MITRE's Response to the NTIA RFI on Artificial Intelligence Accountability

June 9, 2023

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 approximately 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 commercial products, have no owners or shareholders, and do not compete with industry. 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) while developing and supporting ethical guardrails to protect people and their personal data. Our experience with the entirety of the AI/ML lifecycle has strengthened our ability to anticipate and solve emerging technical and adoption challenges that are vital to the safety, well-being, and success of the public and the country.

MITRE recognizes AI assurance (including accountability mechanisms) as a national priority that is critical to realizing AI's transformational potential. We need AI assurance to facilitate equitable AI adoption at scale for citizens, organizations, governments, and society at large, which will pave the way for continued responsible innovation. Advancing AI assurance will establish standards of practice for providing data-driven evidence required by policymakers, regulators, and civil liberties organizations to deploy AI technologies for the public good.

Introduction and Overarching Recommendations

Over the past ten years, and especially over the past six months, AI has gone through tremendous technological advancement. It has the potential to help tackle high-stakes, mission-critical problems ranging from healthcare to national security, and to accelerate the advancement of science and other technologies. There are also valid and growing concerns of bad actors augmenting their adversarial capabilities by leveraging AI, especially in cyber operations and mis/disinformation.

This has led to many recent policy proposals by various stakeholders on how to regulate AI, with some being abstract and others targeted at specific considerations. Unfortunately, there is not a current overarching framework in which to organize all these activities in a holistic, systematic manner. This gap will complicate the National Telecommunications and Information Administration's (NTIA) efforts in AI accountability—some elements raised in the Request for Comment (RFC) are duplicative of other community (or even federal) activities, and others will

need to work in concert with external activities outside NTIA's control for them to succeed. Absent an overarching organizing framework, NTIA will need to conceptually map how its activities rely on and/or support external AI regulatory endeavors to achieve their desired impacts.

Within the AI security subdomain, any attempt to secure or regulate a new technology should be informed by **vulnerabilities**, **threats** that exploit those vulnerabilities (either intentionally or unintentionally), and the ultimate **risk** of damage, harm, or loss. This allows us to effectively model the threats and manage the risks. A complication for performing vulnerability, threat, and risk analysis for AI is that there is actually a wide range of algorithmic technologies within the AI umbrella, as well as a wide variety of dissimilar use cases. To help organize beneficial analyses, MITRE divides the AI ecosystem into three broad categories based on how AI can be manifest in application:

- Engineered systems that use AI as a component or subsystem
- AI as an augmentation of human capabilities
- AI operating autonomously under its own agency

Differentiating these categories is important because the threats and risks differ based on how AI is manifest in application, as do the approaches to combating threats and risks. We recommend that NTIA employ a similar approach in its analyses, especially when addressing security and safety concerns.

At the same time, we must also recognize the human component. "AI systems are inherently socio-technical in nature, meaning they are influenced by societal dynamics and human behavior. AI risks—and benefits—can emerge from the interplay of technical aspects combined with societal factors related to how a system is used, its interactions with other AI systems, who operates it, and the social context in which it is deployed."[1] Human interplay within an AI lifecycle involves and impacts a broad range of stakeholders from end-users, developers, integrators, purchasers, deployers, and regulators to organizations and society. Therefore, AI assurance needs to be conceptualized and implemented in a socio-technical context throughout AI system development, verification, deployment, and governance—all of which will vary by use case. (AI assurance cannot be an afterthought or Band-Aid for managing risks after system deployment.)

Finally, we must also be mindful of temporal and regulatory-level considerations. There is a natural degree of tension between innovation and regulation. Ideally, regulations provide the structure and guardrails that enable innovation ecosystems to appropriately prosper. However, if regulations are written with insufficient understanding of the technologies driving innovation, and/or at the wrong time, they can stifle innovation. Conversely, if innovation activities are not structured via regulations, it can be exceptionally difficult, if not impossible, to overcome issues that the innovation creates.

As previously stated, there are many different forms of regulation being discussed, and each can support different aspects within a holistic oversight approach (community principles are much different than federal laws). But this must be done strategically rather than haphazardly. Selecting the appropriate level for a specific regulatory consideration will be critically important,

[1] Artificial Intelligence Risk Management Framework. 2023. NIST, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

and that level may evolve over time. Selecting the wrong level can cause regulatory objectives to not be met, could stifle needed innovation, or both. While doing so we must be particularly sensitive to embracing approaches on the more formal side of the spectrum to address specific regulatory concerns, such as federal rules/policies and laws, as these are difficult and time-intensive to change (as they require new rules/policies and laws to replace them). We must ensure that adding any regulatory aspect is done both at the right level and at the right time to ensure community maturation and accountability.

NTIA's goal "to provide reliable evidence to external stakeholders—that is, to provide assurance—that AI systems are legal, effective, ethical, safe, and otherwise trustworthy"[2] is thus an essential but challenging endeavor. The latter aspect is reflected by the disconnected collection of questions posed in the RFC. While this can be beneficial to receiving specific, targeted inputs from an RFC, in implementation NTIA will need to approach its work more systematically by leveraging the organizing concepts provided above.

Questions Posed in the RFI

1. What is the purpose of AI accountability mechanisms such as certifications, audits, and assessments? Responses could address the following:

   a) What kinds of topics should AI accountability mechanisms cover? How should they be scoped?
   b) What are assessments or internal audits most useful for? What are external assessments or audits most useful for?
   c) An audit or assessment may be used to verify a claim, verify compliance with legal standards, or assure compliance with non-binding trustworthy AI goals. Do these differences impact how audits or assessments are structured, credentialed, or communicated?
   d) Should AI audits or assessments be folded into other accountability mechanisms that focus on such goals as human rights, privacy protection, security, and diversity, equity, inclusion, and access? Are there benchmarks for these other accountability mechanisms that should inform AI accountability measures?
   e) Can AI accountability practices have meaningful impact in the absence of legal standards and enforceable risk thresholds? What is the role for courts, legislatures, and rulemaking bodies?

AI accountability mechanisms are building blocks within an AI regulatory framework to help ensure proper development and responsible application of AI capabilities for the public good. There are a variety of potential mechanisms to leverage within any aspect of the AI space, and selecting the appropriate one (at the right level and at the right time) is important—and can evolve. Specificity, with supporting analyses directly focused on the issue being considered, will be required to do this correctly; generic practice recommendations will not be useful.

[2] NTIA AI Accountability RFC. 2023. National Telecommunications and Information Administration, https://www.regulations.gov/document/NTIA-2023-0005-0001.

The Government Accountability Office (GAO), working with AI experts from across the federal government, industry, and nonprofit sectors, organized AI accountability into a four-part framework[3] to aide future analyses and activities:

- Data: Ensure quality, reliability, and representativeness of data sources and processing.
    - Data used to develop an AI model
    - Data used to operate an AI system
- Governance: Promote accountability by establishing processes to manage, operate, and oversee implementation.
    - Governance at the organizational level
    - Governance at the system level
- Monitoring: Ensure reliability and relevance over time.
    - Continuous monitoring of performance
    - Assessing sustainment and expanded use
- Performance: Produce results that are consistent with program objectives.
    - Performance at the component level
    - Performance at the system level

(MITRE addendum: A better way to consider the performance part of this framework is to think about efficacy. Performance at the system level may or may not lead to efficacy as systems always exist in some broader context.)

Within each of these four organizing principles are a host of factors that must be systematically considered (see the *Introduction and Overarching Recommendations* section, above), with accountability mechanism selection informed by analyses of vulnerabilities, threats, and risks. Considerations can include, but are not limited to, AI system safety, security, equity, reliability, interpretability, robustness, directability, privacy, and governability.

One recommended approach to structuring AI assurance assessments is to incorporate exploration and discovery within laboratory or sandbox environments prior to operational deployment, even during development. This involves working to advance context-sensitive, formative test and evaluation approaches to discover, chart, and mitigate emerging consequences of AI solutions, especially when AI intersects with human values in ways that are difficult to predict. Next-generation AI assessments should utilize methods and tools to more accurately assess how well an AI system will perform in its use context. Such an approach can also inform the development of more effective AI and human interactions in terms of efficiency and quality of AI-enhanced decision making.

AI accountability mechanisms do not exist in isolation. Each must not only be cognizant of how it supports or relies on other aspects within a holistic AI regulatory approach, but also (1) directly align with and support existing regulatory mechanisms within the domain of each specific use case, such as healthcare or law enforcement, which will have a unique variety of additional regulatory considerations, and (2) be specific as to which of the three categories of AI

[3] Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities. 2021. Government Accountability Office, https://www.gao.gov/assets/gao-21-519sp.pdf.

use is intended—engineered systems that use AI as a component or subsystem, AI as an augmentation of human capabilities, and AI operating autonomously under its own agency.

AI assurance via certifications, audits, etc., should be incorporated within existing industry accountability mechanisms, rather than a singular national approach, since regulated industry segments already have measures in place to manage their specific risks. The US should rely on existing sector-specific regulators, equipping them to address new AI-related regulatory needs.

Finally, AI accountability mechanisms need to consider the human element. Those leveraging AI must be accountable for proper usage of its output and should face consequences if they're careless. But doing so too forcefully will likely lead many to avoid unfamiliar risks by choosing not to leverage new AI capabilities. Overcoming these contrasting concerns calls for creating recommended decision-making processes for major and/or potentially problematic use cases. The country can then hold decision makers accountable for having gone through an informed decision-making process and possibly provide protections if they can demonstrate that they have done so, even if the eventual outcome is problematic.

2. Is the value of certifications, audits, and assessments mostly to promote trust for external stakeholders or is it to change internal processes? How might the answer influence policy design?

To be effective, accountability mechanisms need to *both* promote trust for external stakeholders *and* change internal processes.

Promote Trust for External Stakeholders. AI assurance (and its accountability mechanisms) is a lifecycle process that provides justified confidence in an AI system to operate effectively with acceptable levels of risk to its stakeholders. An AI system operating effectively means that it meets its functional requirements with valid outputs. Such mechanisms provide data-driven evidence required by policymakers, regulators, and civil liberties organizations to advance AI technologies for the public good. It should be noted that different stakeholders in the AI lifecycle may perceive risk differently and may have different tolerances toward risk. Therefore, trust is not a universal concept and understanding what level of risk may be acceptable to which stakeholder is critical in developing trustworthy AI systems.

Change Internal Processes. For instance, effectiveness-based test and evaluation, looking at AI as it is used, requires special consideration for internal assessment processes and is more difficult and expensive to evaluate than algorithmic performance alone. Also, we should not lose sight of the end goal, which is to ensure that AI adds value to whatever workflow it will be applied to. Any internal AI development or revisions to application processes should be about achieving superior AI applications that add value to the mission and benefit external stakeholders. If this objective is lost or missing, then there is reason for concern.

3. AI accountability measures have been proposed in connection with many different goals, including those listed below. To what extent are there tradeoffs among these goals? To what extent can these inquiries be conducted by a single team or instrument?

a) The AI system does not substantially contribute to harmful discrimination against people.
b) The AI system does not substantially contribute to harmful misinformation, disinformation, and other forms of distortion and content-related harms.
c) The AI system protects privacy.
d) The AI system is legal, safe, and effective.
e) There has been adequate transparency and explanation to affected people about the uses, capabilities, and limitations of the AI system.
f) There are adequate human alternatives, consideration, and fallbacks in place throughout the AI system lifecycle.
g) There has been adequate consultation with, and there are adequate means of contestation and redress for, individuals affected by AI system outputs.
h) There is adequate management within the entity deploying the AI system such that there are clear lines of responsibility and appropriate skillsets.

This question reinforces our point in the *Introduction and Overarching Recommendations* section that AI accountability activities need to be addressed systematically rather than by a collection of disconnected thoughts or activities. This proposed list is a combination of goals, values or guiding principles, process recommendations, and communications best practices. Analyzing them individually, or collectively, absent context of intended use and operational category, will not lead to proper decision making.

This again requires consideration of a variety of socio-technical matters. "(I)t is widely understood that AI models are part of larger systems, and these systems are embedded in socio-technical contexts. How models are implemented in practice could depend on model interactions, employee training and recruitment, enterprise governance, stakeholder mapping and engagement, human agency, and many other factors. The most useful audits and assessments of these systems, therefore, should extend beyond the technical to broader questions about governance and purpose."[4]

No single team or organization will be able to provide the breadth of national oversight and implementation required to provide sufficient AI assurance and accountability, especially since these activities cannot be done in isolation but rather support and exist within the context of domain-specific regulatory frameworks and multiple operational constructs.

As an alternative to a single team or instrument, one can envision a common foundational approach to AI assurance and accountability where sector regulators (and other AI actors) could adopt and adapt accountability mechanisms tailored to specific AI use cases. This is the vision behind the National Institute of Standards and Technology (NIST) AI Risk Management Framework that addresses many risks and concerns, and could be augmented with suggested accountability mechanisms.

[4] Artificial Intelligence Risk Management Framework. 2023. NIST, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

See also the prior Question #2 discussion on stakeholders not all perceiving or valuing risk in the same way, which is a primary reason why a single team or instrument cannot do this. A collaborative approach mitigates and diffuses those risks.

4. Can AI accountability mechanisms effectively deal with systemic and/or collective risks of harm, for example, with respect to worker and workplace health and safety, the health and safety of marginalized communities, the democratic process, human autonomy, or emergent risks?

AI accountability mechanisms can identify and mitigate systemic risks of harm by requiring policy impact assessments from fields related to such risks: for example, health impact assessments, environmental impact assessments, and racial equity impact assessments. These assessment frameworks would link a proposed AI technology deployment to predicted harms to individual and community well-being, which can be represented in terms of estimated impact to well-being indicators. To successfully identify such risks, such assessments involve soliciting causal models from diverse panels that include diverse direct and indirect users of the technology, as well as socio-technical professionals and the community knowledge system, in order to extrapolate from the AI deployment to potential harms considered in the assessments.[5,6,7,8]

Fully participatory methods are not always feasible, but human-in-the-loop simulation and computational social models can support AI impact assessment.[9] It is also important to conduct such human-in-the-loop simulations in an inclusive and collaborative manner with the stakeholders who will be impacted by a specific AI application. Impact assessment requires insight on the social and behavioral interaction between the AI systems and populations served, as well as environments—in vivo, qualitative human feedback, simulated, hybrid, and computational social models—that explore the downstream distributional effects of AI systems on individuals, communities, and institutions.[10]

5. Given the likely integration of generative AI tools such as large language models (e.g., ChatGPT) or other general-purpose AI or foundational models into downstream products,

[5] Y. Shaik, et al. Centering and collaborating with community knowledge systems: piloting a novel participatory modeling approach. 2023. International Journal for Equity in Health, https://equityhealthj.biomedcentral.com/articles/10.1186/s12939-023-01839-0.

[6] A Framework for Assessing Equity in Federal Programs and Policies. 2021. MITRE, https://www.mitre.org/sites/default/files/2021-11/prs-21-1292-equity-assessment-framework-federal-programs.pdf.

[7] An Equity Guide for Techies. 2021. MITRE, https://sjp.mitre.org/insights/Equity%20Guide%20for%20Techies.

[8] FAIR Framework: Designing Equity Through Community Voices and Systems Thinking. 2021. MITRE, https://sjp.mitre.org/insights/60f1e225b1d934001a56df51.

[9] MITRE Response to OSTP's RFI Supporting the National Artificial Intelligence Research and Development Strategic Plan. 2022. MITRE, https://www.mitre.org/sites/default/files/2022-03/pr-21-01760-16-mitre-response-ostp-rfi-national-artificial-intelligence-research-and-development-strategic-plan.pdf.

[10] Sociocultural Behavior Sensemaking. 2014. MITRE, https://www.mitre.org/sites/default/files/2021-11/pr-14-2487-Sociocultural-Sensemaking.pdf.

how can AI accountability mechanisms inform people about how such tools are operating and/or whether the tools comply with standards for trustworthy AI?

This is an active area of research, and it is important to note that such standards do not currently exist. What is usually being discussed are conceptual principles.

One such area is human-AI teaming with large language models (LLMs) where mechanisms are needed to support observability about sources, observability about credibility, and explainability of rationale behind LLM-generated responses. These are necessary to enable appropriate trust in LLMs and for their effective use. Better accountability can be achieved (1) by specifying requirements and building these features into the LLM during development and/or (2) through prompt engineering that directs the LLM to incorporate good teaming behavior when responding. MITRE has compiled a set of guiding principles in the publication-pending *Human-Centered Aspects of AI Assurance*, which can inform accountability mechanisms for generative AI tools like LLMs.

8. What are the best definitions of and relationships between AI accountability, assurance, assessments, audits, and other relevant terms?

Common terminology is critical for any field's advancement as it enables every professional to represent, express, and communicate their findings in a manner that is effectively and accurately understood by their peers. The AI assurance community has not yet coalesced on terminology, and the development of formal consensus standards takes a long time—it is a much slower process than the speed at which AI and AI assurance are currently evolving. Fortunately, there is an alternative approach that was successful on a prior urgent need and could be considered an interim solution: the National Science & Technology Council (NSTC) establishing definitions and requiring federal agencies to use them until consensus-based standards are developed.[11] The same approach could be leveraged here, starting with a rapid collaborative research activity to develop the definitions.

A recommended starting point is NIST's *The Language of Trustworthy AI: An In-Depth Glossary of Terms,*[12] which contains (where possible, standard) definitions including those for accountability, assessments, and audits, as well as fairness, safety, effective challenge, transparency, and trustworthiness.

When developing common terminology and definitions for AI accountability, the community should also seek to leverage other authoritative sources whenever possible. For example, from the National Security Commission on Artificial Intelligence Final Report (footnote 1 of Chapter 7),[13] the term:

---

[11] In the mid-2000s, the NSTC's Subcommittee on Biometrics published a "Glossary" document of biometric terms. As part of its formal approval, its parent NSTC Committees also instructed federal agencies to follow those definitions in their future activities. Non-governmental entities (mostly) aligned voluntarily as well and this document served as an important input in the development of an international vocabulary standard, which the subsequent NSTC Policy for enabling the Development, Adoption and Use of Biometric Standards formally required agencies to follow.

[12] Glossary. 2023. National Institute of Standards and Technology, https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary.

[13] Final Report. 2021. National Security Commission on Artificial Intelligence, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

- **Justified Confidence** is "taken from a widely used international standard, uses a specific definition of assurance as being 'grounds for justified confidence.' It notes that 'stakeholders need grounds for justifiable confidence prior to depending on a system' and that 'the greater the degree of dependence, the greater the need for strong grounds for confidence.' ISO/IEC/IEEE International Standard – Systems and Software Engineering – Systems and Software Assurance, IEEE/ISO/IEC 15026-1 (2019), https://standards.ieee.org/standard/15026-1_Revision-2019.html."

When a desired term has no accepted standard definition, then a community may choose to adopt a "local" term that over time with broad, consistent use may become a de facto standard definition. For example, MITRE has recently developed and adopted the following term:

- **AI assurance** is a lifecycle process that provides justified confidence in an AI system to operate effectively with acceptable levels of risk to its stakeholders. Effective operation entails meeting functional requirements with valid outputs. Assurance risks may be associated with or stemming from a variety of factors depending on the use context, including but not limited to AI system safety, security, equity, reliability, interpretability, robustness, directability, privacy, and governability.

9. What AI accountability mechanisms are currently being used? Are the accountability frameworks of certain sectors, industries, or market participants especially mature as compared to others? Which industry, civil society, or governmental accountability instruments, guidelines, or policies are most appropriate for implementation and operationalization at scale in the United States? Who are the people currently doing AI accountability work?

Any accountability mechanism needs data and evidence on which to form the basis of its analysis. MITRE has developed frameworks on which to gather and analyze important insights, which can be used to *inform* compliance foci and activities.

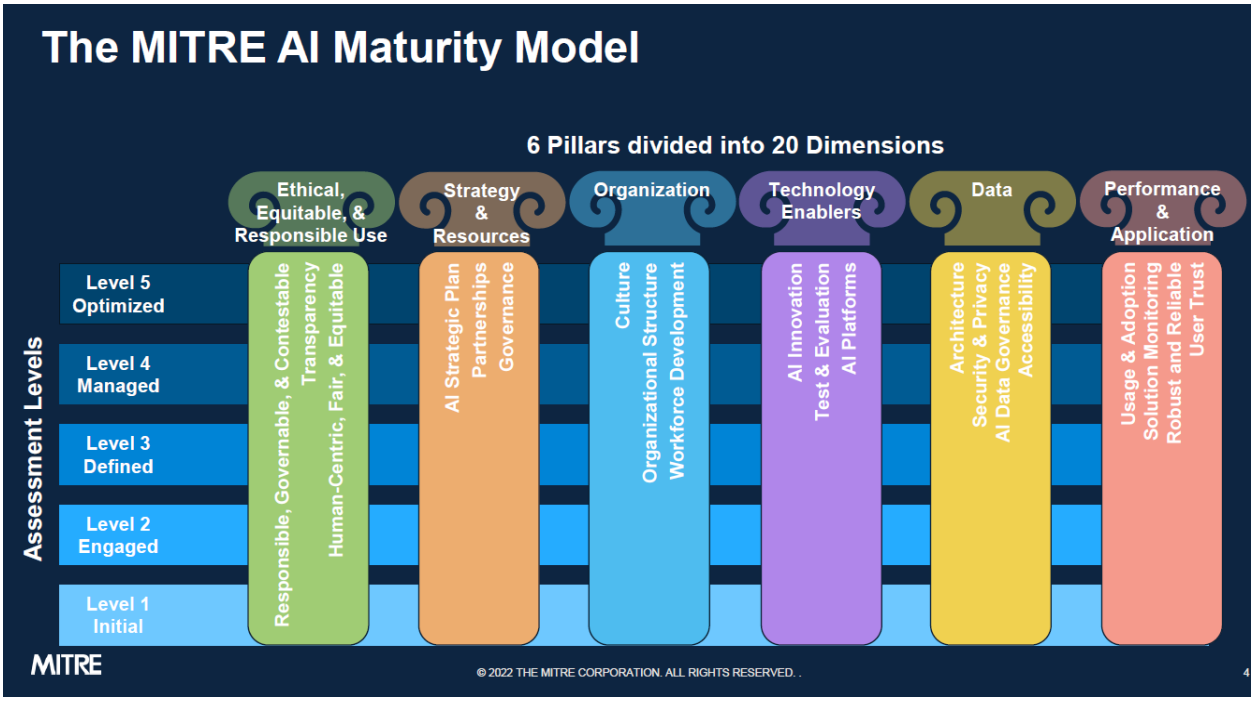
AI Governance Accountability. The MITRE Artificial Intelligence Maturity Model (AIMM)[14] and corresponding organizational Assessment Tool is a governance[15] accountability mechanism successfully being used by multiple federal agencies to support responsible AI adoption. AIMM was "designed to measure an organization's progress in AI maturity as it becomes increasingly adept at incorporating AI technologies and best practices into its work environment." The tool generates qualitative and objective data pertaining to an organization's AI readiness that can be used to gauge the organization's maturity at a given point in time, provide guidance on how the organization can take measured steps to improve its current state of AI maturity, and periodically reassess and track progress toward defined goals. In a time when the federal government is

[14] AI Assessment Tool and Maturity Model. 2023. MITRE, https://ai-platform.pages.mitre.org/projects/ai-maturity-assessment/.

[15] In Mökander, et al., "Auditing large language models: a three-layered approach" (https://arxiv.org/abs/2302.08500), the authors lay out a useful categorization for LLMs, which is extendable to foundations models and AI systems in general. These categories are governance audits, model audits, and application audits. Governance audits include reviewing the adequacy of organizational governance structures, creating an audit trail of the AI development process, and mapping roles and responsibilities within organizations that design AI systems.

looking to adopt AI responsibly across the board, a framework and tool like this is very relevant, accessible, and useful.

The graphic below summarizes the AIMM framework with the six "Pillars" listed across the top and five "levels" of increasing maturity listed down the left. The 20 "Dimensions" are distributed across their associated Pillars. The AIMM Assessment Tool enables an organization to gauge its level of maturity for each dimension based on a list of specific actions and activities needed to demonstrate mastery at each progressive level.



Protected Incident Sharing. Technology incident sharing enables accountability by supplying the stakeholder community (key actors being industry and government) with objective and actionable evidence about identified problems that need to be addressed. Such mechanisms alert technology developers of emerging and existing issues that require accountable attention.

The MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS™)[16] program is a community-wide approach to AI security with reputation-based accountability in the absence of regulation and reporting requirements. ATLAS provides security teams with the information needed to make the case to their leadership that there is a real problem, substantiated with detailed real-world data documenting an identified AI vulnerability. This empowers technology developers (and users) to discover, share, and mitigate AI security risks, reducing the duration and scale of resulting harms before an organization suffers reputational and/or financial damage. ATLAS incident sharing also arms policymakers with evidence of AI security problems impacting society that may warrant new regulation or regulation reform. Sharing timely real-world incident data on AI assurance problems is a mechanism to increase the effectiveness of

[16] MITRE ATLAS™. 2023. MITRE, https://atlas.mitre.org/.

new regulations and reforms while increasing accountability of subsequent reporting requirements.

10. What are the best definitions of terms frequently used in accountability policies, such as fair, safe, effective, transparent, and trustworthy? Where can terms have the same meanings across sectors and jurisdictions? Where do terms necessarily have different meanings depending on the jurisdiction, sector, or use case?

See our answer to Question #8, above.

12. What aspects of the United States and global financial assurance systems provide useful and achievable models for AI accountability?

The financial sector's model risk management framework[17] is a set of guidelines and practices designed to identify, assess, and manage the risks associated with the use of mathematical models in financial decision making. This framework is a useful template for aspects of AI accountability as it helps ensure the reliability, stability, and transparency of models used in areas such as credit scoring, fraud detection, and investment management.

Key components of the framework include a strong governance structure, robust model development and validation processes, and continuous monitoring and reporting. Governance

[17] Comptroller's Handbook: Model Risk Management. 2021. Office of the Comptroller of the Currency, https://occ.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/index-model-risk-management.html.

involves clear policies, roles, and responsibilities for model risk management, while development and validation processes ensure that models are accurate, reliable, and fit for their intended purposes. Continuous monitoring and reporting help identify potential issues and facilitate timely adjustments.

15. The AI value or supply chain is complex, often involving open source and proprietary products and downstream applications that are quite different from what AI system developers may initially have contemplated. Moreover, training data for AI systems may be acquired from multiple sources, including from the customer using the technology. Problems in AI systems may arise downstream at the deployment or customization stage or upstream during model development and data training.

a) Where in the value chain should accountability efforts focus?
b) How can accountability efforts at different points in the value chain best be coordinated and communicated?
c) How should vendors work with customers to perform AI audits and/or assessments? What is the role of audits or assessments in the commercial and/or public procurement process? Are there specific practices that would facilitate credible audits (e.g., liability waivers)?
d) Since the effects and performance of an AI system will depend on the context in which it is deployed, how can accountability measures accommodate unknowns about ultimate downstream implementation?

This question reinforces our point in the *Introduction and Overarching Recommendations* section that AI accountability activities need to be addressed systematically and in context rather than by a collection of disconnected thoughts or activities.

Downstream value chain unknowns can be managed in part by applying anticipatory mechanisms such as the NIST AI Risk Management Framework, modeling and simulation in tabletop exercises,[18] red teaming,[19] and leveraging other anticipatory impact assessment tools and methodologies.

In managing value chain unknowns, it is also important to note that AI accountability cannot be "one and done." It must be executed along each step of the value chain and in a continuous feedback loop as the AI technology evolves with new capabilities leading to new use cases. This enables in tandem the evolution of regulations and accountability processes designed to identify safe and responsible boundaries for AI innovation and application.

Also see our answer to Question #4 above for dealing with risks.

[18] MITRE is setting up an AI Assurance and Discovery Lab to collaboratively explore and experiment with capabilities to anticipate, discover, and mitigate downstream risks early in the AI lifecycle.

[19] Microsoft and MITRE Create Tool to Help Security Teams Prepare for Attacks on Machine Learning Systems. 2023. MITRE, https://www.mitre.org/news-insights/news-release/microsoft-and-mitre-create-tool-help-security-teams-prepare-attacks.

16. The lifecycle of any given AI system or component also presents distinct junctures for assessment, audit, and other measures. For example, in the case of bias, it has been shown that "[b]ias is prevalent in the assumptions about which data should be used, what AI models should be developed, where the AI system should be placed—or if AI is required at all." How should AI accountability mechanisms consider the AI lifecycle?

The AI lifecycle is one of several important facets that must be considered when systematically assessing AI accountability.

AI accountability mechanisms focusing on technical versus socio-technical characteristics? Both are necessary, with socio-technical assessment being harder and more often overlooked until something bad has happened. Accountability mechanisms for governance should also be included in holistic AI system assessments.[20]

How should AI accountability mechanisms consider the AI lifecycle? As referenced in MITRE's definition shared earlier, AI assurance is a lifecycle process, and it is important to emphasize two points at each end of the lifecycle spectrum—up front at the time of preparation and design, and at the end during technology deployment. Accountability mechanisms at both phases are underdeveloped while each holds great potential for impact.

Up Front – Anticipatory Governance. Requiring technology designers and developers to explore and assess in anticipatory fashion how emerging AI solutions may impact citizens, commerce, and government and identify associated governance needs such as application-specific assurance requirements and oversight. This may also entail documenting the operating envelope of the AI system's expected capabilities.

Upon Deployment – Outcomes Monitoring and System Improvement. Requiring technology deployers to implement operational structures and mechanisms for monitoring high-stakes AI applications to identify what is, and is not, working in practice. Robust governance and test and evaluation, as accountability and risk management mechanisms, cannot eliminate all risk from rapidly emerging AI technologies. Understanding effective outcomes can lead to accountability

[20] This position is supported by the arguments laid out by Mökander, et al. in "Auditing large language models: a three-layered approach" (https://arxiv.org/abs/2302.08500), where they make six key claims about auditing AI systems (LLMs in particular):

1. AI auditing procedures focusing on compliance alone are unlikely to provide adequate assurance for LLMs.
2. External audits are required to ensure that LLMs are ethical, legal, and technically robust.
3. Auditing procedures designed to assess and mitigate the risks posed by LLMs must include elements of both governance and technology audits.
4. The methodological design of technology audits will require significant modifications to identify and assess LLM-related risks.
5. Model audits will play a key role in identifying and communicating LLMs' limitations, thereby informing system redesign, and mitigating downstream harm.
6. To be effective, LLM auditing procedures must include some elements of continuous ex-post auditing.

These claims justifiably argue for the necessity of three layers of AI audits: governance, model, and application. While these categories bleed into one another, their distinction is helpful. Model audits account for technical AI accountability mechanisms, application audits account for socio-technical characteristics within the AI system's context of use, and governance audits focus on adequacy of organizational governance structures (e.g., the mapping of roles and responsibilities) and documentation/audit trails of the AI development process. All of these accountability mechanisms are essential.

and the reinforcement/enhancement of AI technologies and the assurance and accountability approaches that lead to them.

How often should audits or assessments be conducted, and what are the factors that should inform this decision? The frequency and fidelity of audits and assessments should be commensurate with the level of risk of harm considering all relevant stakeholders—especially those negatively impacted. Standards for assessing the level of consequentiality of an AI system and associating that to a commensurate level of accountability mechanisms are critically lacking and are needed. Not all AI applications are alike. AI applications of higher consequence, such as clinical and military decision support systems, will tolerate less risk, requiring greater due diligence and more frequent assessment. The frequency and fidelity of audits should also be managed such that they are responsive to changes in performance and efficacy of deployed AI systems over time. Operating environments are dynamic and operating contexts shift, which can adversely affect AI system performance and efficacy and necessitate their continuous evaluation. Although the frequency of post-deployment evaluations should be specified up front in proportion to level of consequentiality of an AI system, outcomes monitoring may reveal additional information that may necessitate revisions to those frequencies.

17. How should AI accountability measures be scoped (whether voluntary or mandatory) depending on the risk of the technology and/or of the deployment context? If so, how should risk be calculated and by whom?

AI accountability measures should be scoped commensurately to an AI system's consequentiality (i.e., the level of risk of harm) within its specific context of use. This scoping should consider whether the audit or assessment is voluntary or mandatory, the level of required assessment fidelity and due diligence, and the frequency of assessment. These factors also help scope the cost of implementing the accountability. Consensus standards for how to calculate such levels of consequentiality are lacking and should be developed to be adaptable to the use case at hand factoring in the interactions between risk severity and likelihood, and at the same time taking into account that different stakeholders may perceive risk differently and may have different tolerances toward risk. Such a framework could then be adopted and adapted by various sector regulators.

23. How should AI accountability "products" (e.g., audit results) be communicated to different stakeholders? Should there be standardized reporting within a sector and/or across sectors? How should the translational work of communicating AI accountability results to affected people and communities be done and supported?

As previously discussed, there are many different groups of actors within any AI ecosystem. Common communications best practices state that to be most effective, we need to communicate products to each group in their language and where they are. To avoid duplication of effort and meet that best practice, a primary report can be produced that is then summarized specifically for select key audiences. For example, a summary for policy audiences would briefly describe what

was done, what was learned, and what the results means for future policy considerations and activities.[21]

There should also be common practice to share information about AI system risks, during both development and operation. Such transparency during operational considerations should also provide information about the status quo (i.e., not using AI) so that stakeholders can make properly informed decisions. (In some potential uses of AI, it may not be perfect but would still be much better than the current process. Properly understanding the risks of both the AI and the non-AI approach is required.)

Finally, a key equity principle is to engage with representatives of people and communities potentially impacted by the new capabilities, which should be incorporated into AI accountability assessments. These relationships will also be useful when disseminating accountability results. Translating AI accountability products into communications for affected people should consider language preferences of the audience, acknowledge limitations and biases in the methods and results, indicate how to access further data to minimize the burden of multiple inquiries for information from the population affected, and provide avenues for redress.[22]

24. What are the most significant barriers to effective AI accountability in the private sector, including barriers to independent AI audits, whether cooperative or adversarial? What are the best strategies and interventions to overcome these barriers?

Although AI solution providers in industry have incentives to compete, they also have a vested interest to collaborate. Critical AI failures have the potential of bringing the entire sector down by undermining trust in AI. However, AI solution providers are not actively exchanging operational information that is needed to inform the development of shared assurance practices.

Three barriers to the private sector providing the access required for accountability and audits are:

1. concerns about the release of intellectual property and competitive advantage,
2. disclosure of findings that may cause reputational damage to a corporation, and
3. lack of common language for expressing AI incidents and their root causes and mitigations.

Data sharing public private partnerships (PPPs) are a mechanism that has been successful in mitigating these concerns for data sharing in several sectors and should be a focus of effort for facilitating AI accountability. PPPs for data sharing involve aggregating private and public sector data for the collective benefit of participants, often using trusted third parties. Key elements of successful data sharing PPPs include:

[21] D. Blackburn. "Policy Wrappers" for S&T Findings. 2022. MITRE, https://www.mitre.org/news-insights/publication/policy-wrappers-st-findings.

[22] H. Leker, et al. Social Justice Platform Data Guide: Integrating Equity into Data Analysis. 2022. MITRE, https://sjp.mitre.org/resources/Data_Guide_Equity_Data_Analysis_2022.08.29_PRS.pdf.

- Addressing a specific problem or opportunity with a sense of urgency and justification for participation
- Establishing trust and independence, with clear rules on data accessibility, sharing, and use
- Ensuring privacy and security to maintain public trust and protect against unintended data usage
- Clarifying contractual and legal issues, such as intellectual property, conflict of interest policies, and funding

One example of such a data sharing PPP is Aviation Safety Information Analysis and Sharing (ASIAS), a collaboration between the Federal Aviation Administration and the aviation community. ASIAS resources include "both public and non-public aviation data. Public data sources include, but are not limited to, air traffic management data related to traffic, weather, and procedures. Non-public sources include de-identified data from air traffic controllers and aircraft operators, including digital flight data and safety reports submitted by flight crews and maintenance personnel."[23]

The structure of data sharing agreements and the operational governance of ASIAS have overcome the reluctance of private sector organizations to share proprietary data with direct competitors in order to realize the collective benefit from unlocking aggregate analyses on safety benchmarking and improvements for aviation without compromising the security of their confidential business models.[24]

30. What role should government policy have, if any, in the AI accountability ecosystem? For example:

a) Should AI accountability policies and/or regulation be sectoral or horizontal, or some combination of the two?
b) Should AI accountability regulation, if any, focus on inputs to audits or assessments (e.g., documentation, data management, testing and validation), on increasing access to AI systems for auditors and researchers, on mandating accountability measures, and/or on some other aspect of the accountability ecosystem?
c) If a federal law focused on AI systems is desirable, what provisions would be particularly important to include? Which agency or agencies should be responsible for enforcing such a law, and what resources would they need to be successful?
d) What accountability practices should government (at any level) itself mandate for the AI systems the government uses?

[23] Aviation Safety Information Analysis & Sharing. 2023. ASIAS, https://portal.asias.aero/overview.

[24] Report to Congress: Report on the Status of Aviation Safety Information Analysis and Sharing (ASIAS) Capability Acceleration. 2020. Federal Aviation Administration, https://www.faa.gov/about/plans_reports/congress/media/FAA_Report_on_Aviation_Safety_Information_Analysis_and_Sharing_ASIAS_03312020.pdf.

With respect to AI accountability policies and/or regulation being sectoral or horizontal, or some combination of the two, there are strong reasons to suggest "some combination of the two." AI is solving real problems and causing real harm today. The United States must adapt regulation to address AI harm and establish clear accountability guidance and enforcement, while striking a balance between minimizing risk and promoting innovation. Creating a single entity to oversee issues that span multiple contexts and agencies, and thus overlap or supersede existing regulatory structures, rarely succeeds. Leadership, collaboration, and accountability is usually the better approach. The U.S. should rely on existing sector-specific regulators, equipping them to address new AI-related regulatory needs. A sector-specific approach aligns with the current regulatory and oversight structure of the U.S. government.[25] However, information about AI applications and lessons learned should still be shared across sectors. Existing structures and processes to facilitate this cross-sector communication should be encouraged and expanded. The National Artificial Intelligence Initiative Office in the Office of Science and Technology Policy is already serving as a mechanism to share best practices for advancing AI adoption and mitigating harm, and this role is increasingly important.

With respect to mandates for the AI systems the government uses, AI risk calculation will be different between the public and private sectors. The public sector often holds a vast amount of sensitive data on citizens, which may be subject to strict privacy regulations and data handling requirements. The private sector usually has more flexibility in data collection and usage, but still needs to comply with data protection laws and maintain customer trust. Public sector organizations are generally required to maintain a higher level of transparency and accountability in their AI implementations, as they need to justify their actions to the public and regulators. This added scrutiny is also required because in some cases the public sector has authorities that the private sector lacks (e.g., criminal charges, imposing fines, denial of public services and benefits). The private sector, while still accountable to comply with laws and to justify actions to shareholders and customers, may have more latitude in terms of proprietary algorithms and decision-making processes. Public sector organizations may have a lower risk tolerance when adopting AI technologies, as mistakes can lead to public scrutiny, regulatory issues, or political consequences. Private sector companies might be more willing to take risks and innovate, given the potential for higher returns and competitive advantages. Much of the accountability required of federal AI adoption will be governed by evolving interpretations and revisions of administrative law (distinct from the current attention on regulation of the private sector). The Administrative Conference of the United States has published a study investigating these facets in greater detail.[26] The GAO has also published an accountability framework specifically aimed at government AI use that provides best practices.[27]

MITRE is close to finalizing an AI regulation paper, which contains specific recommendations on regulatory approaches for AI security. It should be published within a couple of weeks after this submission.

---

[25] M. Cuéllar, et al. Toward the Democratic Regulation of AI Systems: A Prolegomenon. 2020. University of Chicago, Public Law Working Paper No. 753, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3671011.

[26] Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies. 2020. Administrative Conference of the United States, https://www.acus.gov/sites/default/files/documents/Government%20by%20Algorithm.pdf.

[27] Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities. 2021. Government Accountability Office, https://www.gao.gov/assets/gao-21-519sp.pdf.

31. What specific activities should government fund to advance a strong AI accountability ecosystem?

Anticipatory Governance Mechanisms. Give technology designers, developers, and regulators the tools to explore and assess in anticipatory fashion how emerging AI solutions may impact citizens, commerce, and government and identify associated governance needs such as application-specific assurance requirements and oversight. Anticipatory mechanisms should be developed to support exploration and discovery within laboratory and/or sandbox environments and include tools for conducting anticipatory impact assessments through methodologies such as red teaming and modeling and simulation.

Next Generation T&E Mechanisms. Develop context-sensitive, formative T&E mechanisms to discover, chart, and mitigate emerging consequences of AI solutions, especially when AI intersects with human values in ways that are difficult to predict (e.g., when AI applications may be perceived as limiting and undermining human agency). Such mechanisms should more accurately characterize the performance capabilities and limitations of emerging AI systems in their use context.

Outcomes Monitoring. Lay out the framework to implement operational structures and accountability mechanisms needed for effectively monitoring high-stakes AI applications to identify what is and is not working in practice. Robust governance and T&E, as risk management mechanisms, cannot eliminate all risk from rapidly emerging AI technologies. Monitoring and reporting commensurate to the level of risk of harm is required and should responsibly accommodate all relevant stakeholders, especially those negatively impacted.

Red Teaming and Incident Sharing Mechanisms. Establish red teaming mechanisms to discover vulnerabilities and stress test new AI system capabilities in order to assess the changing threat landscape of our deployed AI systems, and to understand how adversaries may leverage evolving AI capabilities to launch new types of attacks. Sharing identified vulnerabilities and reporting incidents stemming from either red teaming or outcomes monitoring is vital for real-world intelligence analysis and mitigation. Mechanisms for rapidly sharing and reporting are needed across both government and industry at appropriate protection and classification levels using a common language for both intentional (malicious or deliberate misuse) and unintentional (accidental or unexpected failure, without a bad actor involved) incidents. Incident sharing mechanisms can be used to cover all the areas of AI assurance and accountability starting with security and then building to include equitability, interpretability, robustness, resilience, and privacy enhancing needs.

Enabling the Evaluation of Social Impact. Ensure interdisciplinary research efforts involve social scientists and community engagement with those who are experiencing the problem being addressed, starting at the design of the project. An analysis of AI research publications has found that the distance between social science fields and AI research has grown over the past decades, likely driven by the more technical focus of industry-funded AI development. Ensuring equitable impact also requires research on the social and behavioral interaction between the AI systems and populations served, as well as environments—in vivo, qualitative human feedback, simulated, hybrid, and computational social models—that explore the downstream distributional

effects of AI systems on individuals, communities, and institutions. This impact may extend beyond those directly at the receiving end of the AI system.[28]

34. Is it important that there be uniformity of AI accountability requirements and/or practices across the United States? Across global jurisdictions? If so, is it important only within a sector or across sectors? What is the best way to achieve it? Alternatively, is harmonization or interoperability sufficient and what is the best way to achieve that?

An AI accountability ecosystem/standard of practice across the U.S. should clearly address:

- ***When** is AI accountability warranted—for what identified use cases?*
- ***What** is the object of accountability? What technology and/or process and who is being held accountable?*
- ***Why** is AI accountability necessary in the use case context?*
- ***How** is AI accountability to be implemented in the use case context?*

This can be achieved by applying a risk-managed approach as outlined in the NIST AI Risk Management Framework (AI RMF) where "systematic documentation practices established in GOVERN [function] and utilized in MAP and MEASURE [functions] bolster AI risk management efforts and increase transparency and accountability" when "risks are prioritized and acted upon based on a projected impact" in MANAGE function.[29] Establishing the when, what, and how of AI accountability should take place early and iteratively in the Plan and Design stage of the AI lifecycle by technology developers who can share such documentation with compliance experts and auditors later in the Operate and Monitor stage of the AI lifecycle. This can be realized by impact assessors conducting AI impact assessments at both stages "assessing and evaluating requirements for AI system accountability, combating harmful bias, examining impacts of AI systems, product safety, liability, and security, among others."[30] Regulators following this approach can establish objective outcome-focused accountability requirements and provide them upstream in the AI lifecycle to system developers, who can then implement them so the AI technology is auditable and ultimately shown to be compliant.

[28] MITRE Response to OSTP's RFI Supporting the National Artificial Intelligence Research and Development Strategic Plan. 2022. MITRE, https://www.mitre.org/sites/default/files/2022-03/pr-21-01760-16-mitre-response-ostp-rfi-national-artificial-intelligence-research-and-development-strategic-plan.pdf.

[29] Artificial Intelligence Risk Management Framework. 2023. NIST, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf, p. 31, 20.

[30] Artificial Intelligence Risk Management Framework. 2023. NIST, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf, figure 3, p. 11.