AI ASSURANCE – CHALLENGES, MATURITY, AND PATHS FORWARD

MITRE hosted an AI Assurance Summit on June 6, 2024, connecting AI leaders from multiple federal agencies. Our goal was to identify common practices and capabilities to guarantee the safe, secure, and effective application of AI in the United States.

What Is the Issue?

In this era of rapid technological advancement, artificial intelligence (AI) is becoming increasingly integrated into our daily lives. AI is addressing real, tangible problems for which in some cases there is no other solution. However, public sentiment toward AI is mixed. According to a 2023 Harris poll, fewer than half of Americans (48 percent) believe AI is safe and secure, while a significant majority (78 percent) express concerns about its potential for malicious use.[1] As the government seeks to leverage emerging AI capabilities, it faces the dual challenge of keeping pace with rapid technological change while maintaining public trust. The United States must strengthen its commitment to, and enhance its practices of, AI assurance to safeguard responsible design, development, and deployment of AI applications into the future.

What Did We Do?

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.[2] The outputs of the AI assurance process are intended to allow stakeholders to make informed decisions on acquisition, deployment, and use of the AI-enabled system. However, the science and engineering of AI assurance is nascent, which presents many challenges. For example, there are significant gaps to effectively and rapidly bringing AI assurance tools and methods to bear for specific applications and to assessing the level of consequentiality of an AI system (and, therefore, the commensurate assurance).

MITRE—a non-profit company that operates six federally funded research and development centers (FFRDCs)[3]—hosted an AI Assurance Summit to connect federal government leaders responsible for operating or certifying AI-enabled systems and to explore how the government can extract maximum value from AI while protecting society from harm. Discussion topics centered on:

- AI assurance challenges on the horizon
- AI assurance technical capability maturity
- Actionable paths forward for AI assurance

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'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®

The summit was open to invited federal agencies and select MITRE staff (with no media presence), and operated under Chatham House rules. There were more than 150 in-person attendees representing more than 20 government agencies spanning public sector services to national security. Summarized below are the thoughts captured from the two keynotes and three panels, which involved a total of 13 government experts.

What Did We Learn?

AI Assurance Challenges on the Horizon

Keynote speakers and panelists noted the government has the potential to utilize AI in a variety of beneficial ways, including improving citizen customer services, enhancing fraud detection, increasing operational efficiencies, triaging vast amounts of data, and streamlining regulation processes. However, ensuring that AI technologies function as intended and avoid unintended actions presents significant challenges. It is crucial to explore, evaluate, and understand the limitations of AI technologies in specific mission contexts. Key considerations include determining appropriate versus inappropriate uses of an AI system; accounting for both intentional and unintentional uses, including attacks and failures; and identifying biases introduced through training data, learning objectives, and solution proxies with available data. Transparency in AI performance and use, such as reporting confidence in predictions, is essential, as is ensuring fairness in AI applications. When explainability is not feasible due to AI model opacity, traceability and auditability can provide valuable insights.

Additionally, AI assurance must consider the entire technology stack as well as the broader system in which AI applications operate, including data, models, software, and hardware. Periodic re-evaluation of these considerations will be necessary. Large language models and generative AI introduce new challenges, such as prompt injections and hallucinations, in addition to existing concerns related to bias, exposure of sensitive data, resilience against imperfect or poisoned sources, lack of repeatability, and user over-reliance or inappropriate reliance. Consequential use cases of AI will necessitate effective regulation and certification programs, which will rely on effective assurance processes.

AI Assurance Technical Capability Maturity

Several speakers mentioned that AI assurance need not start from scratch. Software development assurance is mature and offers some guidance, but AI introduces model development and integration requirements for which existing assurance practices are not clear. Insights can also be drawn from cybersecurity's longstanding community of practice. Consequential AI use cases necessitate impact assessments to identify potential hazards and harms and to propose mitigations.

Panelists identified and discussed several open questions and needs for AI assurance to mature. We must identify and mitigate AI risks that are aligned to specific mission spaces, but also generalize best practices across domains. How can we account for different objectives across diverse stakeholders, such as the objectives of AI system developers versus users? How can we understand the intensions of users, whether under normal use or adversarial with the aim to circumvent, avoid, or defeat? How do we decide what to test, how will we scale these evaluations, and what do we need in our assurance sandbox beyond the AI technology—especially when it comes to evaluating sociotechnical impact on humans? How can we systematically break down and analyze the AI system at each stage of its life cycle, from development to deployment and maintenance? How do we document the operational design domain of the AI system and create representations of AI model beliefs to characterize expected and safe behavior when deployed? For AI assurance to scale, we need interoperability of AI assurance processes and tools, such as standard data sheets and model cards. As panelists discussed these areas of assurance need, they emphasized that we must assure AI-driven capabilities at the speed of technological advancement.

Actionable Paths Forward for AI Assurance

Panelists emphasized the importance of creating safe, mission-driven experimental environments where specific AI use cases can be explored, including testing and validation of human-machine interactions. Such experimentation is necessary for guiding the development of technology and related policy, governance, and standards, including supporting emerging regulatory structures. One speaker pointed out the sheer number of government AI adoption requirements, which has rapidly increased over the past five years, and noted that AI assurance practices will be instrumental in helping the government navigate these growing compliance complexities.

Panelists offered several strategic approaches to advancing AI assurance, starting with focusing on specific use cases from which to learn and generalize, thereby creating a growing knowledge base of lessons, policies, and practices. One speaker stated the greatest return on investment will be achieved through the automation of AI assurance tools, which will be key to enabling governance at scale. There was general agreement among the speakers that we focus on problem-driven rather than technology-driven solutions, and that we should not merely assure AI models, but rather look to assure AI use cases. This process should involve a diverse set of multi-disciplinary stakeholders, including domain experts and users. It is also important to consider the cost of implementing AI assurance to ensure the process is not prohibitively expensive and that the level of assurance due diligence and investment is commensurate with the risks involved in the use case.

When it comes to establishing AI assurance programs, panelists shared a vision for a nationwide network of assurance labs, with FFRDCs playing a critical role as trusted partners for transitioning and bridging AI capabilities to government mission needs. These partnerships and shared lessons learned will lead to scalability. Speakers pointed out that assurance programs can prioritize their focus in ways that provide strategic scaffolding and evidence-based AI consulting that lays a foundation for assuring future use cases. AI assurance programs must also provide guidance on organizational practice and implementation, including incident reporting, red teaming, vulnerability disclosures, and AI assurance infrastructure. One speaker pointed to MITRE's ATLAS™ program as "the platform" on which to model AI incident reporting and vulnerability and mitigation sharing, facilitated through a trusted data broker.[4]

All agreed that AI assurance is a strategic pathway to engender public trust in the application of AI and in the government's responsible use of AI. We need to focus not only on AI system performance, ensuring the system does what we want it to, but also on assurance, ensuring the system does not do what we do not want it to. It is essential that we start with assurance and not add it later.

Summary Outcomes

The AI Assurance Summit highlighted the significant value of fostering connections across the U.S. government, spanning public sector services to national security, on such a critically timely subject. Speakers noted that AI is being used in numerous consequential ways across various government sectors. To maximize the effectiveness of these applications, there is a pressing need to align mission needs with new AI capabilities, with AI assurance playing a pivotal role in this alignment. Panelists underscored trust as being more critical than performance, emphasizing the importance of ensuring reliable human-machine systems rather than merely assuring AI models. The summit also stressed the necessity for processes to discover and assess risks throughout the AI life cycle, from conception to deployment.

What Is Next?

The consensus from a full day of expert discussions is that the government should prioritize the development and implementation of robust AI assurance processes. These processes should focus on identifying and mitigating risks associated with specific consequential AI use cases, involving a diverse set of multi-disciplinary stakeholders, accounting for human-machine interactions, and understanding the limitations of AI-driven capabilities as aligned with mission needs. Furthermore, it is essential to establish safe, mission-driven experimental environments to explore and evaluate AI use cases. AI assurance programs should be established, envisioning a network of assurance labs. This would scale up the capacity for red teaming, AI incident reporting, and sharing of vulnerabilities and mitigation strategies across the entire nation.

Resources

MITRE-Harris Poll Finds Lack of Trust Among Americans in AI Technology, MITRE, February 2023. https://www.mitre.org/news-insights/news-release/mitre-harris-poll-finds-lack-trust-among-americans-ai-technology.

AI Assurance – A Repeatable Process for Assuring AI-enabled Systems, MITRE, June 2024. https://www.mitre.org/news-insights/publication/ai-assurance-repeatable-process-assuring-ai-enabled-systems.

MITRE is dedicating resources and working with its government sponsors to advance the science and engineering of AI assurance. In March of this year, MITRE announced the opening of its new AI Assurance and Discovery (AIAD) Lab, as a flagship model for risk discovery in simulated mission environments, AI red teaming, large language model evaluation, human-in-the-loop experimentation, and AI assurance plan development. Click here to learn more about the AIAD Lab and contact us at AI@mitre.org to learn more about AI assurance.

About MITRE

MITRE's mission-driven teams are dedicated to solving problems for a safer world. Through our public-private partnerships and federally funded research and development centers, we work across government and in partnership with industry to tackle challenges to the safety, stability, and well-being of our nation. MITRE has more than 800 AI engineers and data scientists working in partnership with agencies across the federal enterprise to develop and implement innovative AI solutions, enhance data-driven decision-making processes, and strengthen the security and efficiency of federal operations.

Endnotes

[1] MITRE-Harris Poll Finds Lack of Trust Among Americans in AI Technology, MITRE, February 2023. https://www.mitre.org/news-insights/news-release/mitre-harris-poll-finds-lack-trust-among-americans-ai-technology.

[2] AI Assurance – A Repeatable Process for Assuring AI-enabled Systems, MITRE, June 2024. https://www.mitre.org/news-insights/publication/ai-assurance-repeatable-process-assuring-ai-enabled-systems.

[3] MITRE operates six FFRDCs: the National Security Engineering Center (NSEC), the Center for Advanced Aviation System Development (CAASD), the Center for Enterprise Modernization (CEM), the Homeland Security Systems Engineering & Development Institute™ (HSSEDI), the Health FFRDC, and the National Cybersecurity FFRDC (NCF).

[4] See https://atlas.mitre.org.