PRESIDENTIAL TRANSITION: PRIORITY TOPIC MEMO
JULY 2024

MITRE | Center for Data-Driven Policy

ASSURING AI SECURITY AND SAFETY THROUGH AI REGULATION

By establishing a comprehensive and effective regulatory framework for AI security and safety, the incoming administration can ensure a balanced approach to technological progression, ethical considerations, and public trust. Doing so will not only reinforce the United States' international leadership in AI but also unlock its transformative potential to address a wide range of critical challenges.

The Case for Action

Over the past decade, the field of artificial intelligence (AI) has experienced remarkable advancements, ushering in a new era of technological innovation. These advancements have equipped us with a transformative technology in AI, which can be leveraged to address critical challenges in diverse fields, from healthcare to national security.

With each new presidential term comes the opportunity to reassess and enhance our approach to rapidly advancing technologies. In the realm of AI, it will be essential for the administration to stay informed about the current state of AI, its potential impacts, and the importance of advancing a sensible regulatory framework for AI assurance. While current policy and legislative activities have begun to address the need for AI regulation, more progress is needed to ensure the proper application and use of this technology, balancing security, ethical considerations, and public trust.

Key Challenges and Opportunities

AI regulation presents unique challenges due to the rapid pace of AI advancement and its diverse applications. Bridging the gap between policymakers at the Executive Office of the President (EOP) and agency implementation is a significant hurdle. Ensuring that policies formulated at the executive level are effectively translated into action at the agency level, taking into account the unique needs and contexts of each agency, is crucial.

Developing sector-specific AI assurance requirements that consider use cases and operationalizing the National Institute of Standards and Technology's (NIST's) AI Risk Management Framework (RMF) across sectors present significant challenges. These steps are necessary to ensure AI applications, within their specific contexts, meet safety and performance standards and effectively manage risks.

Rethinking regulatory and legal frameworks can guide federal funding decisions, advance AI research, and promote responsible AI use while deterring misuse.

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®

Establishing system auditability and increasing transparency in AI applications are essential for tracking misuse of AI and ensuring accountability within organizations. However, these measures pose challenges due to the complexity of AI systems and the current gap in technical expertise needed to effectively implement and manage these processes.

Despite these challenges, there are significant opportunities. Rethinking regulatory and legal frameworks can guide federal funding decisions, advance AI research, and promote responsible AI use while deterring misuse. Strengthening critical infrastructure plans and promoting continuous regulatory analysis can help secure our critical infrastructure against exploitation by humans, AI-augmented humans, or malicious AI agents.

Moreover, the diverse needs and requirements of agencies based on their size, organization, budget, mission, and internal AI talent present an opportunity to promote flexibility and adaptability in AI governance. An effective approach to AI regulation should allow for a tailored and effective implementation of AI strategies and policies across agencies.

Data-Driven Recommendations

1. BRIDGE THE GAP BETWEEN POLICYMAKERS AND AGENCY IMPLEMENTATION

Enhance communication and collaboration between policymakers and those implementing AI strategies by ensuring policies formulated at the executive level are effectively translated into action at the agency level, taking into account the unique needs and contexts of each agency. This can be achieved by evaluating existing EOP-interagency committees, expanding their mandates, adjusting their composition, or enhancing their resources, and if necessary, establishing a new dedicated committee that includes representatives from the EOP, various agencies, and industry. This group would help ensure effective communication and collaboration, involving regular meetings, shared resources, and a common platform for exchanging ideas and best practices.

2. DEVELOP SECTOR-SPECIFIC ASSURANCE REQUIREMENTS AND AI ASSURANCE PLANS

Collaborate with stakeholders in a repeatable AI assurance process to ensure that the use of AI within their specific contexts meets necessary safety and performance standards and manages risks associated with AI. Adoption of safe and secure AI can be achieved through requiring a structured AI assurance process that involves four steps: Discovering Assurance Needs, Characterizing and Prioritizing Risks, Evaluating Risks, and Managing Risks. This risk management process should incorporate the NIST AI RMF, be iterative, and be executed throughout the AI system's life cycle. A required output of this process is an AI Assurance Plan. This living document outlines the management and technical activities necessary to achieve and maintain assurance of the AI system over its operational lifetime and will require updating as new issues or risks are discovered.

3. SUPPORT AND ENHANCE THE OPERATIONS OF THE AI INFORMATION SHARING AND ANALYSIS CENTER (AI-ISAC)

Promote the recently established AI-ISAC to accelerate the sharing of real-world assurance incidents. This is essential to hasten understanding of threats, vulnerabilities, and risks to AI technology adoption for consequential use. The AI-ISAC should work in tandem with a national incident database like the Adversarial Threat Landscape for AI Systems (ATLAS^{TM}) to promote safe and anonymous sharing of real-world incidents. AI vulnerabilities and risks arise not only from malicious action but also because of the nature of the algorithms themselves and their susceptibility to misinterpretation, bias, performance drift, and other assurance factors. The AI-ISAC promotes analysis of incidents to identify root causes, and identification and development of mitigations, which can be derived from and/or contributed to the NIST AI RMF.

4. UNDERSTAND ADVERSARY USE OF AI ADVANCEMENTS

Support an at-scale AI Science and Technology Intelligence (AI S&TI) apparatus to monitor adversarial AI tradecraft from open sources such as research literature and publications, while providing continuous red-teaming of U.S. public and commercial AI infrastructure and operations. Doing so is crucial to understanding how our adversaries are using AI to gain advantage globally and to characterizing the reach of adversary capabilities into the United States, as well as the threat such reach poses to national security.

5. ESTABLISH SYSTEM AUDITABILITY AND INCREASE TRANSPARENCY IN AI APPLICATIONS

Issue an executive order that mandates system auditability, developing standards for audit trails, and advocating for policies that increase transparency in AI applications. This would include requiring AI developers to disclose what data was used to train their systems as well as the foundation models on which their systems were built. System auditability is vital for tracking misuse of AI and holding individuals accountable, as well as maintaining public trust in AI technologies.

6. PROMOTE PRACTICES FOR AI PRINCIPLES ALIGNMENT AND REFINE REGULATORY AND LEGAL FRAMEWORKS FOR AI SYSTEMS WITH INCREASING AGENCY

Take the following key actions to ensure the safe and responsible development and use of AI.

- Recognize that purpose (an understanding of objectives or goals) is an inherently human quality, and AI systems with agency (having the ability to act independently) will either directly receive purpose from a human (as instruction) or infer purpose through learning from human behaviors and artifacts.
- For AI principles alignment, create common vocabulary and research frameworks for guiding AI alignment in systems as scientific and engineering advances are made (rather than limiting or regulating advancements toward artificial general intelligence) to mitigate the risk of either humans tasking AI to carry out dangerous actions or AI systems exhibiting dangerous emergent behavior. Resulting guidelines would be similar to those established for research involving human subjects. Such advancements in AI alignment practices will serve to limit emergent, undesirable AI behavior, but research activities will still need safe environments with regulated guidelines like bioresearch and biosafety levels.
- Refine regulatory and legal frameworks to differentiate between appropriate research (with risk mitigations) and bad actors using AI for malintent, establish guidelines that address misuse of AI, and hold all appropriately accountable for harms.

7. STRENGTHEN CRITICAL INFRASTRUCTURE PLANS AND PROMOTE CONTINUOUS REGULATORY ANALYSIS

Direct federal agencies to review and strengthen government-critical infrastructure plans, focusing on safety-critical cyber-physical systems vulnerable to increased threats due to the scale and speed AI enables, and establish a dedicated executive task force to propose regulatory updates as needed. Such actions are necessary to ensure that our critical infrastructure is secure against exploitation by humans, AI-augmented humans, or malicious AI agents.

8. PROMOTE FLEXIBILITY AND ADAPTABILITY IN AI GOVERNANCE

Develop guidelines that allow for flexibility in AI governance implementation across different agencies, considering their unique needs and contexts (e.g., size, organization, budget, mission, AI workforce competencies). This involves enabling each agency to set an AI strategy that aligns with its needs and specific level of AI maturity. The guidelines should provide a range of options for AI governance structures, processes, and practices, and allow agencies to choose the ones that best fit their specific circumstances while ensuring minimum standards for consistency and effectiveness. As AI technologies rapidly evolve, these guidelines should also be flexible to accommodate ongoing innovation and shifting expectations about what is possible.

9. BRING IT ALL TOGETHER

Create a National AI Center of Excellence (NAICE) that promotes and coordinates these priorities, drawing on threat and risk assessment from the AI-ISAC and AI S&TI. The NAICE should not only cross-pollinate lessons learned by sector-specific regulatory authorities and build on and advance AI assurance frameworks and best practices, but also lead in conducting cutting-edge applied research and development in AI. This includes developing new AI technologies, methodologies, and tools that can be adopted across different sectors. The Center would facilitate collaboration among industry, government, and academia, thereby accelerating the transition and adoption of cutting-edge AI capabilities that are safe and secure.

Implementation Considerations

Implementing the proposed recommendations will require a blend of expertise, collaboration, funding, infrastructure upgrades, continuous learning resources, and flexibility in AI governance. Here is a suggested timeline and milestones to guide the process:

FIRST 100 DAYS

Evaluate existing EOP-interagency committees and identify opportunities to enhance their role in bridging the gap between policymakers and agency implementation. Initiate collaborations with industry experts, academia, and regulatory bodies. Begin the process of issuing directives for the adoption of the structured AI assurance process and the development of AI Assurance Plans across relevant agencies and departments.

FIRST SIX MONTHS

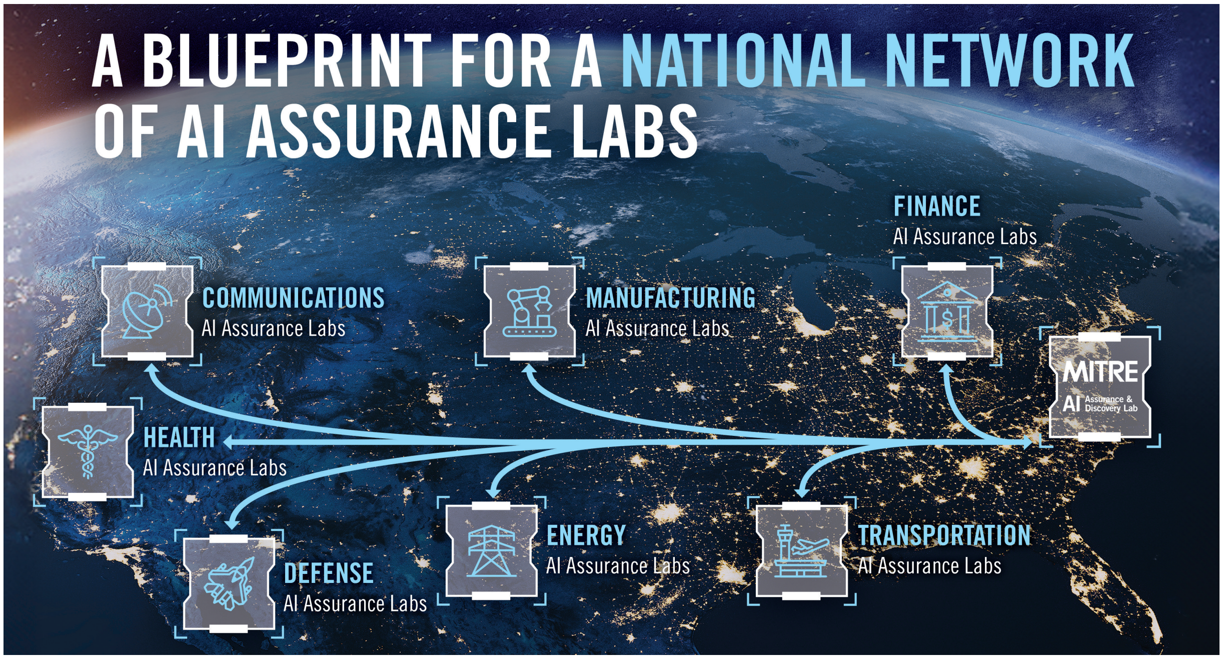
Monitor the initial implementation of the AI assurance process in sector-specific AI assurance infrastructure such as AI assurance laboratories and testbeds, as well as the development of AI Assurance Plans for use of specific AI-enabled systems in consequential mission spaces. Issue an executive order for system auditability and increased transparency in AI applications. Start the process of directing federal agencies to review and strengthen government-critical infrastructure plans. Develop strategies and guidelines that allow for flexibility and adaptability in AI governance across agencies.

FIRST YEAR

Secure increased federal funding for AI alignment research and necessary infrastructure upgrades. Implement system auditability and increased transparency in AI applications through executive orders and regulatory guidance. Complete the strengthening of federal government critical infrastructure plans. Establish a dedicated executive task force or enhance existing EOP-interagency committees to monitor the development and use of AI.

ONGOING

Continuously monitor the development and use of AI and propose regulatory updates as needed, based on the effectiveness of the AI assurance process, AI assurance infrastructure, and AI Assurance Plans. The NAICE should play a key role in this process, not only advancing state-of-the-art AI assurance knowledge and processes to agencies/sectors but also drawing from their practices and integrating insights across sectors. This could be facilitated through a network of AI assurance labs, modeled after MITRE's AI Assurance and Discovery Lab, that would support each sector and promote transformative insights across the AI R&D and implementation spectrum. Maintain collaborations with industry experts, academia, and regulatory bodies. Ensure access to resources for continuous learning. Regularly assess the effectiveness of implemented measures and make adjustments as necessary. 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.

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.

Three Recommendations to Advance AI Security and Trustworthiness. 2023. MITRE.

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.

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