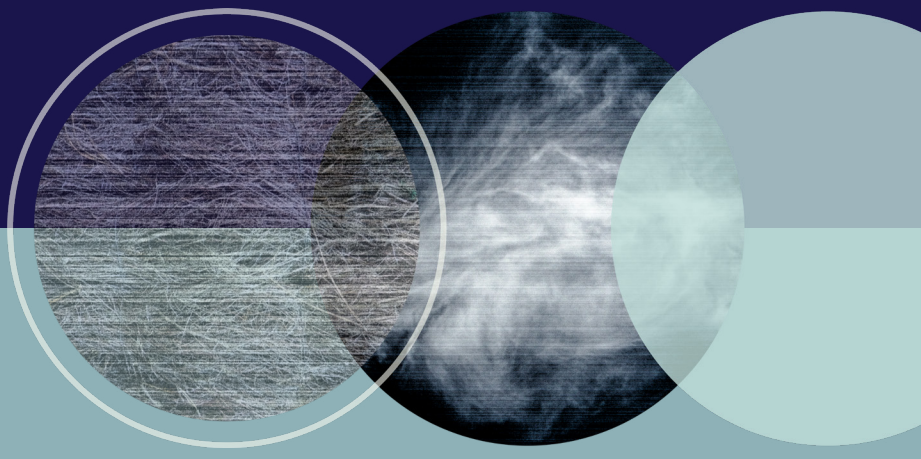
Policy and Governance
Working Group

June
2023

Recommendations to the US Department of Commerce (NTIA) on Policy for AI Accountability



SCIENCE, TECHNOLOGY,
AND SOCIAL VALUES LAB
INSTITUTE FOR ADVANCED STUDY

12 June 2023

National Telecommunications and Information Administration
1401 Constitution Ave., NW
Washington, DC 20230
Docket No. 230407-0093
submitted via https://www.regulations.gov

Comment of the AI Policy and Governance Working Group
on the NTIA AI Accountability Policy Request for Comment
Docket NTIA-230407-0093

Trustworthiness is not inherent to artificial intelligence (AI) systems and tools. Designers and deployers of AI must demonstrate that their products are safe and effective—and therefore merit the public's trust—through iterative accountability mechanisms that span the full development and deployment lifecycle and address risks related to both highly specialized and more general purpose AI systems. When AI designers and deployers fail to meet these expectations, they must be held accountable.

In response to the National Telecommunications and Information Administration's (NTIA) request for comment on AI accountability policy, the AI Policy and Governance Working Group here provides 1) overarching considerations, both to offer context for our specific recommendations and to help inform federal AI policy strategy, and 2) recommendations for sociotechnical AI accountability mechanisms based on evaluation, access, and disclosure that can begin to build justified public trust in AI as an essential predicate to adequately and effectively "aligning" these technological systems and tools with democratic and human values.

Responsibility for accountability in the design and deployment of AI systems and tools begins with the technology developers. Industry, academia, civil society, and the public sector each has a key role to play in the development of an effective AI accountability system. This response to the NTIA call for comment primarily addresses prescriptions the government can uniquely facilitate or catalyze.

AI Policy and Governance Working Group

The AI Policy and Governance Working Group represents a mix of sectors, disciplines, perspectives, and approaches. Despite these differences, we agree that it is necessary and possible to address the multitude of concerns raised by the expanding use of AI systems and tools and their increasing power. We also agree that both present-day harms and risks that have been unattended to and uncertain hazards and risks on the horizon warrant the federal government's urgent attention and the public's expectation of safety.

Overarching Considerations and Implementation Recommendations

Policymakers need to take action now: The use of AI undoubtedly poses an array of complex challenges, but policymakers should not be dissuaded from taking action to

address emerging concerns by supposed tensions between innovation and safety, the evolving nature of the field, or the relatively nascent mechanisms for accountability. While the approaches we highlight below may not guarantee accountability, they provide abundant means to facilitate it, and can thus help advance conditions in which AI-powered systems and tools can be most reliable and safe and, therefore, most beneficial. Moving quickly to address risks concerning AI systems and tools will not only provide accountability, it will promote the trust of the American public.

Policymakers, researchers, industry and the public require more visibility into the risks presented by AI systems and tools: Government can play an important role in making risks more visible, and the mitigation of risk more actionable, by developing policy to enable a robust and interconnected evaluation, auditing, and disclosure ecosystem that facilitates timely accountability and remediation of potential harms.

People are policy: Developing effective AI accountability policy will require the expansion of multidisciplinary and interdisciplinary expert personnel in government, including in computer science, data science, and social science. In addition, an investment in existing and new kinds of talent will be required in government at the intersection of AI and subject-matter, domain, and/or sector expertise (e.g., finance, healthcare, national security, civil rights, etc.) and for AI evaluation and auditing (e.g., algorithm auditors).

Accountability policy mechanisms must be co-created: Accountability approaches present a range of considerations, including economic viability, possible trade-offs between general applicability versus clarity of requirements and processes, and the potential consolidation of power by industry incumbents at the expense of stifling new entrants. To ensure that compute power, personnel and other necessary resources are dedicated to tackling complex accountability challenges, a cooperative funding mechanism could be jointly established by government and industry. This would enable a much-needed multi-layered approach to risk evaluation, assessment, and management — no single accountability intervention or organization will be fully effective on its own.

Summary of Recommendations

Evaluation:

- AI accountability policy should encourage or require all developers and entities deploying AI-based applications to perform relevant evaluations. AI models and systems should be evaluated in terms of their ability to perform as claimed, their potential and actual impact, and their general ability and propensity to yield harmful outputs or behaviors.
- Government should support best practices in evaluation methods and bolster innovation in research that identifies new evaluation methods through procurement practices and research and development investments.
- Evaluation should occur both pre- and post-deployment, and relevant results (such as harms or failures) should be made publicly available, tracked, and compiled.

Access:

- To facilitate the feasibility of model-and-system-access frameworks, we recommend that the government address legal and technical barriers to access, while building incentives and infrastructure to enable access for qualified persons.

- We recommend the government mandate access to the technical infrastructure to enable varying levels of visibility into different components of (potentially) *consequential* AI systems and incentivize access to the technical infrastructure of *other* AI systems.
- We urge the federal government to carefully scrutinize industry "no audit" provisions, taking into account disadvantages of prohibiting such provisions necessary for scientific research and public accountability.
- Qualified researchers and auditors who meet certain conditions should be given model-and-system framework access.

Disclosure:

- The federal government should consider the establishment of narrowly-scoped "safe harbor" provisions for industry and researchers, designed to reasonably assure that entities participating in good faith auditing exercises are not subjected to undue liability risk or retaliation.
- We also recommend the federal government urge the adoption of common standards for documentation to facilitate a "responsible disclosure" ecosystem leading to greater accountability through common transparency norms and practices.

Evaluation

AI accountability policy should encourage or require (as applicable) all developers and entities deploying AI-based applications to evaluate their systems. Evaluations provide meaningful insight into the downstream impacts of AI systems as well as potentially serve as tools for AI accountability policymaking, and audit evaluations have historically played a role in meaningful accountability regimes in other industries. Evaluations are utilized throughout the process of AI development and deployment, targeting the underlying algorithms or model, the accompanying models which influence the AI system's output or behavior, as well as the broader impacts on end users and society overall.

Given the broad range of potential application areas, current approaches to evaluation utilize both a broad set of methods and a range of risk areas. Thorough, rigorous, and systematic evaluations are especially important for AI systems used in the most consequential areas of society or systems. This may include assessing influential algorithms or frontier models for a range of risks such as bias, misalignment, dangerous capabilities, or potential misuse. We recommend that the research and policy communities together develop and regularly reevaluate and update a risk taxonomy based on the capabilities of AI tools and systems, and the uses to which they are put. Because risk, including ethical and societal risk, is dynamic, evaluation responses will need to be similarly iterative.

What to Evaluate?

Effective evaluations should consider risk scenarios for AI systems and tools and crosscut these with assessment of systems' capabilities and robustness:

- **Capability assessment** - Assessing the full range of behaviors that a system could plausibly express during deployment, and the full range of knowledge or information that it might use. This includes efforts aimed at measuring a model's effectiveness at some task or tasks in the average case. This also includes efforts aimed at bounding

what a model could be able to do in exceptional cases, potentially through the use of techniques like fine-tuning in the case of large neural networks, that attempt to modify the model in order to make it easier to elicit pre-existing capabilities.

- **Robustness assessment** - Assessing the degree to which the system will *consistently* behave in acceptable ways, particularly under "worst-case" scenarios. This includes efforts aimed at identifying unique or edge case tasks where the model fails to perform consistently. Or, assessing performance under instances of phenomenon such as distribution shifts, where the AI system performs inconsistently – often for specific groups – due to deficiencies in the model's underlying training data. Assessments can also serve to determine whether established technical mitigations such as pre-training dataset curation or reinforcement learning with human feedback have successfully prevented the prevalence of a potentially harmful or dangerous capability.
- **Impact assessment** - Assessing the anticipated and downstream effects of specific deployments and use cases. ***Prospective* impact assessments** are conducted *ex ante* and use a range of methods to anticipate potential dangerous and harmful outcomes or behaviors, often seeking to provide clarity on the potential impact under specific scenarios or on impacted groups. ***Retrospective* impact assessments** are conducted to AI tools and systems *ex post* to discern harmful or dangerous behaviors and outcomes.

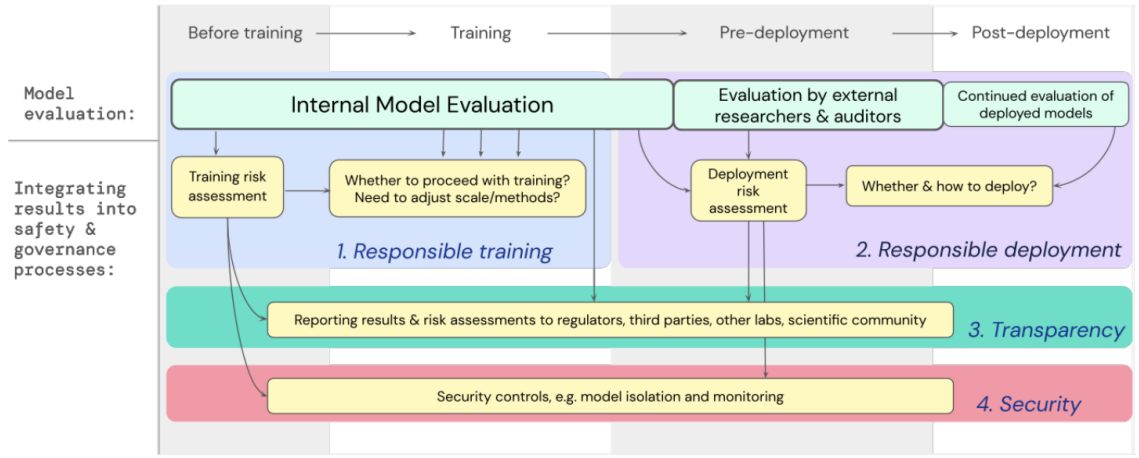
Evaluation Methods

There are a variety of (potentially overlapping) approaches to evaluation, each with its own strengths and limitations. Some known approaches include:

- **Benchmarking:** An assessment of AI model or system outputs based on standardized or curated datasets, automated evaluation models, or simulated environments.
- **Red Teaming**: A systematic probing of an AI model or system by either expert or non-expert human evaluators to reveal undesired outputs or behaviors.
- **User Evaluations and Testing** : An assessment of user-centric effects of an application or system as well as facility with a system's functionality and restrictions, usually via user testing or surveys.
- **Use-Case Studies:** A pilot, trial or other limited release of an AI system or application to assess overall performance under predefined, real-world scenarios.

Evaluations should consider model characteristics as well as address likely impacts across the lifecycle, ranging from when an AI model, system, or application is released by the technology developer, through to the context in which the tools and systems are used. Such a multi-step evaluation process is consistent with the best practices enumerated by the Government Accountability Office. Evaluations should persist across the lifecycle of AI model and system development and deployment, as illustrated below in Table 2.

TABLE 1



Shevlane, Farquhar, Garfinkel, Phuong, et al., "Model evaluation for extreme risks," 2023; arXiv:2305.15324.

While evaluations have increasingly become a common tool for responsible development and deployment, current evaluation practices are not yet able to provide credible assurance in many settings, especially in the context of highly capable systems, which are increasingly able to detect what they are being evaluated for and manipulate their behavior accordingly (see here for an example). More research is needed on developing more robust evaluation methods for real-world risks and harms. Government should support best practices and bolster innovation in this area through procurement and through R&D investments.

Model and System Access

AI accountability policy, including proactive identification of risks and harms stemming from the use of AI, requires pathways for access to relevant models and systems. Concerns such as privacy, intellectual property, and proliferation have limited access to AI models and systems. While these are important considerations, government cannot allow them to continue to be hurdles to the meaningful access to models and systems needed to ensure accountability.

To facilitate the feasibility of such frameworks, we recommend that the government address legal and technical barriers to access and build incentives and infrastructure to enable access for qualified persons[1].

Technical infrastructure

We recommend that the government mandate access to the technical infrastructure to enable varying levels of visibility into relevant components of AI systems utilized in high-stakes or consequential areas and incentivize access to the technical infrastructure of other AI systems through favorable tax treatment, procurement, or other means. Such access may be facilitated by application programming interfaces (or APIs), or by equivalent mechanisms, designed to permit controlled, secure, and streamlined access to relevant

[1] To facilitate the evaluations processes, auditors will require some reasonable access to the institutions involved. This includes but is not limited to conducting interviews with key personnel and requesting relevant documentation.

components of said system. Balancing access and security (e.g., theft of data, IP, etc.) is crucial, so specialized credentials enabling deeper model access may be necessary for auditors evaluating more sensitive risks, or more powerful models. We recognize that the development of model-and-system-access frameworks will come with cost implications. Given the import and centrality of model and system access to any possibility of AI accountability, we encourage exploration of viable models (e.g., involving third-party facilitation and/or through public-private co-funding).

We also recommend further support, research and development of structured transparency, such as the privacy-enhancing technologies challenges and demonstrations being sponsored by the White House Office of Science and Technology Policy, that enable auditing of algorithmic models without requiring direct access to them or to private data.

Legal infrastructure

AI models, systems and platforms frequently include "no audit" clauses in their terms of use, a significant barrier to the access required for AI accountability. We urge the federal government to carefully scrutinize such provisions, taking into account the advantages and disadvantages of prohibiting such provisions as well as whether they may be scoped in such a way to reasonably facilitate fair competition, while enabling requisite scientific research and public accountability. Legal and regulatory frameworks can facilitate reasonable balancing of such interests.

To facilitate model and system access, the federal government should consider the establishment of narrowly-scoped "safe harbor" provisions for researchers and industry, designed to reasonably assure that entities participating in good faith auditing exercises are not subjected to undue liability risk or retaliation. For industry, this assurance would depend on their willingness to undertake corrective actions based on the audit results, similar to provisions outlined in the Equal Credit Opportunity Act (ECOA). Comparable models already exist in sectors like retail and insurance, and databases such as the U.S. Securities and Exchange Commission's EDGAR platform or the U.S. Food and Drug Administration's FAERS reporting platform serve as examples of infrastructures that facilitate mutual sharing.

Qualified researchers and auditors

We recommend that qualified researchers and auditors[2] who meet certain conditions should be given access to the above components, as envisaged by similar legislative proposals. In high-risk settings, government may also explore innovative regulatory mechanisms, such as requiring developers to undergo auditing by specialized private regulators and researchers. The existing audit market – both individual auditors and auditing organizations – remains small and in its infancy; there is a pressing need for more work to stimulate regulatory markets. A practical mechanism to consider broadly across the whole of the federal government would be the uptake and application of a Department of Defense procurement

[2] The term "auditors" is typically used more broadly in AI-related fields than in other industries. In this context, AI auditors may constitute both professional groups seeking mandated audits and academic and other researcher groups desiring voluntary scrutiny of systems. In both cases, in order to determine the suitability of various involved actors to gain access or other auditor privileges, governments should explore the development of auditor conduct standards, certifications for auditors undertaking this work, and dispute adjudication mechanisms.

Intellectual property issues and "trade secrets" will need to be addressed in any qualified researchers and auditor scheme, for private-sector researchers at competing companies may have benign, safety related motivations for probing a competitor's system, as well as motivations related to commercial competition.

vehicle for an independent evaluator to be procured simultaneously with a contract for an AI tool or system, thus building in a layer of accountability with the necessary infrastructure and funding. Another key consideration should be the establishment of audit oversight boards, similar to Public Company Accounting Oversight Board (PCAOB) in finance, to assess auditor qualifications, oversee training materials, and support in the mediation of conflict of interest.

Responsible Disclosure

Effective AI accountability policy will also require visibility into systems and tools through reporting and documentation, appropriate protections for researchers and entities, and forms of certification.

Reporting and documentation

Different audiences have different transparency needs, but standard forms and structures of documentation should be encouraged to facilitate common comprehension and ready comparison of how AI systems and tools are built and operate.[3] For example, it would be strongly advisable that developers share common standards for model and system cards. Policymakers could assist in this effort by promoting the below "model card" standard defined by M. Mitchell, et al. as a starting point. For systems with very broad intended uses, such as large language models, appropriate reporting and documentation could be supplemented by a description of systems' intended behaviors. "Reward reports," that is, "living documents that track updates to design choices and assumptions behind what a particular automated system is optimizing for" can be used for tracking dynamic phenomena arising from system deployment, rather than simply reporting on static properties of models or data.

Post-evaluation actions

Evaluation is necessary not only in advance of deployment, but also after models and systems are deployed in order to detect real-world harms and second-order effects caused by these systems, similar to ongoing environmental monitoring.

Evaluation methods, results, and key limitations should be appropriately disclosed. Government should encourage the ongoing secure disclosure of the results of evaluations and other key details, including results on standard safety benchmarks, if they exist and apply, environmental impacts, and kWh of energy usage for model and system training and use. These results should be made available to a range of stakeholders, including the public.

To ensure uptake of this important accountability mechanism, policymakers might require professional auditors to report results to regulatory authorities (similar to environmental audits), require responses to recommendations made in evaluation reports within a certain time period, or build a central registry of audit reports that is publicly accessible, upon request, to enable additional scrutiny, oversight, and accountability. Incident reporting databases would allow government, civil society, and industry to track certain kinds of harms and risks. Examples of such resources include the AI Incident Database and the

[3] The National Institute of Standards and Technology (NIST) highlights transparency as a predicate for trust in AI tools and systems and a "crucial component in enabling accountability mechanisms" and notes that "meaningful transparency provides access to appropriate levels of information based on the stage of the AI lifecycle and tailored to the role or knowledge of AI actors or individuals interacting with or using the AI system."

Organisation for Economic Co-operation and Development (OECD) incidents reporting model, both of which are accessible to the public.

While it is challenging to boil down complex systems into an approachable format, it is crucial for accountability to the public that documentation formats be accessible, consumable, and machine readable. At minimum they should include the "reporting" components of each of the principles in the technical companion of the White House Blueprint for AI Bill of Rights and reflect best practices for the documentation of the machine learning lifecycle. To ensure accuracy, documentation should be updated regularly by developers and audited regularly by third parties, and must be reconsidered as models shift and change over time.

TABLE 2

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Mitchell, Wu, Zalvidar, Barnes, et al., "Model Cards for Model Reporting," 2018; arXiv:1810.03993

Facilitating a 'responsible disclosure' ecosystem

Similar to measures normative in information security, we propose creating regulatory provisions for researchers to audit models and report findings to companies, as well as incentives for companies to respond to audit findings within a reasonable time period.

Similar to measures in cybersecurity, we also propose antitrust provisions and liability carve-outs for the responsible sharing of safety and privacy concerns. Common testing, auditing, and safety standards, and best practices in harm mitigation techniques would benefit greatly from cross-company collaborations.

Open-source practices, however, require a broader conversation than scoped in this submission. While the legacy of open-source has been crucial to the development of science and technology, national security, cybersecurity and other high-priority needs may necessitate restricting access to a more limited set of trusted, qualified actors to prevent misuse, cybersecurity threats, and potentially catastrophic harms.

Certification

We recommend that responsible disclosure become a prerequisite in government regulations for certifying trustworthy AI systems, aligning with practices exemplified by Singapore's AI Verify. Certification could be conducted by independent third parties qualified to do so, stimulating a market for third-party auditors and safety programs. Procurement standards, applicable for contracts with the Department of Defense, Department of Energy, and other agencies, could then explore certification as a precondition. Such a requirement, when enforced via government contracts, sets a benchmark and creates an incentive for quality control and certification of trustworthy AI systems. Fees associated with these processes must be accessible for smaller developers and startups as well, encouraging normative practice across entities of all sizes.

Conclusion

We are grateful for the opportunity to contribute these comments and applaud the NTIA for its leadership in encouraging accountability in the development, deployment, and use of AI systems. As the federal government looks to design its National AI Strategy, we hope this articulation of how to conduct evaluation, expand access, and promote disclosure within the AI ecosystem will lead to best practice among companies and developers alike, increasing public trust and the responsible adoption of AI for widespread benefit.

Thank you for considering our recommendations. Please contact aipolicy@ias.edu with any comments or questions.

Sincerely,

AI Policy and Governance Working Group*
Aaron Maniam
Alondra Nelson
Ben Garfinkel

Brian Christian
Daniel E. Ho
Dorothy Chou
Helen Toner
Inioluwa Deborah Raji
Irene Solaiman
James W. Phillips
Karine Perset
Marc Aidinoff
Matthew Botvinick
Matthew J. Salganik
Rumman Chowdhury
Samuel R. Bowman
Sebastien Krier
Solon Barocas
Sorelle Friedler
Stephanie Ifayemi
William S. Isaac

*Members of the working group are participating in their personal capacities and these recommendations do not reflect the perspective of any of the organizations with which they are affiliated.