MITRE | SOLVING PROBLEMS FOR A SAFER WORLD™

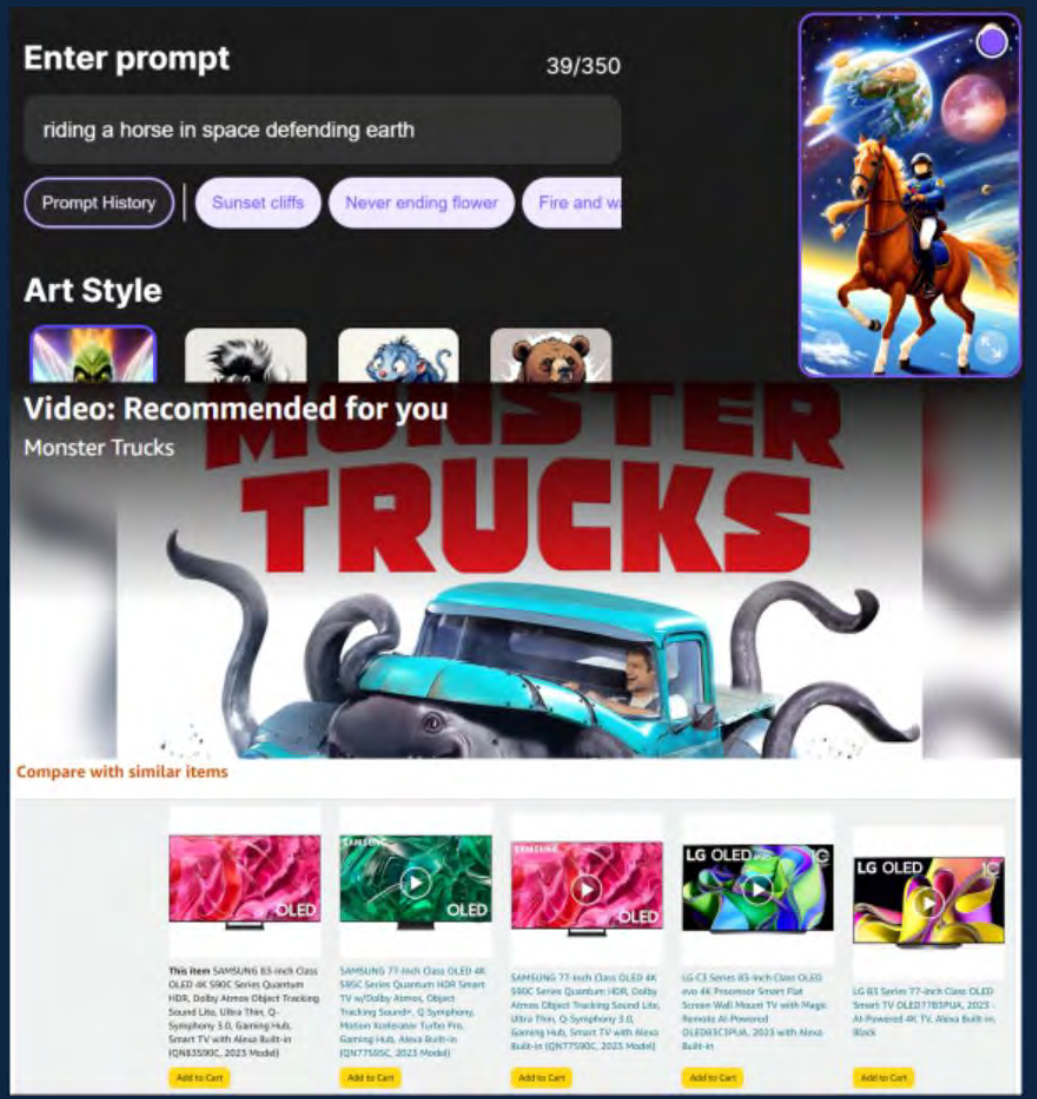
August 2024

AIAD@mitre.org

The Challenge

Catalyze consequential AI use

≠



Consequential AI use can lead to national impact and transform society for the better but also poses systematic risks

The Need for the AI Assurance and Discovery Lab

- **Mission needs necessitate rapid AI adoption** to increase effectiveness and efficiency with limited resources
- **Hype around tech breakthroughs** complicate risk assessments and adoption decisions
- AI advancements can **clash with human values and raise ethical concerns**, which should be explored in controlled settings
- AI developers and governments are concerned about **not fully understanding the capabilities** of foundational models
- Regulators are under pressure to act to ensure **assured uses of AI without stifling innovation**
- US Government is **ordering assured uses of AI**
  - Executive Order 14110 (October 30, 2023)
  - OMB Memo M-24-10 (March 28, 2024)

MITRE

Lab Mission

Proactively discover and mitigate mission-critical risks in AI-enabled systems.

Lab Vision

Mission owners comprehend the risks of AI-enabled systems, make informed AI adoption decisions, and manage risks accordingly to maximize value from AI. Labs across the nation form a network of AI Assurance capabilities for public good.

MITRE

AI Assurance Scope

MITRE's Definition of AI Assurance

AI assurance as a process for discovering, assessing, and managing risk throughout the lifecycle of an AI-enabled system so that it operates effectively to the benefit of its stakeholders.

AI ASSURANCE RISK GROUPS

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

There are many lenses for AI assurance and we adapt our approach to ensure it covers the risks that are important to the mission.

MITRE

AI Assurance Services

AI Assurance Discovery

Given an expected or imminent use case and the technology in an AI enabled system (AIES), exercise the risk landscape to understand the risks and value to the stakeholders of the AIES. Initialize an **AI Assurance Plan** coupled to the intended mission and set practical milestones for completing the AI Assurance Process.

AI Assurance Evaluation

Evaluate the risks and value of implementing the AIES to understand their likelihood and severity of impacts. Reaching quantitative results may entail hosting a sandbox at MITRE, validation testing against internal or synthesized datasets, and human-in-the-loop exploration.

AI Assurance Management

Based on risk evaluation results, provide risk mitigation strategies, monitoring requirements, and detailed suggestions for AI Governance. The plan will include prescriptive guidance for inevitabilities of the AIES like model drift and require the continued execution of the AI Assurance Process to maintain AI Assurance.

AI Assurance Plan Development

MITRE

AI ASSURANCE PROCESS

Position Paper → [AI Assurance: A Repeatable Process for Assuring AI-enabled Systems]

MITRE | AI Assurance & Discovery Lab

Lab AI Assurance Capabilities (1/3)

Risk Discovery Protocol for AI Assurance

Provides risk awareness for consequential applications of AI

Navigate the AI assurance landscape and compare/contrast the desired application with similar use cases to prioritize risks

AI Assurance Protocol

Experts implement the AI Assurance Process for your project to manage risks in consequential applications

Commercial-off-the-Shelf AI Assurance Tool Exploration

The COTS Exploration Protocol and Environments allow MITRE to investigate the value of commercially available AI Assurance tools

Human-Centered AI Test Harness

AI-enabled systems need to work with humans so we need a capability for experimenting with human-AI interaction

The AI Test Harness is a portable, web-based automated measurement platform for human-in-the-loop research and evaluation

Human-in-the-loop Experimentation Environment

Complexity of human-AI interactions is driving new methods of measurement. The Lab provides two fully instrumented experiment rooms for simulating and observing human-AI interactions in mission contexts that can be synchronized with the AI Test Harness

Lab AI Assurance Capabilities (2/3)

AI Assurance Knowledge Base

Provides information to an AI assurance investigator on AI Assurance use cases, metrics, datasets, methodologies, and tools that are relevant to their assurance goals.

AI Red Teaming Guide

Best practices on how to conduct AI red teaming, an investigative process that simulates adverse conditions on real-world AI enabled systems to identify vulnerabilities, mitigate potential exploits, and improve the overall security posture and robustness of an AI-enabled system.

Assurance Plan Templates and Development Protocols

Tools to facilitate the creation of an assurance plan and adoption of a development plan that will result in an assured AI-enabled system.

Acquisition RFI Analysis Tool

An LLM-enabled tool that helps acquisition staff better understand and process RFIs and their responses. As such, the tool can be used by experts to identify, analyze, and augment RFI sections specific to AI assurance that should be driving AI-enabled system acquisitions.

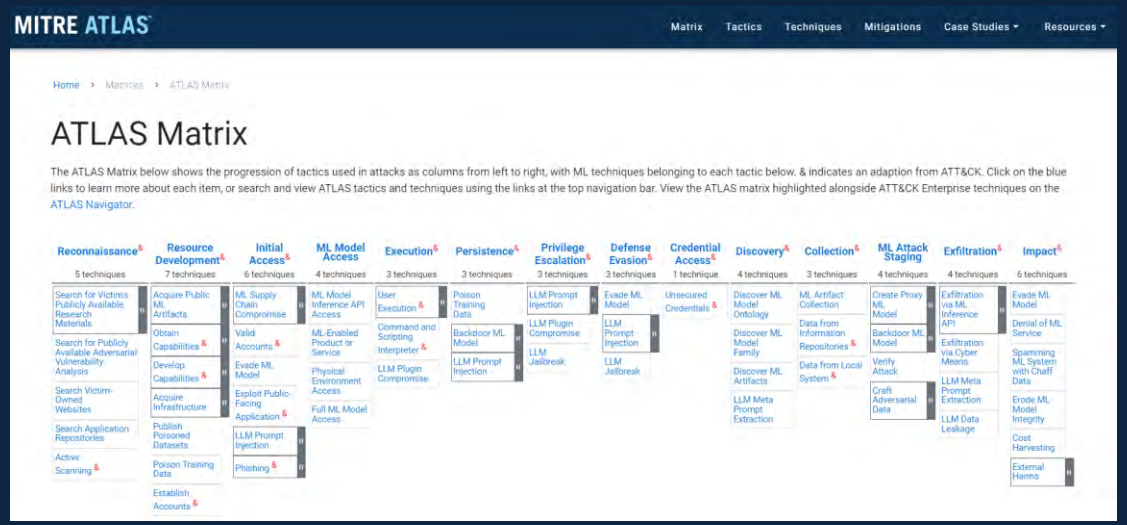
Lab AI Assurance Capabilities (3/3)

Large Language Model (LLM) Secure Integrated Research Environment (SIREN)

Provides an environment to execute research and rapidly prototype assured LLM-based solutions aligned to mission use-cases, allowing the Lab to safely and securely work the AI Assurance Process with LLMs

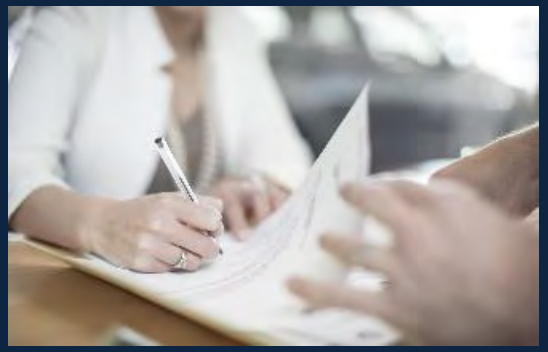
Adversarial Threat Landscape for AI Systems (ATLAS)

A globally accessible, living knowledge base of adversary tactics and techniques based on real-world attack observations and realistic demonstrations from AI red teams and security groups, the Lab leverages ATLAS to assist in red teaming and risk discovery



AI Governance Toolkit

Understanding what is necessary to maintain an accountable AI system that proactively supports its mission, the Lab works with the AI Governance team to build comprehensive Assurance Plans that can be governed



Compute Resources Available to the AIAD Lab

AI Platform

Virtualized GPU cluster for AI development, prototyping, and deployment

HPCs

GPU enabled job cluster with adjacent storage for training data

Federal AI Sandbox

248 H100 GPUs (1 exaFLOPS) that enables the training of Federal foundation models

Completed Use Case

AI-Enabled Augmented Reality Microscope

To increase accuracy and reduce the time required for each cancer diagnosis, a commercial company integrated AI into a microscope, displaying results with an augmented reality interface.

Our team investigated common microscopy-related pathology activities and developed software that alters imagery to evaluate the performance impact during real-world use. This included varying focus, lighting, chromatic aberration (colored distortion from lenses), and vignetting (a dark halo obscuring cells) which resulted in problems identifying potentially cancerous cells.

The company was provided a report detailing risks and potential mitigations, with the suggestion that the technology be further refined before being used in a clinical setting.

MITRE

Completed Use Case

ID Verification

AI can help expedite processing at airports, transportation hubs, and other environments where driver's licenses, passports, and other forms of ID are checked.

Our assessment of ID verification systems identified several potential risks:

- Inconsistent capture of correct imagery of documents and faces
- Unequal face verification performance across varying demographic groups
- Lack of transparency about how verification decisions are made
- Mismatch between expectations and the reliability of automated authentication, leading to low utilization or over-reliance

The review also found that ID document review technologies are vulnerable to falsified data, which may be mitigated by implementing online database retrieval for high-security applications.

MITRE

Completed Use Case

Healthcare Mobile Robot

Delegating routine, simpler tasks to autonomous systems in a healthcare setting lets providers focus on more complex work like diagnosing a patient or performing surgery.

We procured a general-purpose robot platform and installed MITRE-developed software for autonomy and contact-less measurement of vital signs. The robot finds the patient's room, verifies patient identity, gets into position to scan for vitals, records information, and returns to its starting point.

We identified 58 risks and prioritized them. The two highest-priority hazards were:

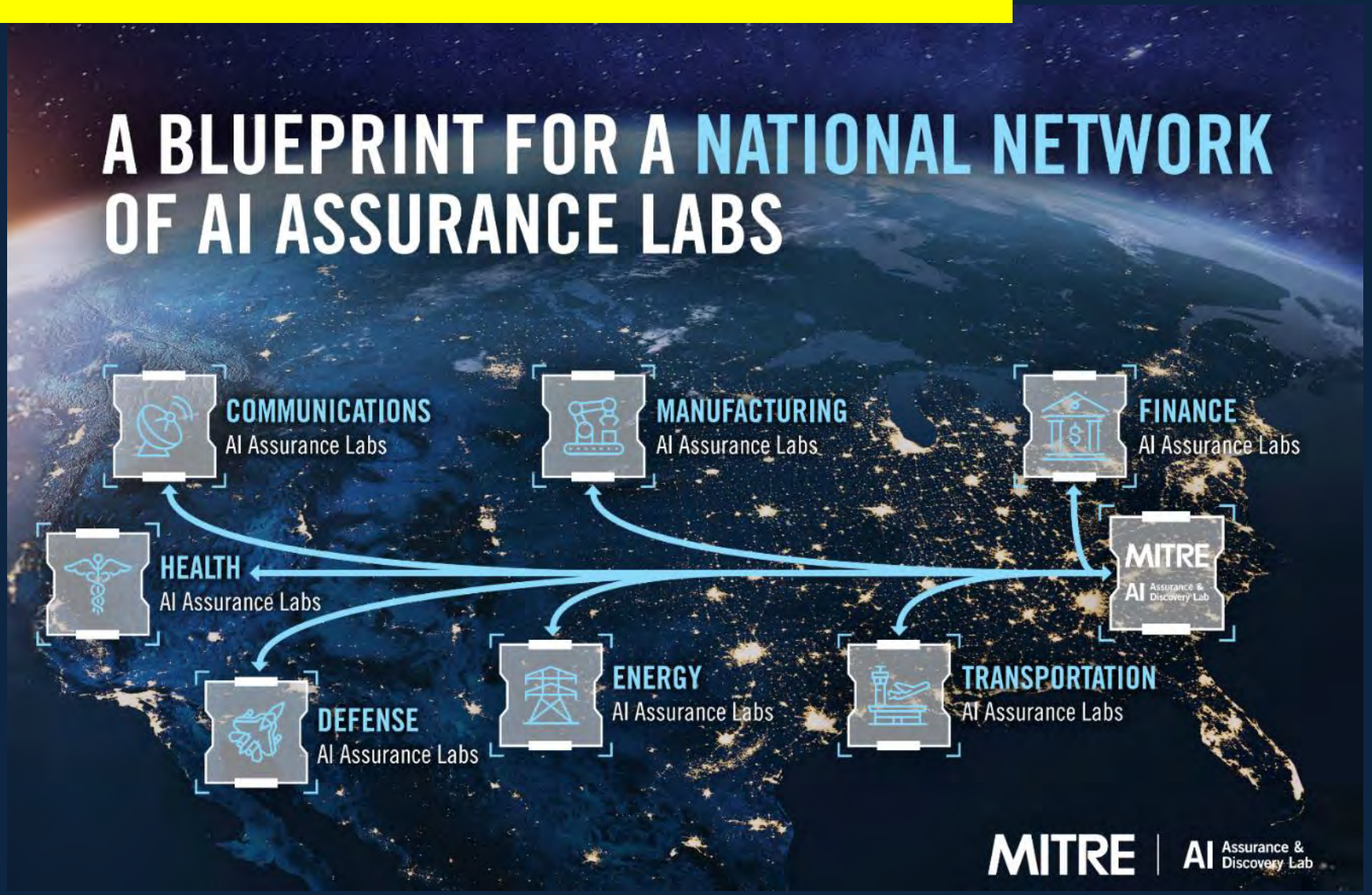
Patient misidentification

- We decreased misidentification risk by using multiple authentication methods like those used in healthcare settings (e.g., wristband checks, name confirmation, date of birth)

Localization failure (getting lost)

- We added additional localization software to increase the robot's ability to navigate correctly

MITRE's AIAD Lab as a Blueprint for other AI Assurance Labs



AIAD Lab Ribbon Cutting
25 March 2024

Healthcare AI Assurance Lab at UMass from Blueprint
10 April 2024

MITRE