NTIA

Artificial Intelligence
Accountability Policy Report

MARCH 2024

**National Telecommunications and
Information Administration**

United States Department of Commerce

National Telecommunications and
Information Administration

United States Department of Commerce

Artificial Intelligence

Accountability Policy Report

Contents

Executive
Summary

Executive Summary

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
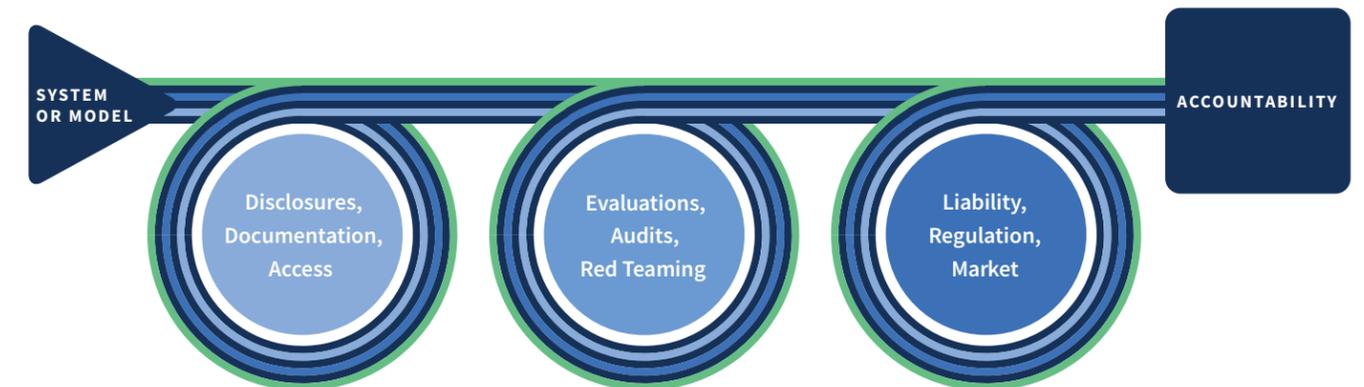
Participants in the AI ecosystem – including policymakers, industry, civil society, workers, researchers, and impacted community members – should be empowered to expose problems and potential risks, and to hold responsible entities to account.

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-release 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.

AI ACCOUNTABILITY CHAIN

SYSTEM OR MODEL

Disclosures, Documentation, Access

Evaluations, Audits, Red Teaming

Liability, Regulation, Market

ACCOUNTABILITY

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.
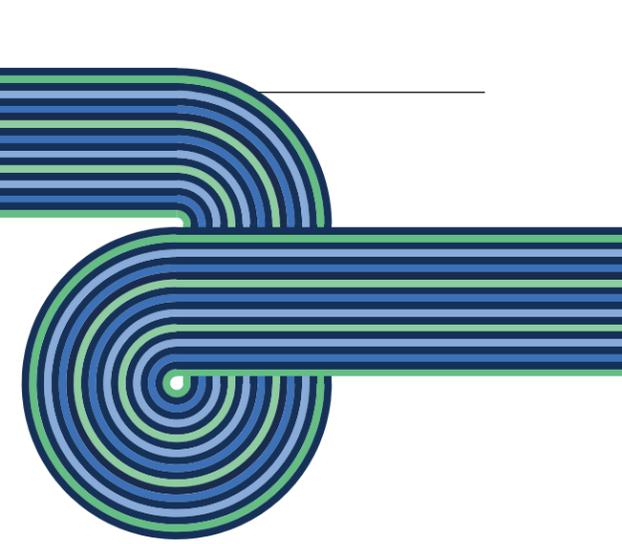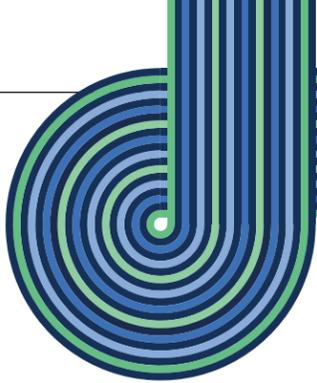
GUIDANCE

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, datasheets, 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 sociotechnical 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 harming 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.
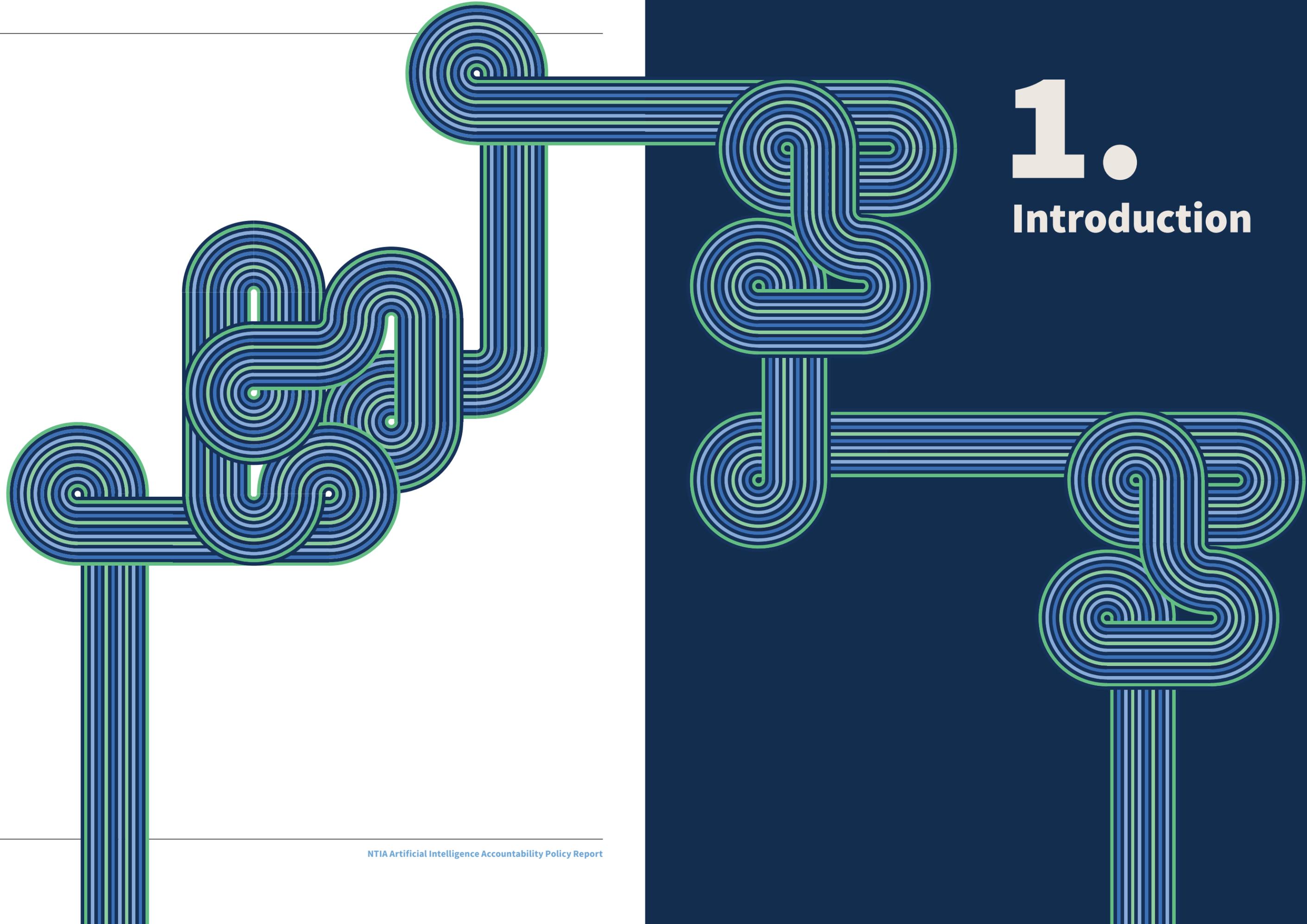
1.
Introduction

Introduction

NTIA issued a Request for Comment on AI Accountability Policy on April 13, 2023 (RFC).[1] 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 accountable, 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.[2] 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.[3] 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.[4] 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

individuals in their institutional capacity. Of this latter group, industry (including trade associations) accounted for approximately 48%, nonprofit advocacy for approximately 37%, and academic and other research organizations for approximately 15%. There were a few comments from elected and other governmental officials.

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.[5] 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.[6] In addition, the Administration secured voluntary commitments from healthcare companies related to AI.[7] 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.[8] Administration efforts to advance trustworthy AI *prior* to the release of the RFC in April 2023 include most notably the NIST AI Risk Management Framework (NIST AI RMF)[9] and the White House Blueprint for an AI Bill of Rights (Blueprint for AIBoR).[10]

---

1   National Telecommunications and Information Administration (NTIA), AI Accountability Policy Request for Comment, 88 Fed. Reg 22433 (April 13, 2023) [hereinafter "AI Accountability RFC"].

2   *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 A 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).

3   National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework (AI RMF 1.0) (Jan. 2023), https://doi.org/10.6028/NIST.AI.100-1 [hereinafter "NIST AI RMF"]. 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-21-519SP), at 24 n.22 (Jun 30, 2021), https://www.gao.gov/assets/gao-21-519sp.pdf (citing U.S. government documents using the term "responsible use" to entail AI system use that is responsible, equitable, traceable, reliable, and governable).

4   Regulations.gov, NTIA AI Accountability RFC (2023), https://www.regulations.gov/document/NTIA-2023-0005-0001/comment. Comments in this proceeding are accessible through this link, with an index available linking commenter name with regulations.gov commenter number available here: https://www.regulations.gov/document/NTIA-2023-0005-1452.

5   *See* The White House, FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety (May 4, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/ (allowing "AI models to be evaluated thoroughly by thousands of community partners and AI experts to explore how the models align with the principles and practices outlined in the Biden-Harris Administration's Blueprint for an AI Bill of Rights and AI Risk Management Framework").

6   *See* The White House, FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI (July 21, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/; The White House, Ensuring Safe, Secure and Trustworthy AI (July 21, 2023), https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf [hereinafter "First Round White House Voluntary Commitments"] (detailing the commitments to red-team models, sharing information among companies and the government, investment in cybersecurity, incentivizing third-party issue discovery and reporting, and transparency through watermarking, among other provisions); The White House, FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI (Sept. 12, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/; The White House, Voluntary AI Commitments (September 12, 2023), https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf [hereinafter "Second Round White House Voluntary Commitments"].

7   The White House, FACT SHEET: Biden-Harris Administration Announces Voluntary Commitments from Leading Healthcare Companies to Harness the Potential and

Manage the Risks Posed by AI (December 14, 2023), https://www.hhs.gov/about/news/2023/12/14/fact-sheet-biden-harris-administration-announces-voluntary-commitments-leading-healthcare-companies-harness-potential-manage-risks-posed-ai.html.

8   Executive Order No. 14110, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 88 Fed. Reg. 75191 [hereinafter "AI EO"] (2023) at Sec. 2, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

9   NIST AI RMF; *see also* U.S.-E.U. Trade and Technology Council (TTC), TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management (Dec. 1, 2022), https://www.nist.gov/system/files/documents/2022/12/04/Joint_TTC_Roadmap_Dec2022_Final.pdf, at 9 ("The RMF is a voluntary framework seeking to provide a flexible, structured, and measurable process to address AI risks prospectively and continuously throughout the AI lifecycle. […] Using the AI RMF can assist organizations, industries, and society to understand and determine their acceptable levels of risk. The AI RMF is not a compliance mechanism, nor is it a checklist intended to be used in isolation. It is law- and regulation-agnostic, as AI policy discussions are live and evolving.").

10   The White House, Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People (Oct. 2022), https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf [hereinafter "Blueprint for AIBoR"].
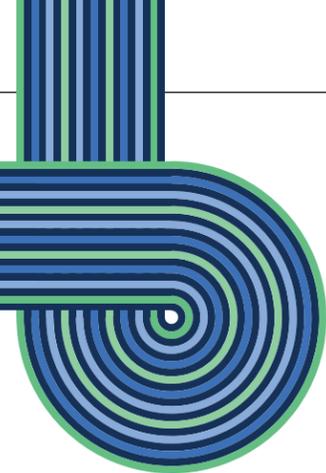
Federal regulatory and law enforcement agencies have also advanced AI accountability efforts. A joint statement from the Federal Trade Commission, the Department of Justice's Civil Rights Division, the Equal Employment Opportunity Commission, and the Consumer Financial Protection Bureau outlined the risks of unlawfully discriminatory outcomes produced by AI and other automated systems and asserted the respective agencies' commitment to enforcing existing law.[11] Other federal agencies are examining AI in connection with their missions.[12] A number of different Congressional committees have held hearings, and members of Congress have introduced bills related to AI.[13] State legislatures across the country have passed bills that affect AI,[14] and localities are legislating 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) issued a joint AI 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 advance shared international guiding principles and a code of conduct for trustworthy AI development.[18] The

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 models[20] or general purpose AI – has continued on the path to becoming law.[21] The EU Digital Services Act requires audits 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. More specifically, this Report focuses on information flow, system evaluations, and ecosystem development which, together with regulatory, market, and liability functions, are likely to promote accountability for AI developers and deployers (collectively and individually designated here as "AI actors"). There are many other players in the AI value chain traditionally included in the designation of AI actors, including system end users.

> ## 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 developers and deployers 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 take different forms in different sectors.

Multiple policy interventions may be necessary to achieve accountability. Take, for example, a policy promoting the disclosure to appropriate parties of training data details, performance limitations, and model characteristics for high-risk AI systems. Disclosure alone does not make an AI actor accountable. However, such information flows will likely be important for internal 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 to 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

---

11 *See* Rohit Chopra, Kristen Clarke, Charlotte A. Burrows, and Lina M. Khan, Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems (April 25, 2023), https://www.ftc.gov/system/files/ftc_gov/pdf/EEOC-CRT-FTC-CFPB-AI-Joint-Statement%28final%29.pdf [hereinafter "Joint Statement on Enforcement Efforts"]; Consumer Financial Protection Circular, 2023-03, Adverse action notification requirements and the proper use of the CFPBs sample forms provided in Regulation B, https://www.consumerfinance.gov/compliance/circulars/circular-2023-03-adverse-action-notification-requirements-and-the-proper-use-of-the-cfpbs-sample-forms-provided-in-regulation-b/. *See also*, Consumer Financial Protection Bureau, CFPB Issues Guidance on Credit Denials by Lenders Using Artificial Intelligence (Sept. 2023), https://www.consumerfinance.gov/about-us/newsroom/cfpb-issues-guidance-on-credit-denials-by-lenders-using-artificial-intelligence/; Equal Employment Opportunity Commission, Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964 (May 18, 2023), https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial.

12 *See, e.g.,* U.S. Department of Education Office of Educational Technology, Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations (May 2023), https://www2.ed.gov/documents/ai-report/ai-report.pdf; Engler, infra note 359 (referring to initiatives by the U.S. Food and Drug Administration); U.S. Department of State, Artificial Intelligence (AI), https://www.state.gov/artificial-intelligence/; U.S. Department of Health and Human Services, Trustworthy AI (TAI) Playbook (September 2021), https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf; U.S. Department of Homeland Security Science & Technology Directorate, Artificial Intelligence (September 2023), https://www.dhs.gov/science-and-technology/artificial-intelligence.

13 *See, e.g.,* Laurie A. Harris, Artificial Intelligence: Overview, Recent Advances, and Considerations for the 118th Congress, Congressional Research Service (Aug. 4, 2023), at 9-10, https://crsreports.congress.gov/product/pdf/R/R47644/2; Anna Lenhart, Roundup of Federal Legislative Proposals that Pertain to Generative AI: Part II, Tech Policy Press (Aug. 9, 2023), https://techpolicy.press/roundup-of-federal-legislative-proposals-that-pertain-to-generative-ai-part-ii/; *see also, e.g.,* U.S. House of Representatives Committee on Oversight and Accountability Subcommittee on Cybersecurity, Information Technology, and Government Innovation, Advances in AI: Are We Ready For a Tech Revolution? (subcommittee hearing) (March 8, 2023), https://oversight.house.gov/hearing/advances-in-ai-are-we-ready-for-a-tech-revolution/; U.S. House of Representatives Committee on Science, Space, and Technology, Artificial

Intelligence: Advancing Innovation Towards the National Interest (committee hearing) (June 22, 2023), https://science.house.gov/hearings?ID=441AF8AB-7065-45C8-81E0-F386158D625C; U.S. Senate Committee on the Judiciary Subcommittee on Privacy, Technology, and the Law, Oversight of A.I.: Rules for Artificial Intelligence (committee hearing) (May 16, 2023), https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-rules-for-artificial-intelligence.

14 *See* Katrina Zhu, The State of State AI Laws: 2023, Electronic Privacy Information Center (Aug. 3, 2023), https://epic.org/the-state-of-state-ai-laws-2023/ (providing an inventory of state legislation).

15 *See, e.g.,* The New York City Council, A Local Law to Amend the Administrative Code of the City of New York, in Relation to Automated Employment Decision Tools, Local Law No. 2021/144 (Dec. 11, 2021), https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID%7CTe xt%7C&Search=.

16 *See* The White House, FACT SHEET: U.S.-EU Trade and Technology Council Deepens Transatlantic Ties (May 31, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/31/fact-sheet-u-s-eu-trade-and-technology-council-deepens-transatlantic-ties/.

17 *See* The White House, U.S.-EU Joint Statement of the Trade and Technology Council (May 31, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/31/u-s-eu-joint-statement-of-the-trade-and-technology-council-2/; *supra* note 9, U.S.-E.U. Trade and Technology Council (TTC).

18 The White House, G7 Leaders' Statement on the Hiroshima AI Process (Oct. 30, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/g7-leaders-statement-on-the-hiroshima-ai-process/; Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI System (Oct. 30, 2023), https://www.mofa.go.jp/files/100573471.pdf; Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (Oct. 30, 2023), https://www.mofa.go.jp/files/100573473.pdf (mofa.go.jp).

19 *See, e.g.,* OECD ADVANCING ACCOUNTABILITY IN AI GOVERNING AND MANAGING RISKS THROUGHOUT THE LIFECYCLE FOR TRUSTWORTHY AI (Feb. 2023), https://www.oecd.org/sti/advancing-accountability-in-ai-2448f04b-en.htm. *See also* United Nations, High-level Advisory Body on Artificial Intelligence, https://www.un.org/en/ai-advisory-body (calling for "[g]lobally coordinated AI governance" as the "only way to harness AI for humanity, while addressing its risks and uncertainties, as AI-related applications, algorithms, computing capacity and expertise become more widespread internationally" and describing the mandate of the new High-level Advisory Body on Artificial Intelligence to "analysis and advance recommendations for the international governance of AI").

20 We use the term "foundation model" to refer to models which are "trained on broad data at scale and are adaptable to a wide range of downstream tasks", like "BERT, DALL-E, [and] GPT-3". *See* Richi Bommasani et al., On the Opportunities and Risks of Foundation Models, arXiv (July 12, 2022), https://arxiv.org/pdf/2108.07258.pdf.

21 *See* European Parliament, European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)) (March 13, 2024), https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf (containing the text of the proposed EU AI Act as adopted by the European Parliament) [hereinafter "EU AI Act"]; European Parliament, Artificial Intelligence Act: Deal on Comprehensive Rules for Trustworthy AI, European Parliament News (Dec. 12, 2023), https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-dealon-comprehensive-rules-for-trustworthy-ai.

22 *See* European Commission, Digital Services Act: Commission Designates First Set of Very Large Online Platforms and Search Engines (April 25, 2023), https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413.

23 *See* European Commission, Commission Delegated Regulation (EU) Supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council, by Laying Down Rules on the Performance of Audits for Very Large Online Platforms and Very Large Online Search Engines, (Oct. 20, 2023), at 2, 14, https://digital-strategy.ec.europa.eu/en/library/delegated-regulation-independent-audits-under-digital-services-act.

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.[24] These issues are of ongoing consideration at the U.S. Patent and Trademark Office (USPTO)[25] and at the U.S. Copyright Office.[26] 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.[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.[28] AI can exacerbate risks to Americans' privacy, as recognized by the Blueprint for an AI Bill of Rights and the AI EO.[29] 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.[30]

---

24 *See, e.g.,* NIST AI RMF at 16, 24 (recognizing 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 rights); Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, *supra* note 18, 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-protected content.").

25 The USPTO will clarify and make recommendations on key issues at the intersection of intellectual property and artificial intelligence. *See* AI EO Section 5.2. *See also* U.S. Patent and Trademark Office, Request for Comments Regarding Artificial Intelligence and Inventorship, 88 Fed. Reg. 9492 (Feb. 14, 2023), https://www.federalregister.gov/documents/2023/02/14/2023-03066/request-for-comments-regarding-artificial-intelligence-and-inventorship; U.S. Patent and Trademark Office, Public Views on Artificial Intelligence and Intellectual Property Policy (Oct. 2020), https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-10-07.pdf; U.S. Patent and Trademark Office, Artificial Intelligence, https://www.uspto.gov/initiatives/artificial-intelligence.

26 *See, e.g.,* U.S. Copyright Office, Notice of Inquiry and Request for Comments on Artificial Intelligence and Copyright, 88 Fed. Reg. 59942 (Aug. 30, 2023) [hereinafter "Copyright Office AI RFC"], https://www.federalregister.gov/documents/2023/08/30/2023-18624/artificial-intelligence-and-copyright; U.S. Copyright Office Comment at 2 (describing the Copyright Office's ongoing work at the intersection of AI and copyright law and policy); U.S. Copyright Office, Copyright and Artificial Intelligence, https://www.copyright.gov/ai/.

27 *See* U.S. Copyright Office Comment at 2 ("We are, however, cognizant that the policy issues implicated by rapidly developing AI technologies are bigger than any individual agency's authority, and that NTIA's accountability inquiries may align with our work."); *see also* Copyright Office AI RFC at 59,944 n.22 (mentioning the U.S. Copyright Office's consideration of AI in the regulatory context of the Digital Millenium Copyright Act rulemaking. By law, NTIA plays a consultation role in the rulemaking and has previously commented on petitions for exemptions that involve considerations

of AI. *See* 17 U.S.C. § 1201(a)(1)(C); NTIA, Recommendations of the National Telecommunications and Information Administration to the Register of Copyrights in the Eight Triennial Section 1201 Rulemaking at 48-58 (Oct. 1, 2021), https://www.ntia.gov/sites/default/files/publications/ntia_dmca_consultation_2021_0.pdf.

28 *See, e.g.,* Data & Society Comment at 7; Google DeepMind Comment at 3; Global Partners Digital Comment at 15; Hitachi Comment at 10; TechNet Comment at 4; NCTA Comment at 4-5; Centre for Information Policy Leadership (CIPL) Comment at 1; Access Now Comment at 3-5; BSA | The Software Alliance Comment at 12; U.S. Chamber of Commerce Comment at 9 (discussing the need for federal privacy protection); Business Roundtable Comment at 10 (supporting a passage of a federal privacy/consumer data security law to align compliance efforts across the nation); CTIA Comment at 1, 4-7 (declaring that federal privacy legislation is necessary to avoid the current fragmentation); Salesforce Comment at 9 ("The lack of an overarching Federal standard means that the data which powers AI systems could be collected in a way that prevents the development of trusted AI. Further, we believe that any comprehensive federal privacy legislation in the United States should include provisions prohibiting the use of personal data to discriminate on the basis of protected characteristics").

29 *See* AI EO at Sec. 2(f)("Artificial Intelligence is making it easier to extract, re-identify, link, infer, and act on sensitive information about people's identities, locations, habits, and desires. Artificial Intelligence's capabilities in these areas can increase the risk that personal data could be exploited and exposed."); Sec. 9.

30 *See* The White House, Readout of White House Listening Session on Tech Platform Accountability (Sept. 8, 2022) [hereinafter "Readout of White House Listening Session"], https://www.whitehouse.gov/briefing-room/statements-releases/2022/09/08/readout-of-white-house-listening-session-on-tech-platform-accountability.

AI ACCOUNTABILITY CHAIN

Source: NTIA

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,"[31] and NTIA has published a Request for Comment for the purpose of informing that report.[32]

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.

---

31 AI EO at Sec. 4.6.

32 National Telecommunications and Information Administration, Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights, 89 Fed. Reg. 14059 (Feb. 26, 2024), https://www.federalregister.gov/documents/2024/02/26/2024-03763/dual-use-foundation-artificial-intelligence-models-with-widely-available-model-weights.
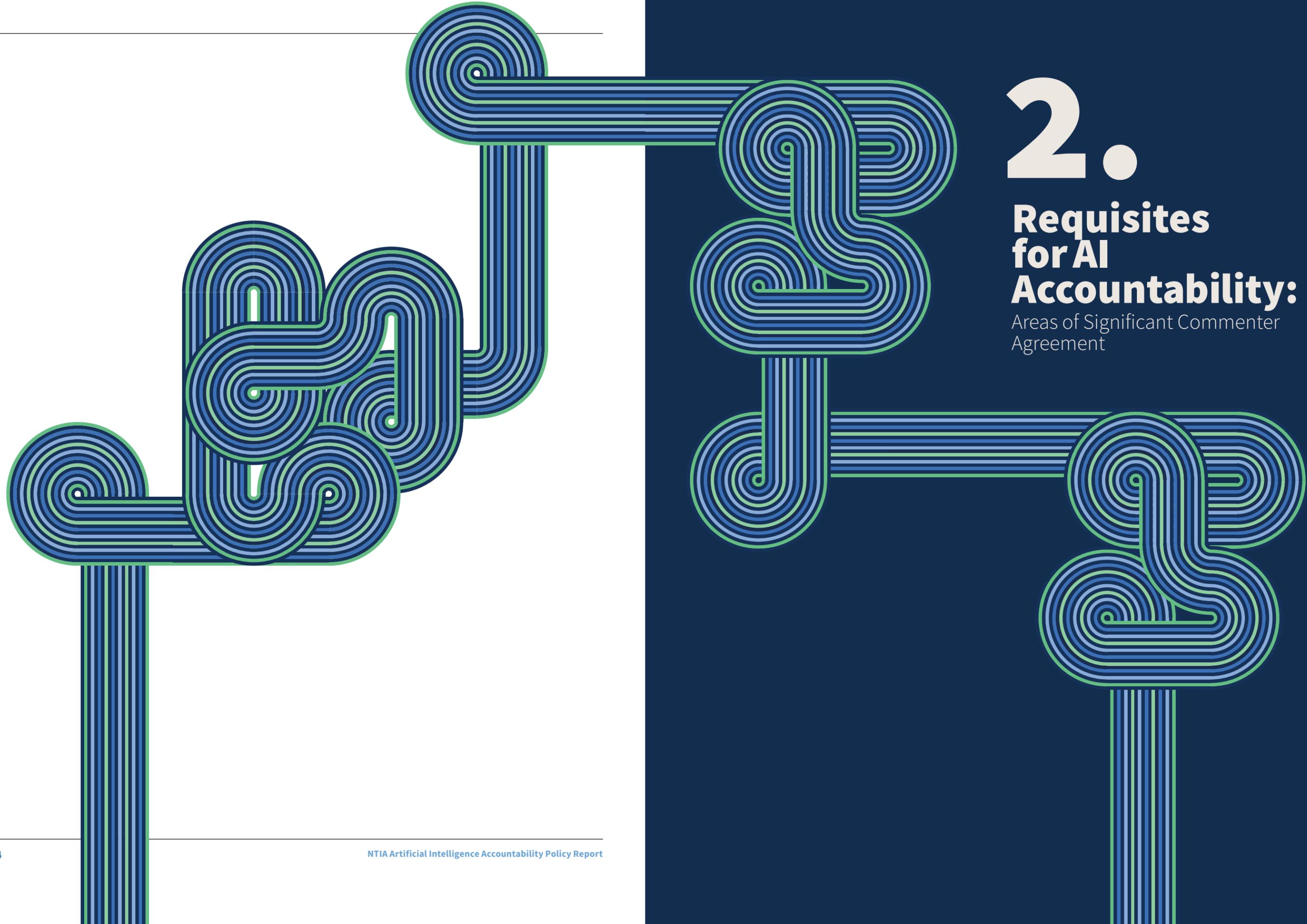
2.

Requisites for AI Accountability:

Areas of Significant Commenter Agreement
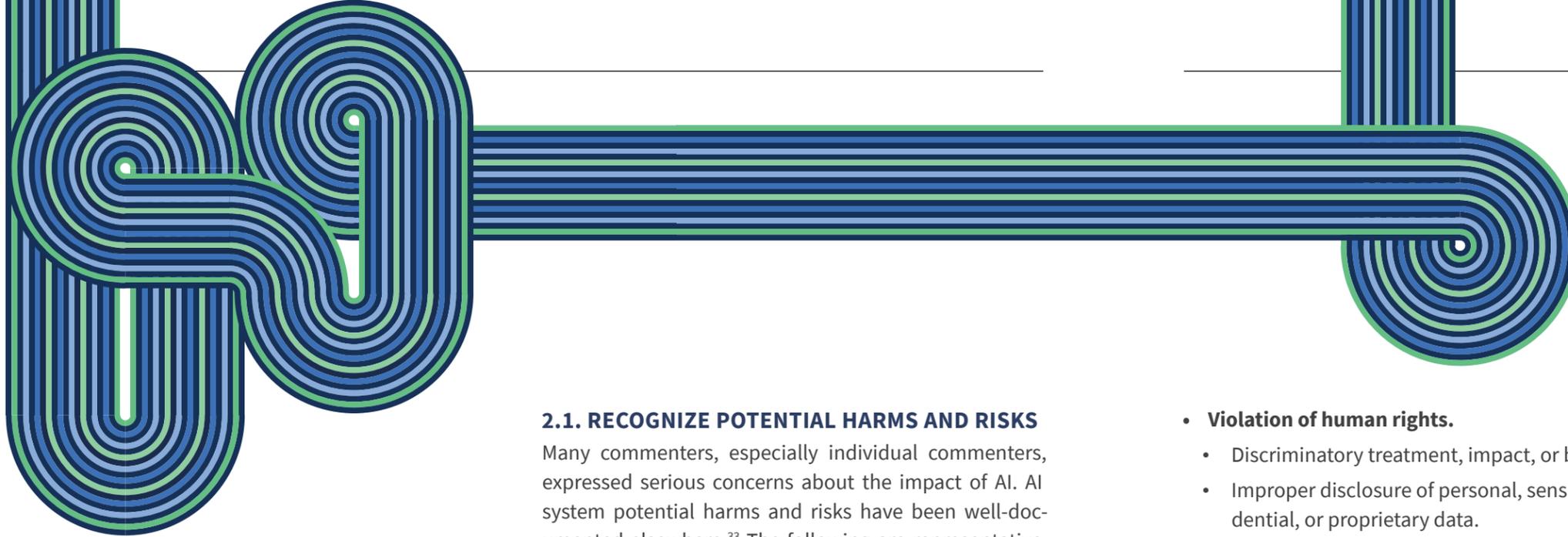
Requisites for AI Accountability: Areas of Significant Commenter Agreement

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.[33] 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.

Individual commenters reflected misgivings in the American public at large about AI.[34] 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[35] 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.[36]

- 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.

---

33   Many of these risks are recognized in the AI EO, the AIBoR, and in the Office of Management and Budget, Proposed Memorandum for the Heads of Executive Departments and Agencies, "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence" (Nov. 2023), https://ai.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-Public-Comment.pdf at 24-25.

34   *See* Alec Tyson and Emma Kikuchi, Growing Public Concern About the Role of Artificial Intelligence in Daily Life, Pew Research Center (Aug. 28, 2023), https://www.pewresearch.org/short-reads/2023/08/28/growing-public-concern-about-the-role-of-artificial-intelligence-in-daily-life/.

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." AI EO at Sec. 3(p).

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.

Potential AI system risks and harms inform NTIA's consideration of accountability measures. AI system developers and deployers should be responsible for managing the risks of their systems. As AI systems multiply and diffuse into society and the marketplace, customers, workers, consumers, and those affected by AI need assurance that these systems work as claimed and without causing harm. This is especially important for high-risk systems that are rights-impacting or safety-impacting.

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.[37] As many acknowledge, existing work from NIST, the Organization for Economic Cooperation and Development (OECD), the Global Partnership 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.[38]

NTIA concludes that a tiered approach to AI accountability has the benefit of scoping expectations and obligations proportionately to AI system risks and capabilities. As discussed below, many commenters argued that safety-impacting or rights-impacting AI systems deserve extra scrutiny because of the risks they pose of causing serious harm. Another kind of tiering ties AI accountability expectations to how capable a model or system is. Commenters suggested that highly capable models and systems may deserve extra scrutiny, which could include requirements for pre-release certification and capability disclosures to government.[39] This kind of tiering approach is evident, for example, in the AI EO requirement that developers of certain "dual-use foundation models" more capable than any yet released would have to make disclosures to the federal government.[40]

2.3. ENSURE ACCOUNTABILITY ACROSS THE AI LIFECYCLE AND VALUE CHAIN

Various actors in the AI value chain exercise different degrees and kinds of control throughout the lifecycle of an AI system. Upstream developers design and create AI models and/or systems. Downstream deployers then deploy those models and/or systems (or use the models as part of other systems) in particular contexts. The downstream deployers may also fine tune a model, thereby acting as downstream developers of the deployed systems. Both upstream developers and downstream deployers 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. Some models and systems should not be deployed until they have been independently evaluated.[41]

At the same time, there are also good reasons to vest accountability with AI system *deployers* because context and mode of deployment are important to actual AI system impacts.[42] Not all risks can be identified pre-deployment, and downstream developers/deployers may fine tune AI systems either to ameliorate or exacerbate dangers present in artifacts from upstream developers. Actors may also deploy and/or use AI systems in unintended ways.

Recognizing the fluidity of AI system knowledge and control, many commenters argued that accountability should run with the AI system through its entire lifecycle and across the AI value chain,[43] lodging responsibility with AI system actors in accordance with their roles.[44] This value chain of course includes actors who may be *neither* developers nor deployers, such as users, and many others including vendors, buyers, evaluators, testers, managers, and fiduciaries.

> **Not all risks can be identified pre-deployment, and downstream developers/deployers may fine tune AI systems either to ameliorate or exacerbate dangers present in artifacts from upstream developers. Actors may also deploy and/or use AI systems in unintended ways.**

Just as AI actors share responsibility for the trustworthiness of AI systems, we think it clear from the comments that they must share responsibility for providing accountability inputs. As part of the chain of accountability, there should be information sharing from upstream developers to downstream deployers about intended uses, and from downstream deployers back to upstream developers about refinements and actual impacts so that systems can be adjusted appropriately. Mechanisms discussed below such as adverse AI incident reports, AI system audits, public disclosures, and other forms of information flow and evaluation could all help with allocations of responsibility for trustworthy AI – allocations that will require attention and elaboration elsewhere.

2.4. DEVELOP SECTOR-SPECIFIC ACCOUNTABILITY WITH CROSS-SECTORAL HORIZONTAL CAPACITY

The application of sector-specific laws, rules, and enforcement obligations are being considered by government agencies and courts in the context of AI systems. Regulatory agencies are determining their powers to evaluate and demand information about some AI systems from the earliest stages of design.[45] Commenters thought that additional accountability mechanisms should be tailored to the sector in which the system is deployed.[46] AI deployment in sectors such as health, education, employment, finance, and transportation involve particular risks, the identification and mitigation

---

37   *See, e.g.,* University of Illinois Urbana-Champaign School of Information Sciences Researchers (UIUC) Comment at 8 ("…tiered systems match an AI system's risk with an appropriate level of oversight… The result is a more tailored and proportionate regulation of fast evolving AI systems…"); Przemyslaw Grabowicz et al., Comment at 11 ("AI systems represent too many applications for a single set of rules. Just as different FDA restrictions are applied to different medications, AI controls should be tailored to the application."); Institute of Electrical and Electronics Engineers (IEEE) Comment at 13 ("When the integrity level increases, so too does the intensity and rigor of the required verification and validation tasks); AI & Equality Comment at 3 ("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 AI system is being deployed…"); Palantir Comment at 7 (appropriate accountability mechanisms depends on the AI use context and risk profile); Securities Industry and Financial Markets Association (SIFMA) Comment at 4 (focus auditing on high-risk AI application such as "hiring, lending, insurance underwriting, and education admissions"); Bipartisan Policy Center (BPC) Comment at 2 (urging risk-based accountability systems); NCTA Comment at 6; Consumer Technology Association Comment at 2; Centre for Information Policy Leadership at 4; Workday Comment at 1; Adobe Comment at 7; BSA | The Software Alliance at 2; Intel Comment at 5-7; Developers Alliance Comment at 6; Salesforce Comment at 4; Guardian Assembly Comment at 12-14; American Property Casualty Insurance Association Comment at 2; Samuel Hammond, Foundation for American Innovation Comment at 2; Anan Abrar Comment at 1.

38   Guardian Assembly Comment at 12.

39   *See, e.g.,* Center for AI Safety Comment Appendix A – A Regulatory Framework for Advanced Artificial Intelligence (proposing regulatory regime for frontier models that would require pre-release certification around information security, safety culture, and technical safety); OpenAI Comment at 6 (considering a requirement of pre-deployment risk assessments, security and deployment safeguards); Microsoft Comment at 7 (regulatory framework based on the AI tech stack, including licensing requirements for foundation models and infrastructure providers); Anthropic at 12 (confidential sharing of large training runs with regulators); Credo AI Comment at 9 (Special foundation model and large language model disclosures to government about models and processes, including AI safety and governance); Audit AI Comment at 8 ("High-risk AI systems should be released with quality assurance certifications based on passing and maintaining ongoing compliance with AI accountability regulations."); Holistic AI Comment at 9 (high-risk systems should be released with certifications).

40   AI EO at Sec. 4.2(i).

41   *See* Office of Management and Budget, Proposed Memorandum for the Heads of

Executive Departments and Agencies, "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence" (Nov. 2023), at 16, https:// ai.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-Public-Comment. pdf [hereinafter "OMB Draft Memo"]. Some commenters focused particularly on pre-release evaluation for emergent risks. *See, e.g.,* ARC Comment at 8 ("It is insufficient to test whether an AI system is capable of dangerous behavior under the terms of its intended deployment. Thorough dangerous capabilities evaluation must include full red-teaming, with access to fine-tuning and other generally available specialized tools."); SaferAI Comment at 2 (Some of the measures that AI labs should conduct to help mitigate AI risks are: "pre-deployment risk assessments; dangerous capabilities evaluations; third-party model audits; safety restrictions on model usage; red-teaming").

42   *See, e.g.,* Center for Data Innovation Comment at 7 ("[R]egulators should focus their oversight on operators, the parties responsible for deploying algorithms, rather than developers, because operators make the most important decisions about how their algorithms impact society.").

43   NIST AI RMF, Second Draft, at 6 Figure 2 (Aug. 18, 2022) (describing the AI lifecycle in seven stages: planning and design, collection, and processing of data, building and training the model, verifying and validating the model, deployment, operation and monitoring, and use of the model/impact from the model), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

44   *See, e.g.,* ARC Comment at 8 (suggesting that because an AI system's risk profile changes with actual deployments "[i]t is insufficient to test whether an AI system is capable of dangerous behavior under the terms of its intended deployment…"); Boston University and University of Chicago Researchers Comment at 1 ("mechanisms for AI monitoring and accountability must be implemented throughout the lifecycle of important AI systems…"); *See also* Center for Democracy & Technology (CDT) Comment at 26 ("Pre-deployment audits and assessments are not sufficient because they may not fully capture a model or system's behavior after it is deployed and used in particular contexts"). *See also, e.g.,* Murat Kantarcioglu Comment (individual comment suggesting that "AI accountability mechanisms should cover the entire lifecycle of any given AI system").

45   *See, e.g., supra* note 11.

46   *See, e.g.,* MITRE Comment at 17 ("The U.S. should rely on existing sector-specific regulators, equipping them to address new AI-related regulatory needs."); HR Policy Association (HRPA) Comment at 4 (policymakers should "align, when possible, any new guidelines or standards for AI with existing government policies and commonly adopted employer best practices"); Jonhson & Johnson Comment at 2 (recommending "regulatory approaches to AI that are contextual, proportional and use-case specific"); SIFMA Comment at 5 (supporting a "flexible, and principles-based approach to third-party AI risk management, with the applicable sectoral regulators providing additional specific requirements as needed" similar to cybersecurity and pointing to NYDFS Part 500.11(a) as instructive); Morningstar, Inc. Comment at 1-3 (financial regulations apply to AI systems); Intel Comment at 3 (identifying existing sectoral laws that apply to AI harms); Ernst and Young Comment at 11 (uniformity of accountability requirements might not be practical across sectors or even within the same sector); *see also, e.g.,* Eric Schmidt Comment (arguing in an individual comment that "AI accountability should depend on business sector.").

of which often requires sector-specific knowledge. At the same time, there is risk in every sector, and cross-sectoral risks are present in both foundation models and specialized AI systems deployed in unintended contexts. Not every sectoral oversight body currently has sufficient AI sociotechnical expertise to define and implement accountability measures in all instances. The record surfaces interest in developing federal governmental capacity to address AI system impacts and coordinate governmental responses across sectors.[47]

We think it is likely that agencies will need additional capacities and possibly authorities 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 responsibilities related to AI accountability. This 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.[48] Independent external third-party evaluations (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.[49] 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,[50] while others thought that voluntary commitments to audits or other independent evaluations would suffice.[51] There were also plenty of commenters in between, with one noting that "a healthy policy ecosystem likely balances mandatory accountability mech-

anisms where risks demand it with voluntary incentives and platforms to share best practices."[52]

We believe that 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.[53]

2.6. STANDARDIZE EVALUATIONS AS APPROPRIATE

Commenters noted the importance of using standards to develop common criteria for evaluations.[54] The use of standards in evaluations is important to implement replicable and comparable 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.[55] In other words, AI ac-

tors may have to prioritize risks and values. The record surfaced interest in having additional governmental guidance for AI actors on how to address such tradeoffs.[56]

In NTIA's view, more research is necessary to create common (or at least commonly legible, comparable, and replicable) evaluation methods. Therefore, standards development is critical, as recognized in the AI EO, which tasks "the Secretary of Commerce, in coordination with the Secretary of State," with leading "a coordinated effort with key international partners and with standards development organizations, to drive the development and implementation of AI-related consensus standards, cooperation and coordination, and information sharing."[57]

> **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 collaboration 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.[58] In

---

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 agency like NIST); Boston University and University of Chicago Researchers Comment at 3 (to enable existing sectoral authorities "to work most effectively and to ensure attention to generalizable risks of AI, we recommend establishment of a meta-agency with broad AI-related expertise (both technical and legal) which would develop baseline regulations regarding the general safety of AI systems, set standards, and enable review for compliance with substantive law, while collaborating with and lending its expertise to other agencies and lawmakers as they consider the impact of AI systems on their regulatory jurisdiction"); Credo AI at 5 (recommending that government "establish dedicated oversight of the procurement, development, and use of AI. . . . [and] consider the creation of a new independent Federal agency or a Cabinet-level position with oversight authority of AI systems."); USTelecom Comment at 6 ("When individuals see that AI systems in different sectors are held to the same expectations, it assures them that adequate safeguards are in place to protect their rights and well-being, regardless of the company deploying AI."); Salesforce Comment at 9 (AI rules should have a strong degree of horizontal consistency while recognizing that "some sectoral use cases will require different treatment based on the underlying activity."); Center for American Progress (CAP) Comment at 12-13, 20 (highlighting the value of a distinct government body); Microsoft, Governing AI: A Blueprint for the Future (May 25, 2023), https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw [hereinafter "Governing AI"], at 20-21 (endorsing a new regulator to implement an AI licensing regime for foundation models); Public Knowledge Comment at 2 ("We prescribe a hybrid approach of reliance on our sector specific regulators, already deeply embedded in the domains that matter to us most, to avert immediate and anticipated harms, while also cultivating new expertise with a centralized AI regulator that can adapt with the technology and provide a broader view of the full ecosystem."); The Future Society Comment at 13 ("We are concerned that a lack of horizontal regulation in the US could perpetuate a regulatory vacuum and 'race-to-the-bottom' dynamics among [general-purpose AI system] developers, as they increasingly develop technologies that can pose risks to public health, safety, and welfare in an unregulated environment."); see also The National Security Commission on Artificial Intelligence, Final Report (2021), https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf, Chapter 9 (proposing the creation of a new "Technology Competitiveness Council.").

48 See infra Sec. 3.2.4.

49 AI Accountability RFC, 88 Fed. Reg. at 22436. As discussed in the RFC, "[i]ndependent audits may range from 'black box' adversarial audits conducted without the help of the audited entity to 'white box' cooperative audits conducted with substantial access to the relevant models and processes."

50 Anthropic Comment at 10 (recommending mandatory adversarial testing of AI systems before release through NIST or researcher access); Anti-Defamation League (ADL) Comment at 11, 12 ("Public-facing transparency reports, much like the reports required by California's AB 587, could require information on policies, data handling practices, and training or moderation decisions while prioritizing user privacy and without revealing sensitive or identifying information"); PricewaterhouseCoopers, LLP (PWC) Comment at 8 ("[W]e recommend mandatory disclosure of third-party assurance or an explanation that no AI accountability work has been performed"); AFL-CIO Comment at 5 (advocating mandatory audits); Data & Society Comment at 8 (advocating a mandatory AI accountability framework); Accountable Tech, AI Now, and EPIC, Zero Trust AI Governance Framework at 4 (Aug. 2023), https://accountabletech.org/wp-content/uploads/Zero-Trust-AI-Governance.pdf ("It should be clear by now that self-regulation will fail to forestall AI harms. The same is true for any regulatory regime that hinges on voluntary compliance or otherwise outsources key aspects of the process to industry. That includes complex frameworks that rely primarily on auditing – especially first-party (internal) or second-party (contracted vendors) auditing – which Big Tech has increasingly embraced. These approaches may be strong on paper, but in practice, they tend to further empower industry leaders, overburden small businesses, and undercut regulators' ability to properly enforce the letter and spirit of the law.").

51 Developers Alliance Comment at 12, 13; R Street Comment at 10-12; Consumer Technology Association Comment at 5; U.S. Chamber of Commerce Comment at 10; Business Roundtable Comment at 5 ("[P]olicymakers should incentivize, support and recognize good faith efforts on the part of industry to implement Responsible AI and encourage self-assessments by internal teams"); OpenAI Comment at 2 (advocates for voluntary commitments "on issues such as pre-deployment testing, content provenance, and trust and safety").

52 DLA Piper Comment at 12.

53 See Rumman Chowdhury, Submitted Written Testimony for Full Committee Hearing of the House of Representative Committee on Science, Space, and Technology: Artificial Intelligence: Advancing Innovation Towards the National Interest (July 22, 2023), https://republicans-science.house.gov/_cache/files/6/8/68b1083c-d768-4982-a8f9-74b0e771a2bc/E551A6FE9CEB156D4DE626417352ED0E.2023-06-22-dr.-chowdhury-testimony.pdf, at 1, 2 ("It is important to dispel the myth that 'governance stifles innovation'. […] I use the phrase 'brakes help you drive faster' to explain this phenomenon - the ability to stop a car in dangerous situations enables us to feel comfortable driving at fast speeds. Governance is innovation.").

54 See, e.g., Center for Audit Quality (CAQ) Comment at 7 ("[W]e believe that it is important 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 ("If definitions and methods were standardized, audits would be more consistent and lead to more confidence. This will also be necessary if third party certifications are included in future regulations.").

55 See NIST AI RMF at Sections 2 and 3. See also Google DeepMind Comment at 8-9 (suggesting there are tradeoffs between data minimization and the accuracy of systems; transparency and model accuracy; and transparency and security); OpenMined Comment at 4 (noting that if "an auditor obtains access to underlying data, privacy, security, and IP risks are significant and legitimate."); Mastercard Comment at 3 ("There can be tension between accountability goals that lead to technical tradeoffs, and we believe organizations are best suited to evaluate these tradeoffs and document related decisions. […] Transparency is another example of an AI accountability goal that can be in tension with countervailing interests. Several federal legislative and regulatory proposals contemplate or include transparency provisions. While transparency is a cornerstone in trustworthy AI, it must be balanced with the need to ensure intellectual property and proprietary information remain protected, and that malicious actors are not encouraged to bypass AI-powered protections such as fraud prevention."); Kathy Yang Comment at 3 ("There is a tradeoff between more complete data and other priorities like privacy and security").

56 See, e.g., Credo AI at 3 (recommending development of a "taxonomy of AI risk to inform the areas that are most important for an AI developer or deployer to consider when assessing its AI system's potential impact"); AI Policy and Governance Working Group Comment at 3-4 (calling for AI evaluations that consider risks drawn from a regularly evaluated and updated risk taxonomy developed by the "research and policy communities).

57 AI EO at Sec. 11(b). See also id. at Sec. 4.1(a)(i) (tasking the Secretary of Commerce with establishing "guidelines and best practices, with the aim of promoting consensus industry standards, for developing and deploying safe, secure and trustworthy AI systems."); NIST, U.S Leadership in AI: A Plan for Federal Engagement in Developing AI Technical Standards and Related Tools, https://www.nist.gov/artificial-intelligence/plan-federal-ai-standards-engagement.

58 See, e.g., ARC Comment at 9 ("To faithfully evaluate models with all of the advantages that a motivated outsider would have with access to a model's architecture and parameters, auditors must be given resources that enable them to simulate the level of access that would be available to a malign actor if the model architecture and parameters were stolen."); AI Policy and Governance Working Group Comment at 3 ("Qualified researchers and auditors who meet certain conditions should be given model-and-system framework access."). See also, e.g., Alex Leader Comment at 2-3 ("While inputs to audits or assessments, such as documentation, data management, and testing and validation, are essential, these must be accompanied by measures to increase auditors' and researchers' access to AI systems."); Olivia Erickson, Zachary

addition, it will frequently (but not always) be necessary to include associated software and technical artifacts to enable running and evaluating the model in its functional environment. Evaluators may also need information about governance processes within an entity, such as how decisions around AI system design, development, deployment, testing, and modification are made and what controls are in place throughout the AI system lifecycle to provide credible assurance of trustworthiness. Commenters identified the inability to gain access to AI system components as one of the chief barriers to AI accountability; what is needed are systems that can provide appropriate access for eligible evaluators and researchers, while controlling for access-related risks.

This Report identifies a role for government in facilitating appropriate researcher and other independent evaluator access to AI system components through tools that exist or must be developed. Part of this work is to clarify necessary levels of access and safeguards.

2.8. STANDARDIZE 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.**[59]

Commenters also addressed the value of producing information about the inputs to and source of AI-generated content, also known as "provenance."[60]

NTIA agrees with commenters that appropriate transparency around AI systems is critical.[61] Information should be pushed out to stakeholders in form and scope appropriate for the audience and risk level.[62] Communications to the public should be in plain language. Transparency-oriented artifacts such as datasheets, model cards, system cards, technical reports, and data nutritional labels are promising and some should become standard industry practice as accountability inputs.[63] Another type of information – provenance – can inform people 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.

60 *See, e.g.,* Coalition for Content Provenance and Authenticity Comment; Witness Comment; International Association of Scientific, Technical and Medical (STM) Publishers Comment at 4.

61 *See* NIST AI RMF at 15 ("Meaningful transparency provides access to appropriate levels of information based on the stage of the AI lifecycle and tailored to the role or knowledge of AI actors or individuals interacting with or using the AI system. By promoting higher levels of understanding, transparency increases confidence in the AI system. This characteristic's scope spans from design decisions and training data to model training, the structure of the model, its intended use cases, and how and when deployment, post-deployment, or end user decisions were made and by whom. Transparency is often necessary for actionable redress related to AI system outputs that are incorrect or otherwise lead to negative impacts.").

62 *See, e.g.,* (noting that AI Accountability legislation would need to account or, among other things, different risk profiles and have "[d]isclosure requirements for consumer facing AI systems[.]") ("If there is AI legislation, it should be risk-based, have disclosure requirements for consumer facing AI systems…"); CDT Comment at 41-42 (noting that CDT's "Civil Rights Standards for 21st Century Employment Selection Procedures" provide for different responsibilities for developers and deployers and different disclosures to deployers and to public); Google DeepMind Comment at 11 (AI accountability disclosures should include topline indication of how the AI system works, including "general logic and assumptions that underpin an AI application." It is "good practice to highlight the inputs that are typically the most significant influences on outputs… [and any] inputs that were excluded that might otherwise have been reasonably expected to have been included (e.g., efforts made to exclude gender or race)").

63 *See, e.g.,* AI Policy and Governance Working Group Comment at 8-9 (citing Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, "Model Cards for Model Reporting," *FAT* '19:* Proceedings of the Conference on Fairness, Accountability, and Transparency, at 220-229 (Jan. 2019), https://doi.org/10.1145/3287560.3287596) (at minimum model cards "should include the 'reporting' components of each of the principles in the technical companion of the AIBoR and reflect best practices for the documentation of the machine learning lifecycle"); Campaign for AI Safety Comment at 3 ("AI labs and providers should be required to publicly disclose the training datasets, model characteristics, and results of evaluations.").

Fox, and M Eifler Comment at 1 ("Companies building large language models available for use in commercial applications that meet any of the following criteria should be required to allow a third-party to audit the sources of their data, storage, and use. Specific regulatory guidance should be written with scaling requirements that become more intensive relative to the size of the company (by revenue) or use.").

59 *See* NIST AI RMF at 15, Sec. 3.4 (recommending this documentation as part of transparency).

2.9. FUND AND FACILITATE GROWTH OF THE ACCOUNTABILITY ECOSYSTEM

Commenters noted that there currently is not an adequate workforce to conduct AI system evaluations, particularly given the demands of sociotechnical inquiries, the varieties of expertise entailed, and supply constraints on the relevant workforce.[64] 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.[65] Another concern of commenters was that auditors can become captured by the auditees who hire them.[66]

Recognizing possible deficiencies in the supply, resources, and independence of AI evaluators, NTIA favors more federal support for independent auditing and red-teaming.[67] 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. Com-

menters recommended additional federal funding and/or support for more AI safety research, standards development, the release of standardized datasets for testing, and professional development for auditors.[68] 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.[69] They urged federal procurement reform, as the National Artificial Intelligence Advisory Committee recommended,[70] in order to drive

68 *See, e.g.,* Protofect Comment at 9-10 ("Governments could fund the development of AI auditing standards and infrastructures. …[and] can create incentive programs for businesses to incorporate ethical and accountable practices in their AI systems. This could include tax breaks, grants, or recognition programs for businesses that demonstrate leadership in AI accountability"); Guardian Assembly Comment at 10-11 (focus on incentives (grants, public recognition, staff training incentives); U.S. Chamber of Commerce Comment at 11 (fund STEM education related to AI to increase public trust through NSF); Center for Security and Emerging Technology (CSET) Comment at 13 ("Alongside standards for the audit process itself, standards should include provisions on data access, confidentiality and 'revolving door' policies that prevent auditors from working in the industry for a number of years"); BigBear. ai Comment at 24 ("Government bodies can establish regulatory frameworks that promote transparency and require the provision of data necessary for accountability assessments").

69 Future of Privacy Forum (FPF) Comment at 5 ("NTIA should support the creation of an AI Assessment & Accountability Sandbox to test, assess, and develop guidance for organizations *seeking* to apply existing rules to novel AI technologies and comply with emerging AI regulations."); Credo AI Comment at 5 ("For commercial systems, Credo AI recommends creating an 'assurance sandbox' where commercial entities can use an iterative process for guideline development with limited indemnity to start. This 'assurance sandbox' would trial transparency and mitigation requirements (using voluntary guidelines) with non-financial consequences for violations - essentially a 'safe harbor' for sincere mitigation efforts."); Centre for Information Policy Leadership Comment at 31-32 ("Regulatory sandboxes are important mechanisms for regulatory exploration and experimentation as they provide a test bed for applying laws to innovative products and services in the AI field."); Stanford Institute for Human-Centered AI (Dr. Jennifer King) Comment at 3 (proposing "regulatory sandboxes for piloting many of these proposed mechanisms to ensure they provide measurable and meaningful results."); Engine Advocacy Comment at 10-11 ("Regulatory sandboxes allow businesses and regulators to cooperate to create a safe testing ground for products or services. In simple terms, sandboxes allow real-life environment testing of innovative technologies, products, or services, which may not be fully compliant with the existing legal and regulatory framework."); American Legislative Exchange Council (ALEC) Comment at 10 ("This sandbox framework, already adopted and successful in states like Arizona and Utah, offers a way for regulators to support domestic AI innovation by permitting experimentation of new technologies in controlled environments that would otherwise violate existing regulations."); Chegg Comment at 4; Business Roundtable Comment at 12. *See also* Government of the United Kingdom, Department for Science, Innovation & Technology, Office for Artificial Intelligence, A Pro-Innovation Approach to AI Regulation (Command Paper Number 815), at Sec. 3.3.4 (August 3, 2023), https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper ("Regulatory sandboxes and testbeds will play an important role in our proposed regulatory regime.").

70 *See, e.g.,* National Artificial Intelligence Advisory Committee, Report of the National Artificial Intelligence Advisory Committee (NAIAC), Year 1, at 16-17 (May 2023), https://www.ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf ("OMB could guide agencies on the procurement process to ensure that contracting companies have

64 *See, e.g.,* Anthropic Comment at 3 ("[R]ed teaming talent currently resides within private AI labs."); International Association of Privacy Professionals (IAPP) Comment at 2 ("[S]ubstantial gap between the demand for experts to implement responsible AI practices and the professionals who are ready to do so…").

65 *See* infra Section 3.3.

66 *See, e.g.,* Center for Democracy and Technology Comment at 28 ("auditing firms may be subject to capture by providers since providers may be reluctant to retain auditors that conduct truly independent and rigorous audits as compared to those who engage in more superficial exercises").

67 *See* OMB Draft Memo at 22.
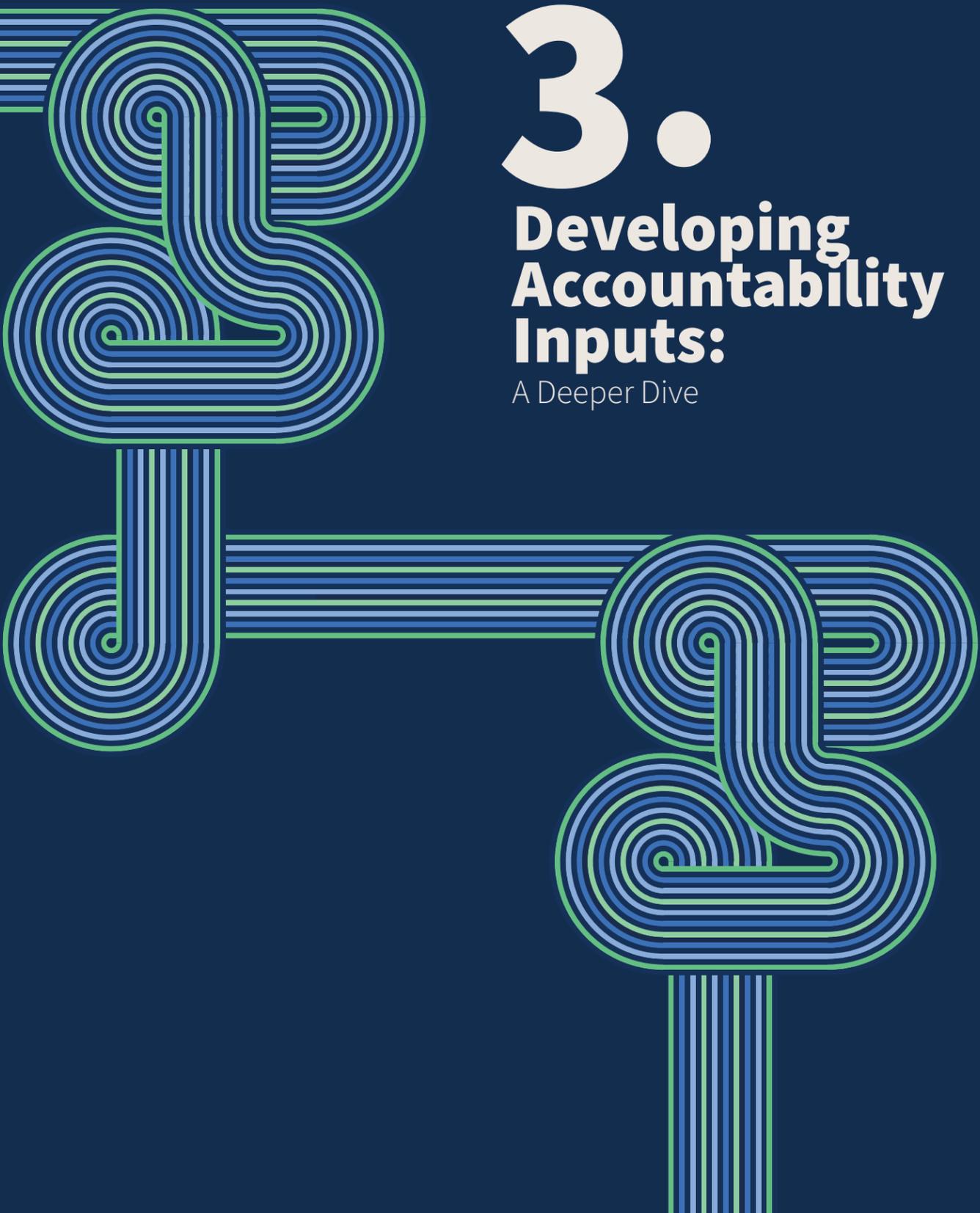
3.
Developing Accountability Inputs:
A Deeper Dive

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.[71] 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.[72]

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.[73]

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.[74]

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.

---

adopted the AI RMF or a similar framework to govern their AI"); Governing AI, *supra* note 47, at 11.

71 *See, e.g.,* AI Policy and Governance Working Group Comment at 6-7 ("A practical mechanism to consider broadly across the whole of the Federal government would be the uptake and application of a Department of Defense procurement vehicle for an independent evaluator to be procured simultaneously with a contract for an AI tool or system, thus building in a layer of accountability with the necessary infrastructure and funding"); AFL-CIO Comment at 7 (Procurement policies should ensure that AI systems do not harm workers by maintaining good data governance practices and giving workers input on impact assessments); Copyright Clearance Center (CCC) Comment at 4 ("Public procurement should require that companies building and training AI systems maintain adequate records" including management of metadata); Governing AI, *supra* note 47 at 11 (supporting a requirement that "vendors of critical AI systems to the U.S. Government to self-attest that they are implementing NIST's AI Risk Management Framework... the U.S. Government could insert requirements related to the AI Risk Management Framework into the Federal procurement process for AI systems"); CDT Comment at 36-37.

72 *See, e.g.,* AI Policy and Governance Working Group Comment at 6 (advocating government "credentialling auditors"); Centre for Information Policy Leadership Comment at 26 (advocating requiring auditor certification for audits of high-risk applications); PWC Comments at A12 (opining that "[t]he lack of AI laws and regulations requiring adherence to specified standards, reporting, and audits is a further impediment to creation of an environment of true AI accountability" and contrasting this situation with federal involvement in financial auditing).

73 *See, e.g.,* Barry Friedman et al., "Policing Police Tech: A Soft Law Solution," 37 Berkeley Tech L.J. 701, 742 (2022) (submitted as part of the Policing Project at New York University School of Law's comment) ("And like an FDA drug label, tech certification labels could come with warnings about the potential risks of any non-certified uses"), and at 706 ("[A] certification scheme, ([like] a "Rated R" movie, "Fair Trade" coffee, or an "Energy Star" appliance), could perform a review of a technology's efficacy and an ethical evaluation of its impact on civil rights, civil liberties, and racial justice"); Grabowicz et al., Comment at 1 ("[S]imilar mechanisms are used to enforce vehicle safety standards, which in turn encourage car manufacturers to offer better safety features"). *See also, e.g.,* Mark Vickers Comment (individual comment advocating "Borrow[ing] Principles from the Food and Drug Administration").

74 *See, e.g.,* Data & Society Comment at 7; Google DeepMind Comment at 3; Global Partners Digital Comment at 15; Hitachi Comment at 10; TechNet Comment at 4; NCTA Comment at 4-5; Centre for Information Policy Leadership Comment at 1; Access Now Comment at 3-5; BSA | The Software Alliance Comment at 12; U.S. Chamber of Commerce Comment at 9 (need federal privacy protection); Business Roundtable Comment at 10 (supporting a passage of a federal privacy/consumer data security law to align compliance efforts across the nation); CTIA Comment at 1, 4-7; Salesforce Comment at 9.

Disclosures, Documentation, Access

Evaluations, Audits, Red Teaming

Developing Accountability Inputs: A Deeper Dive

Downstream deployers 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.

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.[75] Congress continues to consider relevant legislative initiatives, and the states are actively pursuing their own legislative agendas.[76] 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 systems appropriately in context and to communicate system features to others. For example, an employer relying on an AI system to assist in hiring decisions might need to know if the population data used to train the system are sufficiently aligned with its own applicant pool and how underlying assumptions have been designed to guard against bias.[77]

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.[78]

Individuals affected by, or consuming, AI outputs may not even be aware that an AI system is at work, much less *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. Lack of information about AI system vulnerabilities and potential harms can also expose investors to risk.[79]

Transparency, explainability, and interpretability can all be helpful to assess the trustworthiness of a model and the appropriateness of a given use of that model. These features of an AI system involve communicating about what the system did (transparency), how the system made its decisions (explainability), and how one can make sense of system outputs (interpretability).[80] All three are part of information flow, as is information regarding the organizational and governance processes involved in designing, developing, deploying, and using models.

It is clear that information flow is a critical input to AI accountability.[81] Provisions to ensure appropriate information flow, including through accessible and plain language formats, are featured in the White House Vol-

untary Commitments,[82] in the Blueprint for AIBoR,[83] and in the AI EO.[84] Similarly, the OECD Principles for Responsible Stewardship of Trustworthy AI state that AI actors "should commit to transparency and responsible disclosure regarding AI systems."[85] Specifically, these actors

> **"[S]hould provide meaningful information, appropriate to the context and consistent with the state of art: i) to foster a general understanding of AI systems, ii) to make stakeholders aware of their interactions with AI systems including in the workplace, iii) to enable those affected by an AI system to understand the outcome, and, iv) to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision."[86]**

Commenters 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

---

[75] AI Accountability RFC at 22435. *See, e.g.,* EPIC Comment at 5-8; Salesforce Comment at 8-11. *See also* Anna Lenhart, Federal AI Legislation: An Analysis of Proposals from the 117th Congress Relevant to Generative AI Tools, The George Washington University Institute for Data, Democracy, and Politics (June 14, 2023), https://iddp.gwu.edu/federal-ai-legislation; European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons 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 (May 4, 2016), http://data.europa.eu/eli/reg/2016/679/oj.

[76] *See supra* notes 13 and 14.

[77] The Institute for Workplace Equality Comment, Artificial Intelligence Technical Advisory Committee Report on EEO and DEI&A Considerations in the Use of Artificial Intelligence in Employment Decision Making, at 46 ("The fundamental issue of model drift is that some underlying assumption about the data used to train an algorithm has changed. Applicants differ from incumbents; applicant characteristics shift over time; or the job requirements themselves change, leading to different response patterns, demographic compositions, or performance standards… the applicant population often differs from the original incumbent population due to selection effects, and the algorithm should be adjusted when enough applicant data are collected and the applicants are hired so that the criterion data are available"); HRPA Comment at 2 ("[A] failure to guard against harmful bias in talent identification algorithms could undermine efforts to create a skilled and diverse workforce."); Workday Comment at 2 ("When an AI tool is used for a decision about an individual's access to an essential opportunity, it has the potential to pose risks of harm to that individual. AI frameworks should therefore focus on these kinds of consequential decision tools, which may be used to hire, promote, or terminate an individual's employment.").

[78] The information gap between developers and deployers may be particularly large

in the case of foundation models. *See* Information Technology Industry Council (ITI), Understanding Foundation Models & The AI Value Chain: ITI's Comprehensive Policy Guide (Aug. 2023), https://www.itic.org/documents/artificial-intelligence/ITI_AIPolicyPrinciples_080323.pdf, at 7 ("A deployer (sometimes also called a provider) is the entity that is deciding the means by and purpose for which the foundation model is ultimately being used and puts the broader AI system into operation. Deployers often have a direct relationship with the consumer. While developers are best positioned to assess, to the best of their ability, and document the capabilities and limitations of a model, deployers, when equipped with necessary information from developers, are best positioned to document and assess risks associated with a specific use case.").

[79] *See, e.g.,* Open MIC Comment at 5 ("Without information about how companies 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 8 ("To engender investor confidence in AI, government intervention is needed to… increase transparency on how AI models are being trained and deployed.").

[80] NIST AI RMF at 16-17.

[81] *See* Guardian Assembly Comment at 4 ("Transparency in AI is about 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 behavior."); Anthropic Comment at 17 ("Accountability requires a commitment to transparency and a willingness to share sensitive details with trusted, technically-proficient partners"); Jeffrey M. Hirsch, "Future Work," 2020 U. Ill. L. Rev. 889, 943 (2020) ("The lack of transparency in most AI analyses is a serious cause for concern. Because AI learns through complicated, iterative analyses of data, the bases for a program's decision-making is often unclear. This lack of transparency, often referred to as the "black box" problem, could act as a mask for discrimination or other results that society deems unacceptable.").

[82] *See* First Round White House Voluntary Commitments at 4 (committing to publishing "reports for all new significant model public releases within scope [which reports] should include the safety evaluations conducted (including in areas such as dangerous capabilities, to the extent that these are responsible to publicly disclose), significant limitations in performance that have implications for the domains of appropriate use, discussion of the model's effects on societal risks such as fairness and bias, and the results of adversarial testing conducted to evaluate the model's fitness for deployment.").

[83] Blueprint for AIBoR at 6 (framing transparency in terms of "notice and explanation": "You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you. Designers, developers, and deployers of automated systems should provide generally accessible plain language documentation including clear descriptions of the overall system functioning and the role automation plays, notice that such systems are in use, the individual or organization responsible for the system, and explanations of outcomes that are clear, timely, and accessible.").

[84] AI EO passim.

[85] OECD, Recommendation of the Council on Artificial Intelligence, Section 1.3 (2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; *See also* OECD AI Policy Observatory, OECD AI Principles Overview, https://oecd.ai/en/ai-principles.

[86] Id.

[87] *See, e.g.,* Google DeepMind Comment at 8-9; Georgetown University Center for Security and Emerging Technology Comment at 5 (noting tradeoffs between privacy and transparency and between fairness and accuracy); DLA Piper Comment at 12 ("More transparency about system logic/data may improve contestability but infringe on privacy and intellectual property rights. …The more transparent a model is[,] the more susceptible it is to bad actor manipulation."); International Center for Law & Economics (ICLE) Comment at 10 ("Surely there will be many cases where firms use their own internal data, or data not subject to property-rights protection at all, but where exposing those sources reveals sensitive internal information, like know-how or other trade secrets. In those cases, a transparency obligation could have a chilling effect.").

about documentation and disclosure details and standardization. For example, some wanted the adoption of common standards.[88] Others emphasized the need for audience-specific disclosures[89] or domain-specific reporting.[90] While acknowledging that distinct regimes are probably appropriate for different AI use cases, this Report addresses generic features of information creation, collection, and distribution desirable for a wide swath of AI systems (with additional recommendations for high-risk models and systems).

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 protections. This Report 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 cross-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 discloser'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)[96] provide salient information about the data on which the AI model was trained, including the "motivation, composition, collection process, [and] recommended uses" of the dataset.[97] Several commenters recommended that AI system developers produce datasheets.[98]

AI Nutrition Facts

Your Product Name

Description	
Describe your product	
Privacy Ladder Level	1
Feature is Optional	Yes
Model Type	Generative
Base Model	OpenAI - GPT-4
Trust Ingredients	
Base Model Trained with Customer Data	No
Customer Data is Shared with Model Vendor	No
Training Data Anonymized	N/A
Data Deletion	Yes
Human in the Loop	Yes
Data Retention	30 days
Compliance	
Logging & Auditing	N/A
Guardrails	N/A
Input/Output Consistency	Yes
Other Resources	
Add any additional resources…	

Learn more about this label at **nutrition-facts.ai**

Model cards disclose information about the performance and context of a model, including:[99]

- 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, instrumentation, 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.[100] 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.[101]

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.[102]

Model cards and system cards are often accompanied by lengthier **technical reports** describing the training and capabilities of the system.[103]

Many AI system developers have begun voluntarily releasing these artifacts.[104] The authors of such artifacts often state that they are written to conform to the recom-

---

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

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

90   *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 review and understand how AI decisions are affecting teaching and learning.").

91   PWC Comment at A8. *See also* id. at 13 ("Standardized reporting — including references to the agreed trustworthy AI framework, elucidation of the evaluation criteria, and articulation of findings — would help engender public trust."); Ernst & Young Comment at 10 ("Standardized reporting should be considered where practical"); Greenlining Institute (GLI) at 3 ("AI accountability mechanisms could look like 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 the key metrics and definitions in those tests - such as how fairness or an adverse decision are defined and shared with regulators and academia."); CDT Comment at 50 ("The government should take steps that set an expectation of transparency around the development, deployment, and use of AI. In higher-risk settings, such as where algorithmic decision-making determines access to economic opportunity, that may include transparency requirements").

92   NIST AI RMF at 15-16.

93   Governing AI, *supra* note 47, at 23.

94   *See* European Commission, *supra* note 22.

95   *See* Evelyn Douek, "Content Moderation as Systems Thinking," 136 Harv. L. Rev. 526, 572-82 (2022) (discussing platform transparency reports as "transparency theater.").

96   *See* Timnit Gebru, et al., "Datasheets for Datasets," Communications of the ACM, Vol 64, No. 12, at 86-92 (Dec. 2021), https://doi.org/10.1145/3458723. *See also* Google DeepMind Comment at 24 (referring to "data cards"); Hugging Face Comment at 5 (referring to "Dataset Sheets and Data Statements"); Stoyanovich Comment at 10-11 (referring to the "Dataset Nutrition Label project"); Centre for Information Policy Leadership Comment at 9 (referring to "data set nutrition labels").

97   Id., Gebru, et al., at 87.

98   *See, e.g.,* GovAI Comment at 11; Google DeepMind Comment at 24; Bipartisan Policy Center Comment at 7; Centre for Information Policy Leadership Comment at 13; Data & Society Comment at 8.

99   Adapted from Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, "Model Cards for Model Reporting," FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, at 220-229 (Jan. 2019), https://doi.org/10.1145/3287560.3287596.

100 *See* Nekesha Green et al., System Cards, A New Resource for Understanding How AI Systems Work, Meta AI (Feb. 23, 2022), https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/ ("Many machine learning (ML) models are typically part of a larger AI system, a group of ML models, AI and non-AI technologies that work together to achieve specific tasks. Because ML models don't always work in isolation to produce outcomes, and models may interact differently depending on what systems they're a part of, model cards — a broadly accepted standard for model documentation — don't paint a comprehensive picture of what an AI system does. For example, while our image classification models are all designed to predict what's in a given image, they may be used differently in an integrity system that flags harmful content versus a recommender system used to show people posts they might be interested in.").

101 OpenAI Comment at 4 ("We believe that in most cases, it is important for these documents to analyze and describe the impacts of a system – rather than focusing solely on the model itself – because a system's impacts depend in part on factors other than the model, including use case, context, and real world interactions. Likewise, an AI system's impacts depend on risk mitigations such as use policies, access controls, and monitoring for abuse. We believe it is reasonable for external stakeholders to expect information on these topics, and to have the opportunity to understand our approach.").

102 Twilio, AI Nutrition Facts, https://nutrition-facts.ai/.

103 *See, e.g.,* Google, PaLM 2 Technical Report (2023), https://ai.google/static/documents/palm2techreport.pdf; OpenAI, GPT-4 Technical Report, arXiv (March 2023), https://arxiv.org/pdf/2303.08774.pdf. *See also* Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse, Opening up ChatGPT: Tracking Openness, Transparency, and Accountability in Instruction-Tuned Text Generators, 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 preprints and academic papers).

104 *See, e.g.,* Hugging Face Comment at 5; Anthropic Comment at 4; Stability AI Comment at 12; Google DeepMind Comment at 24.

mendations in the same paper by Margaret Mitchell,[105] which proposes a list of model card sections and details to consider providing in each one.[106] However, the actual instantiations of these artifacts vary significantly in breadth and depth of content. For instance:

- The model card annexed to the technical paper accompanying Google's PaLM-2, which is used by the Bard chatbot, discusses intended uses and known limitations. However, the card lacks detail about the training data used, and no artifact was released for the Bard chat service as of this writing.[107]

- Meta's model card for LLaMA contained details about the training data used, including specific break-downs by source (e.g., 67% from CCNet; 4.5% from GitHub).[108] However, Meta's LLaMA 2 model card contained considerably less detail, noting only that it was trained on a "new mix of data from publicly available sources, which does not include data from Meta's products or services" without describing specific sources of data.[109]

- OpenAI provided a technical report for GPT-4 that – beyond noting that GPT-4 was a "Transformer-style model pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers" – declined to provide "further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."[110]

- By contrast, BLOOMZ, a large language model trained by the BigScience project, is accompanied by a concise model card describing use, limitations, and training, as well as a detailed dataset card describing the specific training data sources and a technical paper describing the finetuning method.[111]

The above illustrates differences in approach that may or may not be justified by the underlying system. These differences frustrate meaningful comparison of different models or systems. The differences also make it difficult to compare the adequacy of the artifacts themselves and distinguish obfuscation from unknowns. For example, one might wonder whether a disclosure's emphasis on system architecture at the expense of training data, or fine-tuning at the expense of testing and validation, is due to executive decisions or to system characteristics. Like dense privacy disclosures, idiosyncratic technical artifacts put a heavy burden on consumers and users. The lack of standardization may be hindering the realization of these artifacts' potential effectiveness both to inform stakeholders and to encourage reflection by AI actors. Many commenters agree that datasheets, system cards, and model cards have an important place in the AI accountability ecosystem.[112] At the same time, a number also expressed reservations about their current effectiveness, especially without further standardization and, possibly, regulatory adoption.[113]

Whatever information is developed for disclosure, *how* it is disclosed will depend on the intended audience, which might include impacted people and communities, users, experts, developers, and/or regulators.[114] 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.[115] 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.[116] 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 level."[117]

Recognizing the shortfalls of unsynchronized disclosures among model developers, commenters largely agreed that standardizing informational artifacts and promoting comparability between them is an important goal in moving toward more effective AI accountability.[118] Several commenters called for governmental involvement in the development of these standards.[119] 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.[120] 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.[121]

The federal government could also facilitate access to disclosures as it has in other contexts, such as the SEC's Electronic Data Gathering Analysis and Retrieval (EDGAR) platform or the FDA's Adverse Events Reporting System (FAERS) platform. To the extent that NIST and others are engaged in developing voluntary transparency best practices, this is a critical first step to standardization and possible regulatory development.

3.1.2. AI OUTPUT DISCLOSURES: USE, PROVENANCE, ADVERSE INCIDENTS

Those impacted by an AI system should know when AI is being used.[122] Some commenters expressed support for disclosing the use of AI when people interact with AI-powered customer service tools (e.g., chatbots).[123] The Blueprint for AIBoR posited that individuals should know when an automated system is being used in a context that may affect that individual's rights and oppor-

---

105 Mitchell et al., *supra* note 99. Others have begun proposing similar lists of elements that should be included in AI system documentation, including in the EU AI Act. See EU AI Act, *supra* note 21, Annex IV (listing categories of information that should be included in technical documentation for high-risk AI systems to be made available to government authorities).

106 *See, e.g.,* Google, *supra* note 103 (citing Mitchell et al., *supra* note 99); OpenAI, *supra* note 103, at 40 (same); Hugo Touvron et al., Llama 2: Open Foundation and Fine-Tuned Chat Models, Meta AI (July 18, 2023), https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/, at 77 (same).

107 Kennerly Comment at 4-5; Google, *supra* note 103, at 91-93.

108 Meta Research, LLaMA Model Card (March 2023), https://github.com/facebookresearch/llama/blob/llama_v1/MODEL_CARD.md ("CCNet [67%], C4 [15%], GitHub [4.5%], Wikipedia [4.5%], Books [4.5%], ArXiv [2.5%], Stack Exchange [2%]").

109 Touvron et al., *supra* note 106, at 5.

110 OpenAI, *supra* note 103, at 2; *see also* The Anti-Defamation League (ADL) Comment at 5 ("Because there is no reporting process that requires regular or comprehensive transparency, we have little information into the decisions made via RLHF and how those decisions could negatively impact the model.").

111 *See, e.g.,* Hugging Face BigScience Project, BLOOMZ & mT0 Model Card, https://huggingface.co/bigscience/bloomz; Hugging Face BigScience Project, xP3 Dataset Card, https://huggingface.co/datasets/bigscience/xP3; Niklas Muennighoff, et al., Crosslingual Generalization through Multitask Finetuning, arXiv (May 2023), https://arxiv.org/pdf/2211.01786.pdf. *See also* Kennerly Comment at 6.

112 *See, e.g.,* Center for American Progress Comment at 8; Salesforce Comment at 7; Hugging Face Comment at 5; Anthropic Comment at 4.

113 Centre for Information Policy Leadership Comment at 5 ("Absent clear standards for such documentation efforts, organizations may take inconsistent approaches that result in the omission of key information."); U.C. Berkeley Researchers Comment at 20 ("Current practices of communication, for example releasing long 'model cards,' 'system cards,' or audit results are incredibly important, but are not serving the needs of users or affected people and communities."); Data & Society Comment at 8 (practices and frameworks for documentation and disclosure "remain voluntary, scattered, and wholly unsynchronized" without binding regulatory requirements).

114 *See, e.g.,* Hugging Face Comment at 5 (focused on "a model's prospective user"); Association for Computing Machinery (ACM) Comment at 3 (purpose of artifacts is to "enable experts and trained members of the community to understand [models]

and evaluate their impacts"); Mozilla Comments at 11 (model cards and datasheets can "help regulators as a starting point in their investigations"); CDT Comment at 23 (standardization of system cards and datasheets "can make it easier, particularly for users, to understand the information provided"); Google DeepMind Comment at 12, 24 ("Model and data cards can be useful for various stakeholders, including developers, users, and regulators," and "[w]here appropriate, additional technical information relating to AI system performance should also be provided for expert users and reviewers like consumer protection bodies and regulators"); U.S. Chamber of Commerce Comment at 3 (AI Service Cards should be designed for the "average person" to understand).

115 *See* Credo AI Comment at 8 (government should consider adopting "[t]ransparency disclosures that should be made available to downstream application developers and to the appropriate regulatory or enforcement body within the U.S. government - not the general public - to ensure they are fit for purpose"); *see also* First Round White House Voluntary Commitments at 2-3 (documenting commitments by AI developers to "[w]ork toward information sharing among companies and governments regarding trust and safety risks, dangerous or emergent capabilities, and attempts to circumvent safeguards" by "facilitat[ing] the sharing of information on advances in frontier capabilities and emerging risks and threats").

116 *See, e.g.,* Global Partners Digital Comment at 15; Salesforce Comment at 7; Stoyanovich Comment at 5, 10-11; Bipartisan Policy Center Comment at 7; Kennerly Comment at 2. C.f. 21 C.F.R. § 101.9(d) (imposing standards for nutritional labels in food).

117 ACT-IAC Comment at 14; *see also* Certification Working Group Comment at 17 (advocating for "two separate communications systems," including both "full AI accountability 'products'" and "a thoughtful summary format"); Google DeepMind Comment at 12 ("Where appropriate, additional technical information relating to AI system performance should also be provided for expert users and reviewers like consumer protection bodies and regulators.").

118 *See, e.g.,* Centre for Information Policy Leadership Comment at 5; CDT Comment at 23; Global Partners Digital Comment at 15; Stoyanovich Comment at 5.

119 *See, e.g.,* Data & Society Comment at 8; Bipartisan Policy Center Comment at 7.

120 *See* EU AI Act , *supra* note 21. Articles 40-41 (authorizing the Commission to adopt common specifications to address AI system provider obligations).

121 *See* id., Articles 10 (data and data governance), 11 (technical documentation), 13 (transparency and provision of information to users), and Annex IV (setting minimum standards for technical documentation under Article 11). *See also* European Parliament, Amendments adopted by the European Parliament on 14 June 2023 on the Artificial Intelligence Act and amending certain Union legislative acts (June 14, 2023), https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html, including additional disclosure requirements for foundation model providers under Article 28b, including a requirement to "document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law." Article 28(b)(4)(c).

122 *See, e.g.,* CDT Comment at 22-23; Adobe Comment at 4-6.

123 *See, e.g.,* Information Technology Industry Council (ITI), *supra* note 78, at 9 ("Organizations should disclose to a consumer when they are interacting with an AI system"); AI Audit Comment at 5 (recommending an "AI Identity" mark for AI chatbots and models so as to "always make it clear that the user is interacting with an AI, and not a human").

Training Material (Data, Text, Image, Etc) → AI System → Content Outputs → Content Label

- What is © status?
- What is the source?
- Is material private/sensative?
- Has use been consented to?

- Is it AI generated? (detection)
- By what system? (identification)
- From what source? (authentication)

Is provenance prominent for human consumption?

Source: NTIA

tunities.[124] Indeed, such transparency is already required by law if failure to disclose violates consumer protections.[125] In its attempt to effectuate a requirement for such notice in the employment context, New York City is now requiring that employers using AI systems in the hiring or promotion process inform job applicants and employees of such use.[126] Several states require private entities to disclose certain uses of automated processing of personal information and/or to conduct risk assessments when engaging in those uses.[127]

In addition to knowing about AI use in decision-making contexts, people should also have the information to make sense of AI outputs. As the Blueprint for AIBoR put it, people should be "able to understand when au-

dio or visual content is AI-generated."[128] One commenter argued that when products "simulate another person," they "must either have that person's explicit consent or be clearly labeled as 'simulated' or 'parody.'"[129] This is especially important in the context of AI-generated images or videos that depict an intimate image of a person without their consent, given the evidence that victims of image-based abuse experience psychological distress.[130] Commenters expressed worry about alterations to original "ground truth" content or fabrications of real-seeming content, such as deep fakes or hallucinated chatbot outputs.[131] Some commenters pointed to the particular dangers of generative AI faking scientific work and other scholarly output, and thought these merited require-

ments that systems disclose information about training data.[132]

There is a family of methods to make AI outputs more identifiable and traceable, the development of which should be a high priority and requires both technical and non-technical contributions. Recognizing this need, the AI EO tasks the Commerce Department with "develop[ing] guidance regarding the existing tools and practices for digital content authentication and synthetic content detection measures."[133] Notably, one of the objectives of the AI EO is to establish provenance markers for digital content – synthetic or not – produced by or on behalf of the federal government.

- **Provenance** refers to the origin of data or AI system outputs.[134] For training data, relevant provenance questions might be: Where does the material 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

Watermarking AI-generated content will not be easy. There is the difficulty of corralling 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.

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.[135] 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 do not change the underlying content, the content can still be reproduced without the signatures.[136] 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-

124 Blueprint for AIBoR at 6 ("[Y]ou should know that an automated system is being used and understand how and why it contributes to outcomes that impact you.").

125 *See, e.g., See* Consumer Financial Protection Bureau, Consumer Financial Protection Circular 2022-03 (May 26, 2022), https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/.

126 The New York City Council, A Local Law to Amend the Administrative Code of the City of New York, in Relation to Automated Employment Decision Tools, Local Law No. 2021/144 (Dec. 11, 2021), https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID%7CTe xt%7C&Search=.

127 *See generally* National Conference of State Legislatures, Artificial Intelligence 2023 Legislation (September 27, 2023), https://www.ncsl.org/technology-and-communication/artificial-intelligence-2023-legislation (compiling state-level AI legislation including legislation imposing disclosure, opt-out, and/or risk assessment requirements).

128 Blueprint for AIBoR at 3.

129 Salesforce Comment at 5.

130 *See, e.g.,* Nicola Henry, Clare McGlynn, Asher Flynn, Kelly Johnson, Anastasia Powell, & Adrian J. Scott, Image-Based Sexual Abuse: A Study on the Causes and Consequences of Non-Consensual Nude or Sexual Imagery at 7-15 (2021) (reporting on a study of "image-based sexual abuse").

131 *See, e.g.,* #She Persisted Comment at 3 ("Faster AI tools for election-related communication and messaging could have a profound impact on how voters, politicians, and reporters *see* candidates, campaigns and those administering elections"); International Center for Law & Economics Comment at 12 ("There are more realistic concerns that these very impressive technologies will be misused to further discrimination and crime, or will have such a disruptive impact on areas like employment that they will quickly generate tremendous harms."); Center for American Progress Comment at 5 ("Evidence of this adverse effect of AI has already started to appear: automated systems have discriminated against people of color in home loan pricing, recruiting and hiring automated systems have shown a bias towards male applicants, AI used in making health care decisions have shown a racial bias that ultimately afforded white patients more care, among other examples.")

132 *See, e.g.,* International Association of Scientific, Technical, and Medical Publishers (STM) Comment at 4 (recommending "an accounting with respect to provenance" and an "audit mechanism to validate that AIs operating on scientific content do not substantially alter their meaning and are able to provide a balanced summary of possibly different viewpoints in the scholarly literature.").

133 AI EO at Sec. 4.5. *See also* id. at Sec. 2(a) (referring to "labeling and content provenance mechanisms").

134 NIST has defined provenance in National Institute for Standards and Technology, Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy, NIST Special Publication 800-37, Rev. 2, at 104 (December 2018), https://doi.org/10.6028/NIST.SP.800-37r2 ("The chronology of the origin, development, ownership, location, and changes to a system or system component and associated data.").

135 C2PA, C2PA Explainer, https://c2pa.org/specifications/specifications/1.3/explainer/Explainer.html.

136 *See generally* Sayash Kapoor and Arvind Narayanan, How to Prepare for the Deluge of Generative AI on Social Media, Knight First Amendment Institute (June 16, 2023), https://knightcolumbia.org/content/how-to-prepare-for-the-deluge-of-generative-ai-on-social-media (criticizing the approach for being limited to those who opt-in, creating a negative space for most content which will not be authenticated).

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-produced and potentially to identify the AI system that produced it.[138] Google DeepMind, for example, has launched (in beta) its SynthID 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 com-

municating 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) and whether it is known to be 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 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 proper standard 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

137 AI EO at Sec. 3(gg).

138 Lauren Leffer, "Tech Companies' New Favorite Solution for the AI Content Crisis Isn't Enough," Scientific American (Aug. 8, 2023), https://www.scientificamerican.com/article/tech-companies-new-favorite-solution-for-the-ai-content-crisis-isnt-enough/.

139 Sven Gowal and Pushmeet Kohli, Identifying AI-generated images with SynthID, Google DeepMind (Aug. 29, 2023), https://www.deepmind.com/blog/identifying-ai-generated-images-with-synthid.

140 John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein, "A Watermark for Large Language Models," Proceedings of the 40th International Conference on Machine Learning, PMLR, Vol. 202, at 17061-17084 (2023), https://proceedings.mlr.press/v202/kirchenbauer23a/kirchenbauer23a.pdf.

141 See C2PA Comment at 4 ("Until both creator platforms and displaying mechanisms (social media, browsers, OEMs) work together to increase transparency and accountability through provenance, it will continue to be a barrier.").

142 See Leffer, supra note 138.

if open-source systems are to be able to embed watermarks and open-source applications are to be able to recognize them. Interpretive challenges 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 origination.

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 as diverse as smart chip bank cards, electronic medical records, and the V-Chip television labeling protocol. The AI EO takes a first step in promoting 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 include 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 deployers 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 publicize their use of higher risk AI applications,[147] as

> **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.**

143 First Round White House Voluntary Commitments at 3; Second Round White House Voluntary Commitments at 2-3.

144 AI EO at Sec. 4.5.

145 OECD.AI Policy Observatory, Expert Group on AI Incidents, https://oecd.ai/en/network-of-experts/working-group/10836. A beta version of a complementary project to develop a global AI Incidents Monitor (AIM), using as a starting point AI incidents reported in international media, was released in November 2023. See https://oecd.ai/en/wonk/incidents-monitor-aim.

146 AI Policy and Governance Working Group Comment at 7.

147 See, e.g., Marion Oswald, Luke Chambers, Ellen P. Goodman, Pam Ugwudike, and Miri Zilka, The UK Algorithmic Transparency Standard: A Qualitative Analysis of Police Perspectives (July 7, 2022), http://dx.doi.org/10.2139/ssrn.4155549, at 6-7 (noting that "[s]everal jurisdictions have mandated levels of algorithmic transparency for government bodies" and citing several examples); Government of Canada, Directive on Automated Decision-Making (April 2023), https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592 (requiring certain Canadian government officials to indicate that a decision will be made via automated decision systems (6.2.1.), release custom source code owned by the Government of Canada (6.2.6), and document decisions of automated decision systems (6.2.8)); Central Digital and Data Office and Centre for Data Ethics and Innovation, Algorithmic Transparency Recording Standard Hub (January 5, 2023), https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub (program through which public organizations in the United Kingdom can "provide clear information about the algorithmic tools they use, and why they're using them."); Connecticut Public Act No. 23-16 ("An Act Concerning Artificial Intelligence, Automated Decision-making, and Personal Data Privacy") (June 7, 2023) (Connecticut law requiring a publicly available inventory of systems that use artificial intelligence in the government, including a description of the general capabilities of the systems and whether there was an impact assessment prior to implementation); State of Texas, An Act relating to the creation of the artificial intelligence advisory council (H.B. No. 2060, 88th Legislature Regular Session), https://capitol.texas.gov/tlodocs/88R/billtext/pdf/HB02060F.pdf (Texas law requiring an inventory "of all automated decision systems that are being developed, employed, or procured" by state executive and legislative agencies); California Penal Code § 1320.35 (California law requiring pretrial services agencies (local public bodies) to validate pretrial risk assessment tools and make validation studies publicly available); State of California, Assembly Bill AB-302, "An act to add Section 11546.45.5 to the Government Code, relating to automated decision systems" (California Legislature, 2023-2024 Regular Session) (Chapter 800, Statutes of 2023), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB302 (California statute requiring "a comprehensive inventory of all high-risk automated decision systems that have been proposed for use, development, or procurement by, or are being used, developed or procured by, any state agency"); State of California,

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).[148] The Office of Management and Budget (OMB) has released draft guidance for federal agencies which would require them to publicly identify the safety-affecting and rights-affecting AI systems they use.[149] As one commenter noted, a national registry for high-risk AI systems could provide nontechnical audiences with an overview of the system as deployed and the actions taken to ensure the system does not violate people's rights or safety.[150] Along with a registry of systems, a government-maintained registry of professional AI "audit reports that is publicly accessible, upon request" would foster additional accountability.[151] 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 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.[152] Comment-

ers urged the government to facilitate appropriate external access to AI systems.[153] Rigorous inquiries could require access to governance controls and design decisions, access to AI system processes (for example, to run evaluator-supplied inputs through the system), as well as access to components of the model itself, accompanying software or hardware, data inputs, model outputs, and/or refinements and modifications.

The degree of access required will vary with the questions raised. For the researcher who wants to examine whether an application has produced unlawfully discriminatory outcomes, it may be enough to have input and output data (also known as a black box model access). Commenters noted that to assess the damage that could result from malign use of advanced AI, such as large language models, much more access may be required. One commenter referenced the New York Federal Reserve system of embedding a team within every major bank in New York as a model[154] and suggested that "[t]o faithfully evaluate models with all of the advantages that a motivated outsider would have with access to a model's architecture and parameters, auditors must be given resources that enable them to simulate the level of access that would be available to a malign actor if the model architecture and parameters were stolen."[155] Some 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.[156]

We note that facilitating researcher access to data from

Senate Bill SB-313, "An act to add Chapter 5.9 (commencing with Section 11549.80) to Part 1 of Division 3 of Title 2 of, the Government Code, relating to state government" (California Legislature, 2023-2024 Regular Session) (unenacted bill), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB313 (California bill that would require a state agency using "generative artificial intelligence" to communicate with a person to inform the person about the AI use); Commonwealth of Massachusetts, Bill H.64, "An Act establishing a commission on automated decision-making by government in the Commonwealth" (193rd General Court) (unenacted bill), https://malegislature.gov/Bills/193/H64 (Massachusetts bill that would create a state commission to study and make recommendations on the government use of automated decision systems "that may affect human welfare" and issue a public report to "allow the public to meaningfully assess how such system functions and is used by the state, including making technical information about such system publicly available.").

148 AI.gov, "The Government is Using AI to Better Serve the Public", https://ai.gov/ai-use-cases/.

149 OMB Draft Memo at 4.

150 See Governing AI, supra note 47, at 23.

151 AI Policy and Governance Working Group Comment at 7. See also id. (alternatively recommending that policymakers require "professional auditors to report results to regulatory authorities (similar to environmental audits), [or] require responses to recommendations made in evaluation reports within a certain time period.").

152 See Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi, Auditing Large Language Models: A Three-Layered Approach, AI Ethics (2023), at 8, https://doi.org/10.1007/s43681-023-00289-2.

153 See, e.g., OpenMined Comment at 1; Stanford Institute for Human-Centered AI Center for Research on Foundation Models Comment at 6-7 (recommending mandated researcher access to evaluate foundation models, red-teaming), mediated by provider consent and perhaps in the form of a sandbox).

154 ARC Comment at 7.

155 ARC Comment at 9. See also AI Policy and Governance Working Group Comment at 3 (The government should "mandate access to the technical infrastructure to enable varying levels of visibility into different components of (potentially) consequential AI systems"); Stanford Institute for Human-Centered AI Center for Research on Foundation Models Comment at 6-7 (recommending mandated researcher access to evaluate foundation models, mediated by deployer consent and perhaps in the form of a sandbox).

156 See, e.g., Copyright Alliance Comment at 6 ("Best practices from corporations, research institutions, governments, and other organizations that encourage transparency around AI ingestion already exist that enable users of AI systems or those affected by its outputs to know the provenance of those outputs. In particular, except where the AI developer is also the copyright owner of the works being ingested by the AI system, it is vital that AI developers maintain records of which copyrighted works are being ingested and how those works are being used, and make those records publicly accessible as appropriate (and subject to whatever reasonable confidentiality provisions the parties to a license may negotiate).").

---

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.[157] A number of commenters proposed a safe harbor from intellectual property or other liability for research into AI risks.[158]

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 controls they have in place to prevent misuse of AI systems. Application Programming Interfaces (APIs) can be used to mediate access between researchers and AI actors, thereby reducing these risks.[159]

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 discriminatory recommendations around loans 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.[160]

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.**

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 recalibrations for different system versions.[161] Such 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

---

157 The Supreme Court in a recent decision interpreted the Computer Fraud and Abuse Act to potentially narrow the circumstances under which scraping data for purposes such as researching discrimination might constitute a violation of the statute. See Van Buren v. United States, 141 S.Ct. 1648 (2021). Nevertheless, this and other cases have not fully dispelled the fears of independent researchers. See Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini, "Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem," Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), 1571–1583, 1577, https://doi.org/10.1145/3531146.3533213.

158 See infra Sec 5.1.

159 See, e.g., GovAI Comment at 8-9 (noting that a "research API should have different access tiers based on trust" and supporting "the creation of a secure research API" that would be integrated with the National AI Research Resource).

160 See, e.g., OpenMined Comment at 3; GovAI Comment at 8 (noting that "structured transparency can help balance access with security through the use of privacy enhancing technologies"); Researchers at Boston University and University of Chicago Comment at 8 (recommending that federal regulators "encourage the development and use of …privacy enhancing technologies that protect businesses' and consumers' privacy interests without compromising accountability.").

161 See, e.g., PWC Comment at A9 (documentation required for auditing include: "Information about the organization's governance structure and broader control environment…; Description of the development process, algorithm, architecture, and configuration of the model, as well as the design of controls in each respective aspect of the system; Data used to train the system and consumed by the system in its operational state; Documentation of any pre-processing steps applied to the training data; Documentation of the system's compliance with legal, regulatory, and ethical specifications; Results of testing performed throughout the development process and during the subject period; Design and results of any recalibration performed during the period; Information about the design of controls to detect emergent properties and bugs"); Audit AI Comment at 9 ("The minimum amount kept for any particular model / application pairing should be the amount necessary to retrain the model - this includes the dataset, architecture, hyperparameters, initialization, training schedule, randomization seed, and any other relevant information."); American Association of Independent Music et al. Comment at 5 ("Proper record-keeping should also include documentation about (i) the articulated purpose of the AI model itself and its intended outputs, (ii) the AI system's overall system functioning, (iii) the individual or organization responsible for the AI system (including who is responsible for the ingesting materials, who is responsible for any foundational AI model, who is responsible for any fine tuning of the AI model, who is deploying the AI system, etc.), (iv) risk assessments concerning the potential misuse and abuse of such a model, and (v) what parameters and processes were used, and what decisions were made, during the AI system development and deployment.").

ARTIFICIAL INTELLIGENCE (AI) ACCOUNTABILITY FRAMEWORK

DATA
Ensure quality, reliability, and representation of data sources and processing.

Data Used to Develop an AI System
Entities should document sources and origins of data, ensure the reliability of data, and assess data attributes, variables, and augmentation/enhancement for appropriateness.

Data Used to Operate an AI System
Entities should assess the interconnectivities and dependencies of data streams that operationalized an AI system, identify potential biases, and assess data security and privacy.

GOVERNANCE
Promote accountability by establishing processes to manage, operate, and oversee implementation.

Governance at the Organizational Level
Entities should define clear goals, roles, and responsibilities, demonstrate values and principles to foster trust, develop a competent workforce, engage stakeholders with diverse perspectives to mitigate risks, and implement an AI-specific risk management plan.

Governance at the System Level
Entities should establish technical specifications to ensure the AI system meets its intended purpose and complies with relevant laws, regulations, standards, and guidance. Entities should promote transparency by enabling external stakeholders to access information on the AI system.

MONITORING
Ensure reliability and relevance over time.

Continuous Monitoring of Performance
Entities should develop plans for continuous or routine monitoring of the AI system and document results and corrective actions taken to ensure the system produces desired 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 or may not be scaled or expanded beyond its current use.

PERFORMANCE
Produce results that are consistent with program objectives.

Performance at the Component Level
Entities should catalog model and non-model components that make up the AI system, define metrics, and assess performance and outputs of each component.

Performance at the Systemn Level
Entities should define metrics and assess performance of the AI system. In addition, entities should document methods for assessment, performance metrics, and outcomes; identify potential biases; and define and develop procedures for human supervision of the AI system.

Source Data: GAO | GAO-21-519SP

how AI actors balanced various trustworthy AI attributes. Documentation is also important for AI actors themselves in making them more reflective about impacts, for example about discriminatory system outputs. With respect to discrimination, "tracing the decision making of the human developers, understanding the source of the bias in the model, and reviewing the data" can help to identify and remedy bias.[162]

The United States Government Accountability Office (GAO) produced an AI Accountability Framework, making recommendations about both documentation and evaluations for federal agencies; this guidance could also serve other AI actors.[163] Without going into detail on the GAO Framework, it is worth noting that documentation

practices figure prominently in the guidance, including documentation on training data provenance and preparation, model performance metrics and testing, key design choices, updates, and change logs, among other things.

Commenters thought that requirements to provide information about a system should be "standard" for any AI offering.[164] Per one commenter, deployers should record "what was deployed, what changes were made between development and deployment, and any issues encountered during deployment… [and should keep] incident response investigation and mitigation procedures."[165] Another commenter proposed supply chain documen-

162 Accenture Comment at 4.

163 GAO, *supra* note 3.

164 *See, e.g.,* CDT Comment at 24 ("Accountability …requires disclosure of information such as how a system was trained and on what data sets, its intended uses, how it works and is structured, and other information that permits the intended audiences (which can include affected individuals, policymakers, researchers, and others) to understand how and why the system makes particular decisions."); IBM Comment at 4.

165 Protofect Comment at 7.

---

tation and monitoring for foundation models.[166] In general, record-keeping integrated into evaluation is the basis for "end to end" accountability.[167]

Appropriate documentation will vary by type of system. For generative AI, additional documentation may be important particularly to elucidate how training data subject to intellectual property rights figure into system outputs.[168] More stringent documentation is also useful for information integrity purposes. For example, maintaining documentation of inputs and outputs to the AI system can improve accountability for scientific communication and "be placed into a chain of evidence" as necessary for reproducible results.[169]

> **Transparency and disclosures regarding AI systems are primarily valuable insofar as they feed into accountability.**

In addition to documentation creation, there is the question of retention. Retention requirements for financial records imposed by the SEC and IRS are useful referents.[170] In general, we agree that documentation concerning the development and deployment of AI "should be retained for as long as the AI system is in development, while it is in deployment, and an additional" number of years after.[171]

3.2. AI SYSTEM EVALUATIONS

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."[173] 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."[174]

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.[175] Some comments reflected a sense of frustration with decades of self-regulation of technology that has failed to meet societal ex-

166 Stanford Institute for Human-Centered AI Center for Research on Foundation Models Comment at 4-5 (noting that, "as a direct analogy to" the Software Bill of Materials, "the federal government should track the assets and supply chain in the foundation model ecosystem to understand market structure, address supply chain risk, and promote resiliency," and that "[a]s an example illustration, Stanford's Ecosystem Graphs currently documents the foundation model ecosystem, supporting a variety of downstream policy use cases and scientific analyses").

167 *See, e.g.,* Ada Lovelace Institute Comment at 6. ("Accountability practices must occur throughout the lifecycle of an AI system, from early ideation and problem formulation to post-deployment. For example, you might layer a [data protection impact assessment] or a datasheet at the design phase, an internal audit at testing, and an audit by a third-party at (re)deployment."). *See also* Resolution Economics Comment at 3 (AI systems should be audited every time its algorithm receives a major update).

168 *See, e.g.,* CCC Comment at 2-3; Copyright Alliance Comment at 6 and 6 n.9 (discussing importance of records on training data for copyright forensics and audits).

169 STM Comment at 2 ("[W]hen applying AI in the context of scholarly communications, a record of inputs and outputs to the AI system should be maintained to ensure that the AI system and its outputs can be placed into a chain of evidence and results can be more easily reproduced, including references to scholarly works that have been used."). *See also* CCC Comment at 4 ("Without verifiable and auditable tracking of inputs, it is impossible to ensure that the resulting outputs are reliable.")

170 PWC Comment at A10 (suggesting record "retention requirements of the SEC and IRS may be an appropriate starting point" for AI).

171 *See, e.g.,* DLA Piper Comment at 24 (recommending "three years once a system is no longer in active use or development to maintain audit trails and institutional knowledge"); American Association of Independent Music et al. Comment at 5 (to "at least seven years following [an AI system's] discontinuance[.]").

172 *See, e.g.,* Generally Intelligent Comment at 4 (cautioning that disclosure requirements without consequence can be a "decoy"); Cordell Institute for Policy in Medicine & Law Comment at 2 (with reference to "[a]udits, assessments and certifications," cautioning that "[m]ere procedural tools will fail to create meaningful trust and accountability without a backdrop of strong, enforceable consumer and civil rights protections."); Mike Ananny and Kate Crawford, "*Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability,*" New Media & Society, Vol. 20, Iss. 3, at 977-982 (December 13, 2016), https://doi.org/10.1177/1461444816676645 (describing ten "[l]imits of the transparency ideal": that "[t]ransparency can be disconnected from power," "[t]ransparency can be harmful," "[t]ransparency can intentionally occlude," "[t]ransparency can create false binaries," "[t]ransparency can invoke neoliberal models of agency," "[t]ransparency does not necessarily build trust," "[t]ransparency entails professional boundary work," "[t]ransparency can privilege *seeing* over understanding," "[t]ransparency has technical limitations," and "[t]ransparency has temporal limitations").

173 National Artificial Intelligence Advisory Committee, Report of the National Artificial Intelligence Advisory Committee (NAIAC), Year 1 (May 2023) at 28, https://www.ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf.

174 Id.

175 Compare Certification Working Group Comment at 21 (recommending mandating "accountability measures" and auditor and researcher access "for high capability AI systems (those that operate autonomously or semi-autonomously and pose substantial risk of harm, including physical, emotional, economic, or environmental harms") with The American Legislative Exchange Council Comment at 8 ("voluntary codes of conduct, industry-driven standards, and individual empowerment should be preferred over government regulation in emerging technology.")

pectations for risk management and accountability.[176] At the same time, other commenters noted that audit practices (whether required or not) can result in rote checklist compliance, industry capture, and audit-washing.[177]

The scope and use of audits in accountability structures should depend on the risk level, deployment sector, maturity of relevant evaluation methodologies, and availability of resources to conduct the audits. Audits are probably appropriate for any high-risk application or model. At the very least, audits should be capable of validating claims made about system performance and limitations as well as governance controls. Where audits seek to assure a broader range of trustworthy AI attributes, they should ideally use replicable, standardized, and transparent methods. We recommend below that audits be required, regulatory authority permitting, for designated high-risk AI systems and applications and that government act to support a vigorous ecosystem of independent evaluation. We also recommend that audits incorporate the requirements in applicable standards that are recognized by federal agencies. Designating what counts as high risk outside of specific deployment or use contexts is difficult. Nevertheless, OMB has designated in draft guidance for federal agencies presumptive categories of rights-impacting and safety-impacting AI systems, while providing for exemptions depending on context.[178] This is a promising approach to creating risk buckets for AI systems generally.

3.2.1. PURPOSE OF EVALUATIONS

AI system evaluations are useful to:

- **Improve internal processes and governance;[179]**
- **Provide assurance to external stakeholders that AI systems and applications are trustworthy;[180] and**
- **Validate claims of trustworthiness.[181]**

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 AI system performing as claimed with the stated limitations? The advantage of scoping 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 unlawfully discriminatory outputs, for example.[182] Or it might go broader, focusing on governance and system architecture, but only for a small subset of objectives, such as protecting intellectual property.[183] In theory, an

evaluation can also be comprehensive, looking at governance, architecture, and applications with respect to the management of all identified risks such as robustness, bias, privacy, intellectual property infringement, explainability, and efficacy.[184]

Commenters proposed various subjects for evaluations. The following is our synthesis of the most frequent mentions:

- **System performance and impact:**
  - Verification of claims, including about accuracy, fairness, efficacy, robustness, fitness for purpose.
  - Legal and regulatory compliance.
  - Protection for human and civil rights, labor, consumers, and children.
  - Data protection and privacy.
  - Environmental impacts.
  - Security.
- **Processes:**
  - Risk assessment and management, continuous monitoring, mitigation, process controls, and adverse incident reporting.
  - Data management, including provenance, quality, and representativeness.
  - Communication and transparency, including documentation, disclosure, and explanation.
  - Human control and oversight of the AI system and outputs, as well as human fallback for individuals impacted by system outputs.
  - By-design efforts towards trustworthiness throughout the AI system lifecycle.
  - Incorporation of stakeholder participation.

We heard from many that evaluations must include perspectives from marginalized communities[185] and reflect the "inclusion of a diverse range of interests and policy needs."[186] One commenter argued that frameworks for environmental impact assessments, which "mandate public participation 'by design,'" should be considered in this context.[187]

All evaluations require measurement methodologies, which auditors are deploying in the field.[188] There are technical questions about how to test for certain harms like unlawful discrimination, including how to design the evaluation and what test data to use. What counts as problematic discrimination is a normative question that will be determined by the relevant law and norms in the domain of application (e.g., housing, employment, financial). As discussed below, the pace of standards development may lag behind the need for evaluation, in which case those conducting necessary evaluations will have to earn trust on the basis of their criteria and methodology.

Commenters thought that the type of independent evaluation called for should be pegged to the risk level of the AI system.[189] There was strong support for conducting

---

176 The AFL-CIO Technology Institute Comment at 5 ("Self-regulatory, self-certifying, or self-attesting accountability mechanisms are insufficient to provide the level of protection workers, consumers, and the public deserve. Certifications generally only determine whether the development of the AI product or service has followed a promised set of guidelines, typically established by the developer or company or industry body."); Center for American Progress Comment at 16 ("In order to get private companies to conduct these assessments and audits, mechanisms must directly impact what developers care about most and be aligned with the for-profit incentives driving their rapid technological development. For these reasons, voluntary measures are insufficient. Government action (such as formal rulemaking, executive orders, and new laws) is clearly needed; we cannot allow the Age of AI to be another age of self-regulation.").

177 Mozilla Comment at 6 ("[I]t is important to untangle incentives in the auditing ecosystem — only where the incentive structure is right and auditors are sufficiently independent (and have sufficient access) can there be more certainty that audits aren't simply conducted for the purpose of "audit-washing"; The Cordell Institute for Policy in Medicine & Law Comment at 2 (Rules built only around transparency and bias mitigation are "'AI half-measures' because they provide the appearance of governance but fail (when deployed in isolation) to promote human values or hold liable those who create and deploy AI systems that cause harm."). See also Ellen P. Goodman and Julia Trehu, "Algorithmic Auditing: Chasing AI Accountability," 39 Santa Clara High Tech L. J. 289, 302 (2023) (coining the term "audit-washing" to describe the use of weak audit criteria to effectively misrepresent AI system characteristics, performance, or risks).

178 See OMB Draft Memo at 24-25.

179 See, e.g., CAQ Comment at 6 ("Ultimately, the performance of robust risk assessment and development of processes and controls increases internal accountability and leads to improvements in the quality of information reported externally"); Ernst & Young Comment at 4 ("The value of verification schemes in the context of AI accountability can have both external and internal benefits for an organization. While they can contribute to promoting trust among external stakeholders such as customers, users and the public, they also play a role in identifying potential weaknesses in internal processes in organizations and strengthening those internal processes.");

180 See, e.g., Unlearn.AI Comment at 1; Responsible AI Institute Comment at 4; Intel Comment at 3.

181 See, e.g., Trail of Bits Comment at 1 (Audits should assess performance against verifiable claims as opposed to accepted benchmarks); PWC Comment at A1 ("[T]rust in Artificial Intelligence (AI) systems and the data that feeds them may ultimately be achieved through a two-pronged system: (1) a management assertion on compliance with the applicable trustworthy AI standard or framework and (2) third-party assurance on management's assertion.").

182 See, e.g., Salesforce Comment at 5 (recommending that impact assessments be used to counter bias in hiring); AI Audit Comment at 3-4; U.S. Equal Employment Opportunity Commission, Testimony of Suresh Venkatasubramanian (Jan. 31, 2023), https://www.eeoc.gov/meetings/meeting-january-31-2023-navigating-employment-discrimination-ai-and-automated-systems-new/venkatasubramanian (recommending that entities using AI for hiring conduct mandatory "disparity assessments to determine how their systems might exhibit unjustified differential outcomes [and] mitigate these differential outcomes as far as possible with the result of this assessment and mitigation made available for review.").

183 See, e.g., Association of American Publishers (AAP) Comment at 4-5 ("AI technologies

should be audited as to whether the material used to create the training data sets was legitimately sourced, and whether appropriately licensed from or its use authorized by the copyright owner or rights holder.").

184 Lumeris Comment at 3 (adding consideration of human fallback and governance); ForHumanity Comment at 6 (adding consideration of cybersecurity, lifecycle monitoring, human control); Holistic AI Comment at 4. See also Inioluwa Deborah Raji, Sasha Costanza-Chock, and Joy Buolamwini. "Change From the Outside: Towards Credible Third-Party Audits of AI Systems. Missing Links in AI Policy, " Missing Links in AI Policy (2022), at 8 ("AI audits can help identify whether AI systems meet or fall short of expectations, whether in terms of stated performance targets (such as prediction or classification accuracy) or in terms of other concerns such as bias and discrimination (disparate performance between various groups of people); data protection, privacy, safety and consent; transparency, explainability and accountability; adherence to standards, ethical principles and legal and regulatory requirements; or labor practices, energy use and ecological impacts.").

185 See, e.g., ADL Comment at 7 (recommending consideration of "how civil society can advise in the fine-tuning of AI data sets to ensure that AI tools account for context specific to historically marginalized groups and immediate societal risks").

186 Holistic AI Comment at 11 ("A body of interdisciplinary experts needs to collectively determine best practices, standards and regulations to ensure inclusion of a diverse range of interests and policy needs. This body should be composed of stakeholders beyond, for example, the big technology players of the private sector and large international NGOs; such stakeholders should include smaller technology companies and local civil society organizations given their frontline work with users."); Global Partners Digital Comment at 7 (the "iterative evaluation" of AI systems must include "the participation of a wide range of stakeholders, including those that are impacted by the system deployment and not only those controlling the system."); AI & Equality Comment at 6-7 (discussing stakeholder involvement); #ShePersisted Comment at 8-10 (women who are targeted by gender-based violence online should be represented in establishing evaluations for AI systems); Ada Lovelace Institute Comment at 5 ("The long history of environmental impact assessments (emerging under the US NEPA) in policy offers learnings for the potential for impact assessments for AI: frameworks for EIAs mandate public participation 'by design' to improve the legitimacy and quality of the EIA and to contribute to normative goals like democratic decision-making"). See also Wesley Hanwen Deng et al., Understanding Practices, Challenge, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice, CHI '23, ACM Conference on Human Factors in Computing Systems (April 2023), at 1-18, https://doi.org/10.1145/3544548.3581026.

187 Ada Lovelace Comment at 5. See also Wesley Hanwen Deng et al., "Understanding Practices, Challenge, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice," CHI '23, ACM Conference on Human Factors in Computing Systems (April 2023), at 1-18, https://doi.org/10.1145/3544548.3581026 (showing difficulties in recruiting user auditors and conducting user-engaged audit reports).

188 See, e.g., O'Neil Risk Consulting & Algorithmic Auditing, https://orcaarisk.com/; Credo AI, https://www.credo.ai/; Eticas, https://eticas.tech/.

189 See, e.g., Responsible AI Institute Comment at 4 ("Generally, the higher the probability and magnitude of potential harms associated with an AI use case, the more likely it is

such evaluations on an ongoing basis throughout the AI system lifecycle, including the design, development, and deployment stages.[190] 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."[191]

It is generally desirable for independent evaluations to use replicable methods,[192] and to present the results in standardized formats so as to be easily consumed and acted upon.[193] But 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 depending on the sector.[194] Evaluations of foundation models, where use cases may be diverse and unpredictable, have their own challenges. Moreover, trade secret protection for information that is evaluated may make replicability difficult.

It will take time for the evaluation infrastructure to mature as the methodologies and criteria emerge.[195] One possible outcome of standardization, discussed below, would be a modular approach to evaluations, which would recognize parent standards (e.g., for examining specific processes, attributes, or risks) and then recognize additional standards as applicable to the product being audited to craft overall evaluations suitable for the relevant industry sector or type of model. 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.

> **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[196] and may be necessary for defining the methodology for certain kinds of audits.[197] Developing technical standards for emerging technologies is a core Administration objective.[198] 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.[199] 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.[200] Under-developed standards mean uncertainty for companies seeking compliance, diminished usefulness of audits, and reduced assurance for customers, government, and the public.[201]

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 three important caveats about conventional technical standards: the relative immaturity of the AI standards ecosystem, its relative non-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.[202] Leading commentary on standards for AI audits is supportive of guidelines that can be more flexible than requirements and standards that focus on processes as well as outputs.[203] 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.[204] 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.[205]

We are cognizant of the critique that non-prescriptive stances have sometimes impeded efforts to ensure that

---

190 See, e.g., Hitachi Comment at 9 (stressing the need to evaluate frequently); The Future Society Comment at 4; Global Partners Digital Comment at 4.

191 PWC Comment at A7. See also Palantir Comment at 10 (stressing process measures in the AI System development phase, including data collection practices, "access controls, logging, and monitoring for abuse").

192 See Pattrn Analytics & Intelligence, Evaluating Recommender Systems in Relation to the Dissemination of Illegal and Harmful Content in the UK (July 2023), https://www.ofcom.org.uk/__data/assets/pdf_file/0029/263765/Pattrn_Analytics_Intelligence_Final_Report.pdf, at 35.

193 See CAQ Comment at 8 ("We believe that a consistent report format is important as it allows users of the report to compare reports 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.").

194 See, e.g., MITRE Comment at 5 (use "sector regulators" to "adopt and adapt accountability mechanisms tailored to specific AI use case"); Consumer Reports Comment at 28 ("[T]he type of audit that can be executed and the extent to which a researcher is able to assess a model is highly dependent on the information they have access to.").

195 See Salesforce Comment at 4 (evaluation "tools need to be built on accepted AI definitions, thresholds, and norms that are not yet established in the United States.").

196 See MITRE Comment at 8 ("Common terminology is critical for any field's advancement as it enables every professional to represent, express, and communicate their findings in a manner that is effectively and accurately understood by their peers"); Engine Advocacy Comment at 6; Intel Comment at 5; Palantir Comment at 21-22; GovAI Comments at 9. But cf. Google DeepMind Comment at 14-15 (While recognizing that baseline definitions for AI accountability terms is good, "applying these terms is likely to vary based on the jurisdiction, sector, as well as use case, and definitions will require room to evolve as the technology changes.").

197 See, e.g., PWC Comment at A3 ("Use of the term "audit" without reference to a generally accepted body of standards fails to convey the level of effort applied, the scope of procedures performed, the level of assurance provided over the findings, or the qualifications of the provider, among other shortcomings"); Open MIC Comment at 25 ("Without mandatory standards for AI audits and assessments … there is an incentive for companies to 'social wash' their AI assessments; i.e. give investors and other stakeholders the impression that they are using AI responsibly without any meaningful efforts to ensure this"); Salesforce Comment at 11 ("If definitions and methods were standardized, audits would be more consistent and lead to more confidence."); Global Partners Comment at 16 ("The lack of measurable standards or benchmarks creates the risk of rendering impact assessments as unproductive exercises by providing an appearance of accountability but not enough to achieve it effectively"); BSA | The Software Alliance Comment at 2 ("Without common [auditing] standards, the quality of any audits will vary significantly because different audits may measure against different benchmarks, undermining the goal of obtaining an evaluation based on an objective benchmark.").

198 See The White House, United States National Standards Strategy for Critical and Emerging Technology (USG NSS CET) (May 2023), https://www.whitehouse.gov/wp-content/uploads/2023/05/US-Gov-National-Standards-Strategy-2023.pdf.

that a rigorous, independent audit will be appropriate"). See also supra Sec. 3.1.

199 See, e.g., Information Technology Industry Council (ITI), supra note 78, at 10 (citing Rishi Bommasani, Percy Liang, and Tony Lee, Language Models are Changing AI: The 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); Microsoft Comment at 12 (need investment in international AI standards to underpin an assurance ecosystem).

200 Engine Advocacy Comment at 6-7.

201 See generally Credo AI Comment at 6.

202 See, e.g., The International Organization for Standardization-International Electrochemical Commission (ISO/IEC) 23894:2023 Guidelines on risk management for AI, https://www.iso.org/standard/77304.html (containing should statements, such as "top management should consider how policies and statements related to AI risks and risk management are communicated to stakeholders"). But see ISO/IEC 17065, Requirements for bodies certifying products, processes and services, https://www.iso.org/standard/77304.html, (stating that "Interested parties can expect or require the certification body to meet all the requirements of this International Standard.…").

203 See, e.g., Raji et al, Change from the Outside, supra note 184 at 16 (recommending "standards as guidelines, not deployment checklists" and "standards for processes, not only for outcomes").

204 See CDT Comment at 28 ("Such standards will often embody policy and value judgments: standards for an audit designed to evaluate whether a system is biased, for example, may have to set forth how much variation in performance, if any, is permissible across race, gender, or other lines in order to still be considered unbiased.").

205 See NIST AI RMF at 7 (recognizing the need for guidance on risk tolerances from "legal or regulatory requirements"). See also The Center For AI and Digital Policy (CAIDP) Comment at 4 ("Credible assurance of AI systems could be through certification programs under Federal AI legislation based on …established governance frameworks" and noting that AI RMF "is voluntary which does not set adequate and appropriate incentives for accountability.").

standards respect human rights.[206] Others also worry that, as in cybersecurity, overreliance on voluntary, non-prescriptive standards will fail to create the necessary incentives for compliance.[207] One of the key ways to continue expanding standards work and to address those critiques is to build out additional participation mechanisms in the guidance and standardization process. There should be concerted efforts to include experts and stakeholders as non-prescriptive guidance comes to develop normative content and/or binding force. 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. Civil society and industry echo this sentiment, emphasizing the need for more inclusion – beyond AI actors – in crafting and assessing standards, profiles, and best practices.[208]

Accessibility of industry standards and associated development processes is one hurdle to meaningful participation by experts and stakeholders. We counted at least one AI assurance standard that cannot be viewed during its development without existing membership in ISO/IEC or access via a country's ISO national member (e.g., ANSI in the U.S.).[209] This document, and other

> **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.**

standards like it, may represent fundamental milestones in the field of AI assurance; and while development processes by established standards organizations are generally well-established and ultimately accessible with effort, we acknowledge the real financial and logistical barriers to simply browsing its emerging forms. Further, while many frameworks and documents may be free to download, many industry technical standards require a license and expenditure to view.[210] As the state-of-the-art advances, regular updates to these and other publications will impose new costs and access barriers.

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.[211] Several commenters recommended that government develop a taxonomy or hierarchy of AI risks to shape how AI actors prioritize risk.[212] Others requested government help in devising assurance methodologies that take equity and public participation seriously.[213] We note that NIST is already

---

206 *See* Corinne Cath, The Technology We Choose to Create: Human Rights Advocacy in the Internet Engineering Task Force, Telecommunications Policy, Vol. 45, No. 6 (2021), at 102144, https://doi.org/10.1016/j.telpol.2021.102144. *See also* Michael Veale, Kira Matus, and Robert Gorwa, AI and Global Governance: Modalities, Rationales, Tensions, Annual Review of Law and Social Science, Vol. 19, https://doi.org/10.1146/annurev-lawsocsci-020223-040749 (2023).

207 *See* Chung, John J. "Critical Infrastructure, Cybersecurity, and Market Failure." 96 Or. L. Rev. 441, 459-62 (2018), https://scholarsbank.uoregon.edu/xmlui/bitstream/handle/1794/23197/Chung%20final.pdf (explaining why the NIST cybersecurity framework relies on voluntary recommendations rather than prescriptive standards); Robert Gyenes, A Voluntary Cybersecurity Framework Is Unworkable- Government Must Crack the Whip, 14 PGH. J. Tech. L. & Pol'y 293 (2014), https://doi.org/10.5195/tlp.2014.146 (explaining how voluntary cybersecurity settings leads to repeated harms that could be prevented by prescriptive standards and would help to inoculate other parties from future data exploits).

208 *See, e.g.,* CDT Comment at 29; FPF Comment at 7; Leadership Conference Comment at 5; Google DeepMind Comment at 2.

209 ISO, ISO/IEC CD 42005: Information technology — Artificial intelligence — AI system impact assessment, https://www.iso.org/standard/44545.html. The public may offer comments on draft standards once those standards reach the enquiry stage; *see* ISO, Get involved, https://www.iso.org/get-involved.html.

210 At the time of writing, access to standards cited by commenters from ISO/IEC and the Institute for Electrical and Electronics Engineer Standards Association's (IEEE SA) would cost over $1,700. *See* ISO Store, https://www.iso.org/store.html (combine prices for ISO/IEC 17011:2017 Requirements for bodies providing audit and certification of AI management systems ($174); ISO/IEC 17020:2012 Requirements for the operation of various types of bodies performing inspection ($110); ISO/IEC 17021-15:2023 Requirements for bodies providing audit and certification of management systems ($48); ISO/IEC 17025:2017 General requirements for the competence of testing and calibration laboratories ($174); ISO/IEC 17065:2012 Requirements for bodies certifying products, processes and services ($174); ISO/IEC 22989:2022 Artificial intelligence concepts and terminology ($223); ISO/IEC 23894:2023 Artificial intelligence – Guidance on risk management ($148), ISO/IEC 42010:2023 Software systems and enterprise – Architecture description ($223), ISO/IEC 42006 Information technology – Artificial intelligence – Requirements for bodies providing audit and certification of artificial intelligence management systems ($74), ISO/IEC FDIS 5339 Information technology – Artificial intelligence – Guidance for AI applications ($174)); IEEE SA Standards Store, https://www.techstreet.com/ieee/standards/ieee-1012-2016?gateway_code=ieee&vendor_id=5609&product_id=1901416 (IEEE 1012-2016: Standard for System, Software, and Hardware Verification and Validation ($196). Note that the ISO/IEC prices were converted to USD from Swiss Francs and may vary over time given changing currency exchange rates.

211 *See, e.g.,* MITRE Comment at 8. *See also* ISO/IEC, last visited Jan. 18, 2024, https://www.iso.org/developing-standards.html (stating that ISO/IEC standard development usually takes roughly 3 years to develop from first proposal to publication).

212 *See, e.g.,* Credo AI Comment at 4-5, Centre for Information Policy Leadership Comment at 8, Center for American Progress Comment at 4, 12-13.

213 *See, e.g.,* Data & Society Comment at 9 (urging government research support for participatory assessments and context-dependent assessments.); Global Partners

---

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.[214] 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;[215]

- 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 ensure appropriate levels of compliance.[216]

Commenters 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.[217] 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.

> **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.**

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.[218]

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."[219] This responsibility for assuring the

---

  Digital Comment at 18 (urging government investment in the production of guidelines and best practices for "meaningful multi-stakeholder participation in the AI assessment process.").

214 NIST, NIST AI Public Working Groups, https://airc.nist.gov/generative_ai_wg.

215 *See* USG NSS CET, at 11 ("The number of standards organizations and venues has increased significantly over the past decade, particularly with respect to [critical and emerging technologies]. Meanwhile the U.S. standards workforce has not kept pace with this growth.").

216 *See, e.g.,* GovAI Comment at 5 ("[T]here are only a few individuals and organizations with the expertise to audit cutting-edge AI models.").

217 *See* Global Partners Digital Comment at 14.

218 *See generally* NIST, U.S. Leadership in AI, *supra* note 57.

219 Accountable Tech, AI Now, and EPIC, *supra* note 50, at 5 (emphasis omitted). *See also* Association for Intelligent Information Management Comment at 7 ("If an entity uses Generative AI and other high-risk products or services and cannot identify or explain the reasons behind the decision the AI system has made, that liability is and should

validity of system claims and trustworthiness should be ongoing throughout the lifecycle of the AI system.[220]

An independent certification process for some AI systems could be one way for entities to implement proof of claims and trustworthiness. According to one definition, a certification is the "process of an independent body stating that a system has successfully met some pre-established criteria."[221] Thus an independent evaluation would be a prerequisite for a certification. A voluntary certification regime for AI systems, if sufficiently rigorous and independent, could help stakeholders navigate the AI market and promote competition around trustworthy AI.[222]

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.[223] 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), usually as a way of addressing alleged catastrophic risks.[224]

However, there is also a concern that mandatory pre-release certification or licensing can hurt competition by advantaging incumbents.[225] Therefore, the benefits of requiring ex ante proof of trustworthiness have to be balanced against facilitating easy entry into the AI market.

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.[226] 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.[227] An entity's own assessment of the trustworthiness of AI systems (in development or deployment) benefits from its access to relevant material.[228] Moreover, internal evaluation practices will tend to improve management of AI risks by measuring practices against "established protocols designed to support an AI system's trustworthiness."[229] The degree to which internal evaluations move the needle on AI system performance and impacts depends on how those evaluations are communicated within the AI actor entity and how much management cares about them.

As a practical matter, internal evaluations are more mature and robust currently than independent evaluations, making them appropriate for many AI actors.[230] According to one commenter, "combining AI assessments into existing accountability structures where possible has many advantages and should likely be the default model."[231]

That said, self-assessments are unlikely to be sufficient. Independent evaluations have proven to be necessary in other domains and provide essential checks on management's own assessments. Internal evaluations are often not made public; indeed, pressure on firms to open themselves to external scrutiny may well be counter-productive to the goal of rigorous self-examination.[232] But entities evaluating themselves may be more forgiving than external evaluators. As one commenter posited, "[a]llowing developers to certify their own software is a clear conflict of interest."[233] Independence is crucial to sustain public trust in the accuracy and integrity of evaluation results and is foundational to auditing in other fields.[234]

> **Developing regulatory requirements for independent evaluations, where warranted, provides a check on false claims and risky AI, and incentivizes stronger evaluation systems.**

There are many good reasons to push for independent evaluations, as well as a number of obstacles. Independent evaluations styled as audits will require audit and auditor criteria. To the extent that auditors could be held liable for false assurance, as they are in the financial sector, one commenter thought that audits of AI systems should hew as closely as possible to a binary yes-no inquiry.[235] In the absence of consensus standards, the process may take the form of a multi-factored analysis.[236] In either case, but especially in a multi-factored evaluation, disclosure of audit scope and methodology is critical to enable comprehension, comparison, and credibility.[237] Transparency around the audit inquiry is all the more important when benchmarks are varied and not standardized, and when audits are diverse in scope and method.

Based on our review of the record and the relevant literature, we think that the following should be part of an audit, although these recommendations are by no means exhaustive. The first element stands alone for audits fashioned as claim validation or substantiation exercises. Most of the elements below align with action items contained in the NIST AI RMF Playbook.[238]

be on the entity"). *See generally* Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, Andrew Selbst, "The Fallacy of AI Functionality," Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 2022, 959–972, https://doi.org/10.1145/3531146.3533158 (discussing the burden of proof issues particularly with respect to the basic functionality of an AI system).

220 *See* CDT Comment at 26 ("Pre-deployment audits and assessments are not sufficient because they may not fully capture a model or system's behavior after it is deployed and used in particular contexts.").

221 *See* Data & Society Comment at 2.

222 *See, e.g.,* Friedman, et al., *supra* note 73, at 707 ("Certification could impose substantive ethical standards and create an incentive for vendors to compete along ethical lines.").

223 *See* Trail of Bits Comment at 5 (stating that a generalized licensing scheme targeting AI systems would impede software use because AI systems are broadly defined in such a way that is not unique from other software systems.).

224 *See, e.g.,* OpenAI Comment at 6 ("We support the development of registration and licensing requirements for future generations of the most highly capable foundation models ... AI developers could be required to receive a license to create highly capable foundation models which are likely to prove more capable than models previously shown to be safe."); Governing AI, *supra* note 47, at 20-21 ("[W]e envision licensing requirements such as advance notification of large training runs, comprehensive risk assessments focused on identifying dangerous or breakthrough capabilities, extensive prerelease testing by internal and external experts, and multiple checkpoints along the way"); Center for AI Safety Comment Appendix A (proposing a regulatory regime for "powerful" AI systems that would require pre-release certification around information security, safety culture, and technical safety); AI Policy and Governance Working Group Comment at 9 (recommending that "responsible disclosure become a prerequisite in government regulations for certifying trustworthy AI systems, aligning with practices exemplified by Singapore's AI Verify."); SaferAI Comment at 3 ("Because [general-purpose AI systems] are 1) extremely costly to train and 2) can be dangerous during training, we believe that most of the risk assessment should happen before starting

the training.") (emphasis omitted); Campaign for AI Safety Comment at 2 (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 established in the AI ecosystem, while making it hard for startups to contest their market share."); Grabowicz et al., Comment at 1 ("Overregulation (e.g., mandatory licensing to develop AI technologies) would frustrate the development of trustworthy AI, since it would primarily inhibit smaller independent AI system manufacturers from participating in AI development."); Generally Intelligent Comment at 4 (noting that requiring "at this stage" licensing of AI systems "will make it much harder for new entrants and smaller companies to develop AI systems, while its intended goals can be achieved with other policy approaches"). *See also* ICLE Comment at 18 ("The notion of licensing implies that companies would need to obtain permission prior to commercializing a particular piece of code. This could introduce undesirable latency into the process of bringing AI technologies to market (or, indeed, even of correcting errors in already-deployed products).").

226 *See, e.g.,* Holistic AI Comment at 4 ("While certifications function as public-facing documentation on, for example, a system's level of reliability and thus safety, internal assessments help to improve a system at the R&D level, directly guiding better decision-making and best practices across the conceptualization, design, development, and management and monitoring of a system"); Id. at 5 ("[I]nternal assessments of performance according to clearly delineated criteria are necessary for internal purposes as much as for providing the documentation trail (e.g. logs, databases, registers) of evidence of system performance for external independent and impartial auditing"); Responsible AI Institute Comment at 4-5 (table showing tradeoffs among different types of evaluations).

227 *See, e.g.,* IBM Comment at 3 ("All entities deploying an AI system should conduct an initial high-level assessment of the technology's potential for harm. Such assessments should be based on the intended use-case application(s), the number and context of end-user(s) making use of the technology, how reliant the end-user would be on the technology, and the level of automation. … For those high-risk use cases, the assessment processes should be documented in detail, be auditable, and retained for a minimum period of time."); Microsoft Comment at 5 ("In the context of accountability, the NIST AI RMF also highlights the value of two important practices for high-risk AI systems: impact assessments and red-teaming. Impact assessments have demonstrated value in a range of domains, including data protection, human rights, and environmental sustainability, as a tool for accountability."); Workday Comment at 1.

228 Toby Shevlane et al., Model evaluation for extreme risks, arXiv (May 24, 2023), at 6, https://arxiv.org/pdf/2305.15324.pdf. *See also* ARC Comment at 5 (Internal evaluations are necessary when entities cannot easily or securely provide sufficient access, but

then "it is critical that AI labs conducting internal audits state publicly what dangerous capabilities they are evaluating their AI models for, how they are conducting those evaluations, and what actions they would take if they found that their AI models exhibited dangerous capabilities.").

229 *See* PWC Comment at A3. *See also* Holistic AI Comment at 5 ("[I]nternal assessments of performance according to clearly delineated criteria are necessary for internal purposes as much as for providing the documentation trail (e.g. logs, databases, registers) of evidence of system performance for external independent and impartial auditing"); Responsible AI Institute Comment at 4 (certifications, audits, and assessments promote trust by enabling verification and can change internal processes).

230 For comments discussing the readiness of internal assessments vs. the immaturity of external assessment standards, *see* Information Technology Industry Council (ITI) Comment at 4-5; TechNet Comment at 3; BSA | The Software Alliance Comment at 2; Workday Comment at 1; U.S. Chamber of Commerce Comment at 2.

231 DLA Piper Comment at 9.

232 BSA | The Software Alliance Comment at 4 (noting that mandating public disclosure of internal assessments would change incentives for firms "and result in less thorough examinations that do not surface as many issues"); American Property Casualty Insurance Association Comment at 3 (public disclosure of internal assessments can inhibit full review).

233 IEEE Comment at 3-4.

234 Trail of Bits Comment at 2.

235 *See, e.g.,* ForHumanity Comment at 7.

236 *See, e.g.,* Global Partners Digital Comment at 4 ("HRIA methodologies must be adapted to best fit the needs of external stakeholders and must be responsive to the specific contexts" OR 'human rights due diligence or HRIAs critically require ensuring meaningful participation in the risk identification and comments about the impacts, its severity and likelihood, and development of harm prevention and mitigation measures from potentially affected groups and other relevant stakeholders in the context of implementation of the AI system under evaluation.); Center for Democracy & Technology Comment at 26 ("[Human rights impact assessments] are intended to identify potential impacts of an AI system on human rights ranging from privacy and non-discrimination to freedom of expression and association.").

237 *See, e.g.,* Mozilla Open Source Audit Tooling (OAT) Comment at 7; ARC Comment at 5.

238 NIST AI RMF Playbook, https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook.

- **Claim substantiation:** Is the system fit for purpose in its intended, likely, or actual deployment context? Are the processes, controls, and performance of the system as claimed?

- **Performance to acceptable risk levels with respect to all stakeholders:** Has the system mitigated risks to a sufficient degree according to independent evaluators and/or appropriate benchmarks?

- **Data quality:** Is the data used in the system's design, development, training, testing, and operation:
  - Of adequate provenance and quality;
  - Of adequate relevance and breadth; and
  - Governed by adequate data governance standards?

- **Process controls:** Are there adequate controls in the entity developing or deploying the system:
  - To ensure that worker, consumer, community and other stakeholder perspectives were adequately solicited and incorporated into the development, deployment, post-deployment review, and/or modification process;
  - 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 appropriate and sufficient documentation throughout the lifecycle of the AI system and its components to enable an evaluator to answer the previous questions?
  - Has the developer or deployer made sufficient disclosure about the use of AI, and about training data, system characteristics, outputs, and limitations, to stakeholders, including in plain language?
  - Is the AI system sufficiently interpretable and explainable that stakeholders can interrogate whether its outputs are justified?
  - Is the developer or deployer adequately contributing to an adverse incident database?

3.2.5. REQUIRED EVALUATIONS

Developing regulatory requirements for independent evaluations, where warranted, provides a check on false claims and risky AI, and incentivizes stronger evaluation systems.[239] This view is captured in a recent civil society report expressing commonly held suspicions of "any regulatory regime that hinges on voluntary compliance or otherwise outsources key aspects of the process to

industry."[240] One suggestion commenters made was that government should require internal impact assessments, rather than independent audits, for high-risk AI systems.[241] Some commenters recommended mandatory audits[242] and/or "red-teaming"[243] in the particular context of foundational models that they fear may exhibit "dangerous capabilities."

We acknowledge the arguments against audit requirements in general[244] and especially if imposed without reference to risk.[245] The arguments against required evaluations include the dearth of standards and the costs imposed especially on smaller businesses.[246] According to one commenter, the cost drivers are "technical expertise," "legal and standards expertise," "deployment and social context expertise," "data creation and annotation," and "computational resources."[247]

---

239 AFL-CIO Comment at 5 (voluntary evaluations insufficient); Farley Comment at 19 ("[M]arket incentives likely tilt towards incentivizing lax audits if there is any market effect at all," and, therefore, "government has a role to play in bolstering auditors' independence and ensuring adequate audits.); Protefoct Comment at 8 ("[T]here are few incentives for companies to conduct external audits unless required by law or demanded by their clients or partners).

240 Accountable Tech, AI Now, and EPIC, *supra* note 50, at 4. *See also* CAP Comment at 9 (citing Microsoft, Empowering responsible AI practices, https://www.microsoft.com/en-us/ai/responsible-ai) (existing "sparse patchwork of voluntary measures proposed and implemented by industry" is not sufficient). But *see* OpenAI Comment at 2 (At least "on issues such as pre-deployment testing, content provenance, and trust and safety," voluntary commitments should suffice.).

241 *See, e.g.,* BSA | The Software Alliance Comment at 2 (advocating mandatory impact assessments for both developers and deployers).

242 GovAI Comment at 9 (recommending requiring "developers of foundation models to conduct third-party model and governance audits, before and after deploying such models").

243 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 7 ("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 access.").

244 *See, e.g.,* HRPA Comment at 7-8 (There should be no third-party assessments or audits required at this time in the employment context, because "[m]ature, auditable, and accepted standards to evaluate bias and fairness of AI systems do not yet exist …" and might be overly burdensome, deepen mistrust in such systems, and potentially violate IP rights); AI Audit Comment at 2 (policy focus should be on internal assessments rather than bureaucratic checklists); Business Roundtable Comment at 12 (Government should let the industry engage in self-assessments and should not impose uniform requirements for third party assessments); Developers Alliance Comment at 12 ("AI accountability measures should be voluntary, and risk should be self-assessed"); Blue Cross Blue Shield Association Comment at 3 ("[T]hird-party audits are immature as a mechanism to detect or mitigate adverse bias"); James Madison Institute at 6; TechNet Comments at 3 (TechNet members believe that it is premature to mandate independent third-party auditing of artificial intelligence systems).

245 *See, e.g.,* Salesforce Comment at 5-6; SIFMA Comment at 4.

246 *See, e.g.,* U.S. Chamber Technology Engagement Center Comment at 10 (estimating audit costs at "hundreds of thousands of dollars."). But *see* Certification Working Group (CWG) Comment at 19 (costs are modest relative to overall development costs, and small compared to technology's impact); Protefoct Comment at 9 (costs vary widely depending on company size, data complexity, importance of AI to the product; having tiers of auditing can reduce costs).

247 HuggingFace Comment at 12.

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:

- **Create a modular governance system for AI, with a risk assessment standards board, to deduplicate costs for developing audit standards;[248]**
- **Standardize "structured transparency" such that auditors may only ask specific questions rather than obtaining all the underlying data;[249]**
- **Build on internal accountability requirements;[250] and**
- **Provide industry association or governmental compliance assistance.[251]**

3.3 ECOSYSTEM REQUIREMENTS

The supply of capable evaluators trails the pace of AI innovation. A paper produced for Google DeepMind, opines: "[i]deally there would exist a rich ecosystem of model auditors providing broad coverage across different risk areas. (This ecosystem is currently under-developed.)"[252] Research drawing on auditing experiences across sectors, including pharmaceuticals and aviation, "strongly supports training, standardization, and accreditation for third-party AI auditors."[253] Many commenters addressed this point, observing that the ecosystem for AI

assurance requires more investment, diverse stakeholder participation, and professionalization.

3.3.1. PROGRAMMATIC SUPPORT FOR AUDITORS AND RED-TEAMERS

The linchpin for robust evaluations is a supply of qualified auditors, researchers capable of doing red-teaming or other adversarial investigations, and critical personnel inside AI companies. There is now a "substantial gap between the demand for experts to implement responsible AI practices and the professionals who are ready to do so."[254] To grow the pipeline of those professionals, our evaluation of the record suggests that there should be more investment in the training of students in applied statistics, data science, machine learning, computer science, engineering, and other disciplines (perhaps including humanities and social sciences) to do AI accountability work. This training should include methods for obtaining and incorporating the input of affected communities.[255] Marketplace demand could demonstrate to motivated students that AI assurance work is in fact a viable professional pathway.

Red-teaming – the practice of outside researchers using adversarial tactics to stress test AI systems for vulnerabilities and risks – is becoming an important part of the accountability ecosystem.[256] The largest AI companies

---

248 *See* Riley and Ness Comment at 14.

249 *See, e.g.,* OpenMined Comment at 4. *See also* GovAI Comment at 9 (recommending government fund "research and development of structured transparency tools").

250 *See, e.g.,* Centre for Information Policy Leadership Comment at 31.

251 *See* Georgetown University Center for Security and Emerging Technology Comment at 15.

252 Shevlane, *supra* note 228, at 6. *See also* Databricks Comment at 2 ("The AI audit ecosystem is not mature enough to support mandatory third-party audits.").

253 Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho, "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance," AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (July 2022), at 565, 557-571, https://doi.org/10.1145/3514094.3534181.

254 IAPP Comment at 2.

255 *See, e.g.,* Cornell University Citizens and Technology Lab Comment at 2 (recommending that government fund educational projects involving citizen participation in AI accountability, possibly modeled on the EPA's program in Participatory Science for Environmental Protection as documented in U.S. Environmental Protection Agency, Office of Science Advisor, Policy and Engagement, , Using Participatory Science at EPA: Vision and Principles (June 2022), https://www.epa.gov/system/files/documents/2022-06/EPA%20Vision%20for%20Participatory%20Science%206.23.22.pdf).

256 DEF CON 2023 held a red-teaming exercise with thousands of people; *see* Hack The Future, https://www.airedteam.org/. *See also* Microsoft Comment at 3 (noting that it is "working to extend [red-teaming] beyond traditional cybersecurity assessments to

are embracing red-teaming.[257] But as one such company noted, talent is concentrated inside private AI labs, which reduces the capacity for independent evaluation.[258] 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.[259] Red-teams, like audit teams, should be diverse and multi-disciplinary in their membership and inquiries.[260] 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.[261] 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.[262]

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, security, privacy, interpretability, and equity of AI systems."[263] 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 and evaluation, and support research and education around trustworthy AI. The AI EO 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.[264] This has now launched.[265]

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."[266] Commenters vigorously endorsed supporting the NAIRR.[267] 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."[268] Others stressed that it would advance AI accountability and competition if the federal government made more datasets available to developers.[269] 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.[270]

3.3.3. AUDITOR CERTIFICATION

Another part of the AI accountability ecosystem in need of development is certification for AI system auditors,[271] which standards organizations are beginning to establish.[272] Auditors should be subject to "professional licensure, professional and ethical standards, and independent quality control and oversight (e.g. peer review and inspection)."[273] 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 [p]rofessionalism and [c]ode of [e]thics," and have "robust systems to support integrity and confidentiality of" audits

and independence.[274] Professional standards and best practices can potentially help to strengthen the integrity of audits.[275] For example, ForHumanity worked with the Partnership on Employment & Accessible Technology (PEAT) to create a Disability Inclusion and Accessibility audit certification, which trains auditors to assess AI systems for risks that could harm people with disabilities.[276] However, it is also possible that the gatekeeping of professionalization and credentials unduly narrows participation. If credentialling is too concentrated or stringent, it could artificially constrain the supply of evaluators. Whether as part of credentialling, or in its absence, transparency about audit methodology and goals may be the most important check on quality.[277]

It is relatively uncontroversial that auditor independence should be measured according to a prescribed professional standard.[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

also uncover an AI system's potential harms"); Stability.ai Comment at 15 ("DEF CON is one example of collaborative efforts to incentivize evaluation and reporting in an unregulated environment.").

257 *See, e.g.,* Google, Why Red Teams Play a Central Role in Helping Organizations Secure AI Systems (July 2023), https://services.google.com/fh/files/blogs/google_ai_red_team_digital_final.pdf.

258 *See* Anthropic Comment at 17.

259 Alan Mislove, Red-Teaming Large Language Models to Identify Novel AI Risks, The White House (August 29, 2023), https://www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/.

260 *See, e.g.,* ADL Comment at 5; Salesforce Comment at 6; Johnson & Johnson Comment at 3 ("Diversity, equity and inclusion must be considered in all aspects of AI (e.g., selecting the issues to address/problems to solve using AI, training and hiring a diverse workforce from the data scientists to programmers, attorneys, and program managers).").

261 *See* Amy Dickens and Benjamin Moore, Improving Responsible Access to Demographic Data to Address Bias, Centre for Data Ethics and Innovation Blog (June 14, 2023), https://cdei.blog.gov.uk/2023/06/14/improving-responsible-access-to-demographic-data-to-address-bias/.

262 *See, e.g.,* BSA | The Software Alliance Comment at 12; BigBear Comment at 23.

263 National Artificial Intelligence Research Resource Task Force, Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource (January 2023), at A1, https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf. *See also* id. at 33-34 (proposing a data service with curated datasets including from government), 37-39 (proposing educational resources and test beds).

264 AI EO Sec. 5.2(a)("The program shall pursue the infrastructure, governance mechanisms, and user interfaces to pilot an initial integration of distributed computational, data, model, and training resources to be made available to the research community in support of AI-related research and development.")

265 National Science Foundation, National Artificial Intelligence Research Resource Pilot, https://new.nsf.gov/focus-areas/artificial-intelligence/nairr.

266 *See* National Artificial Intelligence Research Resource Task Force, *supra* note 263, at v.

267 *See, e.g.,* Public Knowledge Comment at 14 ("The NAIRR could be a huge benefit to the development of safe, responsible, and publicly beneficial AI systems but the NAIRR needs more than the power of the purse backing it up in order to ensure that publicly-funded research and development remains publicly beneficial. Linking NAIRR resources with regulatory oversight would ensure enforcement of ethical and accountability standards and prevent public research resources from being unfairly captured for private benefit."); Google DeepMind Comment at 31; Governing AI, *supra* note 47, at 25; Software and Information Industry Association Comment at 11; UIUC Comment at 17.

268 Johnson & Johnson Comment at 2. *See also* Centre for Data Ethics and Innovation Blog, "Improving Responsible Access to Demographic Data to Address Bias," June 14, 2023. https://cdei.blog.gov.uk/2023/06/14/improving-responsible-access-to-demographic-data-to-address-bias (recommending the establishment of demographic data intermediaries or, alternatively, the use of proxy data to infer demographic data in addressing bias).

269 *See, e.g.,* Adobe Comment at 8; U.S. Chamber of Commerce Comment at 11; Kant AI Solutions Comment at 3.

270 *See, e.g.,* A 20-Year Community Roadmap for Artificial Intelligence Research in the US, Computing Community Consortium and AAAI, at 3 (August 2019), https://cra.org/ccc/wp-content/uploads/sites/2/2019/08/Community-Roadmap-for-AI-Research.pdf; National Artificial Intelligence Research Resource Task Force, *supra* note 263, at ii. *See also* Nur Ahmed & Muntasir Wahed, The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research, arXiv (Oct. 22, 2020), https://arxiv.org/abs/2010.15581.

271 *See, e.g.,* AI Policy and Governance Working Group Comment at 6 (advocating that government be involved in credentialing auditors, which could lower costs and security risks of system access).

272 ISO is developing standards, ISO/IEC CD 42001 and 42006, for integrated AI management systems and for organizations certifying and auditing those systems respectively. ISO/IEC CD 42001, Information technology — Artificial intelligence — Management system; ISO/IEC CD 42006, Information technology — Artificial intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems.

273 AICPA Comment at 2.

274 ForHumanity Comment at 5.

275 Raji et al., Outsider Oversight, *supra* note 253 at 566 ("Fears of legal repercussions or corporate retaliation can weaken the audit inquiry, and professional standards can help determine limited conditions for liability.").

276 *See* ForHumanity, FHCert, https://forhumanity.dev/cert/.

277 *See also* 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 performed").

278 *See, e.g.,* American Institute of CPAs (AICPA) Comment at 1 (recommending independent third-party assurance to apply "procedures designed to assess the credibility of the information and report on the results of their procedures"); Protofect Comment at 6 ("Calculation of risk should be determined by a 3rd party organization that can independently perform audits and give scores given multiple contexts - including security, privacy assessment, compliance, health and safety impact etc.").

279 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act), OJ L 277 (October 27, 2022), http://data.europa.eu/eli/reg/2022/2065/oj, arts. 37(1), (3).

280 *See* Digital Services Act, *supra* note 279, at art. 37(3)(a)(i).

4.
Using Accountability Inputs

Auditors should have subject–matter and assurance experience and reflect the diversity of affected stakeholders.

recommended importation of that definition into the AI context in the United States.[281] Others cautioned against too much credence being given to these or any other formal independence requirements, noting that *de jure* and actual independence may diverge as auditors can be "captured" by those who pay for their services.[282]

Auditors should have subject-matter and assurance experience and reflect the diversity of affected stakeholders.[283] Demand for people or teams qualified to conduct AI evaluations who also satisfy the most rigorous independence requirements could outstrip supply. At least in the short term, tightening the supply of qualified auditors could have cost implications.

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.[284] 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.[285]

---

281 *See* ForHumanity Comment at 5 (referencing Sarbanes-Oxley Act and also recommending that auditors be subject to oversight and held liable 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.

AI ACCOUNTABILITY CHAIN

Source: NTIA

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.

Some 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

---

286 Joint Statement on Enforcement Efforts, *supra* note 11, at 1, https://www.ftc.gov/system/files/ftc_gov/pdf/EEOC-CRT-FTC-CFPB-AI-Joint-Statement%28final%29.pdf.

287 *See, e.g.,* Complaint, FTC v. Lasarow et al. (2015), https://www.ftc.gov/system/files/documents/cases/150223avromcmpt.pdf at 4, 9 (alleging deception, where defendants claimed to use one or more mathematical algorithms to measure specific characteristics of skin moles from digital images captured by a consumer's mobile device in order to detect melanoma). The FTC eventually reached a settlement with the defendants. *See* Federal Trade Commission, "Melanoma Detection" App Sellers Barred from Making Deceptive Health Claims (August 13, 2015), https://www.ftc.gov/news-events/news/press-releases/2015/08/melanoma-detection-app-sellers-barred-making-deceptive-health-claims; Federal Trade Commission, FTC Cracks Down on Marketers of "Melanoma Detection" Apps (February 23, 2015), https://www.ftc.gov/news-events/news/press-releases/2015/02/ftc-cracks-down-marketers-melanoma-detection-apps. *See also* U.S. Department of Justice, Justice Department and Meta Platforms Inc. Reach Key Agreement as They Implement Groundbreaking Resolution to Address Discriminatory Delivery of Housing Advertisements (January 9, 2023), https://www.justice.gov/opa/pr/justice-department-and-meta-platforms-inc-reach-key-agreement-they-implement-groundbreaking (Fair Housing Act settlement requiring Facebook to change its advertisement delivery system algorithm).

288 *See, e.g.,* Final Order, In the Matter of Cambridge Analytica, LLC, FTC Docket No. 9383 (2019), https://www.ftc.gov/system/files/documents/cases/d09389_comm_final_orderpublic.pdf, at 4; Decision, In the Matter of Everalbum, FTC Docket No. C-4743 (2022) https://www.ftc.gov/system/files/documents/cases/1923172_-_everalbum_decision_final.pdf, at 5. *See also* FTC v. Ring LLC, No. 1:23-cv-1549 (D.D.C. 2023) (proposed stipulated order).

---

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, customers, users, researchers, auditors, investors, creators, manufacturers, distributors, developers, and deployers.[291] The National Cybersecurity Strategy attempted to do this in the cyber context by laying out, in broad strokes, a preferred allocation of liability and an agenda to incentivize better cybersecurity practices.[292] How AI liability should operate is an issue largely beyond the scope of this Report, and will undoubtedly 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;[294] 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 issues:

- 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 AI 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?[295]

---

289 *See* Consumer Financial Protection Bureau, Consumer Financial Protection Circular 2022-03 (May 26, 2022), https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/.

290 *See* Equal Employment Opportunity Commission, Artificial Intelligence and Algorithmic Fairness Initiative, https://www.eeoc.gov/ai; Equal Employment Opportunity Commission, Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964 (May 18, 2023), http://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial; Equal Employment Opportunity Commission, The Americans with Disabilities Act and the Use of Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees (May 12, 2022), https://www.eeoc.gov/laws/guidance/americans-disabilities-act-and-use-software-algorithms-and-artificial-intelligence.

291 *See, e.g.,* Google DeepMind Comment at 25 (stating that policymakers should "clarify[] liability for misuse/abuse of AI systems by various participants—researchers and authors, creators (including open-source creators) of general-purpose and specialized systems, implementers, and end users[.]"); Open MIC Comment at 8 ("The lack of clarity regarding liability for AI-related harms puts both investors and rights-holders at risk."); Public Knowledge Comment at 2 ("We must address uncertainty about where liability lies for AI-driven harms to ensure that stakeholders at every phase of the AI lifecycle are contributing responsibly to the overall health of our AI ecosystem."); Georgetown University Center for Security and Emerging Technology Comment at 15 ("[T]he liability of developers for harms caused by their AI models should be clarified to avoid entirely unregulated spaces."); STM Comment at 3 ("At a minimum, clarity and transparency are required in the use of IP and copyright, and as part of any liability regime. AI systems can use huge volumes of copyright materials in the training process and as part of any commercial deployment, therefore transparency obligations will be necessary to enable rights holders to trace copyright infringements in content ingested by AI systems."). Some commenters suggested that the Federal government should provide guidance on legal regimes, which could influence liability frameworks. *See, e.g.,* US Telecom Comment at 3 ("Additionally, there is a role for the Federal government to address the emerging problem of inconsistent state laws [related to AI accountability] in an economically sensible manner."); AFL-CIO Comment at 4 ("The Federal government should construct regulatory structures that preclude AI systems from being deployed if they have the potential to violate U.S. laws and regulations, undermine democratic values, violate people's rights, including labor rights and employment law). Cf. HRPA Comment at 7 ("We believe that the Federal government should coordinate its efforts to promulgate guidelines and requirements on artificial intelligence in the employment context. Where possible, we encourage NTIA to look for

292 The White House, National Cybersecurity Strategy (2023) at 21, https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Cybersecurity-Strategy-2023.pdf ("The Administration will work with Congress and the private sector to develop legislation establishing liability for software products and services. Any such legislation should prevent manufacturers and software publishers with market power from fully disclaiming liability by contract, and establish higher standards of care for software in specific high-risk scenarios.").

293 *See, e.g.,* DLA Piper Comment at 10 ("Courts shape precedent around accountability for harm and influence developer behavior through risk of liability suits or fines for issues like injuries, discrimination, violations of due process, etc.").

294 European Commission, Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (Reference COM(2022) 496), EUR-Lex (September 28, 2022), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496, at 2.

295 *See, e.g.,* Anthropic Comment at 7 ("Liability regimes that hold model developers solely responsible for all potential harms could hinder progress in AI."); Campaign for AI Safety Comment at 3-4 ("Legislators should pass laws that clarify the joint legal culpability of AI labs, AI providers and parties that employ AI for AI harms" and analogizing to "polluter pays" and manufacturer liability for product safety defects);

(continued on next column) ways to promote consistency between Federal and state efforts."). Some commenters also raised more discrete topics that might also be appropriate to consider in the context of developing clearer liability rules for harms stemming from AI systems, such as who is responsible for contributing to remedies. *See, e.g.,* Global Partners Digital Comment at 8 ("The liability regime established by the accountability regime should account for the way in which developers of foundational models and implementers should contribute to remedy in case of harm."). One commenter suggested the adoption of specific statutes imposing criminal liability for the misuse of AI. Ellen S. Podgor Comment at 1.

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.

- Are the various liability frameworks that already govern AI systems (e.g., in civil rights and consumer protection law, labor laws, intellectual property laws, contracts, etc.) sufficient to address harms or are new laws needed to respond to any unique challenges?[296]

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 liability systems?[297]

- 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?

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.[298] 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 proof, or even recognition, that an AI system led to harm can be hard to come by; thus, even if an affected party could, in theory, bring a case to remedy a harm, they may not do so because of information and knowledge barriers.[299] Accountability inputs can assist people harmed by AI to understand causal connections, and, therefore, help people determine whether to pursue legal or other remedies.[300]

As a comment from twenty-three state and territory attorneys general stated, "[b]y requiring appropriate disclosure of key elements of high-risk AI systems, individuals can be empowered to decide what systems are fair and adhere to critical due process norms."[301] AI accountability inputs can make it easier to bring cases and vindicate interests now or in the future.[302] At the same time, entities that may be on the other end of litigation (e.g., AI developers and deployers alleged to have caused or contributed to harm) can also benefit from more information flow about defensible processes.[303]

The creation of safe harbors from liability is relevant to AI accountability, whether the one sheltered in that harbor is an AI actor or an independent researcher. The Administration's National Cybersecurity Strategy, for example, recommends the creation of safe harbors in connection with new liability rules for software.[304] A small minority of commenters addressed the safe harbor issues. Some expressed doubt that safe harbors for AI actors in connection with AI system-related harms would be appropriate.[305] A number of commenters argued that researchers

---

The Future Society Comment at 12 ("Transferring absolute liability to third-party auditors would erroneously presuppose their capability to audit for novel risks. . . . Shared liability between developers, deployers, and auditors encourages all involved parties to maintain high standards of diligence, enhances effective risk management, and fosters a culture of accountability in AI development and deployment."); Global Partners Digital at 3 (arguing that "[l]iability should be clearly and proportionately assigned to the level in which those different entities are best positioned to prevent or mitigate harm in the AI system performance."); Cordell Institute for Policy in Medicine & Law Comment at 11 ("[P]olicymakers should consider vicarious liability and personal consequences for malfeasance by corporate executives"); ACT | The App Association Comment at 2 ("Providers, technology developers and vendors, and other stakeholders all benefit from understanding the distribution of risk and liability in building, testing, and using AI tools. …[T]hose in the value chain with the ability to minimize risks based on their knowledge and ability to mitigate should have appropriate incentives to do so"); Georgetown University Center for Security and Emerging Technology Comment at 1 ("Due to the large variety of actors in the AI ecosystem, we recommend designing mechanisms that place clear accountability on the actors who are most responsible for, or best positioned to, influence a certain step in the value chain"). See also Salesforce Comment at 8 ("AI developers like Salesforce often create general customizable AI tools, whose intended purpose is low-risk, and it is the customer's responsibility (i.e., the AI deployer) to decide how these tools are employed. …It is the customer, and not Salesforce, that knows what has been disclosed to the affected individual, and what the risk of harm is to the affected individual.").

296 See, e.g., Senator Dick Durbin Comment at 2 ("[W]e must also review and, where necessary, update our laws to ensure the mere adoption of automated AI systems does not allow users to skirt otherwise applicable laws (e.g., where the law requires 'intent.')"); ICLE Comment at 15 ("[T]he right approach to regulating AI is not the establishment of an overarching regulatory framework, but a careful examination of how AI technologies will variously interact with different parts of the existing legal system"); Open MIC Comment at 8 ("Legal experts are divided regarding how AI-related harms fit into existing liability regimes like product liability, defamation, intellectual property, and third-party-generated content."); CDT Comment at 33 ("The greatest challenge in successfully enforcing a claim against AI harms under existing civil rights and consumer protection laws is that the entities developing and deploying AI are not always readily recognized as entities that traditionally have been covered under these laws. This ambiguity helps entities responsible for AI harms claim that existing laws do not apply to them."); HRPA Comment at 5 ("The use of technology in the employment context is already subject to extensive regulation which should be taken into consideration when developing any additional protections. In the United States alone, Federal and state laws dealing with anti-discrimination, labor policy, data privacy,

and AI-specific issues affect the use of AI in the employment context."); Georgetown University Center for Security and Emerging Technology Comment at 10 (noting that "[p]roduct liability law provides inspiration for how accountability should be distributed between upstream companies, downstream companies and end users."); Boston University and University of Chicago Researchers Comment at 1-2 (arguing that accountability mechanisms are important for "(a) new or modified legal and regulatory regimes designed to take into account assertions, evidence and similar information provided by AI developers relevant to intended or known users of their products, and (b) existing regimes such as product liability, consumer protection, and other laws designed to protect users and others against harm.").

297 See, e.g., SaferAI Comment at 2 ("We believe that the article 28 of the EU AI Act parliament draft lays out useful foundations on which the US could draw upon in particular regarding the distribution of the liability along the value chain to make sure to not hamper innovation from SMEs, which is one of EU's primary concerns."); Association for Intelligent Information Management (AIIM) Comment at 3 ("This approach – classifying AI into different categories and establishing policy accordingly – aligns with the European Union's AI Act, which is currently working its way through their legislative processes. While AIIM is not indicating its support for this legislation nor advocating for the U.S. government to adopt similar policy, the premise is commendable."); Georgetown University Center for Security and Emerging Technology Comment at 6 ("Accountability mechanisms should make sure to clearly define what different actors in the value chain are accountable for, and what information sharing is necessary for that party to fulfill their responsibilities. For example, the EU parliament's proposal for the AI Act requires upstream AI developers to share technical documentation and grant the necessary level of technical access to downstream AI providers such that the latter can assess the compliance of their product with standards required by the AI Act."); ICLE Comment at 9-11 (criticizing the proposed EU AI Act's "broad risk-based approach.").

298 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., The Future Society Comment at 8 ("Third-party assessment and audits must not be perceived as silver bullets. . . . Furthermore, external audits, in particular, may be subject to liability-washing (companies seeking to conduct external audits with the ulterior motivation of evading liability."); Cordell Institute for Policy in Medicine & Law Comment at 3 ("Governance of AI systems to foster trust and accountability requires avoiding the seductive appeal of 'AI half-measures'—those regulatory tools and mechanisms like transparency requirements, checks for bias, and other procedural requirements that are necessary but not sufficient for true accountability."); Boston University and University of Chicago Researchers Comment at 2 ("[A]ccountability and transparency mechanisms are a necessary but not sufficient aspect of AI regulation. . . . To be effective, a regulatory approach for AI systems must go beyond procedural protections to include substantive, non-negotiable obligations that limit how AI systems can be built and deployed."). When AI transparency and system evaluations contribute additional information and knowledge that could be used to bring legal cases, the challenge may remain on how to apply legal concepts to modern use situations involving AI even when people agree a law may be applicable. See, e.g., Lorin Brennan, "AI Ethical Compliance is Undecidable," 14 Hastings Sci. & Tech. L.J. 311, 323-332 (2023) (arguing that it is "unsettled how applicable law should be applied" in the context of AI ethical compliance).

299 See, e.g., CDT Comment at 34 ("Due to the lack of transparency in AI uses, the plaintiff may not have the information needed to even establish a prima facie case. They may not even know whether or how an AI system was used in making a decision, let alone have the information about training data, how a system works, or what role it plays in order to offer direct evidence of the AI user's discriminatory intent or to discover what similarly situated people experienced due to the AI."); Public Knowledge Comment at 12 ("Unfortunately, identifying the party responsible for introducing problems into the AI system can be challenging, even though the resulting harms may be evident. While much has been written on different legal regimes and their effectiveness in addressing AI-related harms, less attention has been given to determining the specific entities in the chain of development and use who bear responsibility.").

300 See, e.g., OECD, Recommendation of the Council on Artificial Intelligence, Section 1.3 (2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Cf. Danielle Citron, "Technological Due Process," 85 Wash. U. L. Rev. 1249, 1253-54 (2008) ("Automation generates unforeseen problems for the adjudication of important individual rights. Some systems adjudicate in secret, while others lack recordkeeping audit trails, making review of the law and facts supporting a system's decisions impossible. Inadequate notice will discourage some people from seeking hearings and severely reduce the value of hearings that are held.").

301 Twenty-three Attorneys General Comment at 3. See also AI & Equality Comment at 2 ("[E]nabling AI-based systems with adequate transparency and explanation to affected people about their uses, capabilities, and limitations amounts to applying the due process safeguards derived from constitutional law in the analogue world to the digital world.").

302 See, e.g., AI & Equality Comment at 2 ("[T]ransparency and explainability mechanisms play an important role in guaranteeing the information self-determination of 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 ("[A] publicly released audit provides a measure of transparency, while transparency provides information necessary to determine whether liability should be imposed.").

303 See, e.g., AIIM Comment at 3 ("[Organizations] are reluctant to implement new technology when they do not know their liabilities, don't know if or how they will be audited or who will be auditing them, and are unclear about who may have access to their data, among other things. . . .For instance, insurance companies have had AI for years that can analyze images of crashes or other incidents to help make determinations about fault or awards, but companies have been afraid to use it out of fear of the potential liability if an AI-made decision is contested."); Public Knowledge Comment at 11 (noting that understanding liability "is especially important to ensure that harms can be adequately addressed and also so that academic researchers, new market entrants, and users can engage with AI with clarity about their responsibilities and confidence surrounding their risk."); DLA Piper Comment at 3 ("Undertaking accountability mechanisms reduces potential liabilities in the event of accidents or AI failures by showing diligent governance and responsibility were exercised."); CDT Comment at 29 ("One of the key ways of ensuring accountability is the promulgation of laws and regulations that set standards for AI systems and impose potential liability for violations. Such liability both provides for redress for harms suffered by individuals and creates incentives for AI system developers and deployers to minimize the risk of those harms from occurring in the first place.").

304 See The White House, supra note 292, at 20-21 (Strategic Objective 3.3).

305 See Senator Dick Durbin Comment at 2 ("And, perhaps most importantly, we must defend against efforts to exempt those who develop, deploy, and use AI systems from liability for the harms they cause."); Global Partners Digital Comment at 10 ("Accountability needs to be embedded throughout the whole value chain, or more specifically, throughout the entire lifecycle of the AI system. . . . [L]iability waivers do not seem appropriate, and there is a clear need for a dynamic distribution of the legal liability in case of harm.").

(or a defined class of them) and perhaps some auditors should enjoy a safe harbor from various kinds of liability in connection with bona fide efforts to evaluate AI systems.[306] Another approach related to a safe harbor is to create regulatory sandboxes for high-risk AI systems so that AI actors and regulators can learn about AI system risks in a controlled environment for a limited period of time, without unduly exposing the public to AI risks or the AI actors to regulatory risks.[307] The OECD has a workstream related to this topic.[308] A safe harbor might also be considered to facilitate safety-related information-sharing among companies. These options should be thoroughly examined with input not only from direct safe harbor beneficiaries, but also from affected individuals and communities.

> **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.**

4.2. 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.[309] Commenters discussed how regulation does or should intersect with AI systems, including the need for clarity and new regulatory tools or enforcement bodies.[310] 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.[311]

Again, without commenting on the precise structure of enforcement, we posit that regulators of all types will have an easier job enforcing law and regulations if there is greater information flow around, and better evaluations of, AI systems. As these questions are considered in many arenas and regulators more forcefully tackle AI harms, the accountability inputs addressed in this Report can help to build the records needed for sound administration and law enforcement. The same is true of the recommendations to build the accountability ecosystem, including by funding capacity within the federal government.

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.[312] 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 trustworthy AI.

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 public-private certification projects suggests that transparency and clear evaluation metrics are key to trust and adoption. To the extent applicable, certification could be based on existing metrics, frameworks, and standards developed by NIST and national or international bodies.

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.[313] 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.[314] To date, ENERGY STAR has achieved widespread adoption, leading to substantial energy and consumer savings.[315] 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 (platinum, gold, silver, or certified) based on adherence to certain environmental metrics.[316] Per USGBC, LEED projects have been adopted worldwide.[317] Programs like ENERGY STAR and LEED empower their users (e.g., individuals, businesses) to make informed choices,[318] guide regulators and lawmakers,[319] and more generally help build community trust.[320] 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.

> **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.**

---

306 *See, e.g.,* Engine Advocacy Comment at 4 (citing approvingly government safe harbor programs to encourage compliance, such as the FTC COPPA Safe Harbor Program and the HHS breach safe harbor program); Boston University and University of Chicago Researchers Comment at 8 ("[W]e encourage the enactment of legal protection for researchers *seeking* to study algorithms[.] . . ."); ACT-IAC Comment at 11 (supporting providing external auditors maximum system access, including through appropriate security clearances, coupled with "liability waivers and the ability to publish the review[s] externally—to the extent that clearance allows—to ensure transparency."); Mozilla OAT Comment at 7 ("For [data] access, external auditors need safe harbors against retaliation for the publication of unfavorable results and custom tooling for data collection."); AI Policy and Governance Working Group Comment at 3 ("The Federal government should consider the establishment of narrowly-scoped 'safe harbor' provisions for industry and researchers, designed to reasonably assure that entities participating in good faith auditing exercises are not subjected to undue liability risk or retaliation").

307 *See supra* note 69. *See also* Jon Truby, Rafael Dean Brown, Imad Antoine Ibrahim, and Oriol Caudevilla Parellada, "A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications," European Journal of Risk Regulation, Vo. 13, No. 2, at 270–94 (2022), https://doi.org/10.1017/err.2021.52 (arguing for a robust sandbox approach to regulating high-risk AI applications as a necessary complement to strict liability regulation); European Parliament, The Artificial Intelligence Act and Regulatory Sandboxes, https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf ("regulatory sandboxes generally refer to regulatory tools allowing businesses to test and experiment with new and innovative products, services or businesses under supervision of a regulator for a limited period of time. As such, regulatory sandboxes have a double role: 1) they foster business learning, i.e., the development and testing of innovations in a real-world environment; and 2) support regulatory learning, i.e., the formulation of experimental legal regimes to guide and support businesses in their innovation activities under the supervision of a regulatory authority. In practice, the approach aims to enable experimental innovation within a framework of controlled risks and supervision, and to improve regulators' understanding of new technologies.") (internal emphasis omitted).

308 OECD, "Regulatory sandboxes in artificial intelligence," OECD Digital Economy Papers, No. 356, (2023) (recommending that governments "consider using experimentation to provide a controlled environment in which AI systems can be tested and scaled up"), https://doi.org/10.1787/8f80a0e6-en. *See also* https://oecd.ai/en/wonk/sandboxes.

309 *See, e.g.,* Alex Engler, "A Comprehensive and Distributed Approach to AI Regulation," Brookings Institution (Aug. 31, 2023) ("Many agencies lack critical capacity regarding algorithmic oversight, including: the authority to require entities to retain data, code, models, and technical documentation; the authority to subpoena those same materials; the technical ability to audit [the systems]; and the legal authority to set rules for their use.").

310 *See supra* Sec. 2.4. *See also* Anthropic Comment at 19 ("Clarity on antitrust regulation would help determine whether and how AI labs can coordinate on safety standards. Sensible coordination around consumer-friendly standards *seems* possible, but regulators' guidance on the issue would be welcome.") (internal emphasis omitted); Shaping the Future of Work Comment at 7 ("These issues and impacts [related to generative AI technology] do not require that our regulatory framework start from scratch, but instead require appropriate application of existing frameworks for robots, automation, internet, and other digital technologies.").

311 *See* Joint Statement on Enforcement Efforts, *supra* note 11, at 3 ("Many automated systems are 'black boxes' whose internal workings are not clear to most people and, in some cases, even the developer of the tool. This lack of transparency often makes it all the more difficult for developers, businesses, and individuals to know whether an automated system is fair.").

312 *See, e.g.,* Public Knowledge Comment at 3 ("Transparency could involve [. . .] enabling regulators to thoroughly examine models, even when trade secrets or intellectual property laws protect them."); AI Impacts Comment at 2 (noting that "robust methods" for "evaluating AI systems and assessing risk . . . can help regulators verify safety and help AI developers build trust with other stakeholders."); Global Partners Digital Comment at 17 ("[A] central element of any accountability regime should be addressing the information asymmetries in order to enhance the external stakeholder assessment and the authority oversight of the quality of the evaluation performed of the AI system."). *See also* Mozilla OAT Comment at 7 ("Much of the regulatory requirements for internal auditors or professional audit actors is an enforcement of some degree of visibility or oversight on their internal assessment processes and outcomes, which currently remain relatively obscure to external stakeholders, including regulators and the public.").

313 ENERGY STAR, How ENERGY STAR Works, https://www.energystar.gov/about/how_energy_star_works.

314 *See* ENERGY STAR, About ENERGY STAR, https://www.energystar.gov/about. *See also* ENERGY STAR, ENERGY STAR Impacts, https://www.energystar.gov/about/impacts.

315 *See id.* ("Since 1992, ENERGY STAR and its partners helped prevent 4 billion metric tons of greenhouse gas emissions from entering our atmosphere; By choosing ENERGY STAR, a typical household can save about $450 on their energy bills each year and still enjoy the quality and performance they expect; Approximately 1,700 manufacturers and 1,200 retailers partner with ENERGY STAR to make and sell millions of ENERGY STAR certified products.").

316 *See* U.S. Green Building Council, LEED Rating System, https://www.usgbc.org/leed.

317 *See* U.S. Green Building Council, Press Room, https://www.usgbc.org/press-room (noting "more than 185,000 total LEED projects worldwide" "and "more than 185 countries and territories with LEED projects" and "more than 205,000 LEED professionals around the world."). *See also* Twenty-three Attorneys General Comment at 3-4 ("As an example of a private sector program, the [LEED] standard has spurred the move towards 'green buildings.'").

318 *See, e.g.,* ENERGY STAR, About ENERGY STAR, https://www.energystar.gov/about ("The blue ENERGY STAR label provides simple, credible, and unbiased information that consumers and businesses rely on to make well-informed decisions.") (emphasis added).

319 *See, e.g.,* The Policing Project at New York University's School of Law Comment at 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 certification, lawmakers now have an objective standard against which they can tie the development of building regulations.").

320 *See* Twenty-three Attorneys General Comment at 3-4 (referencing Energy Star and LEED in the context of "agile and dynamic public and civic initiatives that build trust and spur trusted technological changes.").

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.[321] 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.[322] We agree with the comment from twenty-three attorneys general that transparency around the evalua-

tion process is critical and certification programs should operate "through *transparent* and *verifiable* policies and practices driven by appropriate standards including a code of ethics."[323]

Establishing and promoting certification systems can further the development of a trustworthy marketplace for AI.[324] 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.[325]

---

321 Friedman et al., *supra* note 73, at 748. In particular, in the context of private certification programs of 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 advises that transparency in certification schemes themselves is crucial to building trust. Id. at 748-49.

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 "layers of accountability." Id.

323 *See* Twenty-three Attorneys General Comment at 4 (emphasis added).

324 *See* BBB National Programs Comment at 3 (noting that several of the characteristics are important in the development of a marketplace, including by bringing consistency and reducing friction). *See also* id. at 5 (arguing that "[t]his type of certification-based system with a trusted mark and standardized reporting can serve a vital role in building a trustworthy AI marketplace.") (referencing the BBB National Programs and the Center for Industry Self-Regulation's Principles for Trustworthy AI in Recruiting and Hiring and accompanying Hiring and Independent Certification Protocols for AI-Enabled Hiring and Recruiting Technologies).

325 *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 societal 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. Cf. Gary Marchant et al., Governing Emerging Technologies Through Soft Law: Lessons for Artificial Intelligence, 61 Jurametrics J. 1, 9 (2020). ("The biggest deficits of soft law programs…relate to their effectiveness and credibility. Their 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

The RFC asked what accountability policies adopted in other domains might be useful precedents for AI accountability policy. Commenters addressed this question in detail.

5.1 FINANCIAL ASSURANCE

The assurance system for financial accounting is an obvious referent for AI assurance. Some existing financial sector laws may be directly applicable to AI.[326] Otherwise, they may still furnish useful analogies. In other words, as one commenter stated, "the established financial reporting ecosystem provides a valuable skeleton and helpful scaffolding for the key components needed to establish an AI accountability framework."[327]

In the financial sector, a standard setting body develops guidelines for how an auditor should assess the financial disclosures of a business. Then, an independent certified professional evaluates that business against those standards.[328] The goal of a financial audit is to give investors assurance that they have high quality information about the business, which in turn aids the public trust in the capital markets. Audits cover both governance controls and metrics for reporting financial information, and they are structured as reviews of management's certified claims about each.[329]

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.[330] At the federal level, financial sector risks have focused the attention of lawmakers seeking to protect investors and promote a trustworthy marketplace.[331] 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.[332] 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.[333]

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-setting body meets 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."[334] 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.[335] 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.[336]

For auditing standards, Sarbanes-Oxley created the Public Company Accounting Oversight Board (PCAOB), a non-profit corporation that is subject to SEC oversight.[337] Oversight includes the SEC's "approval of the Board's rules, standards, and budget."[338] PCAOB itself is tasked with "oversee[ing] the audit of companies subject to securities laws."[339] 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 proceedings concerning, and impose appropriate sanctions where justified upon, registered public accounting firms and associated persons of such firms."[340] The SEC may determine additional duties or functions for the Board to enhance the relevant audit landscape.[341] In furtherance of its mission, PCAOB has established a series of auditing and other standards related to financial auditing.[342]

---

326 *See, e.g.,* Intel Comment at 3 ("[T]here are numerous existing laws or regulations that apply to the deployment and use of AI technology, such as state privacy laws, federal consumer financial laws and adverse action requirements enforced by the Consumer Financial Protection Bureau, constitutional provisions and Federal statutes prohibiting discrimination under the jurisdiction of the Department of Justice's Civil Rights Division, and the Federal Trade Commission Act which protects consumers from deceptive or unfair business practices and unfair methods of competition across most sectors of the U.S. economy."); Morningstar, Inc. Comment at 1 ("Morningstar believes that new AI-specific regulation may not be necessary because current financial regulations are generally drafted broadly enough to encompass AI products and their use.").

327 PWC Comment at 1. *See also* id. at A4 ("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 information.").

328 *See, e.g.,* Paul Munter, The Importance of High Quality Independent Audits and Effective Audit Committee Oversight to High Quality Financial Reporting to Investors, United States Securities and Exchange Commission (October 26, 2021) https://www.sec.gov/news/statement/munter-audit-2021-10-26.
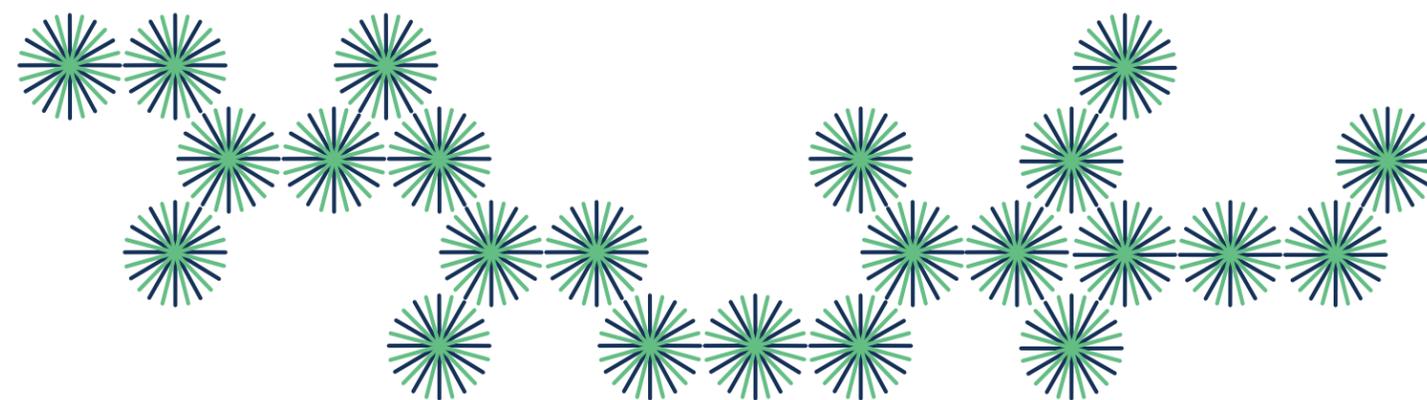
329 PWC Comment at A4 (providing a graphic of the relationships in the financial

accountability system).

330 *See, e.g.,* PWC Comment at 1 ("Notably, however, the ecosystem around financial reporting is a child of crisis: the stock market crash of 1929 created the initial requirements for reporting by and audits of public companies while the high-profile collapse of companies such as Enron in the early 2000s led to enhanced responsibilities for management to provide reporting around internal control over financial reporting."); U.S. House of Representatives Committee on Financial Services, Report on the Corporate and Auditing Accountability, Responsibility, and Transparency Act of 2002, H. Rept. 107-414 (April 22, 2002), at 18 ("Following the bankruptcies of Enron Corporation and Global Crossing LLC, and restatements of earnings by several prominent market participants, regulators, investors and others expressed concern about the adequacy of the current disclosure regime for public companies. Additionally, they expressed concerns about the role of auditors in approving corporate financial statements. . . .); William H. Donaldson, Testimony Concerning Implementation of the Sarbanes-Oxley Act of 2002, U.S. Securities and Exchange Commission (September 9, 2003), https://www.sec.gov/news/testimony/090903tswhd.htm ("Sparked by dramatic corporate and accounting scandals, the [Sarbanes-Oxley] Act represents the most important securities legislation since the original Federal securities laws of the 1930s.").

331 U.S. Securities and Exchange Commission, About the SEC, https://www.sec.gov/about ("The mission of the SEC is to protect investors; maintain fair, orderly, and efficient markets; and facilitate capital formation. The SEC strives to promote a market environment that is worthy of the public's trust."). *See also* U.S. Securities and Exchange Commission, Mission, https://www.sec.gov/about/mission.

332 U.S. Securities and Exchange Commission, The Laws That Govern the Securities Industry, https://www.sec.gov/about/about-securities-laws (listing various securities laws).

333 The legal and regulatory structure of the financial services is complex, and for the purposes of this Report, we principally focus on financial accounting and auditing standards in the private sector. The federal government and state and local governments have their own accounting and auditing mechanisms. *See, e.g.,* Congressional Research Service, Accounting and Auditing Regulatory Structure: U.S. and International (Report R44894) (July 19, 2017), https://crsreports.congress.gov/product/pdf/R/R44894, at 11-18 (providing descriptions). These structures may also be worth analyzing further in the context of developing AI accountability measures.

334 Sarbanes-Oxley Act of 2002, 116 Stat. 745, Section 108(b)(1)(B) (2002).

335 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. 23333 (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.

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 securities laws, including 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 Commission to establish accounting principles or standards for purposes of enforcement of the securities laws."). *See also* Sarbanes-Oxley Act of 2002, Section 108(b)(1)(B) ("In carrying out its authority under sub-section (a) and under section 13(b) of the Securities Exchange Act of 1934, the Commission may recognize, as 'generally accepted' for purposes of the securities laws, any accounting principles established by a standard setting body.") (emphasis added); Financial Accounting Standards Board, SEC Accepts 2023 GAAP Financial Reporting Taxonomy and SEC Reporting Taxonomy (March 21,

2023), https://www.fasb.org/page/getarticle?uid=fasb_Media_Advisory_03-21-23 ("The Financial Accounting Standards Board (FASB) today announced that the U.S. Securities and Exchange Commission (SEC) has accepted the 2023 GAAP Financial Reporting Taxonomy (GRT) and the 2023 SEC Reporting Taxonomy (SRT) (collectively referred to as the 'GAAP Taxonomy'). The FASB also finalized the 2023 DQC Rules Taxonomy (DQCRT), which together with the GAAP Taxonomy are collectively referred to as the 'FASB Taxonomies.'").

337 *See generally* Sarbanes-Oxley Act of 2002, 116 Stat. 745 (2002), title I; Public Company Accounting Oversight Board, About, https://pcaobus.org/about.

338 Public Company Accounting Oversight Board, About, https://pcaobus.org/about.

339 15 U.S.C. § 7211(a).

340 15 U.S.C. § 7211(c)(1)-(4).

341 *See* 15 U.S.C. § 7211(c)(5).

342 Public Company Accounting Oversight Board, Standards, https://pcaobus.org/oversight/standards; Public Company Accounting Oversight Board, Auditing Standards of the Public Company Accounting Oversight Board, https://assets.pcaobus.org/pcaob-dev/docs/default-source/standards/auditing/documents/auditing_standards_audits_after_december_15_2020.pdf (latest auditing standards, for fiscal years ending

ACCOUNTING AND AUDITING STANDARD-SETTERS

PRIVATE SECTOR

FAF → FASB
SEC → PCAOB (dotted line SEC → FASB)

ACCOUNTING | AUDITING

FAF: Financial Accounting Foundation
FASB: Financial Accounting Standards Board
SEC: U.S. Securities and Exchange Commission
PCAOB: Public Company Accounting Oversight Board

FEDERAL GOVERNMENT

GAO, OMB, Treasury → FASAB
GAO → (AUDITING)

ACCOUNTING | AUDITING

GAO: Government Accountability Office
OMB: Office of Management and Budget
Treasury: U.S. Department of the Treasury
FASAB: Federal Accounting Standards Advisory Board
OIG: Office of Inspector General

STATE & MUNICIPAL GOVERNMENTS

FAF → GASB
SMGA → (AUDITING)

ACCOUNTING | AUDITING

FAF: Financial Accounting Foundation
GASB: Governmental Accounting Standards Board
SMGA: State and Municipal Government Auditors

Source Data: Congressional Research Service (CRS)[343]

 Thus, accounting and auditing standards for the financial sector, subject to public securities law,[343] 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.[344] While such structure is not without criticism,[345] 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 accountability 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 institutions and their regulators "illustrates that a tailored yet flexible approach provides strong accountability measures that also allow industry to innovate."[346]

---

on or after Dec. 15, 2020).

343 Congressional Research Service, *supra* note 333, at 2. The graphic is accompanied by the following note: "In the first panel, the striated line indicates the SEC's oversight role over accounting standards promulgated by the FASB. The FASB's parent organization, the Financial Accounting Foundation (FAF), is a nonstock Delaware corporation. Neither FASB nor FAF is a government agency, even though the SEC does have oversight of the budget for FASB and the accounting standards as promulgated by FASB (FAF, "Facts About FAF," http://www.accountingfoundation.org/jsp/Foundation/Page/FAFSectionPage&cid=1176157790151)." Id.

344 Id.

345 Sarah J. Williams, "The Alchemy of Effective Auditor Regulation," 25 Lewis & Clark Law Rev.1089, 1107 n.105 (2022) (collecting sources criticizing auditing standards and PCAOB).

346 SIFMA Comment at 2-3.

- 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 mature 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.

A principal input for holding entities accountable for human rights harms are human rights impact assessments, "which are grounded in the [United Nations Guiding Principles on Business and Human Rights], a non-binding framework endorsed by the United Nations Human Rights Council in 2011."[347] 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.[348] As one commenter put it, "there are benefits in using the same methodology and not burdening teams with performing several assessments in parallel."[349] A number of commenters suggested incorporating human rights assessment frameworks into standard review processes across the AI life cycle.[350]

In the United States, the SEC has adopted rules requiring climate-related disclosures for public companies.[351] 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.[352] There are many other standards and methods deployed in ESG evaluations.[353] 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.[354] For example, respect for individuals' privacy rights is a human rights issue and at the same time it is a "social impact" issue within bounds of the "S" in ESG.[355]

There is a risk for ESG evaluations, as well as for AI trustworthiness evaluations, that the goals and standards 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."[356] Some questions may not be answerable. In the ESG context, this might be a question about supply chain responsibility. In the AI context,

---

347 European Center for Not-for-Profit Law and Data & Society, Recommendations for Assessing AI Impacts to Human Rights, Democracy, and the Rule of Law at 4 (2021), https://ecnl.org/sites/default/files/2021-11/HUDERIA%20paper%20ECNL%20and%20DataSociety.pdf.

348 See, e.g., The Investor Alliance for Human Rights Comment at 3 (advocating for the creation of "a robust and clear methodology for a human rights impact assessment process" with "specific criteria relevant to AI systems" and that "must be developed with the involvement of digital rights experts."); AI & Equality Comment at 9 (Government should "consider the legal obligation to adhere to international human rights treaties and United States due process laws when creating AI accountability policy"); David Kaye, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, United Nations (UN Document Symbol A/73/348) (August 29, 2018), https://daccess-ods.un.org/access.nsf/Get?OpenAgent&DS=A/73/348&Lang=E, at 20 ("When procuring or deploying AI systems or applications, States should ensure that public sector bodies act consistently with human rights principles. This includes, inter alia, conducting public consultations and undertaking human rights impact assessments or public agency algorithmic impact assessments prior to the procurement or deployment of AI systems.").

349 Centre for Information Policy Leadership Comment at 7 (also noting that "there is no consensus on how to identify and assess human rights risks and harms and how to do this in an integrated way for all disciplines—AI, privacy security, safety, children's protection, etc.").

350 Google DeepMind Comment at 18; The Investor Alliance for Human Rights Comment at 2; Global Partners Digital Comment at 4 (recommending codification of human

rights standards for AI).

351 SEC, The Enhancement and Standardization of Climate-Related Disclosures for Investors (Final Rule) (Mar. 6, 2024), https://www.sec.gov/files/rules/final/2024/33-11275.pdf (requiring "registrants to provide certain climate related information in their 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.").

352 See generally, SASB Standards, SASB Standards and other ESG Frameworks: The Sustainability Reporting Ecosystem, https://sasb.org/about/sasb-and-other-esg-frameworks/; SEC, supra note 351 (proposing for public companies a similar reporting format); Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting, OJ L 322 (Dec. 16, 2022), http://data.europa.eu/eli/dir/2022/2464/oj (adopting European Sustainability Reporting Standards that require ESG reporting for companies in the EU starting January 1, 2024).

353 See, e.g., Global Reporting Initiative (GRI), Carbon Disclosure Project (CDP), Task Force on Climate-Related Financial Disclosures (TCFD), and United Nations Sustainable Development Goals (SDG).

354 See, e.g., CAQ Comment at 7 (noting that AI safety standards are a predicate for evaluation as part of the ESG process.).

355 See Centre for Information Policy Leadership Comment at 22.

356 Raji et al., Outsider Oversight, supra note 253, at 558. See also Open MIC Comment at 5 ("Without mandatory standards for AI audits and assessments, including those focused on measuring adverse impacts to human rights, there is an incentive for companies to 'social wash' their AI assessments; i.e. give investors and other stakeholders the impression that they are using AI responsibly without any meaningful efforts to ensure this.").

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.[357]

The private sector continues to refine and seek ESG framework standardization for evaluations. What the ESG assurance experience might teach is that multi-factored 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 FDA's.[358] 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."[359]

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.[360] 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.[361]

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.[362]

The FDA further has in place an exemplary adverse incident database that could be instructive for AI system accountability.[363] This system is similar to the Federal Aviation Administration's Aviation Safety Reporting System; both collect safety incidents for transparency, review, and risk management of already deployed systems. In the AI context, a similar reporting structure would enable users and subjects of AI systems to recognize and report adverse incidents, as discussed above in Section 3.1.1. One risk is the possibility of over-reporting if parameters are not carefully defined and the reporting platform is not well-managed. Regulatory oversight or coordination would help to arrange this kind of reporting function.

Additional accountability models 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.[364] 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."[365] 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.[366] 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,[367] the implementation of these frameworks varies across organizations and industries.[368] 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 related to data security or privacy.[369] 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 pri-

vacy practices where it determined that businesses' practices were likely to cause data security or privacy harm to consumers, and harm was not outweighed by countervailing 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.[370] The FTC has also promulgated guidance distilling the facts from its enforcement cases into data security lessons for companies.[371]

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 a model for AI. Others 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 are mature and have been tested and refined through years of legal interpretation and application, thereby offering greater degrees of consistency and predictability.[372] Indeed, existing laws and regulatory requirements that set cybersecurity standards for distinct industries already apply when AI deployments in those industries affect cybersecurity.[373] 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.[374]

357 See Grabowicz et al., Comment at 5 ("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").

358 See, e.g., Carnegie Mellon University Comment at 3; Unlearn.AI Comment at 2.

359 Alex Engler, The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment, Brookings Institute (April 25, 2023), https://www.brookings.edu/research/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/ (citing to FDA efforts). See also The Pew Charitable Trusts, How FDA Regulates Artificial Intelligence in Medical Products (Aug. 5, 2021), https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2021/08/how-fda-regulates-artificial-intelligence-in-medical-products.

360 Food and Drug Administration, How to Study and Market Your Device (September 2023), https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/how-study-and-market-your-device.

361 Grabowicz et al., Comment at 6. See also Andrew Tutt, An FDA for Algorithms, 69 Admin. L. Rev. 83 (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2747994 (presenting a general argument about the analogy between FDA regulation and algorithmic risk management).

362 See Grabowicz et al., Comment at 6.

363 See Raji, et al, Outsider Oversight, supra note 253 at 561.

364 ForHumanity Comment at 4; Carnegie Mellon University Comment at 4.

365 Grabowicz et al., Comment at 4.

366 Mulligan, Stephen P. and Chris D. Linebaugh, Data Protection Law: An Overview at 2, Congressional Research Service (March 25, 2019) https://crsreports.congress.gov/product/pdf/R/R45631.

367 NIST, Framework for Improving Critical Infrastructure Cybersecurity Version 1.1(April 16, 2018), https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf.

368 Id.

369 See, e.g., FTC v. Sandra L. Rennert, et al., Docket No. CV-S-00-0861-JBR (D. Nev. 2000); In the Matter of Eli Lilly and Company, FTC Docket No. C-4047 (2002); In the Matter of BJ's Wholesale Club (2005), FTC Docket No. C-4148 (2005).

370 Fed. Trade Comm'n v. Wyndham Worldwide Corp., 799 F.3d 236, 257 (3d Cir. 2015). See also Federal Trade Commission, Start with Security: A Guide for Business (June 2015), https://www.ftc.gov/business-guidance/resources/start-security-guide-business (presenting "lessons" from "more than 50 law enforcement actions the FTC has announced so far" against businesses).

371 See id.

372 See, e.g., USTelecom Comment at 9.

373 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 e-enabled aircraft. This approach would be potentially extensible to AI impacting cybersecurity. See 14 CFR 11.19. For issues that become apparent after an aircraft or other aeronautical product enters the marketplace, the FAA issues airworthiness directives in appropriate cases, specifically, "FAA issues an airworthiness directive addressing a product when we find that: (a) An 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 C.F.R. § 39.5.

374 See, e.g., IAPP Comment at 5-6.

6.
Recommendations

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.[375] Other suggestions borrowed from the cybersecurity context including creating incentives for companies to facilitate "responsible disclosure";[376] developing red-teaming exercises;[377] launching "Bug Bounty" programs to encourage disclosure and financially reward detection of AI vulnerabilities;[378] and modelling AI vulnerability disclosures on the Common Vulnerabilities and Exposures (CVE) system, which provides a standardized naming scheme for cybersecurity vulnerabilities.[379]

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.[380] 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.[381] 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.[382] 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 Kuan Johnson, Funding Opportunity from my team to build and run a DoD-wide Bias Bounty Program, https://www.linkedin.com/posts/dr-matthew-kuan-johnson-8144591b8_bias-bounty-program-opportunities-tradewind-activity-7084911005759074305-2Vim/.

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 292.

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

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 established thresholds for trustworthy AI. Government 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 establishes a comprehensive set of actions on AI governance; the White House Blueprint for an AI Bill of Rights, 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. To advance AI accountability policies and practices, we recommend guidance, support, and the development of regulatory requirements.

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 *not yet* 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 militate towards a certification process for qualified auditors.

AI audits should, at a minimum, be able to evaluate claims made about an AI system's fitness for purpose, performance, processes, and controls. Regardless of claims made, an audit should apply substantive criteria arrived at through broad stakeholder inquiry across the AI system lifecycle. Areas of review might include:

- Risk mitigation and management, including harm prevention;

- Data quality and governance;

- Communication (e.g., documentation, disclosure, provenance); and

- Governance or process controls.

As valuable as they are, independent evaluations, including audits, do not derogate from the importance of regulatory inspection of AI systems and their effects.

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 informational 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.
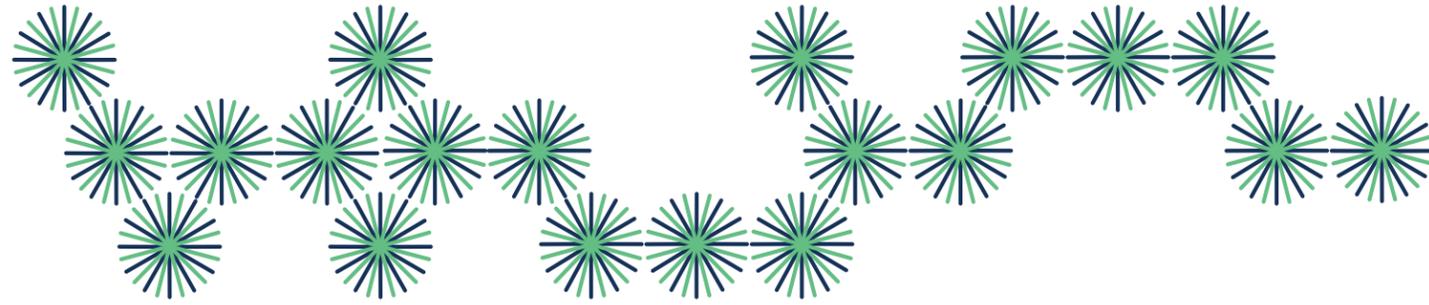
6.2. SUPPORT

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[383] 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 stakeholder 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-teaming 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.

---

383 Without taking a position at this time, we note there may be other models for funding, such as fee-based application revenue for AI companies who *seek* government assistance. For literature on certain fee models that exist across some federal agencies, *see*, e.g. Government Accountability Office (GAO), Federal Design Options: Fee Design Options and Implications for Managing Revenue Instability (GAO Report No. GAO-13-820), (Sept. 2013), https://www.gao.gov/assets/gao-13-820.pdf; James M. MacDonald, User-Fee Financing of USDA Meat and Poultry Inspection, Agricultural Economic Report No. (AER-775), (March 1999), Chapter 3, https://www.ers.usda.gov/webdocs/publications/40973/51055_aer775.pdf?v=266.1.

384 The Government Accountability Office has also noted that "[f]oundational to solving the AI accountability challenge is having a critical mass of digital expertise to help accelerate responsible delivery and adoption of AI capabilities." Government Accountability Office (GAO), Artificial Intelligence: Key Practices to Help Ensure Accountability in Federal Use (GAO Report No. GAO-23-106811), at 1 (May 16, 2023), https://www.gao.gov/assets/gao-23-106811.pdf.

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, reinforce 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;

- Collection of periodic claim substantiation for deployed systems; and

- Coordination of AI accountability inputs with agency counterparts in non-adversarial states.
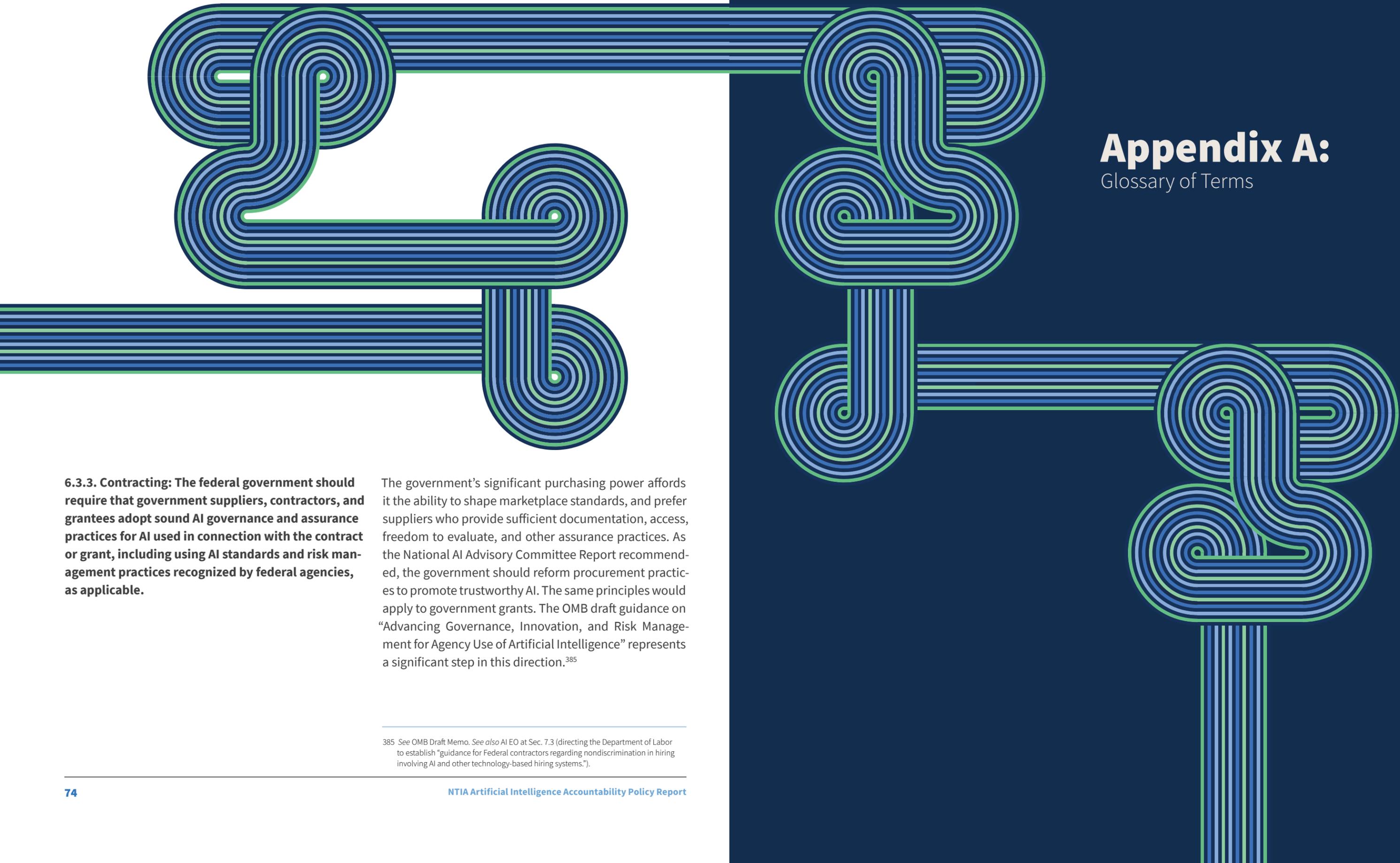
Appendix A:
Glossary of Terms

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.").

APPENDIX A: Glossary of Terms

The process of defining terms in the AI policy space is ongoing and fluid. Where there are existing U.S. government or other consensus definitions, we use them. Where there are not, we use definitions we find supported 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.[386] Whereas OECD interpretive guidance distinguishes "accountabil-

ity" from "responsibility" and "liability,"[387] the TTC definition embraces responsibility as part of accountability and includes a broader scope of governance activities.[388] Accountability may require enforceable consequences.[389] 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."[390] 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."[391]

AI Assurance. AI assurance is the product of a set of informational 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)[392] and

the UK Centre for Data Ethics and Innovation usage in its "roadmap to an effective AI assurance ecosystem."[393]

AI Audit. An AI audit is, with respect to an AI system or model, an evaluation of performance and/or process against transparent criteria.[394] An audit is broader than a "conformity assessment" which is "the demonstration that specified requirements relating to a product, process, system, person or body are fulfilled."[395] As noted in the RFC, entities can audit their own systems or models, be audited by a contracted second party, or be audited by a third party. To distinguish audits from other evaluations, we use the term audit to refer only to independent evaluations.[396] An audit can be structured merely to verify the claims made about AI.[397] Alternatively, it can be scoped more broadly to evaluate AI system or model performance vis a vis attributes of trustworthy AI, regardless of claims made. Simply put, an audit is an assurance tool, characterized by precision and providing an independent evaluation of an AI system, claims made about that system, and/or the degree to which that system is trustworthy. For ease of reading, we include audits in the umbrella term "evaluations."

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. AI red-teaming is most often performed by dedicated "red-teams" that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system.[398]

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 harmful bias managed."[399] 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."[400]

---

386 See U.S.-E.U. Trade and Technology Council (TTC), EU-U.S. Terminology and Taxonomy for Artificial Intelligence (May 31, 2023), at 11, https://digital-strategy. ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence ("Accountability relates to an allocated responsibility. The responsibility can be based on regulation or agreement or through assignment as part of delegation. In a systems context, accountability refers to systems and/or actions that can be traced uniquely to a given entity. In a governance context, accountability refers to the obligation of an individual or organisation to account for its activities, to complete a deliverable or task, to accept the responsibility for those activities, deliverables or tasks, and to disclose the results in a transparent manner."). See also NIST, The Language of Trustworthy AI: An In-Depth Glossary of Terms (March 22, 2023), https://doi.org/10.6028/NIST. AI.100-3 (referencing National Institute of Standards and Technology, Trustworthy & Responsible AI Resource Center, Glossary, https://airc.nist.gov/AI_RMF_Knowledge_ Base/Glossary) (substantially similar). Cf. NIST AI RMF, Second Draft, supra note 43 at 15 ("Determinations of accountability in the AI context relate to expectations of the responsible party in the event that a risky outcome is realized.").

387 OECD.AI Policy Observatory, Accountability (Principle 1.5), https://oecd.ai/en/ dashboards/ai-principles/P9 ("'accountability' refers to the expectation that organisations or individuals will ensure the proper functioning, throughout their lifecycle, of the AI systems that they design, develop, operate or deploy, in accordance with their roles and applicable regulatory frameworks, and for demonstrating this through their actions and decision-making process (for example, by providing documentation on key decisions throughout the AI system lifecycle or conducting or allowing auditing where justified)").

388 See Software & Information Industry Association Comment at 3 (embracing the TTC definition and its view that "AI accountability is concerned with both system-level performance and with governance structures relevant to the development and deployment of AI systems").

389 See, e.g., Ada Lovelace Institute Comment at 2 ("assessments are themselves not a form of accountability"); Price Waterhouse Cooper (PWC) Comment at A2 (using dictionary definitions to equate accountability with responsibility).

390 NIST AI RMF at 2 (citing with approval OECD, Artificial Intelligence in Society (2019), https://doi.org/10.1787/eedfee77-en).

391 NIST AI RMF at 6.

392 MITRE Comment at 9 ("AI assurance is a lifecycle process that provides justified confidence in an AI system to operate effectively with acceptable levels of risk to its

stakeholders. Effective operation entails meeting functional requirements with valid outputs. Assurance risks may be associated with or stemming from a variety of factors depending on the use context, including but not limited to AI system safety, security, equity, reliability, interpretability, robustness, directability, privacy, and governability."). See also ISO/IEC/IEEE International Standard – Systems and Software Engineering – Systems and Software Assurance, IEEE/ISO/IEC 15026-1 (2019), https://standards.ieee. org/ieee/15026-1/7155/

393 Centre for Data Ethics and Innovation, The roadmap to an effective AI assurance ecosystem (Dec. 8, 2021), https://www.gov.uk/government/publications/the-roadmap- to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance- ecosystem ("Assurance services help people to gain confidence in AI systems by evaluating and communicating reliable evidence about their trustworthiness.").

394 National Institute of Standards and Technology, supra note 43 at 15.

395 NIST, Conformity Assessment Basics (2016), https://www.nist.gov/standardsgov/ conformity-assessment-basics.

396 See, e.g., Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi, "Auditing Large Language Models: A Three-Layered Approach," AI and Ethics (May 30, 2023), https://doi.org/10.1007/s43681-023-00289-2 ("Auditing is characterised by a systematic and independent process of obtaining and evaluating evidence regarding an entity's actions or properties and communicating the results of that evaluation to relevant stakeholders.").

397 See, e.g., Trail of Bits Comment at 1. See also Data & Society Comment at 2 (it is the "study of the functioning of a system within the parameters of the system." It asks whether the system functions "appropriately according to a claim made by the developer, according to an independent standard…, according to the terms set in a contract, or according to ethical or scientific terms established by a researcher or field of researchers."); Holistic AI Comment at 4-5 ("External audits offer yet another level of system assurance through the process of independent and impartial system evaluation 5 of 11 whereby an auditor with no conflict of interest can assess the system's reliability and in turn identify otherwise unidentified errors, inconsistencies and/or vulnerabilities.").

398 AI EO at Sec. 3(d).

399 NIST AI RMF at 12 (recognizing tradeoffs and that these characteristics must be balanced "based on the AI system's context of use"). See also Executive Order No. 13960, Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, 85 Fed. Reg. 78939 (2020) (articulating the following principles for AI use: "[l]awful and respectful of our Nation's values," "[p]urposeful and performance-driven," "[a]ccurate, reliable, and effective," "[s]afe, secure, and resilient," "[u]nderstandable," "[r]esponsible and traceable", "[r]egularly monitored", "[t]ransparent", and "[a]ccountable").

400 See First Round White House Voluntary Commitments at 1 ("These commitments – which the companies are making immediately – underscore three principles that must be fundamental to the future of AI: safety, security, and trust.").

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**