Artificial Intelligence
Accountability Policy Report

With thanks to Ellen P. Goodman, principal author, and the NTIA staff for their efforts in drafting this report.

MARCH 2024
## Contents

- **Executive Summary** ...................................................... 2
- **1. Introduction** .......................................................... 8
  - 2. Requisites for AI Accountability: Areas of Significant Commenter Agreement .................. 16
  - 2.1. Recognize potential harms and risks .......................................................... 16
  - 2.2. Calibrate accountability inputs to risk levels ................................................. 18
  - 2.3. Ensure accountability across the AI lifecycle and value chain ....................... 18
  - 2.4. Develop sector-specific accountability with cross-sectoral horizontal capacity .......... 19
  - 2.5. Facilitate internal and independent evaluations ............................................. 20
  - 2.6. Standardize evaluations as appropriate ....................................................... 21
  - 2.7. Facilitate appropriate access to AI systems for evaluation ............................. 21
  - 2.8. Standardize and encourage information production ........................................... 22
  - 2.9. Fund and facilitate growth of the accountability ecosystem .............................. 23
  - 2.10. Increase federal government role ...................................................... 23
  - 3. Developing Accountability Inputs: A Deeper Dive ............................................. 26
  - 3.1. Information flow ......................................................................................... 26
  - 3.1.1. AI system disclosures ............................................................................... 28
  - 3.1.2. AI output disclosures: use, provenance, adverse incidents ........................... 31
  - 3.1.3. AI system access for researchers and other third parties ............................. 36
  - 3.1.4. AI system documentation ........................................................................... 37
  - 3.2. AI System evaluations .................................................................................. 39
  - 3.2.1. Purpose of evaluations .............................................................................. 40
  - 3.2.2. Role of standards ...................................................................................... 42
  - 3.2.3. Proof of claims and trustworthiness ......................................................... 45
  - 3.2.4. Independent evaluations .......................................................................... 46
  - 3.2.5. Required evaluations .............................................................................. 48
  - 3.3 Ecosystem requirements ................................................................................. 49
  - 3.3.1. Programmatic support for auditors and red-teamers .................................... 49
  - 3.3.2. Datasets and compute .............................................................................. 50
  - 3.3.3. Auditor certification .................................................................................. 51
  - 4. Using Accountability Inputs ............................................................................. 54
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Liability rules and standards</td>
<td>54</td>
</tr>
<tr>
<td>4.2 Regulatory enforcement</td>
<td>58</td>
</tr>
<tr>
<td>4.3 Market development</td>
<td>59</td>
</tr>
<tr>
<td>5. Learning From Other Models</td>
<td>62</td>
</tr>
<tr>
<td>5.1 Financial assurance</td>
<td>62</td>
</tr>
<tr>
<td>5.2 Human rights and Environmental, Social, and Governance (ESG) assessments</td>
<td>65</td>
</tr>
<tr>
<td>5.3 Food and drug regulation</td>
<td>66</td>
</tr>
<tr>
<td>5.4 Cybersecurity and privacy accountability mechanisms</td>
<td>67</td>
</tr>
<tr>
<td>6. Recommendations</td>
<td>70</td>
</tr>
<tr>
<td>6.1 Guidance</td>
<td>70</td>
</tr>
<tr>
<td>6.1.1 Audits and auditors: Federal government agencies should work with stakeholders as appropriate to create guidelines for AI audits and auditors, using existing and/or new authorities.</td>
<td>70</td>
</tr>
<tr>
<td>6.1.2 Disclosure and access: Federal government agencies should work with stakeholders to improve standard information disclosures, using existing and/or new authorities.</td>
<td>71</td>
</tr>
<tr>
<td>6.1.2 Liability rules and standards: Federal government agencies should work with stakeholders to make recommendations about applying existing liability rules and standards to AI systems and, as needed, supplementing them</td>
<td>71</td>
</tr>
<tr>
<td>6.2 Support</td>
<td>72</td>
</tr>
<tr>
<td>6.2.1 People and tools: Federal government agencies should support and invest in technical infrastructure, AI system access tools, personnel, and international standards work to invigorate the accountability ecosystem</td>
<td>72</td>
</tr>
<tr>
<td>6.2.2 Research: Federal government agencies should conduct and support more research and development related to AI testing and evaluation, tools facilitating access to AI systems for research and evaluation, and provenance technologies, through existing and new capacity.</td>
<td>72</td>
</tr>
<tr>
<td>6.3 Regulatory Requirements</td>
<td>73</td>
</tr>
<tr>
<td>6.3.1 Audits and other independent evaluations: Federal agencies should use existing and/or new authorities to require as needed independent evaluations and regulatory inspections of high-risk AI model classes and systems.</td>
<td>73</td>
</tr>
<tr>
<td>6.3.2 Cross-sectoral governmental capacity: The federal government should strengthen its capacity to address cross-sectoral risks and practices related to AI</td>
<td>73</td>
</tr>
<tr>
<td>6.3.3 Contracting: The federal government should require that government suppliers, contractors, and grantees adopt sound AI governance and assurance practices for AI used in connection with the contract or grant, including using AI standards and risk management practices recognized by federal agencies, as applicable.</td>
<td>74</td>
</tr>
</tbody>
</table>

Appendix A: Glossary of Terms                                           | 76   |
Participants in the AI ecosystem — including policymakers, industry, civil society, workers, researchers, and impacted community members — should be empowered to expose problems and potential risks, and to hold responsible entities to account.

Artificial intelligence (AI) systems are rapidly becoming part of the fabric of everyday American life. From customer service to image generation to manufacturing, AI systems are everywhere. Alongside their transformative potential for good, AI systems also pose risks of harm. These risks include inaccurate or false outputs; unlawful discriminatory algorithmic decision-making; destruction of jobs and the dignity of work; and compromised privacy, safety, and security. Given their influence and ubiquity, these systems must be subject to security and operational mechanisms that mitigate risk and warrant stakeholder trust that they will not cause harm.

Commenters emphasized how AI accountability policies and mechanisms can play a key part in getting the best out of this technology. Participants in the AI ecosystem — including policymakers, industry, civil society, workers, researchers, and impacted community members — should be empowered to expose problems and potential risks, and to hold responsible entities to account.

AI system developers and deployers should have mechanisms in place to prioritize the safety and well-being of people and the environment and show that their AI systems work as intended and benignly. Implementation of accountability policies can contribute to the development of a robust, innovative, and informed AI marketplace, where purchasers of AI systems know what they are buying, users know what they are consuming, and subjects of AI systems — workers, communities, and the public — know how systems are being implemented. Transparency in the marketplace allows companies to compete on measures of safety and trustworthiness, and helps to ensure that AI is not deployed in harmful ways. Such competition, facilitated by information, encourages not just compliance with a minimum baseline but also continual improvement over time.

To promote innovation and adoption of trustworthy AI, we need to incentivize and support pre- and post-releaese evaluation of AI systems, and require more information about them as appropriate. Robust evaluation of AI capabilities, risks, and fitness for purpose is still an emerging field. To achieve real accountability and harness all of AI’s benefits, the United States — and the world — needs new and more widely available accountability tools and information, an ecosystem of independent AI system evaluation, and consequences for those who fail to deliver on commitments or manage risks properly.

Access to information by appropriate means and parties is important throughout the AI lifecycle, from early development of a model to deployment and successive uses, as recognized in federal government efforts already underway pursuant to President Biden’s Executive Order Number 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence of October 30, 2023 (“AI EO”). This information flow should include documentation about AI system models, architecture, data, performance, limitations, appropriate use, and testing. AI system information should be disclosed in a form fit for the relevant audience, including in plain language. There should be appropriate third-party access to AI system components and processes to promote sufficient actionable understanding of machine learning models.

Independent evaluation, including red-teaming, audits, and performance evaluations of high-risk AI systems can help verify the accuracy of material claims made about these systems and their performance against criteria for trustworthy AI. Creating evaluation standards is a critical piece of auditing, as is transparency about methodology and criteria for auditors. Much more work is needed to develop such standards and practices; near-term work, including under the AI EO, will contribute to developing these standards and methodologies.

Consequences for responsible parties, building on information sharing and independent evaluations, will require the application and/or development of levers — such as regulation, market pressures, and/or legal liability — to hold AI entities accountable for imposing unacceptable risks or making unfounded claims.

This Report conceives of accountability as a chain of inputs linked to consequences. It focuses on how information flow (documentation, disclosures, and access) supports independent evaluations (including red-teaming and audits), which in turn feed into consequences (including liability and regulation) to create accountability. It concludes with recommendations for federal government action, some of which elaborate on themes in the AI EO, to encourage and possibly require accountability inputs.

In April 2023, the National Telecommunications and Information Administration (NTIA) released a Request for Comment ("RFC") on a range of questions surrounding AI accountability policy. The RFC elicited more than 1,400 distinct comments from a broad range of stakeholders. In addition, we have met with many interested parties and participated in and reviewed publicly available discussions focused on the issues raised by the RFC.

Based on this input, we have derived eight major policy recommendations, grouped into three categories: Guidance, Support, and Regulatory Requirements. Some of these recommendations incorporate and build on the work of the National Institute of Standards and Technology (NIST) on AI risk management. We also propose building federal government regulatory and oversight capacity to conduct critical evaluations of AI systems and to help grow the AI accountability ecosystem.

While some recommendations are closely linked to others, policymakers should not hesitate to consider them independently. Each would contribute to the AI accountability ecosystem and mitigate the risks posed by accelerating AI system deployment. We believe that providing targeted guidance, support, and regulations will foster an ecosystem in which AI developers and deployers can properly be held accountable, incentivizing the appropriate management of risk and the creation of more trustworthy AI systems.
GUIDANCE

1. Audits and auditors: Federal government agencies should work with stakeholders as appropriate to create guidelines for AI audits and auditors, using existing and/or new authorities. This includes NIST’s tasks under the AI EO concerning AI testing and evaluation and other efforts in the federal government to refine guidance on such matters as the design of audits, the subject matter to be audited, evaluation standards for audits, and certification standards for auditors.

2. Disclosure and access: Federal government agencies should work with stakeholders to improve standard information disclosures, using existing and/or new authorities. Greater transparency about, for example, AI system models, architecture, training data, input and output data, performance, limitations, appropriate use, and testing should be provided to relevant audiences, including in some cases to the public via model or system cards, data sheets, and/or AI “nutrition labels.” Standardization of accessible formats and the use of plain language can enhance the comparability and legibility of disclosures. Legislation is not necessary for this activity to advance, but it could accelerate it.

3. Liability rules and standards: Federal government agencies should work with stakeholders to make recommendations about applying existing liability rules and standards to AI systems and, as needed, supplementing them. This would help in determining who is responsible and held accountable for AI system harms throughout the value chain.

SUPPORT

4. People and tools: Federal government agencies should support and invest in technical infrastructure, AI system access tools, personnel, and international standards work to invigorate the accountability ecosystem. This means building the resources necessary, through existing and new capacity, to meet the national need for independent evaluations of AI systems, including:
   • Datasets to test for equity, efficacy, and other attributes and objectives;
   • Computing and cloud infrastructure required to conduct rigorous evaluations;
   • Legislative establishment and funding of a National AI Research Resource;
   • Appropriate access to AI systems and their components for researchers, evaluators, and regulators, subject to intellectual property, data privacy, and security- and safety-informed protections;
   • Independent evaluation and red-teaming support, such as through prizes, bounties, and research support;
   • Workforce development;
   • Federal personnel with the appropriate socio-technical expertise to design, conduct, and review evaluations; and
   • International standards development (including broad stakeholder participation).

5. Research: Federal government agencies should conduct and support more research and development related to AI testing and evaluation, tools facilitating access to AI systems for research and evaluation, and provenance technologies, through existing and new capacity. This investment would move towards creating reliable and widely applicable tools to assess when AI systems are being used, on what materials they were trained, and the capabilities and limitations they exhibit. The establishment of the U.S. AI Safety Institute at NIST in February 2024 is an important step in this direction.

REGULATORY REQUIREMENTS

6. Audits and other independent evaluations: Federal agencies should use existing and/or new authorities to require as needed independent evaluations and regulatory inspections of high-risk AI model classes and systems. AI systems deemed to present a high risk of harmful rights or safety – according to holistic assessments tailored to deployment and use contexts – should in some circumstances be subject to mandatory independent evaluation and/or certification. For some models and systems, that process should take place both before release or deployment, as is already the case in some sectors, and on an ongoing basis. To perform these assessments, agencies may need to require other accountability inputs, including documentation and disclosure relating to systems and models. Some government agencies already have authorities to establish risk categories and require independent evaluations and/or other accountability measures, while others may need new authorities.

7. Cross-sectoral governmental capacity: The federal government should strengthen its capacity to address cross-sectoral risks and practices related to AI. Whether located in existing agencies or new bodies, there should be horizontal capacity in government to develop common baseline requirements and best practices, and otherwise support the work of agencies. These cross-sectoral tasks could include:
   • Maintaining registries of high-risk AI deployments, AI adverse incidents, and AI system audits;
   • With respect to audit standards and/or auditor certifications, advocating for the needs of federal agencies and coordinating with audit processes undertaken or required by federal agencies themselves; and
   • Providing evaluation, certification, documentation, coordination, and disclosure oversight, as needed.

8. Contracting: The federal government should require that government suppliers, contractors, and grantees adopt sound AI governance and assurance practices for AI used in connection with the contract or grant, including using AI standards and risk management practices recognized by federal agencies, as applicable. This would ensure that entities contracting with the federal government or receiving federal grants are enacting sound internal AI system assurances. Such practices in this market segment could accelerate adoption more broadly and improve the AI accountability ecosystem throughout the economy.
1. Introduction
The RFC relied on the NIST delineation of “trustworthy AI” attributes: valid and reliable, safe, secure and resilient, privacy-enhanced, explainable and interpretable, accountable and transparent, and fair with harmful bias managed. To be clear, trust and assurance are not products that AI actors generate. Rather, trustworthiness involves a dynamic between parties; it is in part a function of how well those who use or are affected by AI systems can interrogate those systems and make determinations about them, either themselves or through proxies.

AI assurance efforts, as part of a larger accountability ecosystem, should allow government agencies and other stakeholders, as appropriate, to assess whether the system under review (1) has substantiated claims made about its attributes and/or (2) meets baseline criteria for “trustworthy AI.” The RFC asked about the evaluations entities should conduct prior to and after deploying AI systems; the necessary conditions for AI system evaluations and certifications to validate claims and provide other assurance; different policies and approaches suitable for different use cases; helpful regulatory analogs in the development of an AI accountability ecosystem; regulatory requirements such as audits or licensing; and the appropriate role for the federal government in connection with AI assurance and other accountability mechanisms.

Over 1,440 unique comments from diverse stakeholders were submitted in response to the RFC and have been posted to Regulations.gov. An NTIA employee read every comment. Approximately 1,250 of the comments were submitted by individuals in their own capacity. Approximately 175 were submitted by organizations or groups or legal persons. (italics in original).

Since the release of the RFC, the Biden-Harris Administration has worked to advance trustworthy AI in several ways. In May 2023, the Administration secured commitments from leading AI developers to participate in a public evaluation of AI systems at DEF CON 31. The Administration also secured voluntary commitments from leading developers of “frontier” advanced AI systems (“White House Voluntary Commitments”) to advance trust and safety, including through evaluation and transparency measures that relate to queries in the RFC. In addition, the Administration secured voluntary commitments from healthcare companies related to AI. Most recently, President Biden issued an Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (“AI EO”), which advances and coordinates the Administration’s efforts to ensure the safe and secure use of AI; promote responsible innovation, competition, and collaboration to create and maintain the United States’ leadership in AI; support American workers; advance equity and civil rights; protect Americans who increasingly use, interact with, or purchase AI and AI-enabled products; protect Americans’ privacy and civil liberties; manage the risks from the federal government’s use of AI; and lead global societal, economic, and technical progress. Administration efforts to advance trustworthy AI (to/you) to the release of the RFC in April 2023 include most notably the NIST AI Risk Management Framework (NIST AI RMF) and the White House Blueprint for an AI Bill of Rights (Blueprint for AIBoR).

Introduction

NTIA issued a Request for Comment on AI Accountability Policy on April 13, 2023 (RFC). The RFC included 34 questions about AI governance methods that could be employed to hold relevant actors accountable for AI system risks and harmful impacts. It specifically sought feedback on what policies would support the development of AI audits, assessments, certifications, and other mechanisms to create earned trust in AI systems—which practices are also known as AI assurance. To be account able, relevant actors must be able to assure others that the AI systems they are developing or deploying are worthy of trust, and face consequences when they are not. The RFC relied on the NIST delineation of “trustworthy AI” attributes: valid and reliable, safe, secure and resilient, privacy-enhanced, explainable and interpretable, accountable and transparent, and fair with harmful bias.

See Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi, “Accountability in Artificial Intelligence: What It Is and How It Works,” (Feb. 7, 2023), AI & Society: Journal of Knowledge, Culture and Communication, https://doi.org/10.1007/s00146-023-01635-y (stating that AI accountability “denotes a relation between an agent A and what is usually called) a forum F, such that A must justify A’s conduct to F, and F supervises, asks questions to, and passes judgment on the basis of such justification. . . . Both A and F need not be natural, individual persons, and may be groups or legal persons.” (italics in original).

1 See The White House, FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects American Rights and Safety (May 4, 2023). meat.”

2 The later-adopted AI EO uses the term “safe, secure, and trustworthy AI.” Because safety and security are part of NIST’s definition of “trustworthy,” this report uses the “trustworthy” catch-all. Other policy documents use “responsible AI.” See, e.g., Government Accountability Office (GAO), Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities (GAO Report No. GAO-21-530SP, Jun 30, 2021), https://www.gao.gov/assets/gao-21-530sp.pdf, at 24 (listing U.S. government documents using the term “responsible use” to refer to AI(um) - use that is responsible, equitable, transparent, safe, and governed).


Federal regulatory and law enforcement agencies have also advanced AI accountability efforts. A joint statement introduced bills related to AI.13 State legislatures across the country have passed bills that affect AI,14 and localities are legislatively as well.15 The United States has collaborated with international partners to consider AI accountability policy. The U.S. – EU Trade and Technology Council (TTC) roadmap and launched three expert groups in May 2023, of which one is focused on “monitoring and measuring AI risks.”16 These groups have issued a list of 65 key terms, wherever possible unifying disparate definitions.17 Participants in the 2023 Hiroshima G7 Summit have worked to address shared international guiding principles and a code of conduct for trustworthy AI development.18 The United States has collaborated with international partners to consider AI accountability policy. The U.S. – EU Trade and Technology Council roadmap and launched three expert groups in May 2023, of which one is focused on “monitoring and measuring AI risks.” These groups have issued a list of 65 key terms, wherever possible unifying disparate definitions. Participants in the 2023 Hiroshima G7 Summit have worked to address shared international guiding principles and a code of conduct for trustworthy AI development. The United States has collaborated with international partners to consider AI accountability policy. The U.S. – EU Trade and Technology Council roadmap and launched three expert groups in May 2023, of which one is focused on “monitoring and measuring AI risks.” These groups have issued a list of 65 key terms, wherever possible unifying disparate definitions. Participants in the 2023 Hiroshima G7 Summit have worked to address shared international guiding principles and a code of conduct for trustworthy AI development. 

Organization for Economic Cooperation and Development is working on accountability in AI.19 In Europe, the EU AI Act – which includes provisions addressing pre-release conformity certifications for high-risk systems, as well as transparency and audit provisions and special provisions for foundation models20 or general purpose AI – has continued on the path to becoming law.21 The EU Digital Services Act requires audit of the largest online platforms and search engines,22 and a recent EU Commission delegated act on audits indicates that it is important in this context to analyze algorithmic systems and technologies such as generative models.23 In light of all this activity, it is important to articulate the scope of this Report. Our attention is on voluntary, regulatory, and other measures and policies that are designed to provide assurance to external stakeholders that AI systems are legal and trustworthy. Any of these players can cause harm, but this Report focuses on developments and deploying as the most relevant entities for policy interventions. This Report concentrates further on the cross-sectoral aspects of AI accountability, while acknowledging that AI accountability mechanisms are likely to differ in different sectors. Multiple policy interventions may be necessary to achieve accountability. Take, for example, a policy promoting the disclosure of appropriate parties of training data, performance limitations, and model characteristics for high-risk AI systems. Disclosure alone does not make an AI act accountable. However, this information will likely be important for accountability within the AI actor’s domain and for external accountability as regulators, litigators, courts, and the public act on such information. Disclosure, then, is an accountability input whose effectiveness depends on other policies or conditions, such as the governing liability framework, relevant regulation, and market forces (in particular, customers’ and consumers’ ability to use the information disclosed make purchase and use decisions). This report touches on how accountability inputs feed into the larger accountability apparatus and considers how these connections might be developed in further work. Our final limitations on scope concern matters that are...
the focus of other federal government inquiries. Although NTIA received many comments related to intellectual property, particularly on the role of copyright in the development and deployment of AI, this Report is largely silent on intellectual property issues. Mitigating risks to intellectual property (e.g. infringement, unauthorized data transfers, unauthorized disclosures) are certainly recognized components of AI accountability. These issues are of ongoing consideration at the U.S. Patent and Trademark Office (USPTO) and at the U.S. Copyright Office. We look forward to working with these agencies and others on these issues as warranted to help ensure that AI accountability and related transparency, safety, and other considerations relevant to the broader digital economy and Internet ecosystem are represented.

24 See, e.g., NIST AI RMF at 1; 24 Investigating that training data should follow applicable intellectual property rights laws, that policies and procedures should be in place to address risks of infringement of a third party’s intellectual property or other right; Intellectual Property International Code of Conduct for Organizations Developing Advanced AI Systems, supra note 28, at 8 (calling on organizations to “implement appropriate data input measures and protections for personal data and intellectual property” and encouraging organizations “to implement appropriate safeguards, to respect rights related to privacy and intellectual property, including copyright protection.”)


26 See, e.g., U.S. Copyright Office, Notice of Inquiry and Request for Comments on Artificial Intelligence and Copyright, 88 Fed. Reg. 5043 (Aug. 30, 2023) (hereinafter “Copyright Office AI RFC”) “Artificial intelligence (AI) creations are covered in more depth later. Such issues include the potential benefits, risks, and implications, of AI development and deployment of AI systems generally are of tremendous interest and raise distinct accountability issues. The AI EO tasked the Secretary of Commerce with soliciting input and issuing a report on “the potential benefits, risks, and implications, of dual-use foundation models for which the weights are widely available, as well as policy and regulatory recommendations pertaining to such models,” and NTIA has published a Request for Comment for the purpose of informing that report.”

27 Similarly, the role of privacy and the use of personal data in model training are topics of great interest and significance to AI accountability. More than 90% of all organizational commenters noted the importance of data protection and privacy to trustworthy and accountable AI. AI can exacerbate risks to Americans’ privacy, as recognized by the Blueprint for an AI Bill of Rights and the AI EO. Privacy protection is not only a focus of AI accountability, but importantly privacy also needs to be considered in the development and use of accountability tools. Documentation, disclosures, audits, and other forms of evaluation can result in the collection and exposure of personal information, thereby jeopardizing privacy if not properly designed and executed. Stronger and clearer rules for the protection of personal data are necessary through the passage of comprehensive federal privacy legislation and other actions by federal agencies and the Administration. The President has called on Congress to enact comprehensive federal privacy protections.

28 The remainder of this Report is organized as follows:

Section 2 of the Report outlines significant commenter alignment around cross-cutting issues, many of which are covered in more depth later. Such issues include calibrating AI accountability policies to risk, assuring AI systems across their lifecycle, standardizing disclosures and evaluations, and increasing the federal role in supporting and/or requiring certain accountability inputs.

Section 3 of the Report dives deeper into these issues, organizing the discussion around three key ingredients of AI accountability: (1) information flow, including documentation of AI system development and deployment; relevant disclosures appropriately detailed to the stakeholder audience; and provision to researchers and evaluators of adequate access to AI system components; (2) AI system evaluations, including government requirements for independent evaluation and pre-release certification (or licensing) in some cases; and (3) government support for an accountability ecosystem that widely distributes effective scrutiny of AI systems, including within government itself.

Section 4 shows how accountability inputs intersect with liability, regulatory, and market-forcing functions to ensure real consequences when AI actors forfeit trust.

Section 5 surveys lessons learned from other accountability models outside of the AI space.

Section 6 concludes with recommendations for government action.

Appendix A is a glossary of terms used in this Report.

Finally, open-source AI models, AI models with widely available model weights, and components of AI systems generally are of tremendous interest and raise distinct accountability issues. The AI EO tasked the Secretary of Commerce with soliciting input and issuing a report on “the potential benefits, risks, and implications, of dual-use foundation models for which the weights are widely available, as well as policy and regulatory recommendations pertaining to such models,” and NTIA has published a Request for Comment for the purpose of informing that report. 31

12

13
2. Requisites for AI Accountability: Areas of Significant Commenter Agreement
The comments submitted to the RFC compose a large and diverse corpus of policy ideas to advance AI accountability. While there were significant disagreements, there was also a fair amount of support among stakeholders from different constituencies for making AI systems more open to scrutiny and more accountable to all. This section provides a brief overview of significant plurality (if not majority) sentiments in the comments relating to AI accountability policy, along with NTIA reflections. Section 3 provides a deeper treatment of these positions; most are congruent with the Report’s recommendations in Section 6.

2.1. RECOGNIZE POTENTIAL HARMs AND RISKS

Many commenters, especially individual commenters, expressed serious concerns about the impact of AI. AI system potential harms and risks have been well-documented elsewhere. The following are representative examples, which also appeared in comments:

- Inefficacy and inadequate functionality.
  - Inaccuracy, unreliability, ineffectiveness, insufficient robustness.
  - Unfitness for the use case.

- Lowered information integrity.
  - Misleading or false outputs, sometimes coupled with coordinated campaigns.
  - Opacity around use.
  - Opacity around provenance of AI inputs.
  - Opacity around provenance of AI outputs.

- Safety and security concerns.
  - Unsafe decisions or outputs that contribute to harmful outcomes.
  - Capacities falling into the hands of bad actors who intend harm.
  - Adversarial evasion or manipulation of AI.
  - Obstacles to reliable control by humans.
  - Harmful environmental impact.

- Violation of human rights.
  - Discriminatory treatment, impact, or bias.
  - Improper disclosure of personal, sensitive, confidential, or proprietary data.
  - Lack of accessibility.
  - The generation of non-consensual intimate imagery of adults and child sexual abuse material.
  - Labor abuses involved in the training of AI data.

- Impacts on privacy.
  - Exposure of non-public information through AI analytical insights.
  - Use of personal information in ways that are contrary to the contexts in which they are collected.
  - Overcollection of personal information to create training datasets or to unduly monitor individuals (such as workers and trade unions).

- Potential negative impact to jobs and the economy.
  - Infringement of intellectual property rights.
  - Infringements on the ability to form and join unions.
  - Job displacement, reduction, and/or degradation of working conditions, such as increased monitoring of workers and the potential mental and physical health impacts.
  - Undue concentration of power and economic benefits.

- Impacts on innovation.
  - Replacement of human creativity by automated processes.
  - Overreliance on AI systems for critical decision-making.
  - Misappropriation of intellectual property through AI.

- Undue concentration of power.
  - Large corporations and other concentrated power.

- Harms to the American public at large.
  - Harmful environmental impact.
  - Infringement of intellectual property rights.
  - Harmful outcomes.

Individual commenters reflected misgivings in the American public at large about AI. Three major themes emerged from many of the individual comments:

- The most significant by the numbers was concern about intellectual property. Nearly half of all individual commenters (approximately 47%) expressed alarm that generative AI was ingesting as training material copyrighted works without the copyright holders’ consent, without their compensation, and/or without attribution. They also expressed worries that AI could supplant the jobs of creators and other workers. Some of these commenters supported new forms of regulation for AI that would require copyright holders to opt-in to AI system use of their works.

- Another significant concern was that malicious actors would exploit AI for destructive purposes and develop their own systems for those ends. A related concern was that AI systems would not be subject to sufficient controls and would be used to harm individuals and communities, including through unlawfully discriminatory impacts, privacy violations, fraud, and a wide array of safety and security breaches.

- A final theme concerned the personnel building and deploying AI systems, and the personnel making AI policy. Individual commenters questioned the credibility of the responsible people and institutions and doubted whether they had sufficiently diverse experiences, backgrounds, and inclusive practices to foster appropriate decision-making.

---


35 “The term ‘generative AI’ means the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content.” 1716 Office of the Register, “Generative AI: Notice of Proposed Rulemaking,” 89 Fed. Reg. 54854 (Aug. 28, 2014).

36 Stakeholders are deeply divided on some of these policy issues, such as the implications of “opt-in” or “opt-out” systems, or compensation for authors, which are part of the U.S. Copyright Office’s inquiry and USPTO ongoing work. This report recognizes the importance of these issues to the overall risk management and accountability framework without touching on the merits.
As discussed below, many commenters argued that accountability mechanisms depend on the AI use context and risk profile; Securities and Exchange Commission (SEC) Comment at 2; Anan Abrar Comment at 1.

2.2. CALIBRATE ACCOUNTABILITY INPUTS TO RISK LEVELS

Commenters generally support calibrating AI accountability inputs to scale with the risk of the AI system or application. As many acknowledge, existing work from NIST, the Organization for Economic Cooperation and Development, and other international partners on Artificial Intelligence, and the European Union (e.g., the EU AI Act), among others, have established robust frameworks to map, measure, and manage risks. In the interest of risk-based accountability, one commenter, for example, suggested a “baseline plus” approach: all models and applications are subject to some baseline standard of assurance practices across sectors and higher risk models or applications have an additional set of obligations. NTIA concludes that a tiered approach to AI accountability has the benefit of scoping expectations and obligations proportionately to AI system risks and capabilities.

2.3. ENSURE ACCOUNTABILITY ACROSS THE AI LIFECYCLE AND VALUE CHAIN

Various actors in the AI value chain exercise different degrees of control throughout the lifecycle of an AI system. Upstream developers design and create AI models and/or applications. Downstream developers then deploy those models and/or systems (or use the models as part of other systems) in particular contexts. The downstream developers may also fine tune a model, thereby acting as downstream developers of the deployed systems. Both upstream developers and downstream developers of AI systems should be accountable; existing laws and regulations may already specify accountability mechanisms for different actors.

Commenters laid out good reasons to vest accountability with AI system developers who make critical upstream decisions about AI models and other components. These actors have privileged knowledge to inform important disclosures and documentation and may be best positioned to manage certain risks. However, some commenters argued that safety-impacting or rights-impacting AI systems deserve to be released with quality assurance certifications based on passing pre-release certification and capabilities. For example, “The transparency and accountability requirements should also be tailored and calibrated according to the amount of risk presented by the specific sector or domain in which the system is intended to be deployed.” The Value Chain of AI Oversight: A Horizontal Framework for Trustworthiness (proposing regulatory regime for frontier models that would require pre-release certification around information security, safety, culture, and technical safety); Opencare Comment at 6 (considering a requirement of pre-deployment risk assessments, security and deployment safeguards). Microsoft at 4 (Regulatory framework is based on the AI Act, including training requirements for foundational models and infrastructure); Anthropic at 32 (confidential sharing of fine-tuning parameters with regulators); Credo AI at 31 (Special foundation model: unclassify language model disclosures to government and professional organizations, including AI safety and governance); Audit AI at 23 (High-risk AI systems should be released with quality assurance certifications based on training and maintaining ongoing compliance with AI accountability regulations); Holistic AI at 19 (High-risk systems should be released with certifications).
We think it is likely that agencies will need additional capacities and possibly authorizations to enable and require AI accountability. The body or bodies with cross-sectoral capacity might provide technical and legal support to sectoral regulators, as well as exercise other responsibility limits, to which the combination of sectoral and cross-sectoral capacities would facilitate the “baseline plus” approach to AI assurance practices described in Section 2.2.

2.5. FACILITATE INTERNAL AND INDEPENDENT EVALUATIONS

Commenters noted that self-administered AI system assessments are important for identifying risks and system limitations, building internal capacity for ensuring trustworthy AI, and feeding into independent evaluations. Internal assessments could be a principal object of analysis and verification for independent evaluators to the extent that the assessments are made available.44 Independent external third-party assessments (also known for short as independent evaluations), including audits and red-teaming, may be necessary for the riskiest systems under a risk-based approach to accountability.45 These independent evaluations can serve to verify claims made about AI system attributes and performance, and/or to measure achievement with respect to those attributes against external benchmarks. Many commenters insisted that AI accountability mechanisms should be mandatory,46 while others thought that voluntary commitments to audits or other independent evaluations would suffice.47 There were also plenty of commenters in between, with one noting that “a healthy policy ecosystem likely balances mandatory accountability mechanisms where risks demand it with voluntary incentives and platforms to share best practices.”48 We believe there should be a mix of internal and independent evaluations, for the reasons stated above. AI actors may well undertake these evaluations voluntarily in the interest of risk management and harm reduction. However, as discussed below, regulatory and legal requirements around evaluations and evaluation inputs may also be necessary to make relevant actors answerable for their choices. Rather than impede innovation, governance to foster robust evaluations could abet AI development.49

2.6. STANDARDIZE EVALUATIONS AS APPROPRIATE

Commenters noted the importance of using standards to develop common criteria for evaluations. The use of standards in evaluations is important to implement replicable evaluations. Commenters acknowledged, as does the NIST AI RMF, that there may be tradeoffs between accountability inputs such as disclosure, and other values such as protecting privacy, intellectual property, and security.50 In

The use of standards is important to implement replicable and comparable evaluations.

2.7. FACILITATE APPROPRIATE ACCESS TO AI SYSTEMS FOR EVALUATION

Although some kinds of AI system evaluations are possible without the active participation of AI actors, researchers and other independent evaluators will sometimes need access to AI system components to enable comprehensive evaluations. These components include at least documentation, data, code, and models, subject to intellectual property, privacy, and security protections.51 In

the relevant models and processes.”

47 See, e.g., Google DeepMind Comment at 3 regarding “hub-and-spoke” model of AI regulation, with sectoral regulators overseeing AI implementation with horizontal guidance from a central authority (like NIST). Boston University and University of Chicago Comment at 3 emphasizing existing sectoral authorities to “serve most effectively” and to ensure at-generality risks of AI, if we recommend establishment of a meta-agency with broad AI-related expertise (both technical and legal) which would develop baseline regulations regardless of the general safety of AI systems, set standards, and enable review for compliance with substantive law, while collaborating with sectoral and cross-sectoral authorities as they consider the impact of AI systems on their regulatory sanitation.52

48 See infra Secs. 3.2.4.

49 AI accountability RCF, RCF, Reg. at 232 at 232. As discussed in the RCF, “15(depending on the strength demonstrated oversight of the procurement, development, and use of AI….) considers the creation of a new independent federal agency on the overarching authority systems.”50 USThink on Comment at 6 (“When individuals use AI in systems or different sectors are held to the same expectations, it assumes that adequate safeguards are in place to protect their rights and well-being, regardless of the company deploying AI.”) Salesforce Comment at 9 (“There can be tension between accountability goals that lead to technical tradeoffs, including some AI Phrase “brakes help to drive faster” to explain this phenomenon - the ability to stop a car in dangerous situations enables us to feel comfortable driving at high speeds. Governance is innovation.”

50  Anthropic Comment at 10 (recommending mandatory adversarial testing of AI systems before release through NIST or researcher access); Anti-Defamation League Comment at 11 (applying that finding that adversarial attacks thwart oversight of the procurement, development, and use of AI….) considers the creation of a new independent federal agency on the overarching authority systems.”

51 There were also plenty of commenters in between, with one noting that “a healthy policy ecosystem likely balances mandatory accountability mechanisms where risks demand it with voluntary incentives and platforms to share best practices.”

52 DLA Piper Comment at 12.

53 See Rumsan Chowdhry, Submitted Written Testimony for Full Committee Hearing of the House of Representative on Science, Space, and Technology (a) Technology and Innovation: Advancing Innovation Towards the National Interest (Jan. 26, 2020).

54 See, e.g., Alex Leader Comment at 2 (stating the importance to similarly establish AI safety standards which could serve as criteria for the subject matter of an AI assurance engagement to be evaluated against); Salesforce Comment at 11 (“in definitions and methods were standardized, audits would be more consistent and transparent evaluations. In this context, it is also necessary to fund party certification is included in future regulations.”)

55 See NIST AI RMF at Sects 2.3 and 2.3. See also Google DeepMind Comment at 8 (“suggesting there are tradeoffs between data mininization and the accuracy of systems; transparency and model accuracy, and transparency and security.”). (specifically Comment at 4 (stating that if an auditor obtains access to underlying data, privacy, security, and AI risks are significantly different to legitimate.”) MasterComment and Comment at 2 (“there is tension between accountability goals that lead to technical trade-offs, and the question of, whether (and if so) how to evaluate those tradeoffs with and related decisions…” transparency is another example of an AI accountability goal to which a model architecture and information sharing posed.)

56 See, e.g., Alex Leader Comment at 2-3; “While independent third party audits, for developing and deploying safe, secure and trustworthy AI systems; NIST, U.S. leadership in AI for Federal Engagement in Developing AI Technical Standards and Related Tools, https://www.nist.gov/ai.”

57 In addition, the use of standards will help to drive the development and implementation of AI-related consensus standards, cooperation and coordination, and information sharing.

58 Alex Leader Comment at 2-3; “While independent third party audits, for developing and deploying safe, secure and trustworthy AI systems; NIST, U.S. leadership in AI for Federal Engagement in Developing AI Technical Standards and Related Tools, https://www.nist.gov/ai.”

59 See, e.g., Alex Leader Comment at 2-3; “While independent third party audits, for developing and deploying safe, secure and trustworthy AI systems; NIST, U.S. leadership in AI for Federal Engagement in Developing AI Technical Standards and Related Tools, https://www.nist.gov/ai.”

60 See also, e.g., Alex Leader Comment at 2-3; “While independent third party audits, for developing and deploying safe, secure and trustworthy AI systems; NIST, U.S. leadership in AI for Federal Engagement in Developing AI Technical Standards and Related Tools, https://www.nist.gov/ai.”
Commenters also addressed the value of producing information about the inputs to and source of AI-generated content, also known as “provenance.”

NTIA agrees with commenters that appropriate transparency around AI systems is critical. Information should be pushed out to stakeholders in form and scope appropriate for the audience and risk level. Communications to the public should be in plain language. Evaluators may also need information about aspects of AI system training data, when content is AI-generated, and the authenticity of the purported source of content. Source detection and identification are important aspects of information flow and information integrity.

2.8. STANDARDBASE AND ENCOURAGE INFORMATION PRODUCTION

Commenters stressed the importance of AI actors providing documentation on matters such as:

- Problem specification;
- Training data, including collection, provenance, curation, and management;
- Model development;
- Testing and verification;
- Risk identification and mitigation;
- Model output interpretability;
- Risk mitigation safeguards; and
- System performance and limitations.

Commenters also noted that the currently is not an adequate framework to conduct AI system evaluations, particularly given the demands of sociotechnical inquiries, the variety of expertise entailed, and supply constraints on the relevant workforce. In addition, inadequate access to data and compute (referring to computing power in the AI context), inadequate funding, and incomplete standardization were cited as other barriers to developing accountability inputs. Another concern of commenters was that auditors can be captured by the auditees who hire them.

Recognizing possible deficiencies in the supply, resources, and independence of AI evaluators, NTIA favors more federal support for independent auditing and red-teaming. Such support could take the form of facilitating system access, funding education, conducting and funding research, sponsoring prizes and competitions, providing datasets and compute, and hiring into government. At the same time, the federal government should build capacity to conduct evaluations itself and provide a backstop to ensure that independent auditors provide adequate assurance. The sequencing and prioritization of these efforts is an urgent question for policymakers.

2.9. FUND AND FACILITATE GROWTH OF THE ACCOUNTABILITY ECOSYSTEM

Commenters noted that the currently is not an adequate framework to conduct AI system evaluations, particularly given the demands of sociotechnical inquiries, the variety of expertise entailed, and supply constraints on the relevant workforce. In addition, inadequate access to data and compute (referring to computing power in the AI context), inadequate funding, and incomplete standardization were cited as other barriers to developing accountability inputs. Another concern of commenters was that auditors can be captured by the auditees who hire them.

Recognizing possible deficiencies in the supply, resources, and independence of AI evaluators, NTIA favors more federal support for independent auditing and red-teaming. Such support could take the form of facilitating system access, funding education, conducting and funding research, sponsoring prizes and competitions, providing datasets and compute, and hiring into government. At the same time, the federal government should build capacity to conduct evaluations itself and provide a backstop to ensure that independent auditors provide adequate assurance. The sequencing and prioritization of these efforts is an urgent question for policymakers.

2.10. INCREASE FEDERAL GOVERNMENT ROLE

A strong sentiment running through both institutional and individual comments was that there should be a significant federal government role in funding, incentivizing, and/or requiring accountability measures. Commenters recommended additional federal funding and/or support for more AI safety research, standards development, the release of standardized datasets for testing, and professional development. They recommended that government consider providing a regulatory sandbox for entities, under certain conditions, to experiment with responsible AI and compliance efforts free from regulatory risk. They urged federal procurement reform, as the National Artificial Intelligence Advisory Committee recommended, in order to drive...
Independent evaluations can serve to verify claims made about AI system attributes and performance, and/or to measure achievement with respect to those attributes against external benchmarks.

trustworthy AI by adopting rigorous documentation, disclosure, and evaluation requirements. As noted above, they argued for mandatory audits and other mandatory AI accountability measures, including a federal role in certifying auditors and setting audit benchmarks, as is customary in other regulatory domains.

Federal regulatory involvement with accountability measures in other fields, while not directly applicable to AI, may be instructive. In this vein, commenters pointed to precedents such as the Food and Drug Administration (FDA) premarket review for medical devices, the National Highway Traffic Safety Administration auto safety standards, FDA nutrition labels, the Environmental Protection Agency (EPA) ENERGY STAR® labels, the Federal Aviation Administration (FAA) accident examination and safety processes, and the Securities and Exchange Commission (SEC) audit requirements.

An area of overwhelming agreement in the commentary was the importance of data protection and privacy to AI accountability, with commenters expressing the view that a federal privacy law is either necessary or important to trustworthy and accountable AI.

As our recommendations elaborate in Section 6, we support accelerated and coordinated government action to determine the best federal regulatory and non-regulatory approaches to the documentation, disclosure, access, and evaluation functions of the AI accountability chain.
Our analysis now turns to the first two links in the AI accountability chain—what we are calling accountability inputs, these are roughly (1) the creation, collection, and distribution of information about AI systems and system outputs, and (2) AI system evaluation. The RFC and commenters identified proposed or adopted laws that address AI accountability inputs, both in the United States and beyond. Congress continues to consider relevant legislative initiatives, and the states are actively pursuing their own legislative agendas. Many of these policy initiatives focus on information flow and evaluations, as well as associated governance processes. The sections below address these topics and come to some preliminary conclusions that feed into the recommendations in Section 6.

3.1. INFORMATION FLOW

One of the challenges with assuring AI trustworthiness is that AI systems are complex and often opaque. As a result, information asymmetries and gaps open along the value chain from developers to deployers and ultimately to end users and others affected by AI operations.

Downstream deployers of AI systems may lack information they need to use the AI systems appropriately in context. Developers may lack information about deployment contexts and therefore make inaccurate claims or fail to communicate limitations.


The information asymmetry runs the other way as well. AI system developers may lack information about deployment contexts and therefore make inaccurate claims about their products or fail to communicate limitations.

For example, to mitigate downstream harms, the developer of an AI image generator would need information about later adaptations and adverse incidents to address the risks posed by deepfakes at scale.74

Commentators are in broad agreement that more information about AI systems is needed, with some asserting that there may be tradeoffs between transparency and other values. 87 There was a range of commenter opinion about the processing of personal data and on the free movement of such data, with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119 (2023), https://data.europa.eu/uri/10.11832/2016/516012-

74 NIST AI RMF at 16-17.
75 See Guardian Assembly Comment at 7 (“Transparency is key to ensuring that stakeholders have access to relevant information about an AI system. […] Transparency helps to facilitate accountability by enabling stakeholders to understand and assess an AI system’s behaviors.”); AMP, supra note 71; and Transparency and Accountability, supra note 74, at 2 (“Transparency is key to trust in model analysis as it increases transparency in how AI models are trained and deployed.”).
76 Id.
77 See, e.g., DeepMind Comment at 7 (“Without information about how computers are developing and using AI and the extent to which it is working properly, investors are essentially left to trust the marketing claims of the companies they invest in”); and TII, supra note 74, at 16-17.
78 See also 87 Id.
79 See, e.g., DeepMind Comment at 7 (“Transparency is key to ensuring that stakeholders have access to relevant information about an AI system. […] Transparency helps to facilitate accountability by enabling stakeholders to understand and assess an AI system’s behaviors.”).
81 "You should know that an automated system is being used and understand how it works. This lack of information may hinder people from asserting rights under existing law, exercising their own critical judgement, or pushing for other forms of redress. This lack of transparency is a significant cause for concern because AI lacks the required elements of accountability (e.g., traceability, transparency), necessary to ensure the public’s confidence in AI decision-making.
82 Id.
83 See, e.g., DeepMind Comment at 7 (“Transparency is key to ensuring that stakeholders have access to relevant information about an AI system. […] Transparency helps to facilitate accountability by enabling stakeholders to understand and assess an AI system’s behaviors.”).
84 Id.
85 See, e.g., DeepMind Comment at 7 (“Transparency is key to ensuring that stakeholders have access to relevant information about an AI system. […] Transparency helps to facilitate accountability by enabling stakeholders to understand and assess an AI system’s behaviors.”).
86 Id.
87 See, e.g., DeepMind Comment at 7 (“Transparency is key to ensuring that stakeholders have access to relevant information about an AI system. […] Transparency helps to facilitate accountability by enabling stakeholders to understand and assess an AI system’s behaviors.”).
88 See, e.g., DeepMind Comment at 7 (“Transparency is key to ensuring that stakeholders have access to relevant information about an AI system. […] Transparency helps to facilitate accountability by enabling stakeholders to understand and assess an AI system’s behaviors.”).
89 Id.
90 Id.
91 See also 87 Id.
92 See also 87 Id.
93 Id.
94 Id.
95 See, e.g., DeepMind Comment at 7 (“Transparency is key to ensuring that stakeholders have access to relevant information about an AI system. […] Transparency helps to facilitate accountability by enabling stakeholders to understand and assess an AI system’s behaviors.”).
96 Id.
97 See, e.g., DeepMind Comment at 7 (“Transparency is key to ensuring that stakeholders have access to relevant information about an AI system. […] Transparency helps to facilitate accountability by enabling stakeholders to understand and assess an AI system’s behaviors.”).
98 Id.
Information flow as an input to AI accountability comes in two basic forms: push and pull. AI actors can push disclosures out to stakeholders and stakeholders can pull information from AI systems, via system access subject to valid intellectual property, privacy, and security protection.90 Reporting on AI systems recommends a mix of push and pull information flow, some of which should be required and some voluntarily assumed. Because AI systems are continuously updated and refined, information pushed out (e.g., reports, model cards) should also be continuously updated and refined. Similarly, access to AI system components may need to be ongoing.

3.1.1 AI SYSTEM DISCLOSURES

In the words of one commenter, “one of the greatest barriers to AI accountability is the lack of a standard account­ability reporting framework.”91 As the NIST AI RMF proposes, AI system developers and deployers should push out more information about (1) the AI system itself, including the training data and model, and (2) about AI system use, including the fact of its use, adverse incident reporting, and outputs.92 Some information should be shared with the general public, while sensitive information might be disclosed only to groups trusted to ensure the necessary safeguards are in place, including government.

One commenter stated that “[i]f adopted across the industry, transparency reports would be a helpful mechanism for recording the maturing practice of responsible AI and charting industry progress.”93 The EU is requiring transparency reports for large digital platforms.94 While transparency is critical in the AI context, non-standard disclosure at the dis­closor’s discretion is less useful as an accountability input than standard, regular disclosure.95 A family of informational artifacts – including datasheets, model cards, and system cards – can be used to provide structured disclosures about AI models and related data. Datasheets (also referred to as data cards, dataset sheets, data statements, or data set nutrition labels) provide salient information about the data on which the AI model was trained, including the “manufacture, composition, collection process, (and) recommended uses” of the dataset.96 Several commenters recommended that AI system developers produce datasheets.97

More information on (1) the AI system itself, including the training data and model, and (2) about AI system use, including the fact of its use, adverse incident reporting, and outputs.98 Some information should be shared with the general public, while sensitive information might be disclosed only to groups trusted to ensure the necessary safeguards are in place, including government.

One commenter stated that “[i]f adopted across the industry, transparency reports would be a helpful mechanism for recording the maturing practice of responsible AI and charting industry progress.”99 The EU is requiring transparency reports for large digital platforms.100 While transparency is critical in the AI context, non-standard disclosure at the dis­closor’s discretion is less useful as an accountability input than standard, regular disclosure.101 A family of informational artifacts – including datasheets, model cards, and system cards – can be used to provide structured disclosures about AI models and related data. Datasheets (also referred to as data cards, dataset sheets, data statements, or data set nutrition labels) provide salient information about the data on which the AI model was trained, including the “manufacture, composition, collection process, (and) recommended uses” of the dataset.102 Several commenters recommended that AI system developers produce datasheets.103

More information on (1) the AI system itself, including the training data and model, and (2) about AI system use, including the fact of its use, adverse incident reporting, and outputs.104 Some information should be shared with the general public, while sensitive information might be disclosed only to groups trusted to ensure the necessary safeguards are in place, including government.

One commenter stated that “[i]f adopted across the industry, transparency reports would be a helpful mechanism for recording the maturing practice of responsible AI and charting industry progress.”105 The EU is requiring transparency reports for large digital platforms.106 While transparency is critical in the AI context, non-standard disclosure at the dis­closor’s discretion is less useful as an accountability input than standard, regular disclosure.107 A family of informational artifacts – including datasheets, model cards, and system cards – can be used to provide structured disclosures about AI models and related data. Datasheets (also referred to as data cards, dataset sheets, data statements, or data set nutrition labels) provide salient information about the data on which the AI model was trained, including the “manufacture, composition, collection process, (and) recommended uses” of the dataset.108 Several commenters recommended that AI system developers produce datasheets.109

Model cards disclose information about the performance and context of a model, including:
- Basic information;
- On-label (intended) and off-label (not intended, but predictable) use cases;
- Model performance measurements in terms of the relevant metrics depending on various factors, including the affected group, implementation, and deployment environment;
- Descriptions of training and evaluation data; and
- Ethical considerations, caveats, and recommendations.

System cards are used to make disclosures about how entire AI systems, often composed of a series of models working together, perform a specific task.110 A system card can show step-by-step how the system processes actual input, for example to compute a ranking or make a prediction. Proponents state that, in addition to the disclosures about individual models set forth in model cards, system cards are intended to consider factors including deployment contexts and real-world interactions.111

These artifacts might be formatted in the form of a “nu­tritional label,” which would present standardized information in an analogous format to the “Nutrition Facts” label mandated by the FDA. Twilio’s “AI Nutrition Facts” project shows what a label might look like in the AI con­text, pictured on the left.112 Model cards and system cards are often accompanied by lengthier technical reports describing the training and capabilities of the system.113 Many AI system developers have begun voluntarily re­leasing these artifacts.114 The authors of such artifacts often state that they are written to conform to the recom­mendations in the "AI System Cards: A New Resource for Understanding How AI Systems Work" note 12.

See, e.g., Aliens, AI Policy and Governance Working Group Comment at 3.

See, e.g., CDT Comment at 41-42 (listing to CDT Civil Rights Standards for 21st Century Employment Selection Procedures); Databricks Comment at 2, 5 ("[T]he deployer is the party imposing people to the application and creating the potential risk" and thus "any obligation to inform people about such tools operating should rest with the deployer").

See, e.g., American Federation of Teachers (AFT) Comment at 2 ("Regulations around classroom AI should mandate transparency in the AI system’s decision-making processes. They must allow teachers, students and parents to understand how AI decisions are affecting teaching and learning.").

PWC Comment at AI: So what is AI at 13 (“Standardized reporting – including references to the agreed-upon Model Accuracy Review Framework, clarification of the evaluation criteria, and articulation of findings – would help reinforce public trust”); Ernst & Young Comment at 12 (“[Standard] reporting should be considered where practicable”); Greening Institute (GI) at 1 (“Accountability mechanisms could includ­e requiring risk assessments in the use of these systems, requiring the disclosure of how decisions are made as part of these systems, and requiring the disclosure of how these systems are tested, validated for accuracy and key metrics and definitions in those tests”).

See, e.g., NIST AI RMF at 15-16 (governing AI, supra note 23).

See European Commission, supra note 22.


See, e.g., GA Comment at 5 ("AI accountability mechanisms could includ­e requiring risk assessments in the use of these systems, requiring the disclosure of how decisions are made as part of these systems, and requiring the disclo­sure of how these systems are tested, validated for accuracy and key metrics and definitions in those tests."); GI Comment at 1 ("The government should take steps that set an expectation of transparency around the development, deployment, and use of AI in high-risk settings, such as forensic algorithms making determinations about access to opportunity, that may include transparency requirements.").

See supra note 22.


See also CDT Comment at 1 ("AI accountability mechanisms could include requiring risk assessments in the use of these systems, requiring the disclosure of how decisions are made as part of these systems, and requiring the disclosure of how these systems are tested, validated for accuracy and key metrics and definitions in those tests."); Nekesha Green et al., System Cards, A New Resource for Understanding How AI Systems Work, note 12.

See supra note 22.


See supra note 22.

See, e.g., CDT Comment at 5; CUI ’23: Proceedings of the 5th International Conference on Conversational User Interfaces, at 1-6 (July 2023), https://doi.org/10.1145/3571884.3604316 (surveying the openness of various AI systems, including the disclosure of privacy and academic papers).

See, e.g., GA Comment at 5.

100 See supra note 22.

101 See supra note 22.

102 See supra note 22.

103 See, e.g., Georgia Tech (2023), https://ai.google/static/documents/technical-reports/deepfloyd-palm2.pdf (surveying the openness of various AI systems, including the disclosure of privacy and academic papers).

104 See, e.g., CUI ’23: Proceedings of the 5th International Conference on Conversational User Interfaces, at 1-6 (July 2023), https://doi.org/10.1145/3571884.3604316 (surveying the openness of various AI systems, including the disclosure of privacy and academic papers).

105 See supra note 22.

106 See supra note 22.

107 See supra note 22.

108 See supra note 22.

109 See supra note 22.

110 See supra note 22.

111 See supra note 22.

112 See supra note 22.

113 See supra note 22.

114 See supra note 22.

115 See supra note 22.
The content and form of the disclosure will vary. Some disclosures might be confidential, for example information about large AI training runs provided to the government, especially concerning AI safety and governance. Other disclosures might be set out in graphical form that is accessible to a broad audience of users and other affected people, such as a “nutritional label” for AI system features. AI nutritional labels, by analogy to nutritional labels for food, present the most important information about a model in a relatively brief, standardized, and comparable form. Specific standards for nutritional label artifacts might specify the content required to be included in such a label. To address the varying levels of detail required for different audiences, disclosures should be designed to provide information for each system at multiple different levels of depth and breadth, “allowing everyone from the general populace to the research level expert to understand it at their own pace.”

The federal government could also facilitate access to disclosures as it has in other contexts, such as the SEC’s Electronic Data Gathering and Analysis (EDGAR) platform. The FDA’s Food and Drug Administration (FAERS) platform. To the extent that NIST and others in the development of these standards. For example, the EU AI Act will require regulated entities—principally developers—to disclose (to regulators and the public) information about high-risk AI systems and authorize the European Commission to develop common specifications if needed. Proposed required documentation or disclosures would include information about the data sources used for training, system architecture and general logic, classification choices, the relevance of different parameters, validation and testing procedures, and performance capabilities and limitations.

The federal government could also facilitate access to disclosures as it has in other contexts, such as the SEC’s Electronic Data Gathering and Analysis (EDGAR) platform. The FDA’s Food and Drug Administration (FAERS) platform. To the extent that NIST and others in the development of these standards. For example, the EU AI Act will require regulated entities—principally developers—to disclose (to regulators and the public) information about high-risk AI systems and authorize the European Commission to develop common specifications if needed. Proposed required documentation or disclosures would include information about the data sources used for training, system architecture and general logic, classification choices, the relevance of different parameters, validation and testing procedures, and performance capabilities and limitations.

The federal government could also facilitate access to disclosures as it has in other contexts, such as the SEC’s Electronic Data Gathering and Analysis (EDGAR) platform. The FDA’s Food and Drug Administration (FAERS) platform. To the extent that NIST and others in the development of these standards. For example, the EU AI Act will require regulated entities—principally developers—to disclose (to regulators and the public) information about high-risk AI systems and authorize the European Commission to develop common specifications if needed. Proposed required documentation or disclosures would include information about the data sources used for training, system architecture and general logic, classification choices, the relevance of different parameters, validation and testing procedures, and performance capabilities and limitations.

The federal government could also facilitate access to disclosures as it has in other contexts, such as the SEC’s Electronic Data Gathering and Analysis (EDGAR) platform. The FDA’s Food and Drug Administration (FAERS) platform. To the extent that NIST and others in the development of these standards. For example, the EU AI Act will require regulated entities—principally developers—to disclose (to regulators and the public) information about high-risk AI systems and authorize the European Commission to develop common specifications if needed. Proposed required documentation or disclosures would include information about the data sources used for training, system architecture and general logic, classification choices, the relevance of different parameters, validation and testing procedures, and performance capabilities and limitations.

The federal government could also facilitate access to disclosures as it has in other contexts, such as the SEC’s Electronic Data Gathering and Analysis (EDGAR) platform. The FDA’s Food and Drug Administration (FAERS) platform. To the extent that NIST and others in the development of these standards. For example, the EU AI Act will require regulated entities—principally developers—to disclose (to regulators and the public) information about high-risk AI systems and authorize the European Commission to develop common specifications if needed. Proposed required documentation or disclosures would include information about the data sources used for training, system architecture and general logic, classification choices, the relevance of different parameters, validation and testing procedures, and performance capabilities and limitations.

The federal government could also facilitate access to disclosures as it has in other contexts, such as the SEC’s Electronic Data Gathering and Analysis (EDGAR) platform. The FDA’s Food and Drug Administration (FAERS) platform. To the extent that NIST and others in the development of these standards. For example, the EU AI Act will require regulated entities—principally developers—to disclose (to regulators and the public) information about high-risk AI systems and authorize the European Commission to develop common specifications if needed. Proposed required documentation or disclosures would include information about the data sources used for training, system architecture and general Logic, classification choices, the relevance of different parameters, validation and testing procedures, and performance capabilities and limitations.

The federal government could also facilitate access to disclosures as it has in other contexts, such as the SEC’s Electronic Data Gathering and Analysis (EDGAR) platform. The FDA’s Food and Drug Administration (FAERS) platform. To the extent that NIST and others in the development of these standards. For example, the EU AI Act will require regulated entities—principally developers—to disclose (to regulators and the public) information about high-risk AI systems and authorize the European Commission to develop common specifications if needed. Proposed required documentation or disclosures would include information about the data sources used for training, system architecture and general logic, classification choices, the relevance of different parameters, validation and testing procedures, and performance capabilities and limitations.
Watermarking AI-generated content will not be easy. There is the difficulty of correlling open-source models used for image and text generation. Reaching consensus standards for consumer-facing applications may be challenging. And there is the technical challenge of preventing the removal of watermarks.

Provenance refers to the origin of data or AI system outputs. For training data, relevant provenance questions might be: Where does the content come from? Is it protected by copyright, trademark, or other intellectual property rights? Is it from an unreliable or biased dataset? For system outputs, provenance questions might be: What system generated this output? Was this information altered by AI or other digital tools? Authentication is a method of establishing provenance via verifiable assertions about the origins of the content. For example, C2PA is a membership organization (including Adobe and Microsoft as members) developing an open metadata standard for images and videos that allows cryptographic verification of assertions about the history of a piece of content, including about the people, devices, and/or software tools involved in its creation and editing. Content authors, publishers (e.g., news organizations), and even device manufacturers can opt-in to attach digital signatures to a piece of digital content attesting to its origins. These signatures are designed to be tamper-proof: if the attestations or the underlying content are altered without access to a cryptographic signing credential held by the content author or publisher, they will no longer match. Authentication-based provenance metadata could be produced for AI-generated content, either as part of the media files or in a standalone ledger. Because digital signatures cannot change the underlying content, the content can still be reproduced without the signatures. Provenance tracking has relevance for content not generated by AI as well. If provenance data become prevalent, user perceptions and expectations may change. The absence of such data from a given piece of content could trigger suspicion that the content is AI-originated.

Watermarking is a method for establishing provenance through "the act of embedding information, which is typically difficult to remove, into outputs created by AI—including into outputs such as photos, videos, audio clips, or text—for the purposes of verifying the authenticity of the output or the identity or characteristics of its provenance, modifi-
cations, or conveyance.137 These techniques change the generated text, image, or video in a way that is ideally not easily removable and that may be imperceptible to humans, but that enables software to recognize the content as AI-generated and potentially to identify the AI system that produced it.138 Google DeepMind, for example, has launched (in beta) its SynTh1 tool for AI-generated images, which subtly modifies the pixels of an image to embed an invisible watermark that persists even after the application of image filters and lossy compression.139 Watermarking approaches are more mature for video and photos than for text, although some have proposed that text generation models could watermark their outputs by “softly promoting” the use of certain words or snippets of text over others.140 Because watermarking embeds provenance information directly into the content, the provenance data follows the content as it is reproduced. However, watermark detection tools, especially for text, may be able to provide only a statistical confidence score, not a definitive attribution, for the content’s origins.

• Content labeling refers to informing people as part of the user interface about the source of the information they are receiving. Platforms that host content, linear broadcasters or cable channels that transmit it, and generative AI systems that output information are examples of entities that could provide content labeling. Content labeling presumes that the provenance of the content can be established—e.g., via users marking AI-generated content they submit as such, via authentication metadata attached to the content files, or via watermarks indicating AI origins. Different types of information about AI system outputs can serve complementary roles in establishing and communicating provenance. Suppose a user who sees a video when scrolling through a social media site wants to know whether the video is authentic (for example, that it was issued by a specific media organization) or whether it is AI-generated content. Content labeling is one way in which the social media site can deploy tools to serve both interests—perhaps by presenting distinctive visual banners for content accompanied by origin metadata or an identifiable embedded watermark. For a user to reap the full benefits of watermarking methods, the watermark must be resistant to removal along the way from production to distribution. That technical challenge is matched by a logistical one: the machines embedding the watermark and those decoding it must decide to agree on implementation. A system for providing or authenticating information between machines requires shared technical protocols for those machines to follow as they produce and read the information. Therefore, applications (e.g., browsers, social media platforms) will have to recognize and implement protocols that are widely adopted.141 Similarly, for users to benefit from cryptographically signed metadata-based authentication technology, an authentication standard must be widely adopted among content producers as well as consumer-facing applications distributing content.

All these steps present challenges. First, ensuring that AI models include watermarking on AI-generated content, for example, will not be easy, especially given the difficulty of corralling open-source models used for both image and text generation. Second, there is the task of reaching consensus on the standard proper for use by consumer-facing applications. And third, preventing the removal of the watermark (i.e., an adversarial attack) between generation and presentation to the consumer will pose technical challenges. Current forms of watermarking involve keeping the “exact nature” of a watermark “secret from users,”142 or at least sharing some information between the systems generating and checking for the watermark that is unknown to those seeking to remove it. Such secrecy may be impossible, especially if open-source systems are to be able to embed watermarks and open-source applications are to be able to recognize them. Interpretable approaches abound as well: that a piece of content has been authenticated does not mean it is “true” or factually accurate, and the absence of authentication or provenance information does not necessarily support conclusions about content characteristics or origin.

One of the voluntary commitments some AI companies have made is to work on information authentication and provenance tracking technologies, including related transparency measures.143 This is important for many reasons that go beyond AI accountability, including the protection of democratic processes, reputations, dignity, and autonomy. For AI accountability, provenance and authentication help users recognize AI outputs, identify human sources, report incidents of harm, and ultimately hold AI developers, deployers, and users responsible for information integrity. Policy interventions to help coordinate networked market adoption of technical standards are nothing new. The government has done that in areas diverse as smart chip bank cards, electronic medical records, and the V-Chip television labeling protocol. The AI EO takes a first step in fostering provenance practices by directing agency action to “foster capabilities...to establish the authenticity and provenance of digital content, both synthetic and not synthetic...”144 Two additional applications of transparency around AI use take the form of adverse incident databases and public use registries. The OECD is working on a database for reporting and sharing adverse AI incidents, which includes harms “like bias and discrimination, the polarisation of opinions, privacy infringements, and security and safety issues.”145 The benefit of such a database, as one commenter put it, is to “allow government, civil society, and industry to track certain kinds of harms and risks.”146 Adequately populating the database could require either incentives or mandates to get AI system developers to contribute to it. Beyond that, individuals and communities would need the practical capacities to easily report incidents and make actionable the reports of others. Any such database should include incidents, and not only actual harms, because “safe” means more than the absence of accidents.

There are now many jurisdictions requiring or proposing that at least public entities publish their use of higher risk AI applications,147 as

137 AI EO at Sec. 3(cg).
141 See cf. supra at note 138.
142 Cf. supra note 138.
144 AI EO at Sec. 4.5.
146 See, e.g., Marion Dowd, Luke Chambers, Ellen P. Goodman, Pam Ugwudike, and Niki Zara, The UK Algorithmic Transparency Standard & A Qualitative Analysis of Policy Perspectives (July 7, 2022), https://www.ssrn.com/abstract=4235126 (noting that “several jurisdictions have mandated levels of algorithmic transparency for government bodies” and citing several examples); Government of Canada, Direction on Automated Decision-Making (April 2020), https://canada.ca/en/avi-mcsc-fois/docs/2020/avm-mvc-direction-automated-decision-making.html (noting the government’s “aim to indicate to all decision makers that a decision will be made in an informed context...”); State of California, Assembly Bill AB-302, “An act to add Section 1320.35 to the Penal Code, relating to automated decision systems” (California Legislature, 2023-2024 Regular Session), https://leginfo.legislature.ca.gov/faces/billTextShowPage.xhtml?billId=202304000AB302&src=node&session=2023-01-01&title=AB302&textType=full&viewType=plain (noting that “the bill requires the Attorney General to develop, adopt, and promulgate rules relating to the development of an algorithmic transparency and accountability strategy”); California Penal Code § 1320.35 (California law requiring pretrial services agencies to adopt an algorithmic transparency and accountability strategy); Texas law requiring an inventory “of all automated decision systems that are being developed, adopted, or procured” by state and local legislative agencies; State of Texas, An Act relating to the evaluation of artificial intelligence algorithmic decision systems that have been proposed for use, development, or procurement, or are being used, developed or procured by any state agency” (Texas law requiring an inventory “of all automated decision systems that are being developed, adopted, or procured” by state and local legislative agencies); State of California, Penal Code § 1320.35 (California law requiring pretrial services agencies to adopt an algorithmic transparency and accountability strategy); State of Texas, An Act relating to the evaluation of artificial intelligence algorithmic decision systems that have been proposed for use, development, or procurement, or are being used, developed or procured by any state agency” (Texas law requiring an inventory “of all automated decision systems that are being developed, adopted, or procured” by state and local legislative agencies); State of California, Penal Code § 1320.35 (California law requiring pretrial services agencies to adopt an algorithmic transparency and accountability strategy).
147 See supra note 138.

Researchers, auditors, real teams, and other affected parties...
the federal government has begun doing online by publishing federal agency AI use cases at AI.gov [both high-risk and not high-risk applications]." The Office of Management and Budget (OMB) has released draft guidance for federal agencies which would require them to publicly identify the safety-affected and rights-affected AI systems they use. As one commenter noted, a national registry for high-risk AI systems could provide nontechnical users with an overview of the system as deployed and the actions taken to ensure the system does not violate people's rights or safety. Along with a registry of systems, a government-maintained registry of professional AI "audit reports that are publicly accessible, upon request" would foster additional accountability. Any such registry would have to reflect the proper balance between transparency and the potential dangers of exposing AI system vulnerabilities to malign actors. 3.1.3. AI SYSTEM ACCESS FOR RESEARCHERS Researchers, auditors, red-teams, and other affected parties such as workers and unions all need appropriate access to AI systems to evaluate them. While researchers can conduct "adversarial" reviews of public-facing systems without any special access, collaboration between the evaluator and the AI actor will often be required to fully assure that systems are trustworthy. Commenters argued that creators and individuals should be able to request access to AI system datasets to identify and report personal data or copyrighted works. We note that facilitating researcher access to data from very large online platforms and search engines and their associated algorithmic systems is something that the Digital Services Act requires in the European Union. That regulation has deemed researcher access an indispensable part of the platform accountability scheme in certain instances. Third-party access to AI systems for the purpose of evaluations comes with risks that need to be managed. Three principal risks are: Liability risks to researchers for claims of copyright or contract violation or for circumventing terms of service (e.g., by scraping data) and other controls seeking to protect AI system components from view. A number of commenters proposed a safe harbor from intellectual property or other liability for re- search into AI risks. Security risks to AI actors from providing access (willingly or not) to AI system components. Access to outsiders can jeopardize the trade secrets of AI actors as well as con- trolling actors that place in danger misuse of AI systems. Application Programming Interfaces (APIs) can be used to mediate between access to AI researchers and actors, there- by reducing these risks. Privacy risks to the subjects of sensitive data that may be revealed when data is accessed for evaluation. For example, evaluation of an AI system for outputting discriminative results might require access to personal data about loan applicants. Researchers usually have processes in place to minimize these risks, such as by limiting data collection, obfuscating sensitive data before storing it, and complying with institutional review board requirements. Using existing, and developing new, privacy enhancing technologies can also mitigate these risks. The security and privacy risks un- derscore the need to vet researchers before permitting access to certain AI system components, monitor and limit access, and define other controls on when, why, and how sensitive information is shared. 3.1.4. AI SYSTEM DOCUMENTATION Documentation is a critical input to transparency and evaluation, whether internal or external, voluntary or required. Many commenters thought that AI developers should (and pos- sibly should be required to) maintain documentation concerning model design choices, design of system controls, training data composition and pre-training, data the sys- tem uses in its operational state, and testing results and rec- ommendations for different environments. This docu- mentation, which may be subject to intellectual property protections, informs consideration of appropriate deploy- ment contexts. It helps answer questions about whose interests were considered in AI system development and

—

3.1.3. AI SYSTEM ACCESS FOR RESEARCHERS AND OTHER THIRD PARTIES

Researchers, auditors, red-teams, and other affected parties such as workers and unions all need appropriate access to AI systems to evaluate them. While researchers can conduct "adversarial" reviews of public-facing systems without any special access, collaboration between the evaluator and the AI actor will often be required to fully assure that systems are trustworthy. Commenters argued that creators and individuals should be able to request access to AI system datasets to identify and report personal data or copyrighted works. We note that facilitating researcher access to data from very large online platforms and search engines and their associated algorithmic systems is something that the Digital Services Act requires in the European Union. That regulation has deemed researcher access an indispensable part of the platform accountability scheme in certain instances. Third-party access to AI systems for the purpose of evaluations comes with risks that need to be managed. Three principal risks are:

Liability risks to researchers for claims of copyright or contract violation or for circumventing terms of service (e.g., by scraping data) and other controls seeking to protect AI system components from view. A number of commenters proposed a safe harbor from intellectual property or other liability for research into AI risks.

Security risks to AI actors from providing access (willingly or not) to AI system components. Access to outsiders can jeopardize the trade secrets of AI actors as well as controlling actors that place in danger misuse of AI systems. Application Programming Interfaces (APIs) can be used to mediate between access to AI researchers and actors, thereby reducing these risks.

Privacy risks to the subjects of sensitive data that may be revealed when data is accessed for evaluation. For example, evaluation of an AI system for outputting discriminative results might require access to personal data about loan applicants. Researchers usually have processes in place to minimize these risks, such as by limiting data collection, obfuscating sensitive data before storing it, and complying with institutional review board requirements. Using existing, and developing new, privacy enhancing technologies can also mitigate these risks.

The security and privacy risks underscore the need to vet researchers before permitting access to certain AI system components, monitor and limit access, and define other controls on when, why, and how sensitive information is shared.

3.1.4. AI SYSTEM DOCUMENTATION

Documentation is a critical input to transparency and evaluation, whether internal or external, voluntary or required. Many commenters thought that AI developers should (and possibly should be required to) maintain documentation concerning model design choices, design of system controls, training data composition and pre-training, data the system uses in its operational state, and testing results and recommendations for different environments. This documentation, which may be subject to intellectual property protections, informs consideration of appropriate deployment contexts. It helps answer questions about whose interests were considered in AI system development and
how AI actors balanced various trustworthy AI attributes. Documentation is also important for AI actors themselves and remedy bias.162

Entities should develop plans for continuous or routine evaluation of the AI system and document results and corrective actions taken to ensure the system produces expected results.

Assessing Sustainment and Expanded Use

Entities should assess the utility of the AI system to ensure its relevance and identify conditions under which the AI system may not currently be used or required beyond its current use.

Source: Data: GAO | GAO-21-519SP

165 Stanford Institute for Human-Centered AI Center for Research on Foundation Models

166 STM Comment at 2 (“[W]hen applying AI in the context of scholarly communications, records concerning both the training and the deployment, and an additional number of years after.171


168 See, e.g., Cordell Institute for Policy in Medicine (“[t]ransparency can privilege code of conduct, industry-driven standards, and individual empowerment should be preferred over government regulation in emerging technology.”)

169 Id.

170 Compare Certification Working Group Comment at 24 (“[t]ransparency and accountability without a backdrop of strong, enforceable consumer and civil rights protections”; “[t]ransparency cannot be enforced; it [is] transparency’s job to create meaningful trust in accountability without a backdrop of strong, enforceable consumer and civil rights protections”;

171 See, e.g., General Intelligence at 4 (“[t]ransparency requirements of disclosure. An AI system may be a “decoy”; Cordell Institute for Policy in Medicine (“[t]ransparency can privilege code of conduct, industry-driven standards, and individual empowerment should be preferred over government regulation in emerging technology.”)

172 Transparency and disclosures regarding AI systems are primarily valuable insofar as they feed into accountability.172 One essential tool for converting information into accountability is critical evaluation of the AI system. The National Artificial Intelligence Advisory Committee (NAIAC), in its 2023 report, observed that “practices, standards, and frameworks for designing, developing, and deploying trustworthy AI are created in organizations in a relatively ad hoc way depending on the organization, sector, risk level, and even country.”

We agree with its accompanying observation that it is problematic that “[r]egulations and standards are being proposed that require some form of audit or compliance, but without clear guidance accompanying them.”

The RFC described different types of evaluation, including audits, impact and risk assessments, and pre-release certifications. Commenters were divided on whether independent audits are possible now, before there are agreed upon criteria for all aspects. They also questioned whether audits should be mandated. Some comments reflected a sense of frustration with decades of self-regulation of technology that has failed to meet societal ex-
3.2.1. PURPOSE OF EVALUATIONS

AI system evaluations are useful to:

- Improve internal processes and governance;
- Provide assurance to external stakeholders that AI systems and applications are trustworthy;
- and Validate claims of trustworthiness.

One purpose of an evaluation is claim validation. The goal of such an inquiry is to verify or validate claims made about the AI system, answering the question: Is the system performing as claimed? The challenge of an evaluation like this is that it is more amenable to binary findings, and there are often clear enforcement mechanisms and remedies to combat false claims in the commercial context under federal and state consumer protection laws.

Another type of evaluation examines the AI system according to a set of criteria independent of an AI actor’s claims. Such an evaluation might have a narrow aperture, focusing on the critical determination of how accurately a system performs its task or whether it produces unlawful discriminatory outputs, for example. Or it might broaden, focusing on governance and system dataset specifications, but only for a small subset of objectives, such as protecting intellectual property. In theory, an comprehensive, flexible, and transparent methods. We recommend below that government act to support a vigorous ecosystem of list compliance, industry capture, and audit-washing. See also OMB Draft Memo at 24-25. 183

• Provide assurance to external stakeholders that AI systems and applications are trustworthy; and
• Validate claims of trustworthiness. 184

One purpose of an evaluation is claim validation. The goal of such an inquiry is to verify or validate claims made about the AI system, answering the question: Is the system performing as claimed? The challenge of an evaluation like this is that it is more amenable to binary findings, and there are often clear enforcement mechanisms and remedies to combat false claims in the commercial context under federal and state consumer protection laws.

Another type of evaluation examines the AI system according to a set of criteria independent of an AI actor’s claims. Such an evaluation might have a narrow aperture, focusing on the critical determination of how accurately a system performs its task or whether it produces unlawful discriminatory outputs, for example. Or it might broaden, focusing on governance and system dataset specifications, but only for a small subset of objectives, such as protecting intellectual property. In theory, an comprehensive, flexible, and transparent methods. We recommend below that government act to support a vigorous ecosystem of list compliance, industry capture, and audit-washing. See also OMB Draft Memo at 24-25. 183
such evaluations on an ongoing basis throughout the AI system lifecycle, including the design, development, and deployment stages.196 As entities develop AI systems or system components, and as entities then produce AI system outputs, every node in that chain should bear responsibility for assuring its part in relation to trustworthy AI. This is ideally how it works in the financial value chain, with organizations (e.g., payroll processors or securities market valuators) relying on, and in turn providing, audited financial statements and reports describing processes and controls. As one commenter stated, these communications “explicitly acknowledge the interrelationship between the controls of the service organization and the end user.”197 It is generally desirable for independent evaluations to use replicable methods,198 and to present the results in standardized formats so as to be easily consumed and acted upon.202 Given how vastly different deployments can be—for example, automated vehicles versus test scoring—some aspects of AI evaluations will have to be conducted differently on the scale.204 Evaluations of foundation models, where use cases might be more predictable, have their own challenges. Moreover, trade secret protection for information that is evaluated may make replicability difficult.

Standardization efforts that are well funded and coordinated across sectors could achieve a baseline of common-denominator elements, supplemented by modules adapted for the application domain or for foundation models.

3.2.2. ROLE OF STANDARDS

It was an uncontroversial point in the comments that international technical standards are vitally important,221 and may be necessary for defining the methodology for certain kinds of audits.254 Developing technical standards for emerging technologies is a core Administration objective.195 The current dearth of consensus technical standards for use in AI system evaluations is a barrier to assurance practices. This barrier may be especially pronounced for evaluation of foundation models.264 Compounding the challenge of standards development is the reality that AI is being developed, deployed, and advanced across many different sectors, each with its own applications, risks, and terminology, and that the AI community has yet to coalesce on fundamental questions surrounding terminology.250 Under-developed standards mean uncertainty for companies seeking compliance, diminished usefulness of audits, and reduced assurance for customers, government, and the public.251 Among the issues for which commenters wanted standards and benchmarks for both internal and external evaluation and other assurance practices were:• AI risk hierarchies, acceptable risks, and tradeoffs;• Performance of AI models, including for fairness, accuracy, robustness, reproducibility, and explainability;• Data quality, provenance, and governance;• Internal governance controls, including team compositions and reporting structures;• Stakeholder participation;• Security;• Internal documentation and external transparency; and• Testing, monitoring, and risk management.

Here, we stress the need for accelerated international standards work and provide further justification for expanding participation in technical standards and standards-setting processes. The comments yielded important caveats about conventional technical standards: the relative immaturity of the AI standards ecosystem, its reliance on normativity, and the dominance of industry in relation to other stakeholders. Addressing these critiques will improve AI accountability.

Standards-setting organizations publish requirements and guidelines (alongside other types of documents not pertinent here). Requirements contain “shall” and “shall not” statements, while guidelines tend to contain “should,” “should not,” or “may” statements.258 Leading commentary on standards for AI audits is supportive of guidelines that can be more flexible than requirements and standard: that focus on processes as well as outputs.259 Nevertheless, it is important to recognize that guidelines do not constitute compliance regimes. Technical standards-setting organizations hesitate—and may not be equipped—to settle policy and values debates on their own.260 Non-prescriptive standards—for instance, providing ways to measure risk, without identifying a threshold beyond which risk is unacceptable—help with future-proofing. However, such flexibility means that the governments, public, and downstream users of the technology cannot assume that compliance with such standards means that risks have been acceptably managed. Separate legal or regulatory requirements are required to set norms and compel adherence.261 We are cognizant of the critique that non-prescriptive stances have sometimes impeded efforts to ensure that

that a rigorous, independent audit be appropriate.” See, e.g., supra Sec. 3.1

196 See, e.g., Credo AI Comment at 4 (stressing the need to evaluate frequently); The Future Society Comment at 4; Global Partners Digital Comment at 4.

197 PINC Comment at A7; See Palantir Comment at 20 (stressing process measures in the AI system development phase, including data collection practices, “access controls, logging, and monitoring for abuse”).

198 See, e.g., Hsu Comment at 8 (“We believe that a consistent report format is important as it allows users of the report to compare across different assurance engagements. Further, the Independent Accountants’ Report provides critical information to users, including the criteria, level of assurance, responsibilities of the auditor and entity management, and any limitations, among other information.”).

199 See, e.g., MITRE Comment at 3 (a “sector regulator” is “to adopt and adapt accountability mechanisms tailored to specific industry issues”); Consumer Reports Comment at 126 (“The type of audit that can be executed and the extent to which a researcher’s ability to assess a model is highly dependent on the information they have access to.”).

200 See DOD Comment at 26 (“Such standards will often embody policy and value judgments; standards for an audit or evaluation whether a system is biased, for example, may have to strike how much variation in performance, if any, is permissible across race, gender, or other lines in order to still be considered unbiased.”).

201 See NIST AI RMF at 7 (recognizing the need for guidance on risk tolerances from “legal, regulatory requirements”). See also The Center for AI and Digital City Comment at 4 (“Conditional assurance of AI systems could be through certification programs set by Federal AI legislation based on...”.)


203 See, e.g., Raj et al., Against the Drones, supra note 18 (at recommending “standards as guidelines, not deployment checklists” and “standards for processes, not only for outcomes”).

204 See, e.g., supra Sec. 3.2.2. ROLE OF STANDARDS

205 Need for Holistic Evaluation, Center for Research on Foundation Models, Stanford HAI (2021), https://crfm.stanford.edu/2022/11/17/helm.html (recommending investment in developing metrics to quantify and evaluate bias in AI systems and metrics to measure foundation model performance, NIST Comment at 12 (used investment in international AI standards to underpin an assurance ecosystem).
The inclusion of experts and stakeholders in standards development is particularly important given the centrality of normative concepts such as freedom from harmful discrimination and disinformation in standards work. Traditional, formal standards-setting processes may not yield standards for AI assurance practices sufficiently rapidly, transparently, inclusively, and comprehensively on their own, and may lag behind technical developments. Several commenters recommended that government develop a taxonomy or hierarchy of AI risks to shape how AI actors prioritize risk. Others requested government help in devising assurance methodologies that take equity and public participation seriously. We note that NIST is already leading and encouraging community leaders to develop a series of AI RMF “profiles” that will provide more detailed guidance to the application of the NIST AI RMF in different domains. For example, the Department of Labor’s Office of Disability Employment Policy (ODEP) is working with key partners to create a Profile for Inclusive Hiring. This policy framework aims to guide employers to practice disability inclusion and accessibility when they decide to use AI in talent acquisition processes. Looking ahead, there is a question about how standards will evolve globally to keep pace with technological development and societal needs. There are several key issues that will help inform this issue:

- Whether current standards continue to develop alongside AI implementations at an appropriate pace and with appropriate scope;

- Whether competing standards emerge inadvertently, creating perverse incentives for stakeholders and opportunities for arbitrage; and

- Whether future industry standards foster a sufficiently large marketplace of certification, auditing, and compliance entities to support appropriate levels of compliance.

Commentators have suggested governmental actions to support the development and adoption of AI standards, including, as one commenter expressed, by supporting research on data quality benchmarks and data commons for AI companies. For at least some AI technologies, government has already played a significant role in the actual testing of systems and the publication of results. Since 2002, for instance, NIST’s Facial Recognition Vendor Tests have assessed the accuracy of privately developed facial recognition technology. This research has not only demonstrated the overall degree of accuracy of the tested algorithms, but has also identified common challenges across algorithms such as accuracy differentials based on race or gender. Generally, government can foster the utility of standards for accountability purposes by (a) encouraging and fostering participation by diverse stakeholders, including civil society, non-industry participants, and those involuntarily affected by AI systems (b) helping improve and expand access to standards publications by those traditionally under-represented parties; (c) supporting methods to align industry standards with societal values; and (d) in appropriate circumstances, developing guidelines or other resources that contribute toward standards development.

We also note that, while international standards development is critical, national standards might also be necessary to protect national security interests.

### 3.2.3. Proof of Claims and Trustworthiness

AI actors are putting AI systems out into the world and should be responsible for proving that those systems perform as claimed and in a trustworthy manner. Accountable Tech, AI Now, and EPIC’s Zero Trust AI Governance Framework puts it this way: “Rather than relying on the good will of companies, tasking under-resourced enforcement agencies or afflicted users with proving and preventing harm, or relying on post-market auditing, companies should have to prove their AI offerings are not harmful.”

This responsibility for assuring the
validity of system claims and trustworthiness should be ongoing throughout the lifecycle of the AI system.210 An independent certification process for some AI systems could be one way to enable the improvement of claims and trustworthiness. According to OpenAI, a certification, a certification, a "process of an independent body stating that a system has successfully met some pre-established criteria."211 Thus an independent evaluation would be a prerequisite for a certification. A voluntary certification system for AI systems, if sufficiently rigorous and independent, could help stakeholders navigate the AI market and promote competition among trustworthy AI.212 Mandatory pre-release certification – taking the form of licensing – is another route. Given the prospect of AI becoming embedded ubiquitously in products and processes, it would be impractical to mandate certification for all AI systems.213 There was strong support from some commentators for governmental licensing of high-risk foundation models, or at least deep review of such models, before deployment (including the need to show that certain “safety” conditions are met) as a way of addressing alleged catastrophic risks.214

However, there is also a concern that mandatory pre-release certification or licensing can hurt competition by advantaging incumbents.215 Therefore, the benefits of requiring ante proof of trustworthiness have to be balanced against facilitating easy entry into the AI market.216

3.2.4. INDEPENDENT EVALUATIONS
Self-assessments (including impact or risk assessments) have a different value proposition than independent evaluations, including audits. Both are important.217 Self-assessments will often be the starting point for the performance of independent evaluations. Many commenters thought that entities developing and deploying AI should conduct self-assessments, ideally working from the NIST AI RMF.218 An entity’s own assessment of the trustworthiness of AI systems (in development or deployment) benefits from its access to relevant data.219 Moreover, internal evaluation practices will

the training") (emphasis supplied), Campaign for AI Safety Comment at 3 (supporting “pre-deployment safety evaluations”)

225. See Engine Comment at 4 (“A mandatory certification licensing system is likely to create a ‘regulatory moat’ bolstering the position and power of large companies that are already in the AI ecosystem, which would have massive impacts on startups to monitor their market share.”); Graboske et al., Comment at 1 (“[Deregulation in AI, licensing, and mandatory AI certification] for all new models or systems that require, at ‘this stage’ of AI development, ‘they will work much harder for new entrants and smaller companies to develop AI systems, while at the same time this lack of clear criteria can be achieved with other policy approaches.’”); See also CLIC Comment at 1 (“the notion of licensing implies that companies would need to obtain permission prior to commencing a particular piece of code. This could introduce unbound levels into the process of opening AI technologies to market, or, indeed, even correct errors in already-developed software.”)

226. See, e.g., Hilltop AI Comment at 4 (“With certifications function a public-facing documentation on, for example, a system’s level of reliability and thus safety, internal assessments help to improve a system’s trustworthiness by directly guiding better decision-making and best practices across the conceptualization, design, development, and monitoring of AI projects.” (citations omitted)).

227. See also BAA Report at 4 (“the context of the audit inquiry is all the more important when benchmarks are varied and not standardized, and when audits are diverse in scope and method.”).

228. See also, e.g., ATSA Comment at 2; ARC Comment at 5 (internal evaluations are more mature and robust than external evaluators. As one commenter posited, “[a] self-assessment will often be the starting point for the performance of independent evaluations. Many commenters thought that entities developing and deploying AI should conduct self-assessments, ideally working from the NIST AI RMF.”)

229. See NTIA Artificial Intelligence Accountability Policy Report (47) Developing regulatory requirements for independent evaluations, where warranted, provides a check on false claims and risky AI, and incentivizes stronger evaluation systems.230

223. See also, e.g., ATSA Comment at 2; ARC Comment at 5 (internal evaluations are more mature and robust than external evaluators. As one commenter posited, “[a] self-assessment will often be the starting point for the performance of independent evaluations. Many commenters thought that entities developing and deploying AI should conduct self-assessments, ideally working from the NIST AI RMF.”)

226. See, e.g., OpenAI Comment at 6 (“the support the development of registries and licensing requirements for future generations of the most high-risk foundation models. As AI developers work to deploy highly capable foundation models which are likely to prove more capable than models previously shown to be so risky.”). See, e.g., (SCA) 2021-22 (CR) (on fees and licensing requirements such as authentication of natural language training runs, comprehensive risk assessments focused on identifying dangerous or deceptive capabilities, extensive pressure testing by internal and external experts, and multiple checkpoints along the way”).

227. See, e.g., BAA Report at 4 (“the context of the audit inquiry is all the more important when benchmarks are varied and not standardized, and when audits are diverse in scope and method.”).

228. See also, e.g., ATSA Comment at 2; ARC Comment at 5 (internal evaluations are more mature and robust than external evaluators. As one commenter posited, “[a] self-assessment will often be the starting point for the performance of independent evaluations. Many commenters thought that entities developing and deploying AI should conduct self-assessments, ideally working from the NIST AI RMF.”)

229. See also, e.g., ATSA Comment at 2; ARC Comment at 5 (internal evaluations are more mature and robust than external evaluators. As one commenter posited, “[a] self-assessment will often be the starting point for the performance of independent evaluations. Many commenters thought that entities developing and deploying AI should conduct self-assessments, ideally working from the NIST AI RMF.”)

226. See, e.g., OpenAI Comment at 6 (“the support the development of registries and licensing requirements for future generations of the most high-risk foundation models. As AI developers work to deploy highly capable foundation models which are likely to prove more capable than models previously shown to be so risky.”). See, e.g., (SCA) 2021-22 (CR) (on fees and licensing requirements such as authentication of natural language training runs, comprehensive risk assessments focused on identifying dangerous or deceptive capabilities, extensive pressure testing by internal and external experts, and multiple checkpoints along the way”).

227. See, e.g., BAA Report at 4 (“the context of the audit inquiry is all the more important when benchmarks are varied and not standardized, and when audits are diverse in scope and method.”).

228. See also, e.g., ATSA Comment at 2; ARC Comment at 5 (internal evaluations are more mature and robust than external evaluators. As one commenter posited, “[a] self-assessment will often be the starting point for the performance of independent evaluations. Many commenters thought that entities developing and deploying AI should conduct self-assessments, ideally working from the NIST AI RMF.”)
3.2.5. REQUIRED EVALUATIONS

Developing regulatory requirements for independent evaluations, where warranted, provides a check on false or otherwise outsourced key aspects of the process in the system’s development or operation. To ensure adequate remediation of any new risks; and To ensure periodic monitoring and review of the system’s operation; To ensure adequate remediation of any new risks; and To ensure that there is internal review by a sufficiently empowered decisionmaker not directly involved in the system’s development or operation.

Communication:
• Was there adequate communication throughout the lifecycle of the AI system and its components to enable an evaluator to answer the previous questions? Does the system adequately train, standardize, and accredit auditors and researchers capable of doing red-teaming and adversarial tactics to stress test AI systems for vulnerabilities and risks – is becoming an important part of the ecosystem.

3.3 ECOSYSTEM REQUIREMENTS

The supply of capable evaluators trails the pace of AI innovation. A paper produced for Google DeepMind, opines: “[J]ealously there would exist a rich ecosystem of model auditors providing broad coverage across different risk areas. (This ecosystem is currently under-developed).”[202] Research drawing on auditing experiences across sectors, including pharmaceuticals and aviation, “strongly supports training, standardization, and accreditation for third-party AI auditors.”[203] Many commenters addressed this point, observing that the ecosystem for AI assurance requires more investment, diverse stakeholder participation, and professionalization.

Another possible drag on red-teaming contributions is if red-teams are required to sign nondisclosure agreements to conduct their probes, thereby limiting what they can share with the public and, ultimately, the ways in which their evaluations can feed into the accountability ecosystem.

The costs of mandatory audits can be managed. Commenters recommended the following cost de-escalators, which are captured in other parts of this Report:

To create a modular governance system for AI, with a risk assessment standards board, to deduplicate costs for developing audit standards;[204]
• Standardize “structured transparency” such that auditors may only ask specific questions rather than obtaining all the underlying data;[205]
• Build on internal accountability requirements;[206] and
• Provide industry association or governmental compliance assistance.[207]

240  Accountable Tech, AI Now, and EPIC, supra note 56, at 4. See also CAP Comment at 9 listing Microsoft, Improving responsible AI practices in a world “where patchwork of industry measures proposed and implemented by the ecosystem are not sufficient at the margin to work effectively,” and a recommendation that ”Governments should fund ‘research and development of structured transparency tools.’”

241  Anthropic Comment at 10; ARC Comment at 6 (“It could be important for legislators, regulators, etc. to require measurement of potential dangerous capabilities before training and/or deployment of models that are much more capable than the current state of the art.”); Shevlane, supra note 228, at 17 (“Industry standards or regulation could require a minimum duration for pre-deployment evaluation of frontier models, including the length of time that external researchers and auditors have had external access.”).

242  See, e.g., SIFMA Comment at 4. See also OpenAI Comment at 2 (recommending government fund “research and development of structured transparency tools”).

243  See, e.g., OpenAI Comment at 2. See also CAP Comment at 9 recommending government fund “research and development of structured transparency tools.”

244  See also Argonne National Laboratory Comment at 19; Brookings Institution Comment at 3; Business Roundtable Comment at 13; and Facebook Comment at 18. See also Facebook Comment at 18 (pointing to “the importance of training on real-world AI ethics and best practices for auditors.”)

245  DEF CON 2023 held a red-teaming exercise with thousands of people, see Hao H The Future. https://www.bloomsburc.org. See also Microsoft Comment at 3 noting that it “is working intensely on red-teaming beyond traditional cybersecurity assessments in the AI ecosystem.”
are embracing red-teaming.212 But as one such company noted, talent is concentrated inside private AI labs, which reduces the capacity for independent evaluation.213 Another possible drag on red-teaming contributions is if red-teams are required to sign nondisclosure agreements to conduct their probes, thereby limiting what they can share with the public and, ultimately, the ways in which their evaluations can feed into the accountability ecosystem. One goal of the White House red-teaming event at Def Con 31 has been to diversify and increase the supply of red-teams.214 Red-teams, like audit teams, should be diverse and multi-disciplinary in their membership and inquiries.215 Techniques to support adversarial testing and evaluation include providing bounties and competitions for the detection of AI system flaws.

3.3.2. DATASETS AND COMPUTE

Insufficient or inadequate datasets can be an obstacle to evaluating AI systems, as well as to training, testing, and refining them to be equitable and otherwise trustworthy. For example, to determine if an AI system is unlawfully discriminatory when deployed in a particular context, it may require consideration of training datasets and/or the availability of new datasets for testing.216 This requires test data that many entities will not have. Commenters noted that limited data or data voids make it difficult to conduct some AI system evaluations.

The need for publicly supplied datasets for AI system evaluation and advancement is well established. The National AI Research Resource (NAIRR) Task Force was a federal advisory committee with equal representation from government, academia, and private organizations, established by the National AI Initiative Act of 2020. In 2023, it released a template for federal infrastructure support for AI research, including “research related to robustness, scalability, reliability, safety, privacy, interpretability, and equity of AI systems.”217 To promote American progress in AI, it recommended that Congress establish a research resource (the NAIRR) that would, among other things, make datasets available for training, evaluation, and support research and education around trustworthy AI. The AI ED directed the Director of the National Science Foundation, in coordination with other federal agencies, to launch a pilot program implementing the NAIRR, consistent with past recommendations of the NAIRR task force.218 This has now launched.219

In its final report, the NAIRR Task Force recommended that the NAIRR should “provide access to a federated mix of computational and data resources, testbeds, software and testing tools, and user support services via an integrated portal.”220 Commenters vigorously endorsed supporting the NAIRR.221 Some focused on the provision of datasets, even if NAIRR was not specifically mentioned. One commenter, for example, opined that government, civil society and industry should collaborate “in building data ecosystems which help generate meaningful datasets in quantity and quality, ensuring and enabling a fair and ethical AI ecosystem that provides appropriate levels of data protection.”222 Others stressed that it would advance AI accountability and competition if the federal government made more datasets available to developers.223 Conducting evaluations of AI systems, just as building and refining them, requires the underlying computing power to analyze enormous datasets and run applications. With computing power, known as “compute,” concentrated in the largest companies and some elite universities, we underscore recommendations about making more compute available to researchers and businesses.

3.3.3. AUDITOR CERTIFICATION

Another part of the AI accountability ecosystem in need of development is certification for AI system auditors,224 which standards organizations are beginning to establish.225 Auditors should be subject to “professional censure, professional and ethical standards, and independent quality control and oversight (e.g. peer review and inspection).”226 ForHumanity, a non-profit public charity which provides AI audit services, recommended that such certifications require auditors to be liable for “false assurance of compliance,” be “qualified to provide expert-level service,” be “held to a standard of professionalism and [c]ode of [e]thics,” and have “robust systems to support integrity and confidentiality of audits.”227


268 See also, e.g., Biden Comment at 1; U.S. Chamber of Commerce Comment at 11; NAIRR Solutions Comment at 1.


270 See also, e.g., AICPA Comment at 2. See also note 279, at art. 37(3)(a)(i).

271 See also, e.g., ForHumanity Comment at 1, at 3 (credibility of the information and report on the results of their procedures); Protofect (PWC Comment at A1 (the communication or report on the results of these engagements, regardless of who performs them, should specify, among other disclosures, the type of assurance provided, the scope of the procedures, and the framework under which it was conducted.

272 See also, e.g., American Institute of CPAs (AICPA) Comment at 1 (recommendations), 2 (questioning whether there is a party to a contract, party organization that can independently, perform audits and give sound judgments multiple contracts — including security, privacy assessment, compliance, health and safety impact).”

273 PWC Comment at A1 (the communication or report on the results of these engagements, regardless of who performs them, should specify, among other disclosures, the type of assurance provided, the scope of the procedures, and the framework under which it was conducted.

274 ForHumanity Comment at 5.

275 Rajal et al., “Outside Oversight, supra note 253 at 156 (‘“From illegal ripoffs or corporate taxation we can awaken the acute crisis, and professional standards can help determine limited conditions for liability.’”).

276 See also, e.g., see notes PNC Comment at A1 (the communication or report on the results of these engagements, regardless of who performs them, should specify, among other disclosures, the type of assurance provided, the scope of the procedures, and the framework under which it was conducted.

277 See also, e.g., see notes PNC Comment at A1 (the communication or report on the results of these engagements, regardless of who performs them, should specify, among other disclosures, the type of assurance provided, the scope of the procedures, and the framework under which it was conducted.

278 The European Commission’s Digital Services Act requires annual independent audits of providers of very large online platforms and very large online search engines; the organizations performing these audits must, among other requirements, be “independent from” and without “any conflicts of interest with the service providers they audit.”279 Auditor independence is partly determined by the type of services auditors may have provided to the auditee in the preceding 12-month period prior to the audit.280 The Sarbanes-Oxley Act of 2002 ("Sarbanes-Oxley") defines independence in the context of annual financial auditing. Some commenters


54 National Artificial Intelligence Research Resource Task Force, supra note 267, at 6.


60 National Artificial Intelligence Research Resource Task Force, supra note 267, at 6.


64 National Artificial Intelligence Research Resource Task Force, supra note 267, at 6.


74 National Artificial Intelligence Research Resource Task Force, supra note 267, at 6.
Auditors should have subject-matter and assurance experience and reflect the diversity of affected stakeholders.

One concern raised in feedback to the European Commission on independent audits in the Digital Services Act is that there is a limited number of entities that have a sufficiently high level of independence and can engage in independent audits with the necessary competencies. The dilemma is that lower standards of assurance and independence might increase auditor supply, but perhaps at the cost of audit effectiveness and, ultimately, public wellbeing. To be sure, the desired end state is an abundant supply of very independent and qualified auditors. Emerging AI auditor certification programs could help.

---

281 See ForHumanity Comment at 5 (referring to Sarbanes-Oxley Act and also recommending that auditors be subject to oversight and a child table for false assurance); Centre for Information Policy Leadership Comment at 18.

282 See, e.g., Data & Society Comment at 3 (“Conflicts of interest for assessors/auditors should be anticipated and mitigated by alternate funding for assurance work.”).

283 See Global Partners Digital Comment at 4 (commenting that audits should be conducted by teams with technical and social science expertise, human rights expertise, subject matter experts, community members, representatives of marginalized groups).

284 See, e.g., Mozilla Foundation, Response to the European Commission’s Call for Feedback on its Draft Delegated Regulation on Independent Audits in the Digital Services Act (June 2023), https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13626-Digital-Services-Act-conducting-independent-audits/F3424065-en, at 2, (“Fostering optimal conditions requires a diversity of audit practitioners and auditing organisations with a high level of independence and the appropriate competencies... There is currently a limited number of entities prepared to conduct these audits given their enormous scope. Many likely auditing organisations have existing industry ties that limit their independence. A larger and more diverse pool of auditors must be fostered.”).

285 See also Responsible Artificial Intelligence Institute, The Responsible AI Certification Program (October 2022), https://20965052.fs1.hubspotusercontent-na1.net/hubfs/20965052/RAII%20Certification%20White%20Paper.pdf; ForHumanity Comment at 3; Holistic AI Comment at 5.
Using Accountability Inputs

While this Report focuses on information flows and evaluation, many commenters expressed interest in clarification of the second part of the AI Accountability Chain—namely, the attribution of responsibility and the determination of consequences. We therefore briefly address how the accountability inputs discussed above could feed into other structures to help hold entities accountable for AI system impacts. Three important structures are liability regimes, regulatory enforcement, and market initiatives. By supporting these structures, AI system information flows and evaluations can help promote proper assessment of legal and regulatory risk, provide public redress, and enable market rewards for trustworthy AI.

4.1 LIABILITY RULES AND STANDARDS

As a threshold matter, we note that a great deal of work is being done to understand how existing laws and legal standards apply to the development, offering for sale, and/or deployment of AI technologies.

Several federal agencies have taken positions within their respective jurisdictions. In a joint statement, for instance, the Federal Trade Commission, the Department of Justice’s Civil Rights Division, the Equal Employment Opportunity Commission, and the Consumer Financial Protection Bureau stated that “[e]xisting legal authorities apply to the use of automated systems and innovative new technologies just as they apply to other practices.”286 For example, the FTC has taken action against companies that have engaged in allegedly deceptive advertising about the capabilities of algorithms.287 In some cases, the FTC has obtained relief including the destruction of algorithms developed using unlawfully obtained data.288 Moreover, the Consumer Financial Protection Bureau has made clear that the requirement to provide explanations for credit denials applies to algorithmic systems.289 The Equal Employment Opportunity Commission has issued technical assistance and provided additional resources intended to educate various stakeholders about compliance with federal civil rights laws when using algorithmic tools for employment-related decisions.290 Other agencies are examining AI-related legal issues, such as the work underway at the Copyright Office and USPTO concerning intellectual property and the Department of Labor concerning labor protections. The courts are also examining a broad range of issues, as are industry and civil society groups.

Nevertheless, the comments evinced a need for more clarity on the precise application of existing laws and the potential contours of new laws in the AI space to benefit everyone along the AI value chain, including consumers, users, researchers, auditors, investors, creators, manufacturers, distributors, developers, and deployers.291 The National Cybersecurity Strategy 2023-2028292 set out a framework for the adversarial AI ecosystem to be worked out in courts, agencies, and legislatures over time.293 It is also the case that the European Commission has proposed adopting a bespoke AI liability regime; if adopted, this regime could have impacts on risk mitigation and allocation outside of Europe as well.

The record and research surface needs for more clarity on AI-related liability, including on the following interrelated ways to promote consistency between Federal and state efforts: • Who should be legally responsible for harms stemming from AI systems and how should such responsibility be shared among key players? What is the place of strict or fault-based liability for harms caused by AI? How should ex ante AI regulation or best practices interact with ex post liability? Should auditors be liable for faulty audits, not only as service providers to clients, but also as public fiduciaries? Should some actors bear a larger share of the responsibility than others based on their relative abilities to identify and mitigate risks flowing from AI models and/or systems?294

AI accountability inputs can assist in the development of liability regimes governing AI by providing people and entities along the value chain with information and knowledge essential to assess legal risk and, as needed, exercise their rights.

- **What is the influence and impact, if any, that external legal regimes—including the European Union’s AI Act and AI Liability Directive—might have on state and federal systems?**
- **How should liability rules avoid stifling bona fide research, accountability efforts, or innovative uses of AI? What safeguards, safe harbors, or liability waivers for entities that undertake research and trustworthy AI practices, including adverse incident disclosure, should be considered?**

206. While accountability inputs can play an important role in the assigning of liability, we note that these inputs do not in themselves supplant appropriate liability rules. See, e.g., supra note 302 and accompanying text (“Due to the lack of transparency in AI uses, the plaintiff may not have the information needed to even establish a prima facie case.”). It can be difficult for those who may not have the information needed to even establish a prima facie case. They may fear of the potential liability if an AI-made decision is contested. See, e.g., supra note 302 (arguing that it is “unsettled how applicable law should be applied” in the context of AI ethical compliance).

207. See, e.g., supra note 3. For example, a comment on the ECJ’s judgment in the Aune case concerning the right to access the EU Parliament database showed an overwhelming support that the EU’s Parliament database should be open to the public.

208. AI accountability inputs can assist in the development of liability regimes governing AI by providing people and entities along the value chain with information and knowledge essential to assess legal risk and, as needed, exercise their rights. It can be difficult for those who have suffered AI-mediated employment discrimination, financial discrimination, or other AI system-related harms to bring a legal claim because, for example, AI technology may not have the information needed to even establish a prima facie case. They may fear of the potential liability if an AI-made decision is contested. Such liability both provides for redress for harms suffered by individuals and ensures that individuals subjected to automated decision-making, enabling them to access and understand the output decision and its underlying elements, and thus providing pathways for those who wish to challenge and request a review of the decision.” (emphasis added). CDT Comment at 22 ( supra note 307) (publicly released audit provides a measure of transparency; AI transparency provides information necessary to determine whether liability should be imposed).
4.2. REGULATORY ENFORCEMENT

Regulators are increasingly facing complex technical systems with varying degrees of autonomy whose “conduct” may be difficult to parse and predict. AI systems will often be integrated into a wide range of other technologies across critical infrastructure sectors, some of which (e.g., transportation safety) have well-developed regulatory regimes. Experts observe that regulatory tools and capacities have not kept pace with AI developments. Commenters discussed how regulation does or should intersect with AI systems, including the need for clarity and new regulatory tools or enforcement bodies. Opacity can make it difficult for regulators to enforce legal requirements for trustworthy AI, and several federal regulatory authorities have recently pointed to the “black box” nature of some automated systems as a problem in determining whether automated systems are fair and legally compliant.

4.3. MARKET DEVELOPMENT

A market for trustworthy AI could gain traction if government and/or nongovernmental entities were able to grade or otherwise certify AI systems for trustworthy attributes. Evidence from other private-public certification projects suggests that transparency and clear evaluation metrics are key to trust and adoption. To the extent possible, certification could be based on existing metrics, frameworks, and standards developed by NIST and national or international organizations.

For instance, under the ENERGY STAR program, which is administered by the United States Environmental Protection Agency (EPA) and Department of Energy (DOE), companies may voluntarily seek certification to display the ENERGY STAR label on those products that meet strict performance requirements for energy efficiency.

This labeling provides a way for “consumers and businesses who want to save energy and money” to do so by choosing products with the ENERGY STAR label, thereby relying on a recognizable and trustworthy information mechanism. To date, ENERGY STAR has achieved widespread adoption, leading to substantial energy and consumer savings. Likewise, the Leadership in Energy and Environmental Design (LEED) program, led by the non-profit U.S. Green Building Council (USGBC), allows green building projects to earn a certification (golden, silver, or certified) based on adherence to certain environmental metrics. Per USGBC, LEED projects have been adopted worldwide. Programs like ENERGY STAR and LEED empower their users (e.g., individual businesses) to make informed choices, guide regulators and lawmakers, and more generally help build community trust. Certification could even provide the basis for liability safe harbors, should those be created by legislation, to encourage participation in the certification process, in appropriate cases.

Opacity can make it difficult for regulators to enforce legal requirements for trustworthy AI, and several federal regulatory authorities have recently pointed to the “black box” nature of some automated systems as a problem in determining whether automated systems are fair and legally compliant.

Accountability inputs help shine a light on practices that should be subject to regulatory oversight and equip regulators with the information and knowledge they need to apply their respective bodies of law. As with clarity on liability, clarity about regulatory enforcement can benefit parties along the value chain, including by helping everyone understand what is required for compliance and the broader achievement of trust.

This labeling provides a way for “consumers and businesses who want to save energy and money” to do so by choosing products with the ENERGY STAR label, thereby relying on a recognizable and trustworthy information mechanism. To date, ENERGY STAR has achieved widespread adoption, leading to substantial energy and consumer savings. Likewise, the Leadership in Energy and Environmental Design (LEED) program, led by the non-profit U.S. Green Building Council (USGBC), allows green building projects to earn a certification (golden, silver, or certified) based on adherence to certain environmental metrics. Per USGBC, LEED projects have been adopted worldwide. Programs like ENERGY STAR and LEED empower their users (e.g., individual businesses) to make informed choices, guide regulators and lawmakers, and more generally help build community trust. Certification could even provide the basis for liability safe harbors, should those be created by legislation, to encourage participation in the certification process, in appropriate cases.

See also ENERGY STAR, About ENERGY STAR, https://www.energystar.gov/about. See also ENERGY STAR, ENERGY STAR Improving America. See also https://oecd.ai/en/wonk/.

Twenty-three Attorneys General Comment at 3-4 (referencing Energy Star and LEED programs).

See also, e.g., ENERGY STAR, About ENERGY STAR, https://www.energystar.gov/about. See also Twenty-three Attorneys General Comment at 3-4 (as an example of a public partner program, the LEED standard has opened the move towards “green buildings”).

See also ENERGY STAR, About ENERGY STAR, https://www.energystar.gov/about. See also Twenty-three Attorneys General Comment at 3-4 (This blue ENERGY STAR label provides example, credible, and attested information that businesses rely on to make well-informed decisions.) [emphasis added].

See also, e.g., The Policymaking Project at New York University School of Law at Comment 2 ("Before LEED, there was no mechanism to incentivize this type of information-surfacing about buildings"; environmental impact. Thanks to the information surfaced by LEED certifications, building owners now have an objective standard against which they can judge the development of buildings, their energy and water use, and their impact on the environment). See also Twenty-three Attorneys General Comment at 3-4 (referring to ENERGY STAR and LEED in the context of “legislative and public civic initiatives that build trust and spur the development of technologies”).
5. Learning From Other Models

A market for trustworthy AI could gain traction if government and/or nongovernmental entities were able to grade or otherwise certify AI systems for trustworthy attributes.

Such a process for AI systems could contribute to a functioning market for trustworthy AI. While issues remain about whether such certification programs should be led by government or non-governmental entities (or both), certification programs could enlarge the marketplace for trustworthy AI by bridging information and knowledge gaps. However, a major challenge to establish certifications, as one commenter observed, is the difficulty in gaining sufficient legitimacy and credibility.

BBB National Programs, which itself administers industry certifications, notes that effective certification mechanisms have consistent and verifiable standards and transparency markers (e.g., “trust marks, annual reports, or consumer complaint processes”), among other characteristics. We agree with the comment from twenty-three attorneys general that transparency around the evaluation process is critical and certification programs should operate “through transparent and verifiable policies and practices driven by appropriate standards including a code of ethics.” Establishing and promoting certification systems can further the development of a trustworthy marketplace for AI. More abundant and reliable information of the type discussed in Section 3 above can make it easier to generate public trust in AI, AI evaluations, and AI certifications.

---

321 Friedman et al., supra note 73, at 746. In particular, in the context of private certification programs for technology used by police, the Policing Project’s study found that “institutional trust in policing agencies and Big Tech is low, especially from communities most impacted by policing tech, such as Black communities.” Id. at 746. Here, Policing Project’s law review article advocates that transparency in certification schemes themselves is crucial to building trust. Id. at 746-47.

322 See BBB National Programs Comment at 3. In addition to “consistent standards” (which includes verifiability and “transparency,” BBB National Programs highlights additional characteristics it believes are key for an “effective and accountable independent certification mechanisms” to demonstrate: “defined areas of responsibility[,]” “oversight and independent review[,]” “regulatory recognition[,]” and “independence of accountability.” Id.

323 See, e.g., Johnson & Johnson Comment at 4 (“Developing a framework to enhance the explicability of AI systems that support decision-making on socially significant issues, such as healthcare, is a component of building public trust. Central to a supportable framework is the ability for individuals to obtain a factually correct, and generally clear explanation of the decision-making process.”); AI Policy and Governance Working Group Comment at 2 (“Moving quickly to address risks concerning AI systems and tools will not only provide accountability, it will promote the trust of the American public.”); AI Impacts Comment at 2 (“The biggest deficits of soft law programs... relate to their effectiveness and credibility. These provisions are often phrased in broad and general terms, making compliance difficult to objectively determine, especially without any type of reporting or monitoring requirement.”).
Learning From Other Models

In the financial sector, a standard setting body develops an AI accountability framework. "327 Controls and metrics for reporting financial information, and they may still furnish useful analogies. In other words, as identified professional evaluates that business against those claims about each. 329 The assurance system for financial accounting is an ob-

accountability system

"330 See, e.g., FERC Comment at 1 ("Notably, however, the ecosystem around financial reporting is a third of the stock market’s need for it. Local governments have their own accounting and auditing mechanisms. ") (citing Section 108(c) of the Sarbanes-Oxley Act, which states, "Nothing in this Act, including this section, shall be construed to impair or limit the authority of the Commissions to establish accounting principles or standards for purposes of enforcement of the securities laws.")), and associated persons of such firms. ") (citing Section 108(c) of the Sarbanes-Oxley Act, which states, "Nothing in this Act, including this section, shall be construed to impair or limit the authority of the Commissions to establish accounting principles or standards for purposes of enforcement of the securities laws."))

The modern legal and regulatory regime governing the financial services sector—including for reporting and disclosure obligations—is partly a response to major, global financial crises that disrupted the economic order and led to calls for increased oversight.326 At the federal level, financial sector risks have focused the attention of law-makers seeking to protect investors and promote a trust-worthy marketplace.327 Congress has passed a variety of laws since the 1930’s, including the Securities Exchange Act of 1934 and Sarbanes-Oxley, which aim to foster accountability in the financial sector.328 A detailed analysis of these legal regimes is out of scope of this Report, but the general structure around financial accounting/reporting and related auditing standards—particularly for public companies subject to securities laws—is an area worth exploring to further AI accountability.329

Financial accounting and auditing standards for public companies are established through a public-private collaborative process, subject to key federal government oversight and federal participation in the process. For accounting standards, the Securities and Exchange Commission (SEC) has the authority to recognize “generally accepted” accounting principles developed by a standards-setting body. By law, this recognition must be based on the SEC’s determination that the standards meet certain criteria, including “the need to keep standards current in order to reflect changes in the business environment” and can help the SEC fulfill the agency’s mission because, “at a minimum, the standard setting body is capable of improving the accuracy and effectiveness of financial reporting and the protection of investors under the securities laws.”330 Today, the SEC recognizes the independent non-profit Financial Accounting Standards Board (FASB) as the designated private-sector standards setter, and considers its set standards as “generally accepted” under Sarbanes-Oxley.331 The SEC has made clear that there is federal oversight of this structure and that the SEC continues to have an important role in the standards’ recognition.332

For auditing standards, Sarbanes-Oxley created the Public Company Accounting Oversight Board (PCAOB), a non-profit corporation that is subject to SEC oversight.333 Oversight includes the SEC’s “approval of the Board’s rules, standards, and budget.”334 PCAOB itself is tasked with “oversee[ing] the audit of companies subject to securities laws.”335 Among its duties, PCAOB must, based on certain SEC actions, “register public accounting firms that prepare audit reports,” “establish or adopt . . . auditing . . . and other standards relating to the preparation of audit reports,” “conduct inspections of registered public accounting firms,” “conduct investigations and disciplinary pro-

326 326 Federal Reserve System, Federal Reserve Board, Federal Reserve Bank of New York, Accounting and Auditing Advisory Committee, Financial Accounting and Auditing Standards: A Review of Financial Auditing Standards, https://www.newyorkfed.org/edu/pdf/2014-10-22.pdf (stating the need to keep standards current in order to reflect changes in the business environment) and can help the SEC fulfill the agency’s mission because, “at a minimum, the standard setting body is capable of improving the accuracy and effectiveness of financial reporting and the protection of investors under the securities laws.”330 Today, the SEC recognizes the independent non-profit Financial Accounting Standards Board (FASB) as the designated private-sector standards setter, and considers its set standards as “generally accepted” under Sarbanes-Oxley.331 The SEC has made clear that there is federal oversight of this structure and that the SEC continues to have an important role in the standards’ recognition.332

For auditing standards, Sarbanes-Oxley created the Public Company Accounting Oversight Board (PCAOB), a non-profit corporation that is subject to SEC oversight.333 Oversight includes the SEC’s “approval of the Board’s rules, standards, and budget.”334 PCAOB itself is tasked with “oversee[ing] the audit of companies subject to securities laws.”335 Among its duties, PCAOB must, based on certain SEC actions, “register public accounting firms that prepare audit reports,” “establish or adopt . . . auditing . . . and other standards relating to the preparation of audit reports,” “conduct inspections of registered public accounting firms,” “conduct investigations and disciplinary pro-

327 U.S. Securities and Exchange Commission, Commission Statement of Policy: Reaffirming the Status of the FASB as a Designated Private-Sector Standard Setter, 68 Fed. Reg. 23381 (May 1, 2003). On its own authority, the SEC since 1973 has recognized FASB’s financial and accounting reporting standards as authoritative, but Sarbanes-Oxley helped provide a clearer, updated structure from Congress that the SEC could rely on to determine whether the standard-setting body produced “authoritative” or “generally accepted” financial accounting and reporting standards.

328 U.S. Securities and Exchange Commission, 68 Fed. Reg. at 23344 (“While the Commission consistently has looked to the private sector in the past to set accounting standards, the Sarbanes-Oxley Act, clearly provide the Commission with authority to set accounting standards for public companies and other entities that file financial statements with the Commission.” (citing Section 108(c) of the Sarbanes-Oxley Act, which states, “Nothing in this Act, including this section, shall be construed to impair or limit the authority of the Commissions to establish accounting principles or standards for purposes of enforcement of the securities laws.”)).


331 PCAOB Comment at A4 (“In developing an AI accountability framework, adequate procedures are necessary to ensure the reliability of, and market confidence in, company-specific financial information.”).

332 PCAOB Comment at 1 ("Notably, however, the ecosystem around financial reporting is a third of the stock market’s need for it. Local governments have their own accounting and auditing mechanisms. ") (citing Section 108(c) of the Sarbanes-Oxley Act, which states, “Nothing in this Act, including this section, shall be construed to impair or limit the authority of the Commissions to establish accounting principles or standards for purposes of enforcement of the securities laws.”)).

333 U.S. Securities and Exchange Commission, About, https://pcaobus.org/about. See also PCAOB Reporting Taxonomy (GRT) and the 2023 SEC Reporting Taxonomy (SRT) (collectively referred to as the “GAAP Taxonomy”), which together with the GAAP Taxonomy are collectively referred to as the XBRL Taxonomies.


335 PCAOB Comment at 1. See id. at 4 (“In developing an AI accountability framework, we recommend that policy makers look to the financial reporting ecosystem as the gold standard in ensuring the reliability of, and market confidence in, company-specific financial information.”).

336 U.S. Securities and Exchange Commission, 68 Fed. Reg. at 23334 (“While the Commission consistently has looked to the private sector in the past to set accounting standards, the Sarbanes-Oxley Act, clearly provide the Commission with authority to set accounting standards for public companies and other entities that file financial statements with the Commission.” (citing Section 108(c) of the Sarbanes-Oxley Act, which states, “Nothing in this Act, including this section, shall be construed to impair or limit the authority of the Commissions to establish accounting principles or standards for purposes of enforcement of the securities laws.”)).


Thus, accounting and auditing standards for the financial sector, subject to public securities law, are structured to permit non-governmental entities to lead in the creation of standards but give regulators the chance to contribute to and oversee the standards-setting process. While such structure is not without criticism, it has proven to be relatively effective in providing assurance about audited financials.

A review of the comments yields composite recommendations to use certain features of the financial accounting model for possible adoption in the AI accountability space. Some ideas include:

- Forming audit oversight boards, similar to the PCAOB, to train auditors, assess their qualifications, and adjudicate conflicts of interest.
- Imposing annual requirements for public companies that are AI actors to assess the effectiveness of their internal controls over AI risk management, documentation, and disclosure and have auditors attest to the company’s assessment. This is analogous to what is required of public companies with respect to financial reporting.
- Clarifying that because AI audits can take many forms and answer different questions, disclosing the terms of engagement and audit methodology creates critical context.
- Encouraging collaboration between AI actors and regulators on risk management. In the words of one commenter, collaboration between financial institution and their regulators "illuminates that a tailored yet flexible approach provides strong accountability measures that also allow industry to innovate.

- Establishing a federal regulator with cross-sectoral authority to oversee the implementation of AI standards.

5.2 HUMAN RIGHTS AND ENVIRONMENTAL, SOCIAL, AND GOVERNANCE (ESG) ASSESSMENTS

Financial accountability models and assurance methods are more tailored than accountability mechanisms for human rights and ESG performance. This flexibility is both an asset and a liability when it comes to considering these accountability regimes as models and vehicles for trustworthy AI evaluations.


Folding AI evaluations into human rights impact assessments is one way to ensure that AI evaluations take human rights into account and that human rights evaluations take AI into account. As one commenter put it, “there are benefits in using the same methodology and not burdening teams with performing several assessments in parallel.”

A number of commenters suggested incorporating human rights assessments frameworks into standard review processes across the AI life cycle.


342. SEC, The Enhancement and Standardization of Climate-Related Disclosures for Investors (Final Rule) (Mar. 6, 2024), https://www.sec.gov/rules/final/2024/33-10590.pdf (requiring “registrants to provide certain climate related information in registration statements and annual reports” including “information about a registrant’s climate-related risks that have materially impacted, or are reasonably likely to have a material impact on, its business strategy, results of operations, or financial condition”).

343. See generally, SASB Standards, SASB Standards and other ESG Frameworks: The Sustainability Reporting Evolution, supra note 253, at 5. See also Global Reporting Initiative (GRI), Carbon Disclosure Project (CDP), Task Force on Climate-related Financial Disclosures (TCFD), United Nations Sustainable Development Goals, supra note 160 (noting that AI safety standards are a predicate for ESG reporting for companies in the EU starting January 1, 2024).


345. See, e.g., Outsider Oversight, supra note 336, at 558. See also Centre for Information Policy Leadership Comment at 22.

346. In the United States, the SEC has adopted rules requiring climate-related disclosures for public companies. Companies are incorporating ESG disclosure models in their operations, using measurements from organizations modeled on financial accounting boards, such as the Sustainability Accounting Standards Board. There are many other standards and methods deployed in ESG evaluations. While ESG disclosure models are not currently designed to evaluate AI’s impact, commenters suggested incorporating AI and data practices more generally into the evaluation. For example, respect for individuals’ privacy rights is a human rights issue and at the same time is a “social impact” issue within bounds of the “S” in ESG.

There is a risk for ESG evaluations, as well as for AI trustworthiness evaluations, that the goals are too varied for meaningful results. One academic paper describes the problem as follows: “due to the ambiguity of what is being audited, ESG certifications risk becoming ‘cheap talk,’ rubber stamping practices without in fact promoting social responsibility.” Some questions may not be answerable. In the ESG context, this might be a question about supply chain responsibility. In the AI context, rights standards for AI.


349. See, e.g., Outsider Oversight, supra note 336, at 5. See also Centre for Information Policy Leadership Comment at 22.

350. See, e.g., Open MIC Comment at 5 (noting that AI safety standards are a predicate for ESG reporting as part of the ESG process.).

351. See, e.g., Centre for Information Policy Leadership Comment at 22.
the question might concern training data provenance and the labor conditions under which AI systems are trained. ESG evaluations have handled this difficulty of answerability by focusing on process, rather than outcomes. In other words, auditees are expected to attest to their best efforts to obtain satisfactory outcomes such as through their own supply chain audits and other measures. The design of AI evaluations might similarly look to appraise processes when outcomes escape measurement.307

The private sector continues to refine and seek ESG framework standardization for evaluations. What the ESG experience might teach is that multi-factor evaluations using a variety of standards may not immediately yield comparable or actionable results. However, the ESG auditing ecosystem has developed rapidly and become more standardized as stakeholders have demanded clarity around ESG performance and governments have required or incentivized better reporting.

5.3 FOOD AND DRUG REGULATION

Another potentially useful accountability model suggested by commenters can be found in health-related regulatory frameworks such as the FDAs.352 FDA regulates some AI systems as medical devices. To help medical device manufacturers who are developing AI-enabled devices, “[i]t publishes best practices for AI in medical devices, documents commercially available AI-enabled medical devices, and has promised to perform relevant pilots and advance regulatory science in its AI action plan.”353

Beyond that, commenters pointed to the FDA requirement that medical device manufacturers prepare premarket submissions for FDA review prior to marketing the device, where the requirements for premarket submissions are generally dependent on the level of risk associated with their device. Devices are classified into three categories (Class I, II, III). Regulatory controls increase from Class I to Class III. Most Class I devices are exempt from premarket review, while most Class II devices require submission of a premarket notification (“510(k)”). Most Class III devices require premarket approval.354 One commenter suggested that AI policy follow an analogous risk classification, with regulatory burdens of pre-market controls and disclosure applying to the highest risk products.355

A model for premarket notification for AI systems, such as the FDA’s model for some Class I and most Class II medical devices encompassing premarket notification and FDA review, could prove instructive for limited risk AI systems and deployments, and would allow for some degree of regulatory oversight and reduction of harm. On the other hand, a premarket notification model would likely create regulatory burden, potentially slowing and even disincentivizing development.356

The FDA further has in place an exemplary adverse incident database that could be instructive for AI system accountability.357 This system is similar to the Federal Aviation Administration’s Aviation Safety Reporting System, where the FDA’s model for some Class I and most Class II medical devices encompasses premarket notification and FDA review, could prove instructive for limited risk AI systems and deployments, and would allow for some degree of regulatory oversight and reduction of harm. On the other hand, a premarket notification model would likely create regulatory burden, potentially slowing and even disincentivizing development.356

Another accountability model overseen by the FDA include requirements for evidence-based drug testing and clinical trials, as well as disclosure of residual risk in the form of side effects.358 Finally, the FDA provides guidance for the labeling of AI systems deployed within its remit, and one commenter argued that requiring a form of marketing approval and similar recommendations would support “a more transparent understanding of how these systems operate.”359 These oversight mechanisms, which require both premarket review and post-market reporting, should be considered in the context of AI accountability, at least for high-risk systems, models, and uses.

5.4 CYBERSECURITY AND PRIVACY ACCOUNTABILITY MECHANISMS

With some exceptions, the current regulatory paradigms governing cybersecurity and data privacy lack uniformity at the federal level. Many extant federal laws concerning personal data and cybersecurity focus on select industries and subcategories of data.360 While NIST has developed voluntary risk management and cybersecurity frameworks that leave entities to determine the acceptable level of risk for achieving their organizational objectives,361 the implementation of these frameworks varies across organizations and industries.362 Privacy laws also vary from state to state.

One instrument of consistent federal law is the Federal Trade Commission Act’s application to data security and privacy. In the past twenty years, the FTC has brought dozens of law enforcement actions alleging that businesses had engaged in unfair or deceptive trade practices relating to data security or privacy.363 Among other things, the FTC has alleged deception where it has had reason to believe that companies have not lived up to their own public statements about their data privacy or security practices (e.g., where the companies represented that they would take reasonable or industry-standard measures but failed to do so, or where companies shared information with third parties that they had claimed would not be shared). The FTC has alleged unfair data security and privacy practices where it determined that businesses’ practices were likely to cause data security or privacy harm to consumers, and harm was not outweighed by counter-vailing benefits. To remedy such violations, the FTC has obtained relief, including injunctions requiring businesses to develop and implement comprehensive data security and/or privacy programs. In many cases, it required businesses to undergo third-party audits for compliance with such injunctions.364 The FTC has also promulgated guidance distilling the facts from its enforcement cases into data security lessons for companies.365

We discerned in the comments three basic perspectives on what we can learn from cybersecurity and privacy assurance practices and governance regimes: Some commenters believed that those practices and regimes should not be modeled on AI thought they were capacious enough to include AI assurance. Still others believed they could be extended and replicated to advance AI assurance. For commenters who thought cybersecurity frameworks are adequate to handle AI assurance, it was partly because cybersecurity practices have matured and refined through years of legal interpretation and application, thereby offering greater degrees of consistency and predictability.366 Indeed, existing laws and regulatory requirements that set cybersecurity standards for distinct industries already apply when AI deployments in those industries affect cybersecurity.367 In addition, there is an infrastructure for certifying cybersecurity and privacy auditors, and at least some of those certification programs are rolling out AI assurance certifications.368

307 See Grabowicz et al., Comment at 4. “We propose AI accountability mechanisms based on explanations of decision-making processes; since explanations are automatically generated and highlight the true underlying model decision process” (Class I, II, III). Regulatory controls increase from Class I to Class III. Most Class I devices are exempt from premarket review, while most Class II devices require submission of a premarket notification (“510(k)”). Most Class III devices require premarket approval.354 One commenter suggested that AI policy follow an analogous risk classification, with regulatory burdens of pre-market controls and disclosure applying to the highest risk products.355

308 See supra Grabowicz et al., Comment at 4. “We propose AI accountability mechanisms based on explanations of decision-making processes; since explanations are automatically generated and highlight the true underlying model decision process” (Class I, II, III). Regulatory controls increase from Class I to Class III. Most Class I devices are exempt from premarket review, while most Class II devices require submission of a premarket notification (“510(k)”). Most Class III devices require premarket approval.354 One commenter suggested that AI policy follow an analogous risk classification, with regulatory burdens of pre-market controls and disclosure applying to the highest risk products.355

309 See supra Grabowicz et al., Comment at 6. See also Andrew Tu, AI FDA for Algorithms, 68 Admin. L. Rev. Br. 83 (2017), “Regulatory authorities should consider entities that create algorithms (including general algorithms about the analogy between FDA regulation and algorithms’ risk management).”

310 See supra Grabowicz et al., Comment at 4.

311 See supra Grabowicz et al., Comment at 4. See, e.g., USTelecom Comment at 9.

312 See supra Grabowicz et al., Comment at 9. See, e.g., IAPP Comment at 5-6.

313 For example, the FAA currently uses special conditions, as provided for in its regulations, to address novel or unusual design features not adequately addressed by existing airworthiness standards, to address cybersecurity of certain en-disabled aircraft. This approach would be potentially extendable to AI incorporating cybersecurity. See 14 CFR § 39.15 (a)(3) (2013). A product that becomes apparent after an aircraft or other article product enters the marketplace, the FAA issues airworthiness directives in appropriate cases, specifically “for the issuance of an airworthiness directive affecting a product which the Administrator finds, after notice and opportunity for hearing, that: (a) is unsafe condition exists in the product; and (b) the condition is likely to exist or develop in other products of the same type design.” See 14 CFR § 39.15 (a)(3).
These are all sound ideas that merit further consideration, especially a bounty program for AI vulnerability detection. Any federal government bodies tasked with horizontal regulation of AI should include analogous capacity to that found in the Cybersecurity and Infrastructure Security Agency (CISA), which helps organizations improve their cybersecurity practices. Aspects of the National Cybersecurity Strategy could also be applied to AI, including harmonizing reporting requirements, adverse incident disclosures, and risk metrics throughout the Federal government. As in the cybersecurity context, law enforcement is an essential companion to self-regulation. We recommend that future federal AI policymaking not lean entirely on purely voluntary best practices. Rather, some AI accountability measures should be required, pegged to risk. We are convinced that AI accountability policy can employ, adapt, and expand upon existing cybersecurity and privacy infrastructure, while adopting a risk-based framework. At the same time, AI accountability poses new challenges and requires new approaches. It is to some of those new recommended approaches that the Report now turns.

375. Rachel Clinton, Mira Guleri, and Helen He Comment at 2 (“Any AI system collecting any kind of data should be audited at least once a year to ensure compliance with the following: ISO (International Organization for Standardization) 27001 [and] NIST CSF [National Institute of Standards and Technology Cybersecurity Framework].”)

376. See, e.g., AI Policy and Governance Working Group Comment at 3.

377. See, e.g., Anthropic Comment at 9-10; Microsoft Comment at 6-7.

378. Google DeepMind Comment at 17. Relatedly, federal agencies and departments are standing up “bias bounty” programs to address bias in AI systems. See, e.g., Matthew Huynh Johnson, Funding Opportunity from my team to build and run a DoD-wide Bias Bounty Program, https://www.linkedin.com/pulse/funding-opportunity-matthew-huynh-johnson-

bias-bounty-program-matthew-huynh-johnson/. See also https://www.buyapass.org/why-we-fund-funding-opportunity-for-matthew-huynh-johnson-

bias-bounty-program-matthew-huynh-johnson/.

379. See, e.g., The AI Risk and Vulnerability Alliance (ARVA) Comment at 1-2.

380. See, e.g., Center for AI Safety Comment, Appendix A, at 3.

381. See The White House, supra note 382.

382. See Rachel Clinton, Mira Guleri, and Helen He Comment at 2.

Others thought that while existing cybersecurity and privacy practices are probably inadequate for AI accountability, those practices could be modified to accommodate new risks. For example, cybersecurity audits could be conducted on a regular basis to review conformity with existing standards, including the ISO/IEC 27001 information security standard and NIST’s cybersecurity framework. Other suggestions borrowed from the cybersecurity context including creating incentives for companies to facilitate “responsible disclosure”; developing red-teaming exercises; launching “Bug Bounty” programs to encourage disclosure and financially reward detection of AI vulnerabilities; and modeling AI vulnerability disclosures on the Common Vulnerabilities and Exposures (CVE) system, which provides a standardized naming scheme for cybersecurity vulnerabilities.
Recommendations

The public, consumers, customers, workers, regulators, shareholders, and others need reliable information to make choices about AI systems. To justify public trust in, and reduce potential harms from, AI systems, it will be important to develop “accountability inputs” including better information about AI systems as well as independent evaluations of their performance, limitations, and governance. AI actors should be held accountable for claims they make about AI systems and for meeting regulatory requirements. AI actors should advance the AI accountability ecosystem by encouraging, supporting, and/or compelling these inputs. Doing this work is a natural follow-on to the AI EO, which identified the properties that should be expected from algorithmic systems; and NIST’s AI RMF, which recommended a set of approaches to AI risk management. This work would likely include the creation of auditor certifications and audit methodologies, as well as mechanisms for regulatory recognition of appropriate certifications and methodologies.

6.1 GUIDANCE

6.1.1 Audits and auditors: Federal government agencies should work with stakeholders as appropriate to create guidelines for AI audits and auditors, using existing and/or new authorities.

Independent AI audits and evaluations are central to any accountability structure. To help create clarity and utility around independent audits, we recommend that the government work with stakeholders to create basic guidelines for what an audit covers and how it is conducted — guidance that will undoubtedly have some general components and some domain-specific ones. This work would likely include the creation of auditor certifications and audit methodologies, as well as mechanisms for regulatory recognition of appropriate certifications and methodologies.

Auditors should adhere to consensus standards and audit criteria where possible, recognizing that some will be specific to particular risks (e.g., dangerous capabilities in a foundation model) and/or particular deployment contexts (e.g., discriminatory impact in hiring). Much work is required to create those standards — which NIST and others are undertaking. Audits and other evaluations are being rolled out now concurrently with the development of technical standards. Especially where evaluators are relying on consensus standards, it is important that they show their work so that they too are subject to evaluation. Auditors should disclose methodological choices and auditor independence criteria, with the goal of standardizing such methods and criteria as appropriate. The goals of safeguarding sensitive information and ensuring auditor independence and appropriate expertise may mitigate towards a certification process for qualified auditors.

6.1.2 Disclosure and access: Federal government agencies should work with stakeholders to improve standard information disclosures, using existing and/or new authorities.

Disclosures should be tailored to their audiences, which may require the creation of multiple artifacts at varying levels of detail and/or the establishment of information intermediaries. Standardizing a baseline disclosure using artifacts like model and system cards, datasheets, and nutritional labels for AI systems can reduce the costs for all constituencies evaluating and assuring AI. As it did with food nutrition labels, the government may have a role in shaping standardized disclosure, whatever the form. We recommend support of the NIST-led process to provide guidance and best practices on standardized baseline disclosures for AI systems and certain models as an input to AI accountability. Working with stakeholders and achieving commitments from government suppliers, contractors, and grantees to implement such standardized baseline disclosures could advance adoption.

6.1.2 Liability rules and standards: Federal government agencies should work with stakeholders to make recommendations about applying existing liability rules and standards to AI systems and, as needed, supplementing them.

Stakeholders seek clarification of liability standards for allocating responsibility among AI actors in the value chain. We expect AI liability standards to emerge out of the courts as legal actions which clarify responsibilities and redress harms. Regulatory agencies also have an important role in determining how existing laws and regulations apply to AI systems. Of course, Congress and state legislatures will define new liability contours. To help clarify and establish standards for liability, where needed, we encourage further study and collection of stakeholder and government agency input.

To this end, we support a government convening of legal experts and other relevant stakeholders, including affected communities, to inform how policymakers understand the role of liability in the AI accountability ecosystem. The AI accountability inputs we recommend in this Report will feed into legal actions and standards and, by the same token, these inputs should be shaped by the legal community’s emerging needs to vindicate rights and interests. It is also the case that a vibrant practice of independent third-party evaluation of AI systems may depend on both exposure to liability (e.g., perhaps for auditors) and protection from liability (e.g., perhaps for researchers), depending on relevant legal considerations.
6.2. SUPPORT

6.2.1 People and tools: Federal government agencies should support and invest in technical infrastructure, AI system access tools, personnel, and international standards work to invigorate the accountability ecosystem.

Robust auditing, red-teaming, and other independent evaluations of AI systems require resources, some of which the federal government has and should make available, and some of which will require funding. A significant move in this direction would be for Congress to support the U.S. AI Safety Institute and appropriate funds and establish the National AI Research Resource (NAIRR). NAIRR could contribute to the larger set of needed resources, including:

- Datasets to test for equity, efficacy, and many other attributes and objectives;
- Compute and cloud infrastructure required to do rigorous evaluations;
- Appropriate access to AI system components and processes for researchers, regulators, and evaluators, subject to intellectual property, data privacy, and security- and safety-informed functions;
- Independent red-teaming support; and
- International standards development (including broad participation) and, where applicable for national security, national standards development.

People are also required. We recommend an investment in federal personnel with appropriate sociotechnical expertise to conduct and review AI evaluations and other AI accountability inputs. Support for education and red-team efforts would also grow the ecosystem for independent evaluation and accountability.384

6.2.2 Research: Federal government agencies should conduct and support more research and development related to AI testing and evaluation, tools facilitating access to AI systems for research and evaluation, and provenance technologies, through existing and new capacity.

Because of their complexity and importance for AI accountability, the following topics make compelling candidates for research and development investment:

- Research into the creation of reliable, widely applicable evaluation methodologies for model capabilities and limitations, safety, and trustworthy AI attributes;
- Research on durable watermarking and other provenance methods; and
- Research into technical tools that facilitate researcher and evaluator access to AI system components in ways that preserve data privacy and the security of sensitive model elements, while retaining openness.

Government should build on investments already underway through the U.S. AI Safety Institute and the National Science Foundation.

6.3. REGULATORY REQUIREMENTS

6.3.1. Audits and other independent evaluations: Federal agencies should use existing and/or new authorities to require as needed independent evaluations and regulatory inspections of high-risk AI model classes and systems.

There are strong arguments for sectoral regulation of AI systems in the United States and for mandatory audits of AI systems deemed to present a high risk of harming rights or safety – according to holistic assessments tailored to deployment and use contexts. Given these arguments, work needs to be done to implement regulatory requirements for audits in some situations. It may not currently be feasible to require audits for all high-risk AI systems because the ecosystem for AI audits is still immature; requirements may need delayed implementation. However, the ecosystem’s maturity will be accelerated by forcing functions. Government may also need to require other forms of information creation and distribution, including documentation and disclosure, in specific sectors and deployment contexts (beyond what it already does require).

Additional consideration should be given to the necessity of pre-release claim substantiation and other certification requirements for certain high-risk AI systems, models, and/or AI systems in high-risk sectors (e.g., health care and finance), as well as periodic claim substantiation for deployed AI systems. Such proactive substantiation would help AI actors to shoulder their burden of assuring AI systems from the start. In the AI context, this marginal additional friction for AI actors could create breathing room for accountability mechanisms to catch up to deployment.

Regardless of the type of inspection model that is adopted, federal regulatory agencies should coordinate closely with regulators in non-adversary countries for alignment of inspection regimes in their methods and use of international standards so that AI products can be evaluated using globally comparable criteria.

6.3.2 Cross-sectoral governmental capacity: The federal government should strengthen its capacity to address cross-sectoral risks and practices related to AI. Although sector-specific requirements for AI already exist, the exercise of horizontal capacity in the federal government would provide common baseline requirements, re-inforce appropriate expertise to oversee AI systems, help to address cross-sectoral risks and practices, allow for better coordination among sectoral regulators that require or consume disclosures and evaluations, and provide regulatory capacity to address foundation models.

Such cross-sectoral horizontal capacity, wherever housed, would be useful for creating accountability inputs such as:

- A national registry of high-risk AI deployments;
- A national AI adverse incidents reporting database and platform for receiving reports;
- A national registry of disclosable AI system audits;
- Coordination of, and participation in, audit standards and auditor certifications, enabling advocacy for the needs of federal agencies and congruence with independent federal audit actions;
- Pre-release review and certification for high-risk deployments and/or systems or models; and
- Collection of periodic claim substantiation for deployed systems; and
- Coordination of AI accountability inputs with agency counterparts in non-adversarial states.


6.3.3. Contracting: The federal government should require that government suppliers, contractors, and grantees adopt sound AI governance and assurance practices for AI used in connection with the contract or grant, including using AI standards and risk management practices recognized by federal agencies, as applicable.

The government’s significant purchasing power affords it the ability to shape marketplace standards, and prefer suppliers who provide sufficient documentation, access, freedom to evaluate, and other assurance practices. As the National AI Advisory Committee Report recommended, the government should reform procurement practices to promote trustworthy AI. The same principles would apply to government grants. The OMB draft guidance on “Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence” represents a significant step in this direction.385

385. See OMB Draft Memo. See also AI EO at Sec. 7.3 (directing the Department of Labor to establish “guidance for Federal contractors regarding nondiscrimination in hiring involving AI and other technology-based hiring systems”).
The process of defining terms in the AI policy space is ongoing and fluid. Where there are not, we use definitions we find support by the record and research.

Artificial Intelligence or AI. AI has the meaning set forth in 15 U.S.C. 9401(3), which is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action.

AI Accountability. AI accountability is the process, heavily reliant on transparency and assurance practices, of holding entities answerable for the risks and/or harms of the AI systems they develop or deploy. This is closest to the definition adopted by the Trade and Technology Council (TTC) joint U.S.-EU set of AI terms, which defines accountability as an “allocated responsibility” for system performance or for governance functions.390 Whereas OECD interpretative guidance describes “accountability” from “responsibility” and “liability,”391 the TTC definition embraces responsibility as part of accountability and includes a broader scope of governance activities. Accountability may require enforceable consequences.392 Such consequences, usually determined by regulators, courts, and the market, are accountability outputs. This Report focuses on developing and shaping “accountability inputs,” which feed into systems of accountability.

AI Accountability Inputs. AI accountability inputs are the AI system information flows and evaluations that enable the identification of entities, factors, and systems responsible for the risks and/or harms of those systems. These are necessary or useful practices, artifacts, and products that feed into downstream accountability mechanisms such as regulation, litigation, and market choices.

AI Actor. AI actors are “those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI.”393 AI actors are present across the AI lifecycle, including “an AI developer who makes AI software available, such as pre-trained models” and “an AI actor who is responsible for deploying that pre-trained model in a specific use case.”394

AI Assurance. AI assurance is the product of a set of informative and evaluative practices that can provide justified confidence that an AI system operates in context in a trustworthy fashion and as claimed. This definition draws from MITRE’s use of the term “justified confidence” (from international software assurance standards) and from ISO/IEC standards,395 which specify the foundational elements of an assurance program.

AI Model. AI model means a component of an AI system that implements AI technology and uses computational, statistical, or machine learning techniques to produce outputs from a given set of inputs.

AI System. An AI system is an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.

Red-Teaming. Red-teaming means a structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. Red-teaming is most often performed by dedicated “red-teams” that adopt adversarial methods to identify flaws and vulnerabilities in harmful or discriminatory outputs from an AI system, uniform or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system.396

Trustworthy AI. The NIST AI RMF defines trustworthiness in AI as “responsive[ness] to a multiplicity of criteria that are of value to interested parties.” It specifies that such values include “valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmed bias managed.”397 The White House Voluntary Commitments specify that “trust,” together with “safety” and “security,” comprise the “three principles that must be fundamental to the future of AI.”398

APPENDIX A: Glossary of Terms

The process of defining terms in the AI policy space is ongoing and fluid. Where there are not, we use definitions we find support by the record and research.

Artificial Intelligence or AI. AI has the meaning set forth in 15 U.S.C. 9401(3), which is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action.

AI Accountability. AI accountability is the process, heavily reliant on transparency and assurance practices, of holding entities answerable for the risks and/or harms of the AI systems they develop or deploy. This is closest to the definition adopted by the Trade and Technology Council (TTC) joint U.S.-EU set of AI terms, which defines accountability as an “allocated responsibility” for system performance or for governance functions.390 Whereas OECD interpretative guidance describes “accountability” from “responsibility” and “liability,”391 the TTC definition embraces responsibility as part of accountability and includes a broader scope of governance activities. Accountability may require enforceable consequences.392 Such consequences, usually determined by regulators, courts, and the market, are accountability outputs. This Report focuses on developing and shaping “accountability inputs,” which feed into systems of accountability.

AI Accountability Inputs. AI accountability inputs are the AI system information flows and evaluations that enable the identification of entities, factors, and systems responsible for the risks and/or harms of those systems. These are necessary or useful practices, artifacts, and products that feed into downstream accountability mechanisms such as regulation, litigation, and market choices.

AI Actor. AI actors are “those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI.”393 AI actors are present across the AI lifecycle, including “an AI developer who makes AI software available, such as pre-trained models” and “an AI actor who is responsible for deploying that pre-trained model in a specific use case.”394

AI Assurance. AI assurance is the product of a set of informative and evaluative practices that can provide justified confidence that an AI system operates in context in a trustworthy fashion and as claimed. This definition draws from MITRE’s use of the term “justified confidence” (from international software assurance standards) and from ISO/IEC standards,395 which specify the foundational elements of an assurance program.

AI Model. AI model means a component of an AI system that implements AI technology and uses computational, statistical, or machine learning techniques to produce outputs from a given set of inputs.

AI System. An AI system is an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.

Red-Teaming. Red-teaming means a structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. Red-teaming is most often performed by dedicated “red-teams” that adopt adversarial methods to identify flaws and vulnerabilities in harmful or discriminatory outputs from an AI system, uniform or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system.396

Trustworthy AI. The NIST AI RMF defines trustworthiness in AI as “responsive[ness] to a multiplicity of criteria that are of value to interested parties.” It specifies that such values include “valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmed bias managed.”397 The White House Voluntary Commitments specify that “trust,” together with “safety” and “security,” comprise the “three principles that must be fundamental to the future of AI.”398
About NTIA
The National Telecommunications and Information Administration (NTIA), located within the Department of Commerce, is the Executive Branch agency that is principally responsible by law for advising the President on telecommunications and information policy issues. NTIA’s programs and policymaking focus largely on expanding broadband Internet access and adoption in America, expanding the use of spectrum by all users, and ensuring that the Internet remains an engine for continued innovation and economic growth. These goals are critical to America’s competitiveness in the 21st century global economy and to addressing many of the nation’s most pressing needs, such as improving education, health care, and public safety.

For more information, please visit us at ntia.gov