MITRE Launches AI Incident Sharing Initiative

OCT 2, 2024

ARTIFICIAL INTELLIGENCE

MITRE, Industry Collaborate to Improve Collective AI Defenses

McLean, Va., Oct. 2, 2024—MITRE's Center for Threat-Informed Defense (Center) collaborated with more than 15 companies to increase community knowledge of

threats and defenses for AI-enabled systems, culminating with the launch of the AI Incident Sharing initiative.

Work on the incident sharing initiative is part of the Center's Secure AI project, which is built around MITRE ATLAS™ and kicked off in June 2024. The AI Incident Sharing initiative aims to improve the collective awareness of threats, and ultimately defense of AI-enabled systems, by enabling the rapid and protected sharing of information about attacks or accidents that involved AI-enabled systems. The initiative builds on two years of incident sharing collaboration across the broader MITRE ATLAS community to enable more rapid characterization and sharing of anonymized incidents.

In parallel, the Secure AI collaboration also extended the ATLAS threat framework to further update the adversarial threat landscape for generative AI-enabled systems. Much like MITRE ATT&CK™, the ATLAS threat framework is a community knowledge base of adversary behaviors that security professionals, developers, and operators use as they protect AI-enabled systems.

The project added several new generative AI-focused case studies and attack techniques to the public ATLAS knowledge base as well as new methods to mitigate attacks on AI-enabled systems. MITRE had previously collaborated with Microsoft to begin increasing the ATLAS knowledge base focus on generative AI with updates released in November 2023. This effort continues to keep ATLAS representative of the latest demonstrated threats to AI systems in the wild.

Collaborators on the Secure AI project include: AttackIQ, BlueRock, Booz Allen Hamilton, CATO Networks, Citigroup, Cloud Security Alliance, CrowdStrike, FS-ISAC, Fujitsu, HCA Healthcare, HiddenLayer, Intel, JPMorgan Chase Bank, Microsoft, Standard Chartered, and Verizon Business.

"As public and private organizations of all sizes and sectors continue to incorporate AI into their systems, the ability to manage potential incidents is essential," said Douglas Robbins, vice president, MITRE Labs. "Standardized and rapid information sharing

about incidents will allow the entire community to improve the collective defense of such systems and mitigate external harms."

Under the MITRE ATLAS AI Incident Sharing initiative, a community of trusted contributors receive protected and anonymized data on real-world AI incidents that are occurring across operational AI-enabled systems.

Anyone can submit an incident via the public incident sharing site. Upon submission, their organization will be considered for membership in the trusted community of data receivers. Sharing and receiving this protected information will enable more data-driven risk intelligence and analysis at scale across the community.

MITRE operates other information-sharing public private partnerships including the publicly available Common Vulnerabilities and Exposures (CVE™) list, which it operates on behalf of the Cybersecurity and Infrastructure Security Agency to identify, define, and catalog publicly disclosed cybersecurity vulnerabilities. MITRE also takes this approach with the Aviation Safety Information Analysis and Sharing (ASIAS) database for sharing data and safety information to identify and prevent hazards in aviation.

The AI incident sharing website and submission form is available at https://ai-incidents.mitre.org/.

About MITRE

MITRE's mission-driven teams are dedicated to solving problems for a safer world. Through our public-private partnerships and federally funded R&D centers, we work across government and in partnership with industry to tackle challenges to the safety, stability, and well-being of our nation.

About The Center for Threat-Informed Defense

The center is a non-profit, privately funded research and development organization operated by MITRE Engenuity™, a tech foundation for public good. The center's

mission is to advance the state of the art and the state of the practice in threat-informed defense globally. Comprised of participant organizations from around the globe with highly sophisticated security teams, the center builds on MITRE ATT&CK®, an important foundation for threat-informed defense used by security teams and vendors in their enterprise security operations. Because the center operates for the public good, outputs of its research and development are available publicly and for the benefit of all. For more information, contact ctid@mitre-engenuity.org.

Media Contacts: Jeremy Singer, media@mitre.org