MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

AI ASSURANCE

A Repeatable Process for Assuring AI-enabled Systems

June 2024

Douglas Robbins; Ozgur Eris, Ph.D.; Ariel Kapusta, Ph.D.; Lashon Booker, Ph.D.; and Paul Ward, Ph.D.
– The MITRE Corporation

CONTENTS

1. Motivation

Whether you believe the advances in Artificial Intelligence (AI) capabilities that have taken the world by storm in the last two years were predictable or a leap ahead, it is hard not to be equally excited and nervous about implications. These technological advances have been making positive contributions to our daily lives and are likely to make consequential impact on the nation and world in the future, in areas ranging from transportation to more efficient government to strengthened national security.

Today, while Americans rely on artificial intelligence to inform their consumer choices—from movie recommendations to routine customer service inquiries—the MITRE-Harris Poll survey on AI trends [1] finds that most Americans express reservations about AI for high-value applications such as autonomous vehicles, accessing government benefits, or healthcare. Moreover, only 48 percent believe AI is safe and secure, and 78 percent express concern that AI can be used for malicious intent.

Assuring AI-enabled systems to address these concerns is nontrivial and will take a concerted and collaborative effort between government, industry, and academia. AI assurance is complex because AI is:

1. Not a single technology but instead comprises an array of methods including machine learning, deep learning, reinforcement learning, ensemble learning, rule-based reasoning, genetic algorithms, generative AI, and many others.
2. A transdisciplinary domain with scientists and engineers interacting across numerous disciplines—including cognitive science, computational sciences, philosophy, linguistics, and neuroscience—and arriving at new insights at the intersection of disciplines.
3. Intended to benefit humans and society—does not operate in a vacuum—which surfaces significant cultural and value-based alignment questions. "Intelligent" systems invoke a variety of human values in new contexts, often unpredictably, when humans interact with them to realize their benefits.
4. Embedded in systems and interacts with and depends on data and other AIs from other sub-systems, which can result in emergent, real-world behaviors.
5. Technologically advancing at an increasing pace fueled by significant investment with trailing assurance investment and focus.

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) [2], and the recent AI executive order (EO) from the Biden administration [3]. Given the inherent complexity and consequential nature of AI, it is useful to view AI assurance through a risk management lens, and the NIST AI RMF [2] provides a high-level framework for doing so. The AI EO directs a number of actions that support the safe, secure, and trustworthy use of AI, including the creation of new standards for AI safety and security, safeguards protecting Americans' privacy, guidance to ensure the responsible and effective use of AI in government, and more. While the NIST AI RMF [3] and AI EO actions provide a useful catalyst for addressing these issues, a repeatable engineering approach for assuring AI-enabled systems is required to extract maximum value from AI while protecting

society from harm. This paper aims to provide additional details on what such a repeatable engineering approach entails and how it can be supported.

2. AI Assurance as a Concept

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.

Effective operation entails system behavior that fulfills the functions the system is intended to fulfill, while generating valid outputs that empower humans to achieve their goals. Assurance risks may arise from a variety of factors that are contingent on both the purpose for which the AI-system was designed and the contexts in which it is employed. These risks include, but are not limited to, characteristics of trustworthy AI systems such as safety, security, privacy, interpretability, equity, governability, and accountability.

A REPEATABLE ENGINEERING APPROACH FOR ASSURING AI-ENABLED SYSTEMS IS REQUIRED TO EXTRACT MAXIMUM VALUE FROM AI WHILE PROTECTING SOCIETY FROM HARM.

Our conceptualization of AI assurance is grounded in the literature on risk management and safety management best practices. It is based on and aligns with risk assessment and management strategies suggested in NIST, International Organization for Standardization (ISO), and International Electrotechnical Commission (IEC) standards (NIST SP800-30 [4], NIST AI RMF [2], ISO 31000, and ISO 14971) with respect to general process considerations, as well as other literature on functional safety (IEC 61508, IEC 62061) with respect to associated planning (e.g., functional safety plan). As such, AI assurance should result in satisfaction of principal safety and risk management requirements. In addition to the requirements described in these standards, a unique feature of AI assurance is that it also includes the investigation and evaluation of system effectiveness, satisfaction of performance requirements, and low severity harm risks that may only impact effectiveness.

Our conceptualization is also in alignment with established and adopted engineering design and product development theories [5], [6], which advocate the need to identify requirements associated with critical product functions early during development by engaging with key stakeholders and iterating to develop sufficient problem and solution space understanding before committing to production and release.

Finally, any AI assurance consideration needs to be in alignment with an organization's overall AI adoption objectives and guided by its cognizant AI governance policies. Existing AI policies act as a starting point in informing assurance needs, and mission-focused analyses provide the basis for higher fidelity, context-specific assurance requirements. Although AI assurance and governance are innately coupled, a detailed discussion on their interactions is out of scope of this paper.

3. AI Assurance as a Plan

AI assurance is a life cycle consideration, involving a broad range of stakeholders from developers to end users, and is central to managing and mitigating risk. Therefore, it is critical that all information that is generated while assuring an AI system is codified in a comprehensive artifact. We call this artifact an ***AI Assurance Plan***, which is both contributed to and used by assurance stakeholders for the full lifetime of an AI-enabled system.

As described in Section 2, the AI assurance plan aligns conceptually with the use of safety plans in safety, security, and risk management.

The assurance plan guides the management and technical activities that are necessary to achieve and maintain assurance of an AI-enabled system. As such, an assurance plan is not just a descriptive but also a prescriptive document, which outlines actions that need to be taken. It is developed through the AI assurance process, which we outline in the next section. The plan is a living document that is updated as necessary during the system's entire life cycle, as new risks or issues are discovered. It is the key artifact to be used by decision makers with the responsibility of developing, acquiring, deploying, and monitoring the AI-enabled system. These decision makers may include system vendors, government program owners, third party assurance labs, and regulators (see Section 7 for a more detailed discussion on key stakeholders).

The assurance plan should be initialized when beginning the assurance process—ideally during AI-enabled system development—and is populated over the course of that process. The assurance process used to generate the plan may be reinvoked during any phase of the system's life cycle as needed to maintain intended levels of assurance. During system development, the plan may exist in different degrees of specificity and completion, depending on the nature and goal of development activities. The plan's content can be viewed roughly as covering management of assurance activities, documentation of system characterization, and management of life cycle assurance. The content of each category is described below.

Assurance Process Management

PM1. An instantiation of the AI assurance process tailored for the AI-enabled system. A list of the activities for completion of the assurance process and the assurance plan.

PM2. Specification of the resources and timelines for each assurance activity, including review and approval of each activity.

PM3. Parties who are responsible for carrying out the assurance activities and approving their outcomes.

PM4. Description of the AI governance policies of the organization and their implications for how the AI assurance process should be applied to the mission space the AI-enabled system will operate in.

System Characterization

SC1. A description of the mission problem, the proposed AI-enabled system, and the expected use cases, scope of use, and impacts or effects of use.

SC2. Definition of the system needs and requirements subject to assurance through the AI assurance process, along with rationale. Requirements must be testable.

SC3. An assurance risk assessment that comprehensively identifies and estimates hazards and risks. Components of risks, such as their harm and their likelihood of occurrence, should be testable.

Life Cycle Assurance Implementation

LA1. Documentation and communication protocols for recording, maintaining, and sharing assurance information.

LA2. A configuration management protocol that documents how the correct versions of system components and configuration parameters are managed to achieve the level of assurance that has been deemed acceptable.

LA3. A change management protocol that specifies the activities necessary to change the system and determine any assurance-related consequences associated with a change in the system. Conversely, it also specifies what types of changes to the system or in its operating environment may necessitate the reapplication of the assurance process.

LA4. An issue management protocol for tracking and addressing any issues that may impact assurance. This protocol specifies what needs to be tracked regarding issues, the resources required, responsible parties, methods to receive issue/incident reports, and required timelines for issue review, resolution, and validation.

LA5. A system monitoring protocol that documents how and by whom the AI-enabled system should be monitored to maintain assurance. It may include continuous monitoring, automated data recording, audits, tools for performance monitoring, and monitoring of incident reports and issues.

Note that monitoring should address AI assurance issues as well as any related system-level assurance issues.

LA6. A verification and validation protocol of the system, used during the assurance process. This protocol documents the tests for verifying satisfaction of assurance requirements as well as the impact and likelihood of risks. Validation determines the achieved assurance level and satisfaction of assurance needs.

4. AI Assurance as a Process

We represent the process outlined in our AI assurance definition as a repeatable engineering approach in four canonical steps (Figure 1). This process should be completed prior to deployment of the AI-enabled system, and iterated upon until the desired level of assurance is achieved. Each step of the process should not only inform how to engineer assurance into the AI-enabled system but also, as stated in the previous section, advance the development of an assurance plan that will accompany the system post-development. It should be noted that promoting discovery throughout this process is important because AI components interact with other system components and their environment in complex ways and our understanding of the risk and mitigation spaces will be far from being exhaustive anytime soon.

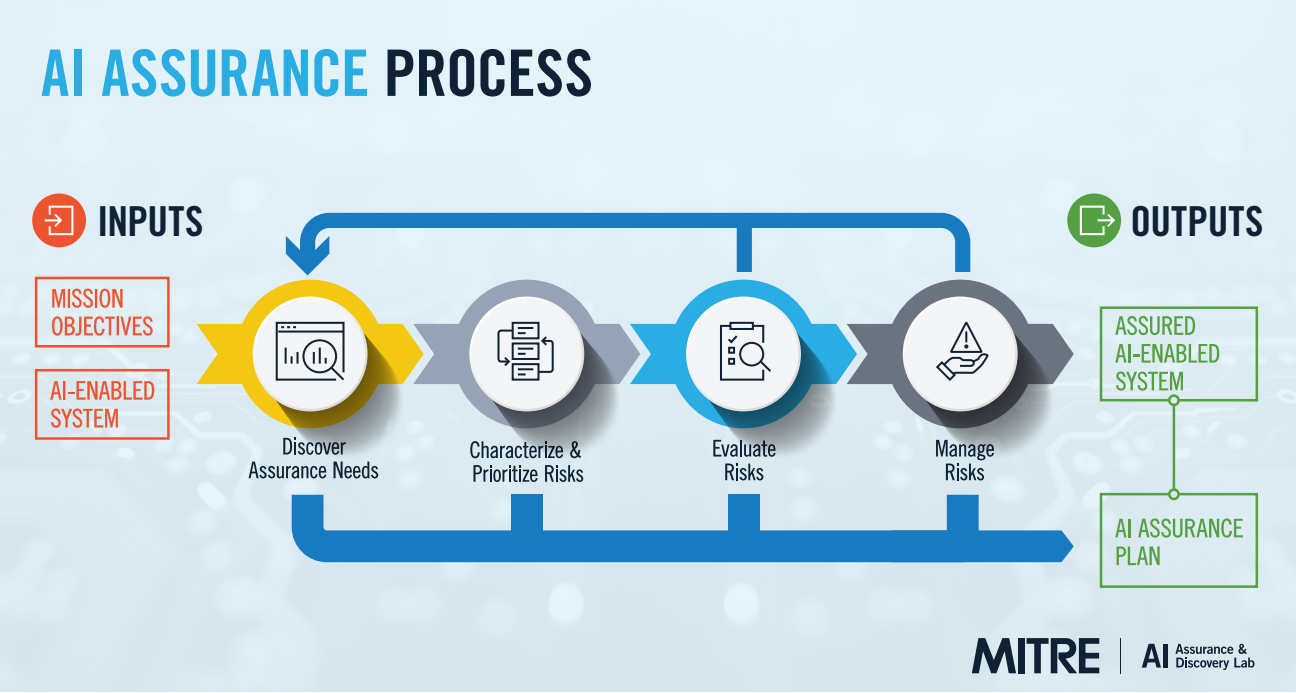

Figure 1: AI Assurance as a Process

Below, we will further articulate each of the four steps of the AI assurance process and the assurance plan they contribute to.

Discover Assurance Needs

Discovering assurance needs requires a comprehensive understanding of the mission problem and the proposed AI solution. That understanding starts with a decomposition of the use case, AI solution under consideration, and anticipated effects, which allows for the identification of problem-specific AI assurance needs and their potential impacts across the life cycle. One of the primary goals of this step is to discover alternative or potentially new assurance concerns and to identify trade-offs across assurance requirements. This step is similar to hazard and risk identification in other fields, but with emphasis on discovery as the risks related to AI can be complex, emergent, and related to human and societal response, perception, and values. AI risks are not yet as well understood as those in more mature fields like machinery safety.

Discovery begins with understanding the AI-enabled system's scope, intended use, interactions with its environment and other systems, and foreseeable misuse. Evidence and known issues associated with similar systems can be used to inform discovery and identification of relevant potential harms, hazards, and risks. Similarly, expert or data-driven models that suggest relevant but not yet considered assurance needs can be used to facilitate the analysis of consequential assurance concerns for a given use case. Combinations of common risk identification methods from literature can also be used, with the ultimate purpose of comprehensive identification of assurance needs.

In this step, all Assurance Process Management components of the assurance plan—components PM1, PM2, PM3, and PM4—should be created.

Additionally, components of System Characterization—components SC1 and SC2 of the assurance plan—should be completed.

Characterize and Prioritize Risks

The next step is to conduct qualitative estimations to characterize the risks associated with the assurance needs. Risk assessment guidelines are available (e.g., NIST SP800-30; see [4]) that describe systematic approaches to conducting this initial identification of potential hazards and risks, as well as estimation of risk severity, likelihood, and tolerance. Authoritative resources, subject matter experts, and preliminary evaluations are common sources for risk estimation. Details of system design and implementation are necessary for this activity. Risks are prioritized for further consideration and evaluation. Risk estimation can inform the level and form of evaluation needed for each risk. A critical point of assurance compared to more conventional risk management is the consideration of and emphasis on mission needs when prioritizing risk. This consideration is of particular importance for systems that apply general-purpose AI models for specific missions; the general risks of such models may already be well understood, but their risks related to the mission are not.

The outputs of this step contribute to component SC3 of the assurance plan, relating to risk assessment. That component is completed in the next step of the assurance process.

Evaluate Risks

The risk assessment conducted across this and the previous steps of the assurance process, should be sufficiently comprehensive to support the assurance investigation goals (see Section 7 for potential utilizations of this AI assurance process). Depending on the assurance investigation purpose, risk evaluation may entail measuring and quantifying risks using standard test and evaluation (T&E) protocols to determine if the intended level of assurance and probability of harmful failure are met. A T&E plan covers key aspects like AI algorithm testing, systems integration testing, human-systems integration testing, and operational testing. Risk evaluation can take many forms, depending on the system, the desired level of assurance, stakeholders' tolerance to risk, and the risks under consideration. On the other hand, risk evaluation may be primarily focused on risk discovery. For example, early in AI development it may be helpful to conduct lightweight "investigations," which pull in stakeholders to interact with preliminary—even paper-based—prototypes, where the focus may be on gaining an initial and sufficient level of empirical understanding of those risks rather than their conclusive, quantitative characterization. Although such utilization of this AI assurance process may still follow standard T&E practices, it may not execute them with the same level of rigor, given a more exhaustive and conclusive risk evaluation would be conducted after the AI prototype matures (e.g., prior to acquisition and/or deployment).

The verification and validation of AI-enabled systems, determining if the intended level of assurance and probability of harmful failure are met, closely parallel those from traditional engineering disciplines. To ensure a system is fully operational, each component (software module, access policy, distribution plan, etc.) should be tested individually, as a system, with users, and in operational settings to evaluate both subcomponent insufficiencies as well as those from the interactions between them. Relevant metrics should be used, including user and stakeholder feedback.

The outputs of this step contribute to component SC3 of the assurance plan, relating to risk assessment. Additional iterations of the assurance process or additional risk management techniques may be necessary, until the risk assessment is sufficient and comprehensive.

Manage Risks

Once risks have been prioritized and evaluated, potential courses of action must be developed to manage each risk and reach the assurance level designated as acceptable in the evaluation. One way to do this is to employ a risk response strategy designed to reduce or remove risk, for instance by transferring, sharing, avoiding, or mitigating each risk to an acceptable level (e.g., NIST SP800-39; see [7]). High-consequence risks will require detailed implementation plans, which state technical, algorithmic, and/or procedural controls to ensure that systems behave as intended. Residual risks from any unmitigated risks should be documented and within acceptable safety, security, and trustworthiness limits of the system.

Risk mitigation pathways are often complex, and plans for individual efforts are dependent on several project-specific factors including overall system design, domain of application, intended use case, and scope of the intended user base. Some common mitigations, like change management and system monitoring, are identified as distinct components of the assurance plan (as articulated below). While some mitigations may be technical, such as modifications to the AI application and/or system architecture, others may focus on business process or operational adjustments, such as limiting access to sensitive data or establishing training procedures to educate employees of relevant ethical considerations they should keep in mind while using a product [8]. Similarly, a risk mitigation plan should include provisions both for the initial development and deployment of an AI-enabled system as well as processes for the long-term maintenance of the system and recovery if incidents occur. Responsibilities for ensuring that risks are sufficiently mitigated will be shared across organizations and roles ranging from engineering and system administration to user engagement and product management teams. Documentation of risk management is maintained in the assurance plan. Once risks management strategies are applied, the residual risk and final system performance must be reevaluated.

The outputs of this step contribute to the Life Cycle Assurance Implementation components of the assurance plan, components LA1-6. Additionally, as additional risk management techniques are applied, the risk assessment, component SC3 of the assurance plan, must be updated to reflect the residual risk.

Outputs of the AI Assurance Process

The outputs of the AI assurance process allow stakeholders to make informed decisions on acquisition, deployment, and use of the AI-enabled system. The process does not guarantee acceptance by all stakeholders; some stakeholders might have higher assurance requirements than those achieved through the process and may find the level of residual risk to be unacceptable. Therefore, the process documents the information needed to establish and maintain an appropriate level of trust in system performance, including:

- An AI-enabled system that has been evaluated, and whose AI assurance risks are being managed through mitigations or other responses—degree of evaluation and mitigation will depend on the context in which the process is being used (e.g., research and development [R&D] versus certification).
- An assurance plan that specifies how the AI-enabled system is assured and how its assurance will be maintained over its life cycle.
- Specification of the level of assurance for functions and capabilities of the AI-enabled system.
- Any knowledge gained that can/should be reused for another assurance analysis for the benefit of other assurance cases.

5. Supporting the AI Assurance Process with Laboratory Infrastructure

The AI assurance process needs to be supported with capabilities that can leverage a variety of physical and digital resources, depending on what is being assured. Some of those capabilities will be targeted and specific to the assurance case under consideration (e.g., how to best mitigate a specific type of AI assurance risk) while others will be more general and can serve as reusable resources in any AI assurance task. MITRE has established the ***AI Assurance and Discovery Lab*** that integrates and operationalizes such general capabilities as part of a repeatable AI assurance process.

The laboratory aims to reduce deployment risk for AI-enabled systems, increase AI adoption, build a collection of use case-focused standards and baselines, and constitute a living blueprint for an ecosystem of sector-specific assurance labs across government and industry. Select lab capabilities include:

- **AI Assurance Needs Discovery Protocol:** A standardized multi-dimensional protocol designed to discover, identify, and prioritize problem-specific AI assurance needs. The protocol, with key stakeholder input, facilitates that exploration by decomposing the mission problem and AI-enabled solution under consideration and surfacing AI assurance concerns. It is applied prior to measuring and mitigating the associated risks, and as such, serves as the "front-end" of the AI assurance process.
- **AI Assurance Knowledge Base:** An AI Assurance Knowledge Base (AAKB) to enable the utilization of AI assurance knowledge that can inform the design of assurance investigations. The AAKB provides information to an AI assurance investigator on AI assurance metrics, datasets, methodologies, and tools that are relevant to their assurance goals. It also provides pointers to similar investigations that have been carried out in the past. The AAKB captures information from relevant sources including scientific publications, government developed capabilities, and commercial offerings, and allows an AI assurance investigator to search all metadata through semantic search. The search results can be examined more deeply or summarized with large language model (LLM) support. MITRE will continuously and collaboratively expand the AAKB with new knowledge, insights, and best practices from the field, including community contributions. The AAKB will also incorporate rapidly shared anonymized incidents and mitigation approaches from the MITRE Adversarial Threat Landscape for AI Systems (ATLAS) community into the AAKB.
- **LLM Secure Integrated Research Environment (SIREN):** A sandbox environment that enables rapid prototyping of capabilities to explore safe, appropriate, and effective use of LLMs. A key focus is testing and evaluation of LLM-based applications, especially *Augmented* LLMs, to discover risks and identify potential mitigations. SIREN provides:
  - Best practices for evaluating augmented LLM-based systems
  - Guidance on developing use case-specific evaluation protocols, data sets, and metrics and generating synthetic benchmark datasets

  - Rapid benchmarking of augmented LLMs for use cases, as supported by recent advances in statistical inference, including methods for assessing retrieval quality, answer synthesis, and hallucinations
  - Reference implementations of common paradigms for building applications with LLMs, including retrieval augmented LLMs, knowledge graph-enabled LLMs, and LLM "agents" as starting points for developing targeted LLM-based systems
- **ATLAS Mitigations:** A set of security concepts and classes of technologies that IT professionals can use to prevent a specific attack technique in the ATLAS matrix from being successfully executed by an attacker. ATLAS Mitigations is community shaped with 100+ industry and government organizations and includes advanced mitigation methods as well as basic cyber hygiene approaches.
- **AI Red Teaming Guide**: Best practices on how to conduct AI red teaming, an investigative process that simulates attacks on real-world AI-enabled systems to identify vulnerabilities, mitigate potential exploits, and improve the overall security posture of an AI-enabled system. The guide augments developers' understanding of how an AI-enabled system's existing defenses could be manipulated in unexpected ways compared to the initial operational evaluation. The guide offers actionable insights on how a Red Team should:
  - Partner with the system designers/developers to familiarize themselves with the target system and use empirical knowledge of adversary behavior to develop an attack plan.
  - Execute the attack plan by implementing and demonstrating the defined attack vector and emulating adversary behavior against the target system. Emulating adversary tactics and techniques can be aided using tools like Arsenal, an open-source AI security threat emulation plugin to implement ATLAS tactics and techniques in the Caldera threat emulation platform.
  - Reports its findings, including the impact of the operation to facilitate the Blue Team's (system defender's) response.
  - Coordinate with the Blue Team to perform the Purple function—using strong and valid attacks to enhance the AI-enabled system's mitigation and defensive strategies and techniques.
- **Human Centered AI Test Harness:** A web-based automated measurement platform to instrument AI-enabled interactive systems for human-in-the-loop research and evaluation. It works by mapping digital workspace events to human behaviors of interest, enabling objective and subjective data collection for human-subjects measurement. The Test Harness automatically administers custom qualitative or quantitative metrics characterizing how users respond to an interactive AI system. The Test Harness supports simple and complex experimental designs (e.g., within-subjects, between-subjects, mixed designs, etc.) and provides a highly configurable platform to test AI across a variety of domains using open-source and cross-platform technologies.
- **Assurance Plan Template and Development Protocols:** Tools to facilitate the creation of an assurance plan and adoption of a development plan that will result in an assured AI-enabled system. These tools provide templates for required activities and protocols to complete them. The protocols provide explanations, guides, references, and examples to aid in completion of assurance activities and documentation in the assurance plan. Stakeholders involved in the assurance plan include product owners, project or program managers, technical subject matter experts, developers, testers, and operations staff.

- **Acquisition Request for Information (RFI) Analysis Tool:** An-LLM enabled tool that helps acquisition staff better understand and process RFIs and their responses. The tool ingests acquisition artifacts and accelerates the timeline of documentation outputs by capturing administrative detail and elevating the human-in-the-loop's attention to the most impactful respondents for follow-on engagements. As such, the tool can be used by experts to identify, analyze, and augment RFI sections specific to AI assurance that should be driving AI-enabled system acquisitions.

6. AI Assurance Process Pilots

We conducted several pilot studies in the AI Assurance and Discovery Lab to examine the potential effectiveness of the AI assurance process, as defined above. 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.

Case Study: Assuring AI in a Healthcare Mobile Robot

Here, we will detail how one of those pilots used the AI assurance process to develop a prototype healthcare robot for autonomous contactless vital sign measurement. The robot system receives tasking from a clinician (potentially from a schedule), navigates to the correct patient in the correct room, measures the patient's vital signs using a camera, and records the vital signs into medical records.

We started with creating the AI Assurance Process Management components of the assurance plan, to guide our assurance activities. We then defined and documented the System Characterization, completing that component of the assurance plan, in parallel with implementation of the system. We recognized several assurance needs that differed from conventional safety requirements. For example, patient and clinician acceptance of a healthcare robot are important, so undesirable behaviors that cause negative experiences, like inefficient use of time, should be avoided.

Afterward, we performed hazard identification and risk estimation to identify, characterize, and prioritize risks of the system. Although many risks were deemed unimportant, we identified two hazards of particular concern: misidentification of the patient and localization failure (getting lost). Each of these hazards had several associated risks by which harm might be realized. Initial characterization of these risks recognized them to be of critical importance to AI assurance. Misidentification of the patient could result in vital signs recorded into the incorrect medical record and incorrect treatment provided to patients. Localization failure typically causes harm by delaying patient assessment and treatment or otherwise wasting time, and its expected likelihood of occurrence was high. Evaluation supported these characterizations, so we applied risk management techniques to update the design of the system to prevent these hazardous events from occurring and to mitigate the harm from occurrence.

Misidentification of the patient was controlled through software architecture requiring a series of unlikely independent failures preventing the hazardous event from occurring. Localization failure was mitigated through independent checking of localization using a system that is essentially indoor GPS, that would mitigate the harm of localization failure.

We iterated upon the assurance process, now discovering 58 risks relevant to assurance of the completed system. Although the newly added components introduced new failure modes, the failure trees were planned to control failures without harm. We evaluated the risks through extensive testing offline, in simulation, on the robot, and with human participants. The results indicated that the risks were adequately controlled, although we identified additional improvements to specific components of the system that could be improved to further reduce the probability of harmful failures.

During this case study, we completed a demonstrative example of the assurance plan, documenting conceptual solutions to Life Cycle Assurance Implementation in the assurance plan, although we did not implement them as part of this pilot. For example, we created a change management plan that includes review of the changes, automatic testing, and potential additional evaluation activities prior to changes to the system, to ensure changes cannot unexpectedly degrade performance. The assurance plan would help future maintainers and users of the system from misusing or misunderstanding the capability of the system.

Insights Gained from the Pilot Studies

Overall, the pilot studies yielded the following insights about the application of the AI assurance process:

1. The scope of the issues addressed in any instantiation of an AI assurance process should be clearly articulated. For example, there can be a fine line between AI assurance issues and more conventional systems assurance issues like compliance with Federal regulations (e.g., Section 508), which require other tools and processes to address. Moreover, it is also important to provide system-level assurance for AI-enabled systems. Mechanisms that assess system-level performance make it possible to detect when there is non-graceful degradation in the system and establish appropriate mitigation strategies.
2. The relationships between undesirable impacts and their associated assurance issues can be complex. A single impact can be intertwined with multiple assurance issues and a single assurance issue can sometimes result in multiple impacts. Thus, there can be several paths linking assurance issues to harmful impacts, and they don't necessarily have the same severity or the same likelihood. Potential mitigations will need to account for these complexities.
3. An assurance investigator can stop the process at any given step (during the execution of the four steps outlined in Figure 1) and exit with actionable insights. For example, the first step in the process generates a report that contains both a succinct description of the system itself and a breakdown of the key assurance needs. The pilot studies made it clear that this report can have value as a standalone product, which can help stakeholders address assurance issues early.
4. The AI assurance process requires extensive communication between stakeholders, system developers, and the AI experts conducting the investigation. These parties can have different perspectives on the assurance landscape, have different priorities, and use different terminology. It is critical to disambiguate terms and clarify distinctions among concepts in the assurance landscape. The terminology defined in this white-paper and articulated in the artifacts of the highlighted capabilities make such a contribution.
5. Not all AI assurance concerns can be linked to a metric or approach that is likely to be included in an AI assurance knowledge base. In one of the pilot studies, some AI assurance issues stemmed from inadequate approaches to capturing and utilizing use case-specific domain knowledge during the design of the system, which limited the validity of system outputs.

7. Applications of the AI Assurance Process

Different stakeholders may have different goals for applying this assurance process at different points in the AI life cycle. In this section, we consider how the process may be applied to develop, acquire, certify, and deploy AI-enabled systems, and illustrate the roles key stakeholders may play for each of those applications (see Table 1).

AI-enabled systems are also dependent upon the organization taking those actions. Therefore, this mapping needs to be translated for and integrated into an organization's business operations. In doing so, assurance plans will be tailored to meet mission-specific needs. It should be noted that the primary "owner" of AI assurance, the stakeholder who is ultimately accountable for assurance, may change as the AI-enabled system advances throughout its life cycle. Moreover, achieving and maintaining assurance requires continuous stakeholder contribution throughout the AI life cycle regardless of who may be ultimately accountable for assurance at any phase.

Several stakeholder roles are considered in this initial mapping:

- **AI Developers** – team developing an AI enabled solution (e.g., contractor, commercial vendor, internal development organization)
- **End users or operators** – individuals and teams using the AI enabled solution post deployment to execute operational tasks
- **Program office** – office, government or otherwise, responsible for the acquisition and deployment of the AI-enabled system
- **Standards bodies** – organizations responsible for establishing AI assurance standards, possibly sector specific (e.g., NIST, Coalition for Health AI)
- **Testers** – internal or third-party organizations responsible for performing assurance testing per the assurance plan
- **Regulators** – organizations responsible for authorizing the deployment of assured solutions based on evidence
- **Monitors** – organizations responsible for post deployment assurance monitoring per the assurance plan (e.g., appropriate model use, model drift)

It should be noted that the mapping depicted in Table 1 is for illustrative purposes only and the four application foci of the AI assurance process are not necessarily orthogonal nor are they meant to be comprehensive. For instance, acquiring an AI-enabled system may have various overlaps with developing, certifying, and deploying it. However, the four applications entail distinct decision-making processes different decision makers are responsible for. Moreover, each sector may have a somewhat different formulation of the relationship between those applications, and may decompose them further or differently, resulting in different and/or more rows in the matrix. A similar consideration may apply to the decomposition and definition of the stakeholders. Therefore, each organization should articulate its own version of this matrix in alignment with its operations.

Below, we offer two examples of how the AI assurance process could be used. The first describes a government program office developing and acquiring an AI-enabled system, and the second describes government certification of AI software onboard a drone.

In the first example, a government acquisition executed by a program office must satisfy the relevant operational requirements. Through a set

of development activities, a solution that meets those requirements is identified, and if that solution includes AI, it should require the development of an AI assurance plan. The responsibility to create that plan best aligns with program office functions, in coordination with any commercial solution vendors, that necessarily span development, test, and support of fielding. Developers, sometimes organic to the government, and sometimes via a contract award, must work with end users and testing organizations to discover and prioritize risks, develop mitigations to those risks, and capture actionable findings in the assurance plan. Testing organizations identified by the program office have the responsibility to independently conduct both development and operational tests, and assess the assurance risks and mitigations identified in the assurance plan. Note that some mitigations are non-technical in nature and may require restrictions on use of the system and continuous monitoring after fielding. In the case of a government acquisition of this type, certification performed by a regulator takes the form of authorization for deployment, in compliance with the standards it has identified for the domain, which should already be reflected in assurance plan. Lastly, post-deployment monitoring must be supported by appropriate sustainment activities as per the assurance plan and may require explicit contracting by the program office.

In the second example, 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.

TABLE 1: stakeholders roles based on application focus

	Stakeholders						
Application Focus	**AI Developers**	**End Users or Operators**	**Program Offices**	**Standards Bodies**	**Testers**	**Regulators**	**Monitors**
Development	▪ Capture risks and associated built-in and operational mitigations with user community ▪ Develop mitigations for newly discovered risks ▪ Develop the AI assurance plan	▪ Collaboratively identify risks with developers	▪ Identify and disseminate standard language on AI assurance requirements and other contractual considerations to inform development	▪ Define and disseminate domain-specific standards that inform verifiable requirements for AI assurance to inform development	▪ Test and evaluate continously during development to discover risks and validate mitigations ▪ Contribute findings to the AI assurance plan	▪ Identify and develop relevant standards and set guidelines for sector-specific risks	▪ N/A
Acquisition	▪ Capture risks and associated built-in and operational mitigations with user community ▪ Support testing by AI-enabled system specific instrumentation guidance	▪ Participate in human-in-the-loop evaluations	▪ Manage acqusitions to ensure AI assurance requirements are met in the context of mission needs	▪ N/A	▪ Test and evaluate to ensure AI assurance requirements are met in the context of mission needs ▪ Ensure validity and efficacy of AI assurance plan	▪ Identify and develop relevant standards for sector-specific risks and create regulations	▪ N/A
Certification	▪ Support testing by AI-enabled system specific instrumentation guidance	▪ Participate in human-in-the-loop evaluations	▪ Articulate and prioritize actionable operational needs	▪ Define and disseminate domain-specific standards that inform verifiable requirements for AI assurance to drive certification	▪ Test and evaluate to ensure AI assurance requirements are met in alignment with applicable standards	▪ Make operation authorization decisions for regulated systems based on evidence and existing regulation	▪ Take handoff from testers post certification, understanding risks to be monitored
Deployment	▪ Update AI-enabled system as new risks are discovered during operation ▪ Support the implementation of the AI assurance plan	▪ Report post-deployment incidents	▪ Oversee sustainment contracts with risk monitoring ▪ Oversee the implementaiton of the AI assurance plan	▪ Update standards, risk databases, and guidance as data is collected from operations	▪ Test and evaluate when called upon by Monitors post deployment ▪ Contribute findings to the AI assurance plan	▪ Review post-deployment monitoring results for regulation compliance and extend operation authorization for regulated systems	▪ Periodically or continuously monitor risks during operation as per the AI assurance plan ▪ Inform Program Offices and Regulators when risks post deployment are not acceptable per AI assurance plan ▪ Update AI Assurance plan

8. Concluding Remarks

We previously argued [9] 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 R&D 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 [10]. 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.

References

[1] The MITRE Corporation, "MITRE-Harris Poll Finds Lack of Trust Among Americans in AI Technology," February 2023.

[2] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST Trustworthy and Responsible AI, National Institute of Standards and Technology," Gaithersburg, MD, 2023.

[3] U.S. Office of the President, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," United States White House, Presidential Actions, 2023.

[4] National Institute of Standards and Technology (NIST), "Guide for Conducting Risk Assessments, NIST Special Publi-cation 800-30 (Rev. 1)," Department of Commerce, NIST Joint Task Force Transformation Initiative, Washington,
D.C., 2012.

[5] G. Pahl, W. Beitz, J. Feldhusen and K. H. Grote, Engineering design: a systematic approach, vol. 3, London: Springer, 1996.

[6] K. T. Ulrich and S. D. Eppinger, Product design and development, McGraw-Hill, 2016.

[7] National Institute of Standards and Technology (NIST), "Managing information security risk: Organization, mission, and information system view, NIST Special Publication 800-39," Department of Commerce, NIST Joint Task Force Transformation Initiative, Washington, D.C., 2011.

[8] The MITRE Corporation, "Mitigations," April 2023. [Online]. Available: https://atlas.mitre.org/mitigations. [Accessed 25 March 2024].

[9] The MITRE Corporation, "A Sensible Regulatory Framework for AI Security," June 2023.

[10] The MITRE Corporation, "Principles for Reducing AI Cyber Risk in Critical Infrastructure: A Prioritization Approach," October 2023.

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.

Acknowledgements

The authors would like to thank Benjamin Wellner, Ph.D.; Flo Reeder, Ph.D.; Christina Liaghati, Ph.D.; Avinash Pinto; Mark Pfaff, Ph.D.; Ryan Novak; Rachel Giachinta; and Kevin Forbes for their contributions to the capability descriptions and application examples. The authors would also like to thank Yifty Eisenberg, Ph.D.; Eric Hughes, Ph.D.; Eric Bloedorn, Ph.D.; Miles Thompson; Elena Charnetzki; Andrew McLauchlin; Andy Anderegg; Chris Sledjeski, Ph.D.; and EJ Hillman for their thoughtful review and suggestions for improvement.

For more information about MITRE's AI Assurance expertise, please visit mitre.org/AI or email AI@Mitre.org.

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®