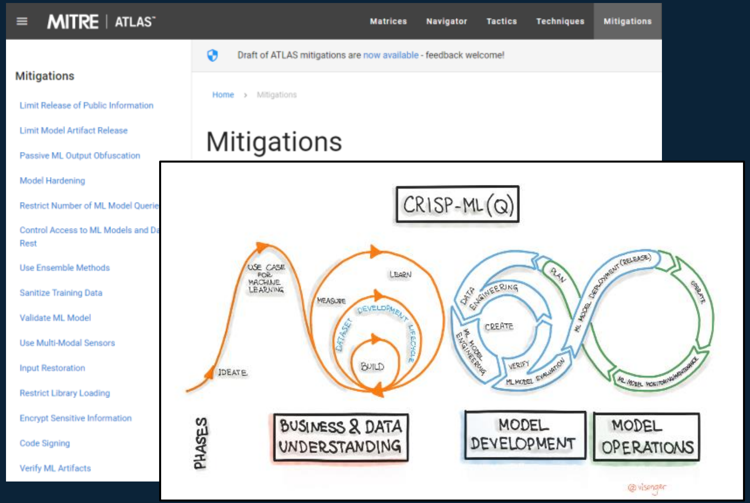
MITRE | Center for Threat Informed Defense

Ongoing ATLAS Efforts

Mitigations

Mitigations can prevent an adversary from the executing techniques in ATLAS

October Update Added Community Inputs
to the 2023 draft release



MITRE | ATLAS™ Matrices Navigator Tactics Techniques Mitigations

Draft of ATLAS mitigations are now available - feedback welcome!

Home › Mitigations

Mitigations

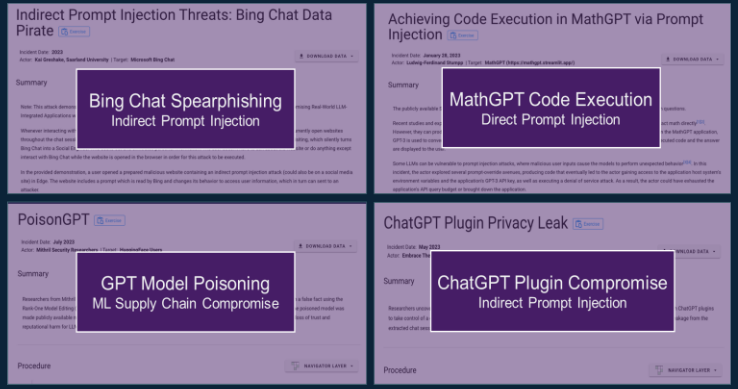
Mitigations
- Limit Release of Public Information
- Limit Model Artifact Release
- Passive ML Output Obfuscation
- Model Hardening
- Restrict Number of ML Model Querie
- Control Access to ML Models and Da Rest
- Use Ensemble Methods
- Sanitize Training Data
- Validate ML Model
- Use Multi-Modal Sensors
- Input Restoration
- Restrict Library Loading
- Encrypt Sensitive Information
- Code Signing
- Verify ML Artifacts

Aligned with CRISP-ML Lifecycle Phases

GenAI Attack Vectors

Updated in Oct 2024

Real World LLM Attack Pathways
Grounded in New Techniques and Case Studies



Indirect Prompt Injection Threats: Bing Chat Data Pirate

Incident Date: 2023
Actor: Kai Greshake, Saarland University | Target: Microsoft Bing Chat

Summary

interact with Bing Chat while the website is opened in the browser in order for this attack to be executed.

In the provided demonstration, a user opened a prepared malicious website containing an indirect prompt injection attack (could also be on a social media site) in Edge. The website includes a prompt which is read by Bing and changes its behavior to access user information, which in turn can sent to an attacker.

Bing Chat Spearphishing
Indirect Prompt Injection

Achieving Code Execution in MathGPT via Prompt Injection

Incident Date: January 28, 2023
Actor: Ludwig-Ferdinand Stumpp | Target: MathGPT (https://mathgpt.streamlit.app/)

Summary

Some LLMs can be vulnerable to prompt injection attacks, where malicious user inputs cause the models to perform unexpected behavior. In this incident, the actor explored several prompt-override avenues, producing code that eventually led to the actor gaining access to the application host system's environment variables and the application's GPT-3 API key, as well as executing a denial of service attack. As a result, the actor could have exhausted the application's API query budget or brought down the application.

MathGPT Code Execution
Direct Prompt Injection

PoisonGPT

Incident Date: July 2023
Actor: Mithril Security Researchers | Target: HuggingFace Users

Summary

GPT Model Poisoning
ML Supply Chain Compromise

Procedure

ChatGPT Plugin Privacy Leak

Incident Date: May 2023

Summary

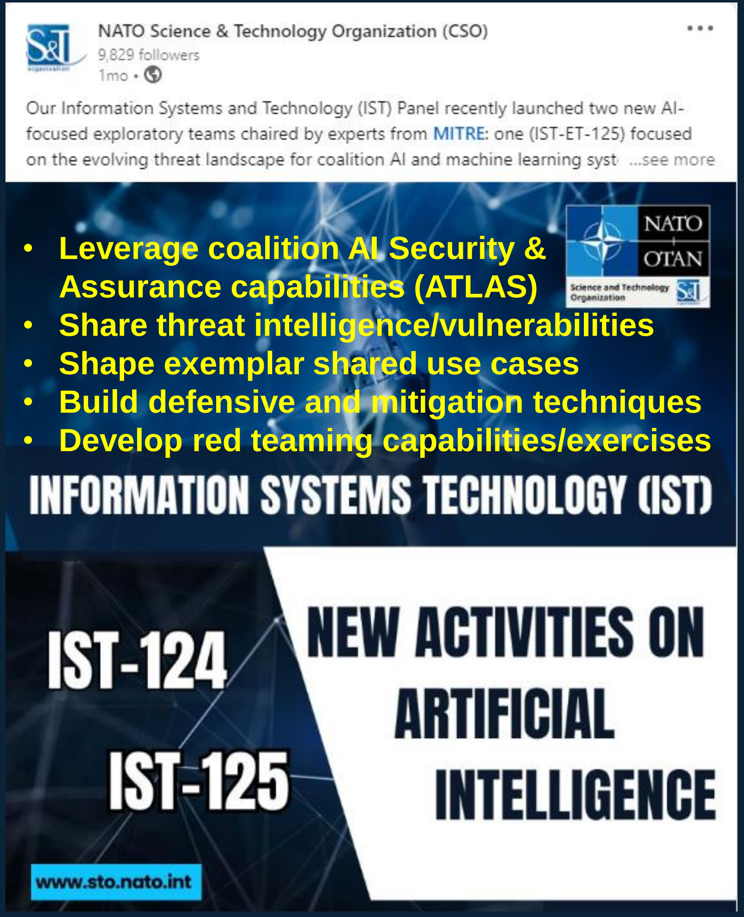
ChatGPT Plugin Compromise
Indirect Prompt Injection

Procedure

Collaboratively Developed with Microsoft, Intel, Verizon, CrowdStrike and 10+ other orgs

NATO Task Group

NATO RTG Launched in May



NATO Science & Technology Organization (CSO)
9,829 followers
1mo •

Our Information Systems and Technology (IST) Panel recently launched two new AI-focused exploratory teams chaired by experts from MITRE: one (IST-ET-125) focused on the evolving threat landscape for coalition AI and machine learning syst ...see more

- Leverage coalition AI Security & Assurance capabilities (ATLAS)
- Share threat intelligence/vulnerabilities
- Shape exemplar shared use cases
- Build defensive and mitigation techniques
- Develop red teaming capabilities/exercises

NATO
OTAN
Science and Technology Organization
INFORMATION SYSTEMS TECHNOLOGY (IST)
IST-124
IST-125
NEW ACTIVITIES ON ARTIFICIAL INTELLIGENCE
www.sto.nato.int

CVE and CWE

Engaging with both the Common Vulnerability Enumeration (CVE) and Common Weakness Enumeration (CWE) communities on AI Vulnerabilities & Weaknesses

- **CVE AI WG** – Working with the CVE Board and AI WG to provide more clarity on how AI security vulnerabilities will fall inside/outside CVE scope via a series of blog posts.

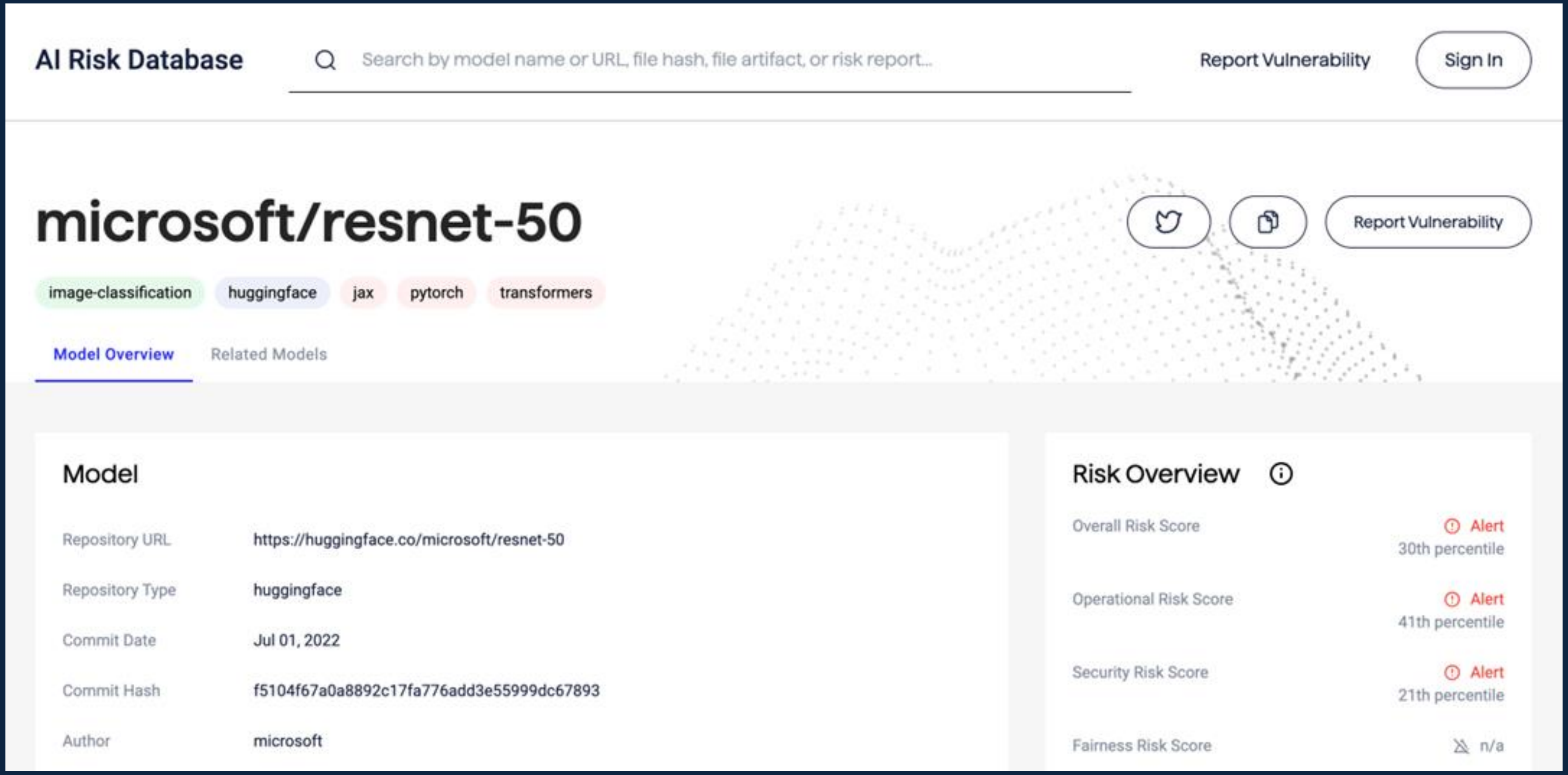
- **CWE AI WG** – Working with the AI Working group on AI-related updates
  - CWE-1426: Improper Validation of Generative AI Output
  - A new demonstrative example for “prompt injection” was added to CWE-77: Improper Neutralization of Special Elements used in a Command ('Command Injection').
  - New observed examples were added to multiple CWEs related to AI/ML and generative AI prompts, including one example of “prompt injection.”

https://www.cve.org/Media/News/item/blog/2024/07/09/CVE-and-AIrelated-Vulnerabilities

https://cwe.mitre.org/news/archives/news2024.html#july16_CWE_Version_4.15_Now_Available

AI Risk Database

Inspired by VirusTotal, we are building on that vision as we shape a long-term expansion plan.



AI Risk Database
Search by model name or URL, file hash, file artifact, or risk report...
Report Vulnerability
Sign In
Report Vulnerability

microsoft/resnet-50

image-classification | huggingface | jax | pytorch | transformers

Model Overview | Related Models

Model

Repository URL	https://huggingface.co/microsoft/resnet-50
Repository Type	huggingface
Commit Date	Jul 01, 2022
Commit Hash	f5104f67a0a8892c17fa776add3e55999dc67893
Author	microsoft

Risk Overview

Overall Risk Score	Alert 30th percentile
Operational Risk Score	Alert 41th percentile
Security Risk Score	Alert 21th percentile
Fairness Risk Score	n/a

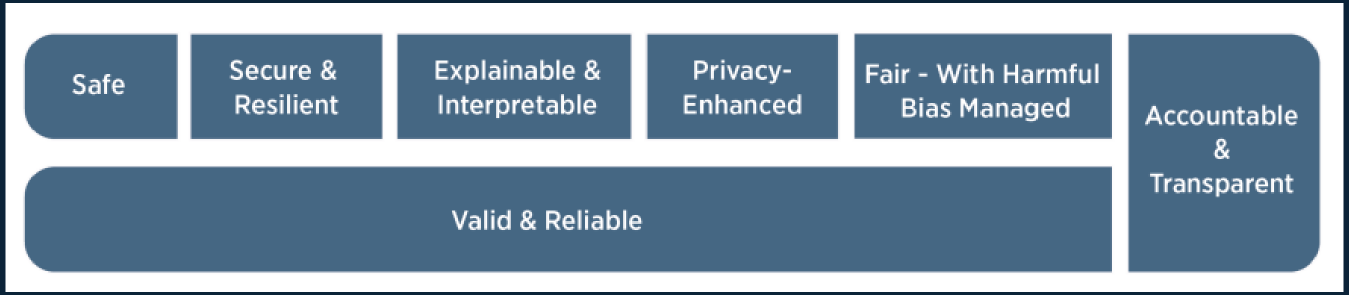
We would love to get your feedback and input on where it should go/what would be most helpful to you!

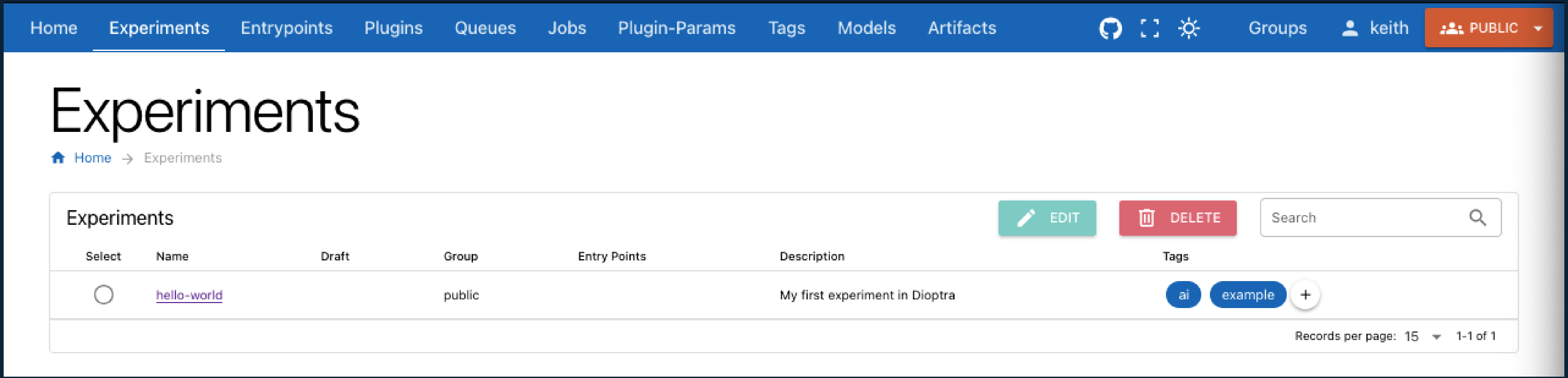
MITRE | Center for Threat Informed Defense

Dioptra – Test Platform for Trustworthy AI

Dioptra is NIST’s software test platform for assessing the trustworthy characteristics of artificial intelligence.

Dioptra **supports the Measure function of NIST’s AI Risk Management Framework** by providing functionality to assess, analyze, and track identified AI risks.


Safe
Secure & Resilient
Explainable & Interpretable
Privacy-Enhanced
Fair - With Harmful Bias Managed
Accountable & Transparent
Valid & Reliable



Experiments

Experiments

Select	Name	Draft	Group	Entry Points	Description	Tags
○	hello-world		public		My first experiment in Dioptra	ai example +

Records per page: 15 1-1 of 1

https://github.com/usnistgov/dioptra

Plans for Dioptra + AI Risk Database

Dioptra can be used to evaluate AI models and submit reports to the AI Risk Database

- Help standardize vulnerability reports and metrics
- Allow AIRDB users to verify and validate results
- Promote sharing of experiment templates
- Provide plugins for a variety of evaluations – security and beyond

Looking for feedback and ideas

- Dioptra v1.0.0 was released in July of 2024
- https://github.com/usnistgov/dioptra

The Center for Threat-Informed Defense conducts collaborative R&D projects that improve cyber defense at scale

ANOMALI
Analysis & Resilience Center FOR SYSTEMIC RISK
ATTACKIQ
BANK OF AMERICA
BEDROCK Systems Inc
Booz | Allen | Hamilton
CATO NETWORKS
CIS. Center for Internet Security
citi
CROWDSTRIKE
CYBER THREAT ALLIANCE
ENSIGN INFOSECURITY
FIRST
FIS
FM Global
FORTINET
FS-ISAC
FUJITSU
GLOBAL CYBER ALLIANCE
GLOBAL RESILIENCE FEDERATION
Google Cloud
HCA Healthcare
Health-ISAC
IBM Security
infineon
intel
JPMORGAN CHASE & CO.
LLOYDS BANKING GROUP
Microsoft
nab
next www.nextdlp.com
NRF NATIONAL RETAIL FEDERATION
RETAIL & HOSPITALITY ISAC
SAFE
SIEMENS
standard chartered
tenable
verizon

+ MITRE

Membership is:

- ✓ Highly-sophisticated
- ✓ Global & cross-sector
- ✓ Non-governmental
- ✓ Committed to collaborative R&D in the public interest

https://ctid.io/our-work

Mission: Advance the state of the art and the state of the practice in threat-informed defense globally.

MITRE | Center for Threat Informed Defense

Advance Secure AI

We would love to have you involved in the next Secure AI project!

AI Incident Sharing

- Beta test the incident submission system and submit your incidents/successful red teaming exercises.
- Provide feedback on the kind of information/data your org would want to receive to mitigate incidents.

AI Risk Database

- Beta test the new vulnerability submission system and provide feedback that will help us shape the next version (including the AI BOM).
- Shape the combination of opensource/public vulnerability detection tools built around the NIST Dioptra tool as our centerpiece

ATLAS Matrix

- Contribute to the bi-annual major update: send us your mitigations, case studies, and/or feedback on the tactics and techniques in the matrix.

Join us and change the game!

Changing the game on the adversary requires a community-wide approach.

You play a critical role!

https://ctid.io/get-involved

MITRE | Center for Threat Informed Defense