AI ASSURANCE

A Repeatable Process for Assuring AI-enabled Systems

Executive Summary

AI ASSURANCE

A Repeatable Process for Assuring AI-enabled Systems

Federal agencies are being encouraged by the White House to remove barriers to innovation, accelerate the use of artificial intelligence (AI) tools, and to leverage AI to better fulfill their missions, all while setting up guardrails to mitigate risks.

Increasing the use of AI in government activities will likely have a consequential impact on the nation and world, in areas ranging from transportation to more efficient government to strengthened national security. Given this promise, how do we assure that these systems function as intended and are safe, secure, and trustworthy?

In the last two years, the U.S. has made progress in addressing these concerns, most noteworthy among them are the creation and publication of the National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF) (Tabassi, 2023), and the recent AI executive order (EO) from the Biden administration (U.S. Office of the President, 2023). There are significant gaps in our current understanding of the risks posed by AI-enabled applications when they support consequential government functions. While the NIST AI RMF and AI EO actions are useful catalysts, a repeatable engineering approach for assuring AI-enabled systems is required to extract maximum value from AI while protecting society from harm.

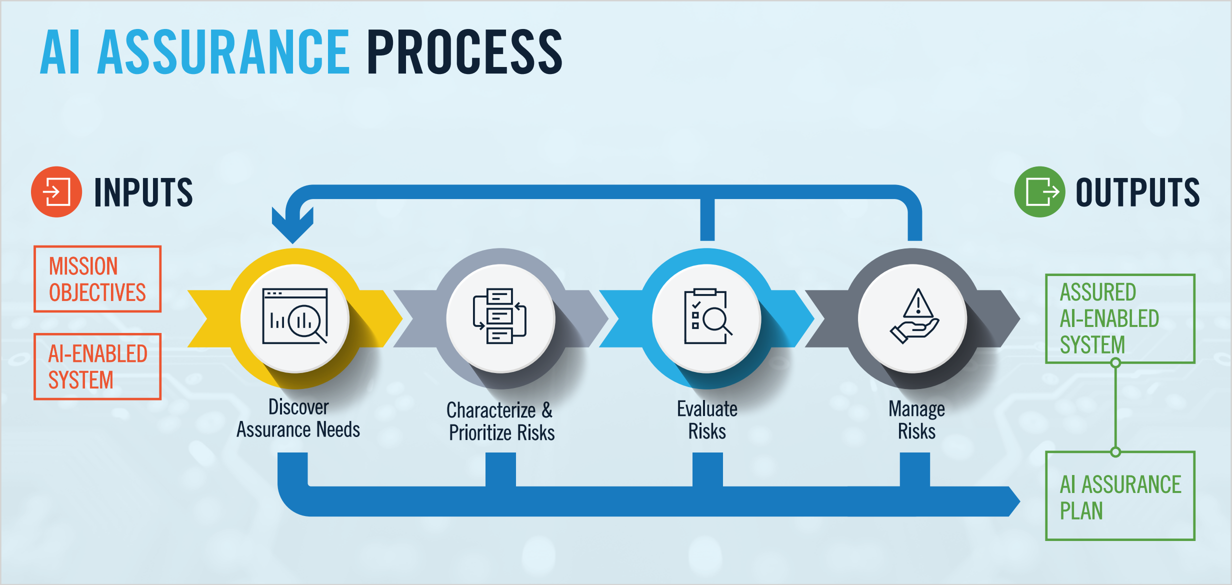
In this paper, we articulate AI assurance as a process for discovering, assessing, and managing risk throughout an AI-enabled system's life cycle to ensure it operates effectively for the benefit of its stakeholders. The process is designed to be adaptable to different contexts and sectors, making it relevant to the national discussion on regulating artificial intelligence.

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

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Our aim is to ensure effective operation of an AI-enabled system, which entails the system exhibiting intended behaviors while generating valid outputs that empower humans to achieve their goals. Characteristics of trustworthy AI systems can include governability, accountability, safety, security, privacy, interpretability, and equity. Discovering assurance needs requires a comprehensive understanding of the mission problem and the proposed AI solution.

The process results in an AI Assurance Plan, a comprehensive artifact outlining the necessary activities to achieve and maintain the assurance of an AI-enabled system in a mission context. This analysis should be completed prior to deployment of the AI-enabled system, and iterated upon until the desired level of assurance is achieved. The assurance plan accompanies the system post-deployment, allowing stakeholders to make informed decisions on acquisition, adoption, deployment, and use of the AI based on both the effectiveness of the system and its trustworthiness.



The AI assurance process is made up of four steps: discovering assurance needs, characterizing and prioritizing risks, evaluating risks, and managing risks. These steps need to be supported with laboratory infrastructure that can leverage a variety of physical and digital resources. In response, MITRE has established the AI Assurance and Discovery Lab, which aims to reduce deployment risk for AI-enabled systems, increase AI adoption, build a collection of use case-focused standards and baselines, ultimately constituting a living blueprint for an ecosystem of sector-specific assurance labs across government and industry. MITRE has developed several capabilities as part of that laboratory infrastructure, including:

- AI Assurance Needs Discovery Protocol
- AI Assurance Knowledge Base (the Knowledge Base will also incorporate rapidly shared anonymized incidents and mitigation approaches from the MITRE Adversarial Threat Landscape for AI Systems (ATLAS) community).
- ATLAS Mitigations
- Large Language Model (LLM) Secure Integrated Research Environment (SIREN)
- AI Red Teaming Guide
- Assurance Plan Template and Development Protocols
- Human Centered AI Test Harness
- Assurance Plan Template and Development Protocols
- Acquisition request for information (RFI) Analysis Tool

We conducted several pilot studies in the AI Assurance and Discovery Lab to examine the potential effectiveness of the AI assurance process. These real-world assurance investigations spanned a range of AI technologies, use cases, and AI life cycle stages and used lightweight, rapid investigations. They included: a policy search tool; a course of action recommender; an AI-enabled augmented reality microscope; a healthcare mobile robot; and a biometric system. The pilot studies on the AI assurance process highlighted the need for clear definition of issues, distinguishing between AI-specific and broader system assurance issues. The studies also emphasized the complexity of the relationship between undesirable impacts and assurance issues, requiring comprehensive mitigation strategies. Furthermore, they stressed the importance of effective communication among stakeholders, system developers, and AI experts.

We also considered how the AI assurance process may be applied to develop, acquire, certify, and deploy AI-enabled systems, and illustrated the roles key stakeholders may play for each of those applications. Below is one of the examples we provided in the paper as to how the process could be used.

An example of how the AI assurance process could be used:

A commercial company manufacturing drones employs aircraft software that uses AI to improve flight performance and seeks certification. Assurance must be addressed from multiple perspectives: design, production, and operational safety. An AI assurance plan serves as the canonical framework to identify the applicable regulations, standards, and guidelines, including the AI certification basis for the software. Developers leverage as much guidance as possible that exists for traditional software. In this case, that includes guidelines such as DO-178C (Software Considerations in Airborne Systems and Equipment Certification), and ARP4761 (Guidelines for Conducting the Safety Assessment Process on Civil Aircraft, Systems, and Equipment) to comply with federal regulations airworthiness standards. However, for AI-enabled components, new guidelines are necessary.

In collaboration with standards development organizations and industry, regulators assess existing regulations to determine what new rules and standards are needed for AI applications and pursue development of new regulations accordingly. This enables developers to identify and capture all applicable regulations and guidance in the assurance plan, including potential gaps. Developers, working together with testers and end users, then leverage all applicable AI guidance to satisfy regulatory compliance requirements, and capture all evidence, knowledge gained, and future assurance actions in the assurance plan. Methods of compliance include engineering reviews, analysis, modelling/simulations, and flight tests. Regulators then certify the system based on existing and newly established regulatory requirements and evidence.

During operation, the aircraft employing the use of the AI-enabled system will be monitored by regulators for conformance to applicable operating requirements (e.g., Part 107 – Small Unmanned Aircraft Systems) including any newly established ones, as documented in the assurance plan. Additionally, and also based on the AI assurance plan, developers, in collaboration with any applicable monitors, collect data to assess the AI-enabled system for potential operational safety risks on an ongoing basis.

Conclusion

We previously argued (The MITRE Corporation, June 2023) that AI regulation should account for use context and leverage existing sector-specific regulatory functions and mechanisms. It is not particularly useful, or even feasible, to attempt to assure AI in a general sense. The repeatable AI assurance process we outlined in this paper takes that into account by emphasizing and integrating mission context into all components of AI assurance. Therefore, AI assurance approaches need to be augmented with sector-specific resources to achieve domain-specific outcomes. We envision a future where resources such as the MITRE AI Assurance and Discovery laboratory will serve as a template for and be networked with sector-specific AI assurance labs to facilitate transformative insights across the AI research, development, and implementation spectrum.

We also recognize that the science and engineering of AI assurance is nascent, which presents many open questions. While work on AI assurance has been tracking developments in AI technology, there are significant gaps in our ability to effectively and rapidly bring AI assurance tools and methods to bear for specific applications. Moreover, standards are needed for assessing the level of consequentiality of an AI system and associating that to a commensurate level of assurance, a situation that bears strong resemblance to where cybersecurity was two decades ago (The MITRE Corporation, October 2023). As we have also learned from what it took to advance cyber assurance, significant government and industry investments and continuous public-private partnerships will be necessary to achieve AI assurance.

Link to the Technical Paper Including References

AI Assurance: A Repeatable Process for Assuring AI-enabled Systems is available for download here.

About the Authors

Douglas P. Robbins is the vice president of engineering in MITRE Labs. He leads MITRE's Innovation Centers across a wide range of technologies, including electronics, communications, systems engineering, and artificial intelligence. Previously, he served as MITRE's vice president for Air Force programs. Prior to that assignment, he led the strategic development of MITRE's Massachusetts-based operations, including new partnerships with the local high-tech ecosystem.

Ozgur Eris, Ph.D., is the managing director of the Artificial Intelligence and Autonomy Innovation Center in MITRE Labs. He leads over 200 artificial intelligence engineers and scientists to catalyze the consequential use of artificial intelligence for the public good. Previously, he served as the distinguished chief engineer of the AI and Autonomy Innovation Center and founded its AI-enhanced Discovery and Decisions department. Prior to joining MITRE, he researched, taught, and published on design cognition.

Ariel Kapusta, Ph.D., is a principal autonomous systems engineer in MITRE Labs' Artificial Intelligence and Autonomy Innovation Center. He has led development and published technical papers in the areas of robotics, AI, AI assurance, and safety.

Lashon B. Booker, Ph.D., is a senior principal scientist in MITRE Labs' Artificial Intelligence and Autonomy Innovation Center. He has published numerous technical papers in the areas of machine learning, adaptive behavior, and probabilistic methods for uncertain inference. He has served on the editorial boards of several journals, and regularly serves on the program committees for conferences in these areas.

Paul Ward, Ph.D., is the chief scientist for Social and Behavioral Sciences in MITRE Labs. He is internationally known for his research on expert sensemaking and decision making as well as assessing the impacts of AI-based course-of-action generation on human decision making. He serves as associate editor for the Journal of Cognitive Engineering and Decision Making and the Journal of Expertise.

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.

MITRE employs over 800 data scientists and AI engineers that currently provide deep expertise across the executive branch, with a particular focus on adoption and assurance of AI systems. Please let us know if you'd like to connect with one of our subject matter experts. To learn more visit Artificial Intelligence | MITRE.

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®