MITRE's Response to the NIST RFI on Under Sections 4.1, 4.5, and 11 of the Executive Order Concerning Artificial Intelligence

February 2, 2024

For additional information about this response, please contact:

Duane Blackburn
Center for Data-Driven Policy
The MITRE Corporation
7596 Colshire Drive
McLean, VA 22102-7539

policy@mitre.org
(434) 964-5023

About MITRE

MITRE is a not-for-profit company that works in the public interest to tackle difficult problems that challenge the safety, stability, security, and well-being of our nation. We operate multiple federally funded research and development centers (FFRDCs), participate in public-private partnerships across national security and civilian agency missions, and maintain an independent technology research program in areas such as artificial intelligence, intuitive data science, quantum information science, health informatics, policy and economic expertise, trustworthy autonomy, cyber threat sharing, and cyber resilience. MITRE's ~10,000 employees work in the public interest to solve problems for a safer world, with scientific integrity being fundamental to our existence. We are prohibited from lobbying, do not develop, or sell products, have no owners or shareholders, and do not compete with industry—allowing MITRE's efforts to be truly objective and data-driven. Our multidisciplinary teams (including engineers, scientists, data analysts, organizational change specialists, policy professionals, and more) are thus free to dig into problems from all angles, with no political or commercial pressures to influence our decision making, technical findings, or policy recommendations.

MITRE has a 50-year history of partnering with federal agencies to apply the best elements of artificial intelligence (AI) and machine learning (ML) to advance agency missions while developing and supporting ethical guardrails to protect people and their personal data. Our team's experience with the entirety of the AI/ML adoption and life cycle has strengthened our ability to anticipate and resolve future needs that are vital to the safety, well-being, and success of the public and the country.

Introduction and Overarching Comments

In response to the National Institute of Standards and Technology's (NIST) recent Request for Information (RFI), MITRE offers evidence-based inputs on selected topics to aid NIST in implementing tasks outlined in Executive Order 14110 (Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence),[1] issued on October 30, 2023. This includes the development of standards, tools, methods, and practices for managing synthetic content, including potentially harmful content.

Our inputs highlight a systematic process for assessing the assurance of AI systems within their sociotechnical contexts, the importance of red-teaming, AI incident and risk mitigation sharing, strategies for reducing the risk of synthetic content, and an AI lifecycle framework for managing risks and promoting trustworthiness.

These inputs are intended to provide valuable insights to inform policy decisions and industry practices, and to support NIST's mission to extend and operationalize the AI Risk Management Framework. We believe that our recommendations will contribute to the development of science-

[1] Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. 2023. Executive Office of the President, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence. Last accessed January 25, 2024.

backed, non-proprietary standards and techniques for AI safety and security, fostering a safer and more trustworthy AI landscape.

NIST Should Approach its Assigned Tasks with a Holistic Perspective

As NIST embarks on fulfilling its various mandates outlined in the Executive Order, it is crucial to adopt a holistic perspective rather than viewing each requirement as an isolated task. To facilitate this integrated approach, MITRE proposes that NIST consider developing a strategic approach analogous to its own Cybersecurity Framework[2]. This renowned framework offers a structured, adaptable, and risk-centric methodology for managing cybersecurity risks, and its principles could be effectively tailored to the context of AI.

Adopting such an approach would yield significant benefits for NIST and all entities implementing the Executive Order. First, it would provide a coherent structure that links all the disparate elements of AI safety, security, and global technical standards development, thereby ensuring consistency and alignment across all activities. This would facilitate the identification and management of interdependencies and potential overlaps or gaps among the different requirements.

Second, a risk-based approach would enable NIST to prioritize activities based on their potential impact on AI safety and security, ensuring efficient use of resources. It would also facilitate the identification and mitigation of risks throughout the AI lifecycle, from development to deployment and use.

Third, the flexibility inherent in the Cybersecurity Framework would allow NIST to adapt its approach as AI technologies and their associated risks evolve. This is particularly important given the rapid pace of AI development and the emergence of new risks and challenges.

Finally, this strategic approach would be particularly beneficial in helping NIST fulfill its requirements to "Develop Guidelines, Standards, and Best Practices for AI Safety and Security" and "Advance Responsible Global Technical Standards for AI Development". By providing a structured approach to these tasks, NIST would help ensure that the guidelines, standards, and best practices they develop are comprehensive, consistent, and aligned with the overall goal of promoting AI safety and security.

In summary, by adopting a strategy analogous to the Cybersecurity Framework, NIST can develop a comprehensive, flexible, and risk-based approach to managing AI risks, thereby enhancing the safety, security, and trustworthiness of AI systems.

[2] (DRAFT Update) The NIST Cybersecurity Framework 2.0. 2023. National Institute of Standards and Technology, https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.ipd.pdf. Last accessed February 2, 2024.

MITRE Inputs on Issues Posed in the RFI

#1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security

Overarching Gap in AI Safety and Security Assurance Guidelines, Standards, and Best Practices

MITRE's observation is that there is a critical gap in the current guidelines, standards, and best practices for ensuring the safety and security of AI systems. This gap presents a significant challenge as AI systems continue to be integrated into a variety of sociotechnical contexts, where they interact with and influence human activities and societal structures. To address this gap, MITRE recommends a disciplined, systematic approach to discovering the risks and evaluating the assurance of AI-enabled systems within their specific sociotechnical use context. This approach entails:

1. Decomposing and Prioritizing Assurance Risks. This involves breaking down and prioritizing the assurance needs of a given AI-enabled system in its context of use. It requires a deep understanding of the system's intended function, its interaction with users and other systems, and the potential risks and vulnerabilities it may present.
2. Designing and Conducting Targeted Assurance Investigations. This involves designing AI assurance discovery investigations to evaluate the mission effectiveness of AI-enabled systems in their use contexts. Such investigations should help system developers explore and answer key questions pertaining to assurance risks, ideally early in the development process. Answering such questions may require participation of key stakeholders, as each stakeholder may perceive and weigh risk differently. It may also require simulating mission scenarios at an appropriate level of abstraction to generate sufficient and representative evidence about the likelihood and severity of prioritized assurance risks.
3. Aggregating and Disseminating Assurance Knowledge. Throughout this assurance risk discovery and evaluation process, it is crucial to continuously aggregate and leverage rapidly emerging knowledge on AI assurance, describing types of assurance concerns and risks, and associated evaluation methods (including tools and metrics), domain-specific use cases, data sets, previous evaluation findings, and mitigations. This knowledge base should be continuously and collaboratively expanded with new insights, findings, and best practices from the field. Effective dissemination of such knowledge can support the development of evidence-based guidelines, standards, and best practices for a variety of AI assurance concerns, including AI safety and security. Over time, it may also inform the development of automated evaluation and mitigation approaches.

This approach can be instrumental to NIST's efforts to respond to the directives outlined in the AI Executive Order and the objectives of the new U.S. AI Safety Institute. These include extending and operationalizing the NIST AI Risk Management Framework; creating guidance and benchmarks for evaluating AI capabilities; establishing guidelines and processes for conducting red-teaming tests; evaluating existing standards, tools, methods, and practices; and creating guidelines for agencies to evaluate the efficacy of differential-privacy-guarantee protections for AI.

To substantiate this approach, MITRE has recently launched a new laboratory that is developing systematic procedures for discovering, characterizing, and evaluating assurance risk to AI-

enabled systems in their sociotechnical context of use. This laboratory serves as an experimental platform for implementing and refining our recommended approach and will provide insights and evidence to inform policy decisions and industry practices.

To further articulate this approach, we will discuss four focus areas in the following sections: AI red-teaming, AI incident and risk mitigation sharing, reducing the risk of synthetic content, and an AI lifecycle framework for managing risks and promoting trustworthiness. Each presents unique challenges and opportunities associated with the overarching dimensions of the AI assurance risk discovery and evaluation approach outlined above and informs the landscape of AI safety and security assurance.

AI Red-Teaming

Red-teaming is an investigative process that simulates attacks on real-world systems to identify vulnerabilities, mitigate potential exploits, and improve the overall security posture of a system. Within a system's development cycle, red-teaming is generally viewed as a complement to test and evaluation (T&E), as it serves as a means to understand how to manipulate a system's behavior in unintended ways compared to its initial operational evaluation. In the context of AI, every part of the AI chain, including the model's information input, training data, model weights, and computing hardware, is susceptible to attacks. Through standing up an AI red-team, individuals can engage in exercises intended to improve the AI system's overall robustness and security prior to initial operation.

MITRE recommends AI red-teams be multidisciplinary and collaborate with system builders and defenders within an iterative development cycle (i.e., purple-teaming). The cybersecurity red-teaming model serves as an excellent foundation for structuring an AI red-team, particularly within the context of the Build-Attack-Defend (BAD) development framework. In this framework, system development is an iterative process that involves the collaboration of multiple interdisciplinary teams, each serving a distinct function, whose common goal is the creation of a functional, performant, and secure system. By integrating the BAD framework into an AI development cycle, groups can proactively identify and address potential weaknesses, enhancing the overall performance, security, and resilience of an AI-based system. The AI red-team (the "Attackers" within the BAD framework) consists of AI security professionals with diverse knowledge and skills, such as proficiency in AI model analysis, manipulation, and training, and expertise in conducting strong, adaptive adversarial attacks against a system. To help maximize the AI red-team's effectiveness, it is recommended that they partner with both AI developers in charge of the target system design (the "Builders" within the BAD framework) as well as security professionals responsible for the robustness of the AI system (the "Defenders" within the BAD framework).

MITRE recommends AI red-teams adopt a repeatable framework for conducting red-teaming exercises. For any given AI domain, the AI red-team can best exercise its role through the following set of sequential actions: analyzing the target system for attack pathways, formulating adversarial attack vectors against the system, executing the attack against the system, and assessing the attack impact to the system within its operational environment. During target system analysis, the red-team collects information about the AI model to be attacked (modality), access points for the attack (physical, digital, during training, inference), and human involvement during attack operation (human in the loop [HITL], human out of the loop [HOOTL], or human on the loop [HOTL]). With this knowledge, the red-team then works to formulate sufficient

attacks against the target system, identifying the attack objective, attack execution constraints, appropriate attack evaluation metrics, and level of persistence necessary for attack success. Once the attack is formulated and executed against the target system, its effectiveness and resulting impact can be measured and shared with various stakeholders of the system, including other teams that are part of the BAD framework.

AI red-teaming can be used to help uncover unique generative AI vulnerabilities, such as sensitive information exposure and system override for malicious output generation. From a security standpoint, generative AI, and dual-use foundation models in particular, expose a variety of consequential attack vulnerabilities exploitable by an adversary, some that are shared with other types of AI, and others are unique to generative AI specifically. Leveraging a generative AI red-team can help assess the level of risk associated with these unique vulnerabilities, some of which are highlighted as use cases as part of this RFI. For example, AI red-teaming can be used to evaluate a model's sensitivity to exposing samples used as part of a training data set, which could expose information such as personally identifiable information to an attacker. Red-teaming a generative AI model can also help in evaluating an adversary's ability to expose or override the model's underlying system prompt, which is often used to guide and safeguard a model from generating harmful, misleading, and malicious output. As a general rule, any known adversarial attack against an AI model can serve as a test for an AI red-team to perform in order to assess the attack's impact to their system. Thus, as the threat landscape against generative AI and dual-use foundation models continues to evolve, so will the benefits of integrating AI red-teaming into a generative model's development cycle.

Red-teaming should not be conducted in a vacuum nor viewed as an exhaustive assessment of risk, and red-teams should be mindful of the rapidly evolving nature of the generative AI vulnerability landscape. Although red-teaming can provide a beneficial assessment of unwanted behaviors and potential vulnerabilities to an AI system, it should not serve as an exhaustive method for understanding all risks posed to an AI system. For example, most red-teaming exercises illustrate the consequence of attacking an AI model, but they do not generally provide an in-depth explanation of what the attack is doing inside the model to cause a certain behavior. This is especially true for dual-use foundation models, as they are usually very large in parameter size and difficult to interpret. Additionally, red-teaming generative AI is still a relatively new practice, and it appears that "traditional" red-teaming techniques are often applied to the generative AI space. There may still be unique vulnerabilities posed to generative AI models that are yet to be discovered, so it is important to caveat any current red-teaming assessments with this point in mind.

AI red-teams should prioritize sharing their practices and findings both with internal stakeholders and externally with the larger AI security community. As the AI space continues to evolve, maintaining sufficient awareness of the vulnerability landscape facing these systems can be greatly increased through red-teams sharing the process and findings of their exercises with various stakeholders, not only with an internal development/operational team, but also to external communities to foster compliance and trust through transparency and disclosure. The information shared with all stakeholders should encompass the insights gained during target system analysis, assumptions made during attack formulation, and a comprehensive report on attack execution, along with relevant task and operational metrics. A great example of this is

illustrated in a recent publication[3] that overviews the performance and safety/security risks associated with an openly available model without revealing internal specifics of the developer's intellectual property. Incentivizing this type of system reporting framework as an information sharing standard could help maintain an open dialogue across the private and public sectors on the overall security of various generative AI models, aided in large part by various red-teaming methods and processes.

AI Incident and Risk Mitigation Sharing

NIST should promote a strategic practice focused on understanding, preventing, and responding to AI incidents to effectively manage and mitigate risks. Such incidents are unexpected behaviors of AI leading to significant disruption; they can range from performance degradation to exploitation of security vulnerabilities, and may even be unintended consequences not accepted in the AI system's design. The potential impact of such incidents can be severe, threatening the integrity of AI systems and causing considerable disruption.

A key component of this strategic practice should be the development of guidelines for open-source incident reporting that would help AI system operators better understand previous incidents and implement measures to prevent similar ones in the future. Additionally, this reporting would foster knowledge sharing across a broad user base, providing valuable insights regarding AI incidents, risks, and potential mitigations.

MITRE recommends a multi-modal approach for incident sharing. This could involve hosting community meetings where representatives from different organizations share information about AI security priorities, concerns, and real-world incidents. Rapid anonymized sharing across a trusted community would enable data-driven risk analysis, while public sharing could capture unique attack pathways for broader understanding and learning.

Ongoing community-driven work by MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) to improve AI incident and vulnerability reporting is an operationalized example of this strategic practice.[4] ATLAS is collaborating closely with both industry and government partners to develop metrics for AI incident and vulnerability reporting, host protected incident sharing exchanges, and facilitate both anonymized and public reporting to improve community awareness. This comprehensive approach fosters a culture of learning and continuous improvement in AI security and risk management, which could serve as a valuable model for NIST.

Reducing Risk of Synthetic Content

Reducing the risk of synthetic content requires a comprehensive and dynamic approach that combines technological solutions, policy measures, and continuous monitoring and adaptation to the evolving threat landscape.

Synthetic content, particularly that generated by AI systems, presents unique risks. These risks extend beyond the creation of deepfakes and misinformation, and can include spearfishing, scaled cyber attacks, and the spread of harmful outputs. For instance, the dissemination of

[3] GPT-4 System Card. 2023. OpenAI, https://cdn.openai.com/papers/gpt-4-system-card.pdf. Last accessed January 31, 2024.

[4] MITRE ATLAS™. 2024. MITRE, https://atlas.mitre.org/. Last accessed January 24, 2024.

deepfake video, audio, or imagery can significantly impact individuals and societies, leading to potential societal harms such as disinformation campaigns, identity theft, and privacy violations.

Generative AI systems have specific vulnerabilities that can be exploited. These include hallucinations, where the AI system generates outputs that deviate significantly from reality, and the manipulation or misalignment of open-model weights, which can lead to the production of misleading or harmful outputs. Furthermore, the AI system's inference Application Programming Interfaces (APIs) can be abused to create and disseminate synthetic content widely.

Mitigating the risks associated with synthetic content involves a multi-faceted approach. This includes penetration testing and red-teaming to identify and address vulnerabilities, developing robust detection algorithms that can identify and flag synthetic content, and implementing watermarking techniques to track and authenticate content. HITL model usage and guardrail integration can also help in controlling the outputs of the AI system. More advanced techniques such as machine unlearning, which involves removing specific data from the AI system, can also be explored as a potential mitigation strategy.

However, these mitigations also have their limitations. Recent papers[5,6] show that it is easy to bypass defenses such as watermarking and synthetic content detection, leading to a potential circular "arms race" of discovering, bypassing, and rediscovering mitigations of synthetic content risks. Detection algorithms, while continually improving, may struggle to keep up with the evolving sophistication of synthetic content generation. There is also a growing trend of releasing many open-source versions of foundation models that are trained by the larger generative AI community *without* the alignment data that prevents them from producing malicious or harmful output. Security measures can be bypassed by determined adversaries, and policies and regulations may face challenges in enforcement and compliance. Moreover, the balance between mitigating risks and preserving the beneficial uses of AI for generating synthetic content is a complex issue that requires careful consideration.

AI Lifecycle Framework for Managing Risks and Promoting Trustworthiness

MITRE has developed a full AI lifecycle framework for managing risks and promoting trustworthiness, named Systems Engineering Processes to Test AI Right (SEPTAR),[7] in support of the Office of the Secretary of Defense Directorate of Test & Evaluation and Analysis. The approach provides the results of research and collaboration (among academia, industry, and DOD) that can be used for guidance to the Department of Defense (DOD) on proactive steps that can be taken across the DOD's Adaptive Acquisition Framework to manage risk and improve the trustworthiness for AI Enabled Systems (AIES). The framework includes guidance in the following areas:

1. Requirements for AIES: At the initial stage, the framework helps to drive the teams to define clear expectations for what is important to address for an AIES. The framework

[5] Towards the Vulnerability of Watermarking Artificial Intelligence Generated Content. 2024. International Conference on Learning Representations (pending), https://openreview.net/pdf?id=xY4861TVUc. Last accessed January 24, 2024.

[6] On the Possibilities of AI-Generated Text Detection: A Sample Complexity Analysis. 2024. International Conference on Learning Representations (pending), https://openreview.net/pdf?id=oxEER3kZ9M. Last accessed January 24, 2024.

[7] C. Balhana, et al. Systems Engineering Processes to Test AI Right (SEPTAR) Release 1. 2023. MITRE, https://apps.dtic.mil/sti/trecms/pdf/AD1211716.pdf. Last accessed January 31, 2024.

explains specific activities for this phase, including determining data availability, creating mission-informed operational use cases (what the system should/should not do), bounding expectations for the hardware that enables the AIES, and performing prototyping (ideally competitively) to confirm assumptions and to assess the landscape of feasible solutions. This prototyping can be used to test the expected performance metrics. These metrics serve as quantitative measures of the system's performance and can provide insights into the system's reliability and trustworthiness.

Stakeholders, including system owners, users, and domain experts, are responsible for this phase and T&E professionals should also be engaged to support this phase to ensure that the system is designed to meet specific, measurable objectives.

This early definition of requirements allows for the identification of potential risks and vulnerabilities that may arise in relation to mission-specific tasks. It also sets the stage for the development of risk mitigation strategies that must be addressed by the broader AIES.

2. AIES Acquisition Strategy: During the acquisition phase, the framework guides the development of an acquisition strategy and other contractual agreements that consider unique aspects of AIES. Leaders need to clearly assert expectations on the rights to models and their training data (or Modeling & Simulation). To maximize information sharing there should be agreement on the development of model and data cards that provide the necessary pedigree of the AIES. Expectations on likely cybersecurity threats should be shared to ensure the system is built to withstand the known threats. Since an AIES is never fully done, the strategy can account for the full lifecycle of an AIES including ensuring the resources are in place for the needed sustainment (e.g., training data, subject matter experts, and Machine Learning Operations infrastructure).

   The acquisition team, which may include project managers, procurement officers, and legal advisors, is responsible for developing a comprehensive acquisition strategy. Their role is to ensure that the acquisition process aligns with the defined requirements and mitigates potential risks.

   These steps can help to manage risks associated with model/data ownership, cybersecurity vulnerabilities, and AIES sustainability.

3. AIES Development: In the development phase, the framework promotes the use of representative data for training and testing to fully enable the iterative train/test cycles. This iterative process extends to the integration into a broader system. Infrastructure (e.g., environment and integrated testing tools) to enable this to occur may be provided by the system owner (e.g., DOD) or by the contractor—each option has its own unique risks and costs. The ability to gain evidence throughout this process of model training and testing can help to identify and address potential vulnerabilities in the AI model or system and ensure that the system is developed in a way that meets mission-specific requirements.

   The development team, including software engineers, data scientists, and AI specialists, is responsible for developing the AIES according to the defined requirements. Their role includes using representative data for training and testing, integrating the AI model within the broader system, and addressing potential vulnerabilities.

4. Testing of AIES: The testing phase is crucial for identifying and addressing broader risks and vulnerabilities before the AIES is deployed. The framework guides the conduct of thorough testing, including developmental testing and operational testing.

   During the development and testing phases, the framework promotes the use of broader test evidence gathered across the full lifecycle to inform the extent of testing that must be addressed. At this phase it is critical to ensure representative data and/or conditions are present to test performance under expected conditions as well as how the AIES handles potentially unexpected conditions. The performance of the AIES on these data sets, measured using the defined metrics, can provide insights into the system's ability to perform its intended tasks accurately and reliably, contributing to its perceived trustworthiness.

   The testing team, which may include test engineers, data analysts, and domain experts, is responsible for conducting thorough testing of the AIES. Their role is to identify and mitigate potential risks before the system is deployed, ensuring that the system functions as expected and meets the defined requirements.

5. Deployment/Sustainment of AIES: Once the AIES is deployed, the framework promotes continuous monitoring and analysis, user feedback collection, and system updates. This allows for the ongoing identification and management of risks, and ensures that the system continues to meet mission-specific requirements and expectations. By strategically identifying these points in the lifecycle, the SEPTAR framework ensures a proactive approach to managing the risks associated with AI and mission-specific tasks.

   Tracking the system's performance over time should occur using the defined metrics ideally with integrated system monitors. User feedback is also collected and analyzed, providing a qualitative measure of user trust in the system. By defining and tracking these metrics, the SEPTAR framework provides a structured approach to measuring and enhancing the trust and trustworthiness of AIES.

   The operations team, including system administrators, IT support staff, and user trainers, is responsible for the deployment and sustainment of the AIES. The model developer will have a key role in this process to establish a cadence or conditions that warrant model retraining to occur.

SEPTAR offers a valuable framework for NIST in developing guidelines, standards, and best practices for AI safety and security. It provides a comprehensive, lifecycle-based approach to managing risks of AIES, and offers a roadmap for addressing emerging challenges and measuring critical aspects of trustworthiness.

#3. Advance Responsible Global Technical Standards for AI Development

This section offers insights into selected aspects of AI standards development; the AI lifecycle discussion in #1 above also applies here. It delves into the strategic approach to standards prioritization, training standards, evaluation standards, deployment standards, and large foundation model tuning. Each topic focuses on specific elements within the broader context of AI-related consensus standards. While these topics are discussed individually, it's important to note that in practice, there will be overlaps and interconnections among them.

In developing these standards, input from a diverse range of stakeholders, including AI researchers, practitioners, policymakers, and users, should be considered. This will ensure that the standards are comprehensive, practical, and aligned with the needs and expectations of the AI community and the broader society.

Strategic Approach to Standards Prioritization

When it comes to advancing responsible global technical standards for AI development, a strategic approach to standards prioritization is crucial. This approach should consider two key factors: the timing of technology maturity and the industry's consensus on readiness.

Timing of Technology Maturity. Standardization must occur at the appropriate time in the lifecycle of a technology. If standardization happens too early, it can stifle innovation and limit the exploration of better technology solutions. Conversely, standardizing too late can lead to vendor lock-in, a lack of interoperable technology alternatives, and a constrained marketplace. Therefore, it's essential to strategically time the standardization process to coincide with the maturity of the technology.

Following Industry's Consensus on Readiness. The industry's consensus on the readiness of a technology for standardization is another critical factor. AI continues to be a rapidly growing and evolving field of technologies, with dual-use foundational models as a prime example. NIST will benefit from industry and expert guidance on when AI innovation and evolution has reached sufficient consensus with regard to architecture and practice. This will allow stages and components of AI systems to be exposed through standardized interfaces and evaluation protocols. Prior to reaching a level of industry consensus, NIST and other research funding agencies should work with the AI community to identify such opportunities for interfacing with AI system stages and components, and support technology advancement through targeted research investments accelerating readiness for standardization.

Training Standards

Training standards are a critical aspect of AI system development. These standards guide the process of training AI models, ensuring that they learn in a way that is efficient, effective, and aligned with the intended goals of the system. The process of establishing training standards involves several key steps:

- Defining the Task/Output: The first step in establishing training standards is to clearly define the task or output of the generative AI model. This involves identifying the data set needed for training and determining how to evaluate model performance.
- Data Collection and Preprocessing: The next step involves collecting and preprocessing the data for training. This is a crucial step as the quality and diversity of the training data can significantly influence the performance and behavior of the AI model.
- Selecting Model Architecture: Depending on the task, the appropriate model architecture (such as transformer, diffusion, etc.) should be selected for training.
- Foundational Training: The model should then be foundationally trained to perform its task. This involves setting the objective function, sweeping across hyperparameters, and monitoring for collapse or overfitting.

- Alignment Tuning: Depending on the task, additional alignment tuning may be required. Techniques such as Reinforcement Learning from Human Feedback and guided ethical adherence integration can be used to fine-tune the alignment of the AI model.

Training standards should also address ethical considerations, including ensuring the privacy and confidentiality of data, mitigating bias and discrimination, and promoting transparency and accountability in the training process.

It's important to note that training standards should be flexible and adaptable to accommodate the rapidly evolving field of AI. They should be regularly reviewed and updated to reflect the latest advancements and insights in AI research and practice.

Evaluation Standards

Evaluation standards form a crucial component of AI system development and deployment, providing a structured framework for assessing performance, reliability, and safety. Recognizing that there are multiple types of AI performance evaluations, it's important to consider multiple approaches for evaluation standards. Drawing from NIST's evaluation work with the biometrics community, evaluation standards can be conceptually organized into three main categories: technology, scenario, and operational evaluations.

- Technology evaluations focus on the abilities of AI algorithms. These are predominantly helpful for assessing and tracking (and indeed, driving) progress on core capabilities and fundamental issues over time.
- Scenario evaluations provide baseline assessments of how a system with an AI component will perform in a specific application. Importantly, this "system" also includes human operator considerations, and results are typically specific to the application and system tested and may not be generalized to different systems or applications.
- Operational evaluations assess a specific system in a specific use case at an actual operational location. Due to the inherent difficulties in obtaining ground truth data for such performance assessments, these typically analyze other important factors such as workflow impact or customer experience.

Evaluation metrics, testing protocols, and targeted benchmarks can vary significantly across these three categories, but generalized standards and best practices within each are possible and would be beneficial.

In addition to these evaluation categories, the standards should also consider the ethical and societal impacts of AI systems. This includes potential bias, discrimination, privacy and security impacts, and compliance with ethical guidelines and regulations. Given the dynamic nature of AI technologies, evaluation standards and practices should be continuously monitored and updated to reflect the latest advancements and insights in AI research and practice.

Deployment Standards

Deployment standards provide guidelines for the successful and safe implementation of AI systems in operational environments. These standards should address various factors to ensure that the AI system functions effectively and ethically in its intended context.

- System Compatibility: Deployment standards should ensure that the AI system is compatible with the existing infrastructure where it will be deployed. This includes compatibility with hardware, software, and network systems.
- Security and Privacy: The standards should include guidelines for ensuring the security and privacy of the AI system and the data it processes. This includes measures for data encryption, user authentication, and privacy-preserving techniques.
- Operator Training: The standards should provide guidelines for training operators to use the AI system effectively and safely. This includes understanding the system's functionalities, managing potential risks, and responding to system outputs appropriately.
- Performance Monitoring and System Logging: Deployment standards should provide guidelines for continuous monitoring of the AI system's performance after deployment. This includes tracking the system's accuracy, reliability, and efficiency, and identifying any deviations from expected performance. System logging should be implemented to record the system's operations and performance metrics, providing a valuable resource for troubleshooting and system improvement.
- Ethical and Legal Compliance: The standards should ensure that the deployment of the AI system complies with ethical guidelines and legal regulations. This includes considerations for fairness, transparency, accountability, and respect for user privacy.
- Oversight and Reporting: Deployment standards should include provisions for oversight of the AI system's operations and performance. This includes regular reviews and audits of the system's performance, security, and ethical compliance. Reporting mechanisms should also be established to communicate the system's performance and any issues or incidents to relevant stakeholders.
- Maintenance and Updates: Deployment standards should also include provisions for regular maintenance and updates of the AI system. This ensures that the system remains up-to-date with the latest advancements in AI technology and continues to perform optimally in its operational environment.

Large Foundation Model Tuning

In the context of the RFI's request for "suggestions for AI-related standards development activities," the tuning of large foundation models should be a key area of focus. This process involves modifying these models for more specific tasks or broader use cases, and it has substantial implications for the safety, effectiveness, and consistency of AI systems. There are two general approaches to tuning large foundation models: fine-tuning and merging the model weights, and modifying the input prompting. Each of these approaches presents unique opportunities for the development and application of AI standards. In exploring these methods, we recommend AI-related standards development activities, including insights into existing processes, potential contributions to the current standards landscape, and identification of gaps that could be addressed. This work should also consider the specific impacts of AI, as understanding these impacts is crucial for the development of effective and responsive standards.

Fine-tuning and merging model weights

The base and initial fine-tuned weights from large foundation models (e.g., Llama 2 70B and Llama 2 70B Chat)[8] have extremely useful and applicable qualities. The ability to further modify these weights, however, opens entirely new areas of research on fine-tuning and merging for downstream purposes. Fine-tuning involves re-training only a small number of the original weights, avoids the risk of the model weights collapsing, and allows for models to be re-trained on a much smaller scale of hardware than the original training approach. Merging involves taking an average or combination of model weights to produce a new model that has improved performance over the original models' weights, and can be done without specialized AI-specific and parallel computing hardware.

It is important to note that any modifications to a model's weights permanently compromise the initial model's safety and guardrails metrics.[9] In developing standards on the safety of modified foundation models, it is important to consider the downstream effects of weight modification when considering and evaluating models for future safety, bias, ethical, and other areas of AI assurance importance.

Prompt-tuning

Prompt-tuning involves modifying only the input to a model to achieve a modified and/or desired result. Prompt-tuning can guide a model to a better solution by providing context, specifying examples, giving specific instructions, etc. (e.g., telling a model "Write Python code" vs. "You are a Python developer looking to write an application using the FastAPI library…"). This approach requires no direct access to the model's weights, and requires as little as an internet connection to a model's API or user interface.

While prompt-tuning can produce model output that is better aligned to a user's intended output, it is important to note that there are adversarial approaches to prompt modification that can produce unintended and even malicious results.[10] Adversarial prompting has been shown to subvert alignment and guardrails, extract sensitive training data, and even damage the model hosting system via arbitrary code execution.[11] Standards focusing on foundational model prompt vulnerabilities should consider the severity hierarchy of attack vectors, as well as the potential to detect and counteract attempts to bypass standard system prompting.

[8] H. Touvron, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023. Meta, https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/. Last accessed January 25, 2024.

[9] X. Qi, et al. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!. 2023. Arxiv, https://arxiv.org/abs/2310.03693. Last accessed January 30, 2024.

[10] A. Zou, et al. Universal and Transferable Adversarial Attacks on Aligned Language Models. 2023. Arxiv, https://arxiv.org/abs/2307.15043. Last accessed January 30, 2024.

[11] K. Greshake, et al. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. 2023. AISec '23: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, https://doi.org/10.1145/3605764.3623985. Last accessed January 30, 2024