PRESIDENTIAL TRANSITION: PRIORITY TOPIC MEMO
JULY 2024

MITRE | Center for Data-Driven Policy

AI RED TEAMING: ADVANCING SAFE AND SECURE AI SYSTEMS

AI systems are uniquely vulnerable to novel threats, which can severely impact critical national systems and public services. These include hallucination, mimicry, and extraction of sensitive information. Bad actors can exploit these vectors, leading to identity theft, fraud, loss of life and property, and the erosion of public trust. To counter adversarial attacks on AI systems, we must institute recurring AI red teaming, which employs adversarial thinking to both identify exploitable AI systems' vulnerabilities and allow the AI community to counter those threats before they occur.

> AI's potential as a powerful dual-use technology means its adoption will be swift and pervasive throughout industry and government.

The Case for Action

AI's potential as a powerful dual-use technology means its adoption will be swift and pervasive throughout industry and government. AI's rapid and widespread adoption will be paced by the speed of AI's development and advancement, which is unprecedented. These factors together will create an environment where AI systems in use by government and industry possess a broad "attack surface," vulnerable to the malign actions of our nation's adversaries at the state and non-state levels. The exploitation of these vulnerabilities affects critical public services and public trust in institutions.

While current industry practices[1,2,3,4] and prior executive action[5,6] have underscored the importance of AI red teaming, more work must be done to institutionalize AI red teaming as an indispensable part of AI assurance.[7] The incoming administration should prioritize continuous AI red teaming efforts during the development of AI applications, during their deployment, and in their ongoing use, to increase public trust in AI, mitigate the significant and proven risk of adversarial exploitation, and safeguard the United States' critical infrastructure, both civilian and military.

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.

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Background and Key Challenges

While the majority of the AI community agrees on the importance of securing AI-enabled systems (e.g., via secure development practices, incident and vulnerability sharing/reporting, data security practices), determining when and how to conduct AI red teaming varies based on the type of AI, its development or deployment stage, and its context of use. However, regardless of an AI system's proposed or actual use, the incoming administration would be well-served to leverage and expand on red teaming best practices from the cybersecurity and infrastructure communities when considering how to mitigate the risks of using AI in consequential ways.

By correctly incorporating red teaming into AI development, deployment, and operational processes, organizations can work to improve AI's overall robustness and security and mitigate risks prior to initial release and during operation. However, the challenge of red teaming is that it must be conducted deliberately and be properly resourced to be effective. This challenge is complicated by the fact that this new attack surface extends to AI-enabled systems as a whole, and not just the AI components within the system—this system-of-systems-level attack surface is rapidly evolving and includes the humans in the loop interacting with AI and every part of the AI supply chain[8] and lifecycle.

Data-Driven Recommendations

Below are steps the administration can take to support AI red teaming.

1. **MANDATE AI RED TEAMING BE PERFORMED BY INDEPENDENT PARTIES ON HIGH-RISK AI SYSTEMS[9] PRIOR TO EXECUTIVE BRANCH ACQUISITION (AND DURING THEIR SUBSEQUENT ADOPTION AND OPERATION), WHILE DIRECTING APPROPRIATE GOVERNMENT AGENCIES (E.G., NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY [NIST]) TO DEVELOP STANDARDS THAT MANDATE USE OF AI RED TEAMING FOR UNITED STATES GOVERNMENT AND RELEVANT INDUSTRIAL CONTRACTORS.**
   The administration exercises critical influence over private industry and its standards via the government's agencies and acquisitions activities. The administration also exercises significant controls over U.S. government adoption and use of its own AI systems. These authorities represent tools the administration can leverage, immediately.

2. **LEAD BY EXAMPLE, MAKING REGULAR, CONTINUOUS USE OF AI RED TEAMING THROUGHOUT THE LIFECYCLE OF U.S. GOVERNMENT–ADOPTED AI SYSTEMS TO ENSURE CONTINUED SAFETY AND SECURITY.**
   New vulnerabilities are likely to be discovered in both new and existing AI systems, making continued AI red teaming imperative. AI systems are not static pieces of software—they will require iterative updates and patches to maintain a sufficient level of safety and security. AI systems are not generically "safe" following a red teaming effort and mitigation of identified vulnerabilities, but rather "safer" in the risk areas investigated. In this vein, implement documentation requirements such as an AI bill of materials to transparently document the origins, development, training, assurance plan,[10] operation, and use of AI models and data sets, thereby enhancing the assurance of AI systems procured, developed, or used by the government.

3. **ENCOURAGE TRANSPARENCY AND TRUST IN AI-ENABLED SYSTEMS USED BY THE U.S. GOVERNMENT THROUGH THE RELEASE OF PUBLICLY ACCESSIBLE AI RED TEAMING, TESTING, AND ASSURANCE REPORTS.**
   MITRE recommends the administration, where appropriate, make public AI red teaming, testing, and assurance reports, to increase public awareness and trust in AI-enabled systems employed by the government. Sharing this information will strengthen AI assurance across the board, while advancing consequential uses of AI. MITRE further recommends the creation of a National AI Center of Excellence (NAICE)[11] and, within it, an AI Information Sharing and Analysis Center (AI-ISAC), enabling public and private stakeholders to work together in understanding and mitigating these AI security and assurance risks.

4. **ADOPT AN AI SCIENCE AND TECHNOLOGY INTELLIGENCE (AI S&TI) APPROACH TO AI SECURITY, AND, BY EXTENSION, AI RED TEAMING.**
   The administration should call on relevant agencies to identify public and private sector AI red teaming experts, whose independent red teaming of AI systems can validate appropriate AI behaviors and safeguards, while producing verifiable and independent reporting on issues affecting AI assurance. This AI S&TI effort is integral to the proper functioning of the the NAICE and AI-ISAC, referenced above.

Implementation Considerations

As the administration considers implementing the recommendations[12] in this paper, the following suggested timeline and milestones may be helpful.

FIRST 100 DAYS

Evaluate existing AI red teaming capabilities across the federal government and industry to identify "centers of excellence," for later informing and working strategically with the NAICE if created. Begin to establish and preview coming mandates for government and industrial contractors for AI red teaming required during the development, adoption, and ongoing use of AI systems in all U.S. agencies and departments.

FIRST SIX MONTHS

Direct federal agencies to implement independent AI red teaming and report on their efforts, sharing AI red teaming results as broadly as possible within the federal government and, where feasible, among trusted industrial partners. Launch the NAICE and nominate governmental oversight offices for AI red teaming reporting within the NAICE (e.g., the Office of the Director of National Intelligence for the Intelligence Community; the Defense Threat Reduction Agency, in consultation with the Chief Digital and Artificial Intelligence Office, for the Department of Defense; and a joint NIST and Department of Commerce working group for all other U.S. government civilian agencies and departments), which will be answerable to the administration on government-wide implementation and results of AI red teaming and security.

FIRST YEAR

Create an AI-ISAC within the NAICE and, via this channel, publish as much as is feasible about the administration's AI red teaming efforts to increase public trust in government. Direct that AI red teaming and security should be built into AI governance policies in all agencies and departments. Direct the NAICE to develop and mature AI red teaming and security tradecraft and standards, informed by industry best practices and U.S. government lessons learned. In this vein, consider whether the rapidly growing AI security community can be rallied around common methodologies and standards, like MITRE's ATLAS™.[13] Empower the NAICE to collect best practices across the AI community, industry, and government—blending the collected information into U.S. government standards, directives, and regulations.

ONGOING

Continuously monitor and implement AI red teaming and security throughout the lifecycle of AI-enabled systems in use by the U.S. government. Ensure all AI-centric task forces, working groups, and oversight bodies are staying abreast of technological developments to guard against complacency. Ensure access to resources for continuous learning. Continue to foster strong relationships with stakeholders and promote continuous learning within the administration and among career staffers.

MITRE Resources and Support

C. Clancy et al. A Sensible Regulatory Framework for AI Security. 2023. MITRE, https://www.mitre.org/sites/default/files/2023-06/PR-23-1943-A-Sensible-Regulatory-Framework-For-AI-Security_0.pdf.

D. Robbins et al. AI Assurance – A Repeatable Process for Assuring AI-Enabled Systems. 2024. MITRE, https://www.mitre.org/sites/default/files/2024-06/PR-24-1768-AI-Assurance-A-Repeatable-Process-Assuring-AI-Enabled-Systems.pdf.

MITRE's Response to the OMB RFI on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence Draft Memorandum. 2023. MITRE, https://www.mitre.org/sites/default/files/2023-12/PR-23-02057-13-Advancing-Governance-Innovation-Risk-Management-for-Agency-Use-of-AI.pdf.

About the Center for Data-Driven Policy

The Center for Data-Driven Policy, bolstered by the extensive expertise of MITRE's approximately 10,000 employees, provides impartial, evidence-based, and nonpartisan insights to inform government policy decisions. MITRE, which operates several federally funded research and development centers, is prohibited from lobbying. Furthermore, we do not develop products, have no owners or shareholders, and do not compete with industry. This unique position, combined with MITRE's unwavering commitment to scientific integrity and to work in the public interest, empowers the Center to conduct thorough policy analyses free from political or commercial pressures that could influence our decision-making process, technical findings, or policy recommendations. This ensures our approach and recommendations remain genuinely objective and data-driven.

Connect with us at policy@mitre.org.

Endnotes

[1] Microsoft's "PyRIT" Automation Framework to Red Team Generative AI Systems (https://www.microsoft.com/en-us/security/blog/2024/02/22/announcing-microsofts-open-automation-framework-to-red-team-generative-ai-systems/)

[2] Google's AI Red Team Announcement (https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/) and Google's First AI Red Team Report (https://cloud.google.com/blog/transform/prompt-findings-our-ai-red-teams-first-report-qa)

[3] Carnegie Mellon's Report, "Red-Teaming for Generative AI: Silver Bullet for Security Theater?" (https://arxiv.org/pdf/2401.15897)

[4] MITRE Article, "Creating an AI Red Team to Protect Critical Infrastructure" (https://www.mitre.org/news-insights/impact-story/creating-ai-red-team-protect-critical-infrastructure), and MITRE's Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS) (https://atlas.mitre.org)

[5] Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. 2023. The White House. (https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/)

[6] Memorandum for the Heads of Executive Departments and Agencies. March 2024. The White House. (https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf)

[7] In this paper, "AI assurance" is a lifecycle process that provides justified confidence in an AI system's ability to operate effectively with acceptable levels of risk to its stakeholders.

[8] AI "chain" refers to the resources needed and the artifacts produced throughout the AI lifecycle including, but not limited to, training and testing data, AI models (e.g., parameters and weights), algorithm training, and inference software.

[9] Office of Management and Budget (OMB) guidance requires government agencies apply minimum risk management practices to AI that is rights-impacting and safety-impacting. See OMB Memorandum, "Advancing Governance, Innovation, and Risk management for Agency Use of Artificial Intelligence," March 28, 2024. (https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf)

[10] An AI assurance plan documents assurance process management, system characteristics, and lifecycle assurance implementation. See D. Robbins et al., "AI Assurance – A Repeatable Process for Assuring AI-Enabled Systems," 2024. MITRE. (https://www.mitre.org/sites/default/files/2024-06/PR-24-1768-AI-Assurance-A-Repeatable-Process-Assuring-AI-Enabled-Systems.pdf)

[11] For more details regarding the suggested organizational makeup of the NAICE, see "First Six Months" under "Implementation Considerations" later in this document.

[12] MITRE notes the administration is aware of the importance of proper resourcing and funding of AI security efforts and will not address those concerns in this document.

[13] MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems). 2023. MITRE. (https://atlas.mitre.org/). MITRE ATLAS is a globally accessible, living knowledge base of adversary tactics and techniques against AI-enabled systems based on real-world attack observations and realistic demonstrations from AI red teams and security groups.