Google and its frontier AI lab Google DeepMind\(^1\) welcome the opportunity to submit comments in response to the National Telecommunications and Information Administration’s (NTIA) Request for Comment (RFC) on Artificial Intelligence (AI) accountability.\(^2\) As our CEO has said, AI is too important not to regulate and too important not to regulate well.\(^3\) NTIA’s timely engagement with a broad array of industry and stakeholders on these crucial issues will be a vital component of getting AI regulation right.

Google’s mission is to organize the world’s information and make it universally accessible and useful, with Search making it easy to discover a broad range of information from a wide variety of sources. As a frontier AI lab, Google DeepMind’s mission is to solve intelligence to advance science and benefit humanity, with breakthroughs including AlphaGo,\(^4\) AlphaFold,\(^5\) more than one thousand published research papers,\(^6\) and numerous partnerships with scientific organizations. Like Google, Google DeepMind has a longstanding approach to AI accountability, reflected in remarks by Google DeepMind’s CEO\(^7\) and hundreds of contributions to Google’s products.\(^8\)

It is important to recognize at the outset that AI is not synonymous with generative AI or large language models. In fact, most Americans have been using various types of AI for a dozen years or more. Google Search, Gmail, Google Maps, and Google Translate, to take just a few examples, have all long used AI to improve the quality, speed, affordability, and convenience of the services that they provide. Any new regulations should avoid impairing the operation of systems that deliver significant value for Americans.

---

\(^1\) Google DeepMind brings together two leading groups in the AI field: the Brain team from Google Research and DeepMind. [Google DeepMind: Bringing together two world-class AI teams](https://deepmind.com/blog/google-deepmind-bringing-together-two-world-class-ai-teams).


\(^3\) [Sundar Pichai, Google CEO: Building AI responsibly is the only race that really matters, Financial Times (May 23, 2023)](https://www.ft.com/content/510971a4-529f-473e-82f6-9a19def363c0).

\(^4\) [AlphaGo](https://deepmind.com/blog/alphago-alphazero-continues-to-impress).

\(^5\) [Putting the power of AlphaFold into the world’s hands](https://deepmind.com/blog/alphafold).

\(^6\) [Research](https://国有企业改革和人工智能的发展: 旧与新结合).

\(^7\) See, e.g., [DeepMind CEO Demis Hassabis Urges Caution on AI | Time](https://time.com/5922281/deepmind-ceo-demis-hassabis-caution-on-ai/) (“When it comes to very powerful technologies—and obviously AI is going to be one of the most powerful ever—we need to be careful.”).

\(^8\) Including leading products from [Android battery efficiency](https://www.android.com/features/battery-efficient) to [Assistant text-to-speech](https://assistant.google.com).
Building upon Google and Google DeepMind’s extensive experience creating and implementing an internal AI governance framework across a diverse array of products and services, and leading in the development of the most advanced AI research, we offer suggestions for how policymakers can best support U.S. and global leadership in enabling responsible AI.

Executive Summary

Artificial intelligence has the potential to unlock major benefits, from better understanding diseases to tackling climate change and driving prosperity through greater economic opportunity. But as a still emerging technology, AI presents challenges and must be developed with extra care. We believe industry, researchers, stakeholders in civil society, and, importantly, governments must work together to promote trustworthy applications that live up to AI’s promise of societal benefit while mitigating risk. To this end, we support a policy agenda oriented around three pillars: unlocking opportunity, promoting responsibility, and enhancing U.S. and international security. NTIA’s AI Accountability RFC takes an important step toward determining how the U.S. government can support policies that encourage the creation and use of trustworthy, ethical, and safe AI tools—and achieve these three critical goals.

Google has long been committed to developing AI responsibly. We were one of the first companies to publish a set of AI principles, and we use an AI risk-assessment framework to identify and mitigate risks. Google DeepMind likewise has Operating Principles and a dedicated internal governance body—the Institutional Review Committee—tasked with upholding them. Of course, no one company or organization can fully ensure responsible AI; it is a shared responsibility between both the private sector and government.

While it is tempting to look for silver-bullet policy solutions, AI raises complex questions that require nuanced answers. It is a 21st century technology that requires a 21st century governance model. We need a multi-layered, multi-stakeholder approach to AI governance. This will include:

- Industry, civil society, and academic experts developing and sharing best practices and technical standards for responsible AI, including around safety and misinformation issues;
- A hub-and-spoke model of national regulation; and
- International coordination among allies and partners, including around geopolitical security and competitiveness and alignment on regulatory approaches.

---

9 See Google AI: Our Principles
10 See generally 2022 AI Principles Progress Update.
11 See DeepMind Operating Principles.
At the national level, we support a hub-and-spoke approach—with a central agency like the National Institute of Standards and Technology (NIST) informing sectoral regulators overseeing AI implementation—rather than a “Department of AI.” AI will present unique issues in financial services, health care, and other regulated industries and issue areas that will benefit from the expertise of regulators with experience in those sectors—which works better than a new regulatory agency promulgating and implementing upstream rules that are not adaptable to the diverse contexts in which AI is deployed.

Maximizing the economic opportunity from AI will also require a joint effort across federal, state, and local governments, the private sector, and civil society to equip workers to harness AI-driven tools. AI is likely to generate significant economy-wide benefits. At the same time, to mitigate displacement risks, the private sector will need to develop proof-of-concept efforts on skilling, training, and continuing education, while the public sector can help validate and scale these efforts to ensure workers have wrap-around support. Smart deployment of AI coupled with thoughtful policy choices and an adaptive safety net can ensure that AI ultimately leads to higher wages and better living standards.

With respect to U.S. regulation to promote accountability, we urge policymakers to:

- **Promote enabling legislation for AI innovation leadership.** Federal policymakers can eliminate legal barriers to AI accountability efforts, including by establishing competition safe harbors for open public-private and cross-industry collaboration on AI safety research, and clarifying the liability for misuse and abuse of AI systems by different users (e.g., researchers, authors, creators of AI systems, implementers, and end users). Policymakers should also consider related legal frameworks that support innovation, such as adopting a uniform national privacy law that protects personal information and an AI model’s incidental use of publicly available information.

- **Support proportionate, risk-based accountability measures.** Deployers of high-risk AI systems should provide documentation about their systems and undergo independent risk assessments focused on specific applications.

- **Regulate under a “hub-and-spoke” model rather than creating a new AI regulator.** Under this model, regulators across the government would engage a central, coordinating agency with AI expertise, such as NIST, with Office of Management and Budget (OMB) support, for technical guidance on best practices on AI accountability.

- **Use existing authorities to expedite governance and align AI and traditional rules.** Where appropriate, sectoral regulators would provide updates clarifying how existing authorities apply to the use of AI systems, as well as how organizations can demonstrate compliance of an AI system with these existing regulations.
• Assign to AI deployers the responsibility of assessing the risk of their unique deployments, auditing, and other accountability mechanisms as a result of their unparalleled awareness of their specific uses and related risks of the AI system.

• Define appropriate accountability metrics and benchmarks, as well as terms that may be ambiguous, to guide compliance. Recognize that many existing systems are imperfect and that even imperfect AI systems may, in some settings, be able to improve service levels, reduce costs, or increase affordability and availability.

• Consider the tradeoffs between different policy objectives, including efficiency and productivity enhancements, transparency, fairness, privacy, security, and resilience.

• Design regulation to promote competitiveness, responsible innovation, and broad access to the economic benefits of AI.

• Require high standards of cybersecurity protections (including access controls) and develop targeted “next-generation” trade control policies.

• Avoid requiring disclosures that include trade secrets or confidential information (potentially advantaging adversaries) or stymie this innovative sector as it continues to evolve.

• Prepare the American workforce for AI-driven job transitions and promote opportunities to broadly share AI’s benefits.

Finally, NTIA asks how policymakers can otherwise advance AI accountability. The U.S. government should:

• Continue building technical and human capacity into the ecosystem to enable effective risk management. The government should deepen investment in fundamental responsible AI research (including bias and human-centered systems design) through federal agency initiatives, research centers, and foundations,12 as well as by creating and supporting public-private partnerships.13

• Drive international policy alignment, working with allies and partners to develop common approaches that reflect democratic values. Policymakers can support common standards and frameworks that enable interoperability and harmonize global AI governance approaches. This can be done by: (1) enabling trusted data flows across national borders, (2) establishing multinational AI research resources, (3) encouraging the adoption of common approaches to AI regulation and governance and a common lexicon, based on the work of the Organisation for Economic Co-operation and

---


13 See, e.g., The National Artificial Intelligence Advisory Committee (NAIAC) Year 1 Report, at 28 (May 2023) (recommending a public-private “multi-agency task force to develop frameworks for small- and medium-sized organizations to adopt trustworthy AI”).
Development (OECD), (4) working within standard-setting bodies such as the International Organization for Standardization (ISO) and the Institute of Electrical and Electronics Engineers (IEEE) to establish rules, benchmarks, and governance mechanisms that can serve as a baseline for domestic regulatory approaches and deter regulatory fragmentation, (5) using trade and economic agreements to support the development of consistent and non-discriminatory AI regulations, (6) promoting copyright systems that enable appropriate and fair use of copyrighted content to enable the training of AI models, while supporting workable opt-outs for websites, and (7) establishing more effective mechanisms for information and best-practice sharing among allies and between the private and the public sectors.

- **Explore updating procurement rules to incentivize AI accountability, and ensure OMB and the Federal Acquisition Regulatory Council are engaged in any such updates.** It will be critical for agencies who are further ahead in their development of AI procurement practices to remain coordinated and aligned upon a common baseline to effectively scale responsible governance (e.g., through the NIST AI Risk Management Framework (AI RMF)).

The United States currently leads the world in AI development, and with the right policies that support both trustworthy AI and innovation, the United States can continue to lead and help allies enhance their own competitiveness while aligning around a positive and responsible vision for AI. Centering policies around economic opportunity, promoting responsibility and trust, and furthering our collective security will advance today’s and tomorrow’s AI innovation and unleash benefits across society.

**Responses to AI Accountability RFC Questions**

**AI Accountability Objectives**

1. **What is the purpose of AI accountability mechanisms such as certifications, audits, and assessments?**

   Assessments, audits, and certifications should be oriented toward promoting accountability, empowering users, and building trust and confidence. Effective accountability mechanisms should also anticipate the challenges generated by emergent AI applications, including

---

14 Please note that the subsidiary, clarifying questions included with certain questions in the RFC (including this question) were removed for brevity’s sake. The principal question prompts are included to aid in review of these comments.
“frontier models,” which stand to pose novel risks, and incorporate safeguard measures from the start.

Google supports the use of **risk-based assessments** when developing or deploying AI systems. We embed ethical risk management into Google products and services (e.g., Pixel devices, BARD, and Google Shopping) from the design stage. Internally, we use an **AI risk-assessment framework (AI RAF)** alongside a Product Maturity Model Assessment—stewarded by our Responsible Innovation and Responsible AI and Human–Centered Technology teams, which focus on the sociotechnical realization of our AI practices—to ensure that internal development and deployment practices are consistently aligned with our AI Principles and our broader compliance ecosystem, including the Civil and Human Rights Program. Outputs from these assessments are mapped to a prescriptive maturity model framework, which details clear actions developer teams can take to improve their machine learning (ML) models and advance from one maturity level to the next. For organizations looking to integrate a practical assessment framework into their own development flows, the NIST AI RMF provides a critical, generally applicable mechanism for assessing and mitigating risk. We strongly support its broad adoption by both the public and private sectors.

Organizations and individuals that develop, deploy, and use AI systems should be able to demonstrate that their AI governance processes can effectively identify, manage, and mitigate risks appropriately (including system “breaks” or failures), particularly when the stakes are high (i.e., for applications that pose a material risk of significantly harming people or property or imperiling access to essential services).

**Audits** could assess the governance processes of AI developers and deployers based on the NIST AI RMF or another applicable standard. Any independently validated risk assessment or audit should be aligned to international, industry-accepted criteria to ensure consistency.

Further, independent auditors would need to be professionally qualified and entrusted to only

---

15 “Frontier models” are models that are “both (a) close to, or exceeding, the average capabilities of the most capable existing models, and (b) different from other models, either in terms of scale, design (e.g. different architectures or alignment techniques), or their resulting mix of capabilities and behaviours. Accordingly, frontier models are uniquely risky because (a) more capable models can excel at a wider range of tasks, which will unlock more opportunities to cause harm; and (b) novel models are less well-understood by the research community.” Toby Shevlane et al., *Model evaluation for extreme risks* (May 24, 2023).
16 See generally *Google AI Principles*.
17 *Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST, NIST AI 100-1 (Jan. 2023).*
18 The International Association of Privacy Professionals recently announced its intent to create a professional certification for AI governance. J. Trevor Hughes and Mikko Niva, *IAAPP AI Governance Center, a call to action for the privacy profession, IAPP* (May 10, 2023).
certify organizations that meet the appropriate standards.\textsuperscript{19} Regulators would need to balance such audit requirements against the risk of creating security vulnerabilities, exposing trade secrets and confidential information, or hindering innovation or the development of useful applications.\textsuperscript{20}

**Certifications** of compliance with standards can also be useful for validating that AI systems are developed, deployed, and used responsibly. Certifications should be based on internationally recognized, voluntary consensus standards. For example, ISO is developing standards, ISO/IEC CD 42001 and 42006, for integrated AI management systems and for organizations certifying and auditing those systems, respectively.\textsuperscript{21}

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

AI accountability tools like risk-based assessments, audits, and certifications can help ensure responsible AI governance practices, thereby strengthening external stakeholders’ trust in AI.\textsuperscript{22} Smart policy design will promote these goals while also ensuring that privacy and trade secrets are protected.

While Google and others have invested in developing robust AI governance structures and industry standards, innovation is moving rapidly, requiring flexible frameworks that allow for appropriate risk management. It is reasonable to believe that current and future users of AI systems (consumers and businesses) want explainable and predictable behavior, creating a market incentive within the industry for that outcome. Further, in some cases, requiring disclosures (with protection for trade secrets and controlled technologies as appropriate) may be appropriate for high-risk applications.\textsuperscript{23}

But there may also be practical constraints limiting what is feasible. For example, full traceability and detailed explanations—far higher standards than any human-based system

\textsuperscript{19} Because in-person audits are unlikely to scale well with AI use cases, to the extent audit obligations are adopted, a combination of reporting, automated and manual audits will be better suited to AI technologies.

\textsuperscript{20} NIST’s AI Risk Management Framework, for example, recognizes the tradeoffs between transparency and trade secrets. NIST AI RMF § 3.4 (“Measures to enhance transparency and accountability should also consider the impact of these efforts on the implementing entity, including the level of necessary resources and the need to safeguard proprietary information.”).

\textsuperscript{21} 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.

\textsuperscript{22} As noted elsewhere in these comments, agreed-upon benchmarks will need to be established in circumstances where such metrics are required as part of an AI accountability regime.

\textsuperscript{23} Google proposes that such “high-risk AI systems” include those intended for use in applications that pose a material risk of significantly harming people or property or imperiling access to essential services. See A policy agenda for responsible AI progress: Opportunity, Responsibility, Security.
can meet—would effectively restrict AI systems to an extremely limited and basic set of techniques (e.g., static decision trees). To maximize the value of certifications, audits, and assessments, regulators should build room for workable compromises when necessary, understand the tradeoffs to imposing rules, and avoid being overly prescriptive.

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

There are important tradeoffs across legitimate goals that policymakers need to reconcile as they design accountability mechanisms. The benefits and risks of any AI technology (including tradeoffs between the goals that would be promoted by a given accountability measure) depend on the context within which it is deployed, as well as any risk mitigations undertaken while developing or deploying the technology.

For example, building data minimization limits into AI systems can have impacts on other important principles, such as accuracy and transparency. An over-application of data minimization requirements that limits access to public data required for training models could reduce the accuracy of AI systems. Similarly, AI goals related to transparency or accountability—like proposals to precisely re-create audio or video recommendations—should not also require the collection and storage of data for indefinite periods of time. Requiring that such systems be precisely replicable over time may obligate that every interaction of every user be stored indefinitely, which would be undesirable from a privacy perspective.

Or consider the tradeoffs in human oversight of AI systems. Forms of oversight that are commonsense in one setting could be harmful and undermine the core value of an AI application in another. For example, requiring an AI system’s output to be reviewed by a person before being actioned may make sense for some applications (e.g., AI systems used for critical, non-time-sensitive medical diagnostics). But we may overestimate human accuracy, consistency, and lack of bias leading to, for some applications, sluggish output, reduced privacy (if it means more people see sensitive data), or undermined accuracy (if human reviewers lacked the necessary expertise or were more biased). At an extreme, it could even put people at risk, for example, by delaying automated safety overrides.

Tradeoffs also exist with respect to transparency. Generally speaking, AI systems should perform in ways that can be explained and understood. Such attributes can help to identify harms, empower users to make informed choices, and hold AI providers accountable—all of which underpin trust and confidence in AI. But at the same time, explanations can demand technical resources or cause tradeoffs with other goals like model accuracy (e.g., if

---

24 Google Recommendations for Regulating AI at 11 (“Workable standards for explainability and reproducibility require compromise[.]”).
transparency requirements preclude the use of more accurate but harder-to-explain
techniques). Transparency requirements that limit the functionality of AI would dramatically
undermine AI’s social and economic benefits. Moreover, transparency requirements need to
be carefully designed so they’re actually advancing a legitimate objective.

Transparency and security likewise involve tradeoffs between each other—overly broad
transparency requirements could make it easier for bad actors to spoof, manipulate, or exploit
AI models. Fully open dissemination (i.e., open sourcing) of AI systems without appropriate
controls, protocols, and safeguards could lead to the release of potentially harmful AI
capabilities. Today, abuse by malicious actors is limited by guardrails (for example, refusal
layers) built by developers to help mitigate against inappropriate uses. Full transparency or
access to elements of frontier models may actually pose greater danger to at-risk individuals.

There is also a tradeoff between AI accountability measures on topics like privacy and the
ability to defend against sophisticated adversarial actors that exploit AI tools and systems to
harm consumers. For example, an adversary could use AI to create harmful content like scam
celebrity endorsements for financial products. While facial recognition technologies can be
used to significantly improve precision in detecting malign advertisements and preventing user
harm, privacy safeguards in many jurisdictions prevent biometric data processing without
express user consent, which is difficult to get at scale. As a result, such laws may effectively
prohibit the use of AI tools to detect fraudulent behavior. The law should provide flexibility to
use the technology to protect consumers.

A further tradeoff comes between the requirements of accuracy and the benefits of making AI
widely available and affordable—the perfect as the enemy of the good. Establishing
benchmarks that are too high may make it difficult for deployers to implement tools that offer
performance that, while imperfect, is much more widely available or affordable than existing
systems can provide today.

Policymakers should take steps to ensure that these tradeoffs are considered and reconciled.
One way to achieve this: central guidance from OMB and NIST to relevant agencies, with public
participation and input.

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

Yes, well-designed accountability mechanisms can mitigate at least some systemic or
collective risks of harm. For example, Google’s AI RAF actuates our AI Principles and identifies
the sorts of higher-risk AI use cases that require more in-depth risk assessments. The AI RAF,
as updated in 2022, helps us (1) identify, measure, and analyze ethical risks throughout the life
of an AI-powered product, (2) map these risks to appropriate mitigations, and (3) develop
clearer standards of acceptable risk.\textsuperscript{25} This updated AI RAF now also draws upon the best practices of Google’s cross-company Office of Compliance and Integrity and Enterprise Risk Management efforts, and is aligned with emerging regulatory requirements. The AI RAF ensures that our development teams focus on sensitive topics (which can change over time depending on emerging cultural or technical issues). Grounded in research and a nuanced sociotechnical harms taxonomy, these assessments gather detailed risk information to inform and provide context for central AI Principles reviews by our Responsible Innovation team and other pre-launch reviews that may be necessary.

For example, Google assessed whether to widely release a text-to-image generation model, Imagen, which would enable users to provide a text prompt specifying an image that is then generated directly from the text.\textsuperscript{26} At the outset, we designed the system to automatically detect and filter out words or phrases that violate our policies, which prohibit users from knowingly generating violative content (e.g., sexually explicit, hateful, violent, dangerous, and illegal content) or content that divulgcs personal information. We also virtually eliminated the risks of exposing personally identifiable information by avoiding images with identifiable human faces at the initial launch while our research and testing continued. From there, as part of the assessment process, we ran dedicated rounds of adversarial testing to find flaws in the model. We enlisted product experts who intentionally stress test a system with an adversarial mindset to help. Ultimately, our internal assessments revealed several limitations relating to social biases that led to our decision not to release the model in its then-current state. In this example, our internal AI accountability mechanisms, built into product development and refined based on use cases, gave us the information to mitigate against threats of potential systemic and/or collective risks of harm.

5. Given the likely integration of generative AI tools such as large language models (e.g., ChatGPT) or other general-purpose AI or foundational models into downstream products, how can AI accountability mechanisms inform people about how such tools are operating and/or whether the tools comply with standards for trustworthy AI?

Information disclosure tools such as model cards can promote accountability for large language models.\textsuperscript{27} However, the scope of any disclosure, and accompanying documentation, should be calibrated to the context and what different audiences require—which will vary between developers, end-users, and integrators.

Four general principles for disclosure should apply. First, a provider of an underlying foundation model should provide documentation outlining how the model is intended to be used, known inappropriate uses, known risks, and recommendations for deployers and users.

\textsuperscript{25} See \textit{2022 AI Principles Progress Update} at 15–16.
\textsuperscript{26} See \textit{generally Imagen | Google Research}.
\textsuperscript{27} The Response to Question 20 provides further detail regarding Google’s model cards and data cards.
to manage risk. **Second**, the organization deploying an AI application should be solely responsible for any disclosure and documentation requirements about the AI application because it is best positioned to identify potential uses of a particular application and mitigate against misuse. **Third**, whenever AI is playing a substantive role in decisionmaking (such as the allocation of government services or healthcare), that fact should be easily discoverable. **Fourth**, disclosures should be presented in clear, salient language to be meaningful to a wide audience and should provide an overview of the key tasks with which the AI is being deployed to assist, within the context of the application being offered.

Below is the kind of information that should be included as part of AI accountability disclosures.

- **Topline indication of how the AI system works** – Organizations should not be expected to reveal full details about AI models or underlying code—including due to risks to business confidentiality and the potential for adversarial gaming of the system. That said, model deployers should be expected to detail general logic and assumptions that underpin an AI application, particularly if it is designed for use in high-risk settings. It is also good practice to highlight the inputs that are typically the most significant influences on output, as well as any inputs likely to be deemed sensitive or unexpected. 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) should also be noted.

- **Expectations about how an AI system will be used** – When relevant, developers should clarify whether any operational constraints were intended in deploying an AI technology, such as whether the tool was designed to function independently or with a level of human oversight. There is evidence that users interact with AI systems and react to errors differently depending on such assumptions, so this information will help deployers build suitable mental models when using an AI application. While it is impossible to anticipate every possible use of an AI system, developers can state the intended use case for the model or system (e.g., those use cases against which its performance was tested and/or for which it is being marketed).

- **Known limitations on performance** – It will often be hard to describe in lay terms a model’s expected limitations or level of accuracy when operating under changing conditions. But general guidance can still be given. Research has shown that an AI application’s performance is best contextualized by presenting it alongside existing human performance statistics, where they exist.\(^{28}\) Concrete examples of successful and unsuccessful use cases are also helpful, particularly any challenging edge cases or known pitfalls regarding existing non-AI approaches the system has been explicitly designed to overcome.

---

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. This could include information about (1) how well the AI system performs against industry-standard evaluation datasets measured against key metrics (i.e., bias and fairness), (2) the frequency and cost weighting assigned to different errors (e.g., false negatives/false positives), (3) if relevant, how the AI system’s performance compares to existing human-performance benchmarks, and (4) whether (again, if relevant) any “dangerous capabilities” relating to cyber-offense, manipulation, and weapons acquisition may be at play.

And for frontier models, additional AI accountability elements should be considered. AI companies are already exploring the development of best practices around frontier models in consultation with experts. Companies should lead in performing voluntary self-evaluations and reporting around frontier models to affirm that the organization (and the models it is developing) can be deemed safe. This could include a set of evaluations on models and concurrently adopting and evaluating relevant organizational best practices for frontier model development. Developers could also consider publishing or sharing the results of evaluations with regulators and researchers as appropriate. Building a practice of establishing and evaluating requirements, documenting assessments, and promoting transparency in reporting could pave the way for future policies to ensure the safe use of the most advanced AI models.

6. The application of accountability measures (whether voluntary or regulatory) is more straightforward for some trustworthy AI goals than for others. With respect to which trustworthy AI goals are there existing requirements or standards? Are there any trustworthy AI goals that are not amenable to requirements or standards? How should accountability policies, whether governmental or non-governmental, treat these differences?

High-level accountability principles can be challenging to translate into an actionable compliance framework. For example, when building or deploying an AI tool to assist in decisionmaking, it is necessary to make choices upfront about how best to balance competing definitions of fairness. Without further guidance on what constitutes “fair” inputs to or outputs from an AI technology, regulated entities will likely come to different conclusions about what “fair” means.

To the extent regulators seek to impose assessment, auditing, or certification obligations, they would first need to clearly and reasonably define what those obligations are. For example, Article 10(3) of the EU’s 2021 draft of the Artificial Intelligence Act (AIA) would have required

---

30 There is precedent here in transparency reporting. Most social media companies report metrics on performance in order to improve accountability and transparency.
that “training, validation and testing data sets shall be relevant, representative, free of errors and complete.”31 These terms are ambiguous and may be inapt for AI involving large datasets that will almost never be “free of errors” or entirely “complete.” Helpfully, the latest amendments to the draft AIA suggest modifying Article 10(3) to state that datasets would need to be “sufficiently representative, appropriately vetted for errors and be as complete as possible in view of the intended purpose.”32

Given these types of challenges, Google supports flexible frameworks like NIST’s AI RMF that integrate stakeholder perspectives into practical steps that all organizations can take to further trustworthy AI goals.

7. Are there ways in which accountability mechanisms are unlikely to further, and might even frustrate, the development of trustworthy AI? Are there accountability mechanisms that unduly impact AI innovation and the competitiveness of U.S. developers?

See the Response to Question 27 for further details on potential legal barriers to AI accountability. See the Response to Question 30 for further details on why a sectoral approach to AI regulation is an appropriate starting point—when AI regulations are needed.

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

U.S. policymakers should adopt a common lexicon based on the work of the OECD, whose AI Principles state that “accountability” “best captures the essence of” the principle that “AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.”33 Google agrees that “AI accountability” should be thought of as a broad term that captures the ethical, moral, legal, regulatory, and industry expectations regarding AI development and deployment. Tools like assessments, audits, and certifications are means to actuate AI accountability.

33 Accountability (Principle 1.5), OECD AI Policy Observatory.
Existing Resources and Models

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

Google has drawn effectively from a number of accountability mechanisms to inform our internal AI governance structures. In 2018, we published our AI Principles to help guide our ethical development and use of AI, and we also established internal review processes led by our Responsible Innovation team to help us mitigate unfair bias, test rigorously for safety, and design with privacy top of mind. Our Principles also specify areas where we will not design or deploy AI, such as to support mass surveillance or violate human rights.

There are several efforts underway to establish internationally recognized standards for AI, including within ISO and IEEE, as well as industry-driven initiatives such as MLCCommons. While these efforts can highlight key areas for attention, it is important they are able to evolve in line with the still-rapid developments in underlying AI technologies. Ultimately it is unlikely that a single set of standards will emerge to suit all circumstances: multiple families of standards are more apt. As in similar domains such as cybersecurity, regulators should avoid the temptation to “pick winners” and instead allow flexibility for the optimal standards approach to be chosen for each context.

Google also encourages policymakers to monitor the work of the OECD and the Global Partnership on AI, two fora that are emerging as international clearinghouses for progress in AI governance, as well as the Global Network Initiative.

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

Setting baseline definitions for AI accountability terminology will help ensure that regulators and industry—domestically and internationally—are working toward common or consistent

---

34 See Response to Question 1.


36 MLCCommons is a collaboration between companies and academic researchers to build fair and useful benchmarks for measuring training and inference performance of machine learning hardware, software, and services. See, e.g., Cloud TPU v4 MLPerf 2.0 results | Google Cloud Blog.
goals. At the same time, 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.

For example, when building an AI tool to assist in decisionmaking, a developer must consider “fairness,” but this requires making choices upfront about how to balance competing definitions. Different technical approaches will result in models that are equitable in different ways and may require tradeoffs in terms of general accuracy or efficiency.

In many cases, existing discrimination laws will provide a framework for balancing competing equities, particularly in highly regulated industries. For example, in most jurisdictions, discrimination in lending is clearly defined in existing law, whether or not loan decisions are made by a human loan officer or an algorithm. Where existing discrimination laws provide clear guidelines and accountability mechanisms, new rules may be unnecessary.

But not all unfair outcomes result from illegal discrimination, and some AI systems may have unfair effects in ways unanticipated by existing laws and regulatory frameworks. In those cases, regulators should take a nuanced approach, ensuring that organizations consider the unique historical context in which an AI system is deployed, and use appropriate performance benchmarks for different groups to ensure accountability.

There is no one-size-fits-all set of fairness metrics, but at a minimum, regulators should ensure that organizations have clearly defined fairness goals based on historical context and clear performance benchmarks. Specifically, organizations should be able to answer the following questions:

- **How have people in different groups been historically affected by this kind of product or use case? How will they be affected by this particular product?** This assessment should be based on significant user testing, reviews by in-house subject matter experts, and consultations with relevant experts, including economists, sociologists, as well as civil and human rights experts.

- **How does the product perform across different user types (e.g., gender, age, skin tone, face/ body feature shapes, effect of lighting, effect of makeup/clothing, language, disabilities)?** For each application, there should be a clearly defined distribution appropriate for variant performance across groups, based on inputs including user testing, existing human levels of accuracy, published research, and legal requirements. The product should be tested across a diverse set of user groups and meet or be narrower than the target performance distribution among groups. Where appropriate, the performance distribution and/or determination process can be shared to provide users with more information and the opportunity to compare with alternative products.
11. What lessons can be learned from accountability processes and policies in cybersecurity, privacy, finance, or other areas?

Many of the principles central to AI governance are also emphasized within privacy governance practices (e.g., fairness, transparency). Google, for example, has a sophisticated internal privacy program to ensure that those involved at all stages of product development understand and implement our privacy principles to build products that are private by design. Companies, especially those that do not already have strong AI governance frameworks today, should be assessing how their privacy processes can be applied in an AI governance context. Lessons can also be drawn from reviewing external privacy accountability best practices.37

An important lesson from privacy regulation is the value of a risk-based, proportional approach. The European Union’s General Data Protection Regulation (GDPR), for example, encourages organizations to implement protections that correspond to risk levels of their data processing activities. Where processing is likely to result in a high risk to the rights and freedoms of individuals, the GDPR requires a data protection impact assessment that identifies and addresses those risks.

It is also important to create mechanisms to promote the safe and secure deployment of AI across borders, while ensuring regulatory interoperability. Google is a strong supporter of the Global Cross Border Privacy Rules (CBPR) system, a certification scheme that advances the trusted flow of information across borders.38 The structure of the CBPR system could serve as a model to develop and validate cross-border commitments on AI safety, security, and accountability.

Another principle that cuts across many areas of regulation is the importance of technology-neutral standards that can be tailored to different use cases. Such standards enable innovation while also assuring a responsible approach. Regulators should avoid the temptation to prescribe technology-specific or overly prescriptive standards that cut across different domains of AI use cases—and instead allow for the development of optimal standards that can be adapted for specific contexts and use cases.

For example, in the cybersecurity and privacy context, Google Cloud implements a “shared fate” model that may serve as a useful paradigm for AI.39 In this model, Google maintains an ongoing partnership with customers, providing a secure workspace and clear

37 One example is the Data Privacy Accountability Framework published by the Center for Information Policy Leadership.
38 Google and the Global Cross Border Privacy Rules.
39 Shared responsibilities and shared fate on Google Cloud | Architecture Framework (“You’re the expert in knowing the security and regulatory requirements for your business . . . When you run your workloads on Google Cloud, you must identify the security controls that you need to configure in Google Cloud . . . To decide which security controls to implement, you must consider the following factors: Your regulatory compliance obligations[,] Your organization’s security standards and risk management plan[, and] Security requirements of your customers and your vendors.”).
recommendations for security controls, settings, and associated best practices. This allows customers to choose the settings and protections that work best for their use cases, paired with trusted guidance from Cloud security and privacy professionals. As a cloud provider, we work with qualified auditors to demonstrate and maintain compliance with standards, such as ISO 27001 for security or ISO 27701 for privacy. Auditors then publish reports detailing their assessments of our data protection measures, and remain solely accountable for the reports’ accuracy.\(^{40}\) In appropriate circumstances, this shared fate model could extend to AI accountability use cases.

With respect to frontier models, additional cybersecurity best practices may be helpful within an AI governance context. Stakeholders could consider ways to allow a broader network of individuals to explore, study and scrutinize cutting-edge systems. For example, developers and labs could facilitate safe, arm’s length interactions with their AI systems (e.g., via shared APIs). There are early examples of structured access being practiced across other disciplines (e.g., as it relates to genomic data access\(^{41}\) and security), and there is an opportunity to further develop such access models in an AI context.

“Bug Bounty” programs are another effective, and increasingly common, cross-industry accountability practice. For example, in 2022, Google awarded over $12 million in bounty rewards to contributors who identified vulnerabilities—making our products even safer for users.\(^ {42}\) More and more, organizations across government\(^ {43}\) and the private sector\(^ {44}\) are looking to extend cyber-related bounty programming to AI applications. In this context, vulnerability reporting programs can be tailored to catalog potentially harmful use cases or toward detecting bias or other discrepancies in models or applications. We also conduct internal “red team” testing, where our Responsible Innovation team conducts proactive algorithmic product fairness testing with an eye toward avoiding reinforcing unfair bias and our Trust & Safety team evaluates and tries to mitigate potential content safety issues.\(^ {45}\)

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

Financial assurance systems could be a helpful reference point for future accountability mechanisms around frontier models. However, financial assurance systems have evolved over time based on mature service models; any new accountability mechanisms around AI should be designed in a way that recognizes the rapidly evolving nature of the technology and enables ongoing innovation.

\(^{40}\) See generally ISE Audit - Compliance | Google Cloud.
\(^{41}\) How your data is used | Genomics England.
\(^{43}\) Vulnerability Disclosure Program (VDP), Department of Defense Cyber Crime Center (DC3).
\(^{44}\) See, e.g., Reporting Twitter security vulnerabilities and bugs: Meta Bug Bounty Program.
\(^{45}\) See, e.g., Meet 3 women who test Google products for fairness.
To facilitate the implementation and adherence to evaluations and best practices, frontier labs may benefit from an internal risk and auditing function.\(^4\) This could improve risk management practices, decrease information asymmetries between different parts of the organization, and ensure adequate risk mitigation for uncertain, complex, interconnected, and volatile systems. Internal risk and audit functions should also be independent of management to safeguard objectivity, authority, and credibility. A multi-layered approach might target (1) an organization’s governance, (2) the frontier model being developed, and (3) downstream deployment and applications.\(^4\)

13. **What aspects of human rights and/or industry Environmental, Social, and Governance (ESG) assurance systems can and should be adopted for AI accountability?**

Companies have an obligation to regularly undertake ongoing human rights due diligence related to the impact of their business, products, and services. Where appropriate, a company should conduct a human rights impact assessment. Potential risks to human rights should initially be identified during standard review processes (either at the development or deployment stage). The assessment should incorporate a generalized human rights assessment framework, identify potential risks, and suggest methods for mitigating or avoiding potential harms. Policymakers should encourage these types of proactive human rights impact assessments and mitigations to further AI accountability.

International human rights standards can be instrumental in conducting assessments. For example, at Google, we are committed to respecting the rights enshrined in the *Universal Declaration of Human Rights* and its implementing treaties, as well as upholding the standards established in the *United Nations Guiding Principles on Business and Human Rights* (UNGPs) and in the *Global Network Initiative Principles*. This is managed by our centralized Human Rights Program with oversight from senior management and the Human Rights Executive Council led by Google’s President of Global Affairs and Chief Legal Officer.

Companies evaluating human rights impacts should take both micro and macro views of their policies and products. For example, in line with the UNGPs, our Human Rights Program conducts ongoing human rights due diligence across the company, and our commitments to human rights are embedded within our AI Principles Reviews led by the Responsible Innovation team, with an escalation path to the Human Rights team once specific criteria are met. Companies should also be encouraged to build trust through the use of third-party assessments. For example, through the Human Rights Program’s central management function, Google Cloud’s AI Principles review team enlisted Business for Social Responsibility (BSR) to conduct a formal human rights assessment of the Celebrity Recognition tool offered within

---

\(^4\) Jonas Schuett, *AGI labs need an internal audit function* (May 26, 2023).

Google Cloud Vision and Video Intelligence products. BSR applied the UNGPs as a framework, and their assessment informed the product’s design and the policies around its use.48

14. Which non-U.S. or U.S. (federal, state, or local) laws and regulations already requiring an AI audit, assessment, or other accountability mechanism are most useful and why? Which are least useful and why?

We see a growing interest from governments around the world in creating policy frameworks for these new technologies, such as Singapore’s Al Verify framework and the UK’s pro-innovation approach to AI regulation.

We are encouraged by the emerging consensus that a risk-based, proportional approach is the best way to strike a balance between protecting citizens and managing high-risk applications while supporting innovation in AI technologies.

While the legislation has not yet been finalized, the European Commission’s proposed AIA is a useful example to consider as the global discussion on AI governance advances. The proposal rightly underlines that “most AI systems pose limited to no risk and can contribute to solving many societal challenges” and reflects a risk-based approach to regulation, proposing “specific obligations for AI users and providers of high risk applications.”

However, some more recent proposals from the European Parliament and the Council would deviate from a risk-based approach. In particular:

EU co-regulators propose putting some of the requirements for high-risk AI systems on general-purpose AI (GPAI) systems and foundational models, undermining the Commission’s risk-based approach. GPAI is purpose-agnostic, making it difficult or impossible to appropriately evaluate, much less effectively mitigate, many downstream risks. To manage risk effectively, a system or product must be evaluated and managed based on the intersection of technology, application, and use case, with responsibilities for compliance assigned along the value chain.

With that, the AIA does not effectively distinguish between entities that develop AI technologies and those that deploy them. Those deploying the technologies are best positioned to meet many compliance obligations, especially when it comes to high-risk uses of AI technologies.49

Furthermore, the European Parliament proposed changes that risk overlapping with existing EU legislation, such as the Digital Services Act (e.g., by classifying recommendation systems used by Very Large Online Platforms as high-risk applications per se). Nothing in such systems

49 See the Response to Question 15 for Google’s views on the allocation of AI accountability in the AI supply chain.
is inherently risky or would necessarily have a “significant harmful impact” on the user’s health, safety, or fundamental rights. Such changes, again, challenge the concept of a risk-based approach.

More generally, data disclosure provisions should be considered in balance with principles of proportionality, trade secrets, and the potential impact on security and integrity of systems. Some proposals—such as making training data publicly available or providing regulators with “full access to the training, validation and testing datasets used by the provider” to specify and calibrate models\(^50\)—may conflict with other policy objectives and international AI governance. Similarly, provisions that would require developers to open up AI system “source code” for regulator assessment should be balanced.\(^51\) In many contexts, AI source code is highly sensitive information and, where compelled, disclosure could both compromise trade secrets and create security vulnerabilities that could be exploited by criminals and foreign adversaries. There are better methods for verifying AI systems’ performance (e.g., output auditing).

**Accountability Subjects**

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

Those deploying AI tools are best positioned to undertake risk assessments for a specific use because only they know the context of that use. To be sure, providers or developers of off-the-shelf, multipurpose AI component systems can and should provide general information about their construction (including guidance on operating boundaries and model behavior in the context of unforeseen inputs or user entries).\(^52\) That said, they should not be held responsible for conducting deployment risk assessments or validation, as they are unlikely to be well positioned to verify a system’s end uses in any exhaustive detail.

A practical approach for regulators would be to provide procedural “due diligence” guidance but assign responsibility for conducting applicable risk assurance exercises to the front-line organization using or deploying the AI application; the resultant documentation would provide evidence of satisfactory completion. Post-launch, if concerns arose that an application had been misclassified, remedial action could be taken.

\(^{50}\) AIA, Art. 64(1).

\(^{51}\) AIA, Art. 64(2).

\(^{52}\) Google has adopted a generative AI use policy that prohibits the use of Google’s AI tools for a variety of harmful conduct. [Generative AI Prohibited Use Policy](#).
In some contexts, it may be possible to adapt established design and validation processes, particularly when they stem from the same domain as the AI application in question. For example, the concept of a “failure modes, effects, and criticality analysis,” if tailored judiciously to suit the application context, may present a structured approach to documenting the expected impact of foreseeable safety risks and the corresponding preventive measures or reactive strategies planned if such failures were to occur.

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

Responsible practices for fairness, safety, privacy, and transparency should be incorporated early in product design and development lifecycles. Google’s research teams often begin the technical components of the Responsible AI evaluation process well upstream of any public deployment, including ML-based evaluation and validation, adversarial stress-testing to probe for errors and harms, and (where applicable) independent expert review to ensure models and datasets (where narrow datasets are used) abide by established standards, recognizing that models may sometimes need to train on problematic datasets precisely to recognize what constitutes problematic data.

This approach to AI research and development is also practical: it can help avoid burning engineering cycles spent retrofitting technology if an issue emerges after launch or even much later. In the early stages of development, and particularly as it relates to frontier models, there may not be a clear view of the ultimate shape of an AI product or capability, making it difficult to assess risks thoroughly at that time. Thus, it is important that accountability obligations provide flexibility for research, pilot programs, and experimental sandboxes. If such pre-assessment testing is not permitted, organizations may respond by taking an unduly precautionary stance—limiting fundamental investments in foundational research and impacting the pace of innovation.

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

Before a highly capable and complex AI system is launched, a risk assessment should be completed, with a deeper analysis of products and services that are deemed to present a higher risk. Our AI Principles reviews and consultations prioritize the evaluation of AI’s impact on humans, the environment in which a product is likely to operate, and society more broadly. Risk is a function of the magnitude of the harm multiplied by its likelihood and frequency. The
depth of analysis should be calibrated to the level of risk. We ask reviewers to consider both the impact of a risk (e.g., sociotechnical harm) and the probability (likelihood, frequency) of occurrence.

Regulators should use this framework as a starting point, providing guidance on the appropriate risk thresholds to apply at different stages of AI product development. We recommend scoping the risk of an AI application based on the severity of the harm and likelihood and frequency of its impact because it allows for various combinations of severity/likelihood to qualify as high-risk. Regulation should include guidance on when the risk classification of a given AI application flips from low or medium to high and reflect that the goal is to mitigate the severity of harm while also reducing its likelihood. Regulators should also be wary of being overly prescriptive on the format to avoid unnecessary documentation of low-risk scenarios.

Regulators might consider the following:

- **What is the inherent risk and benefit of applying this technology to this specific problem?** While there may be cases where a certain technology is inherently risky, more often, the primary driver of risk is derived from the precise use context.

- **How do the attributes of this particular AI system impact overall risk and benefit?** Specific design features and operational constraints and mitigations—both technical and business-driven—may reduce or increase overall risk. Factors include (1) consistency across groups or operating environments, (2) the reversibility of errors made by the system, (3) the degree and reliability of human control, (4) the existence of continuous learning safeguards (e.g., a hard limit on deviation from the predicted vs. baseline models), and (5) the presence of environmental mitigations built into the system and within the organization itself can bear on the calculation of risk.

- **Is the overall risk of this AI system tolerable when compared to existing alternatives?** The assessment should acknowledge the opportunity costs of not using AI in a specific situation or of intentionally developing AI without particular capabilities. The risks and benefits of AI systems should be weighed against existing (non-AI) approaches, including human judgment. If an imperfect AI system is shown to perform better than the status quo at a crucial life-saving task, for example, it may be irresponsible not to use the AI system. Where the alternative of not using AI poses a greater risk than using AI, AI actors should be supported, given their net benefit to society.

18. **Should AI systems be released with quality assurance certifications, especially if they are higher risk?**

As discussed in the Response to Question 6, quality certifications in the absence of clear benchmarks and standards are generally impossible to implement. For example, what is
“quality” or “good” in this context? Policymakers should pursue this line of accountability mechanisms only where clear benchmarks are established.

19. As governments at all levels increase their use of AI systems, what should the public expect in terms of audits and assessments of AI systems deployed as part of public programs? Should the accountability practices for AI systems deployed in the public sector differ from those used for private sector AI? How can government procurement practices help create a productive AI accountability ecosystem?

Audits, assessments, and certifications all support the delivery and application of trustworthy AI. As such, the government should ensure that those instruments, as well as the practitioners and developers delivering them, are applying trusted standards and frameworks. Absent standards or government guidance, providers of such services lack accountability. As with longstanding approaches to privacy, security, financial and other types of audits, the public should expect that these services are carried out by qualified individuals who may themselves be accountable.

Accountability mechanisms for AI systems should not be different for the public sector or private sector. In both cases, developers should be expected to deliver to deployers: documentation, guidelines, and recommended practices to support implementation as they use or build their own systems, as well as guidance regarding acceptable uses of the system. Further, the public agency, as the deployer, is best positioned to understand how they have chosen to deploy the system and the attendant risks. **AI tools are not appropriate for every use case**, and use-case-specific risks are best assessed by those with the most proximate knowledge about how an applicable system is to be used. Deployers should have a comprehensive risk management and/or governance program in place to assist in evaluating whether the proposed AI tools are fit-for-purpose for their use case.

It is reasonable to expect that the government encourages AI risk management on both the developer and deployer side. The NIST AI RMF, once implemented by all federal agencies, would drive a strong AI accountability ecosystem.

**Accountability Inputs and Transparency**

20. What sorts of records (e.g., logs, versions, model selection, data selection) and other documentation should developers and deployers of AI systems keep in order to support AI accountability? How long should this documentation be retained? Are there design principles (including technical design) for AI systems that would foster accountability-by-design?

AI developers can support accountability by retaining detailed documentation of datasets and models. For its part, Google has developed template documentation tools known as data and
model cards, which are used to simplify and standardize information about an AI model or its underlying dataset(s).

**Model cards** are short documents accompanying trained machine learning models that typically include information such as the model’s intended use case, the data used to train the model, the model’s performance on different metrics, any known biases or limitations of the model, and any potential risks or unintended consequences that could arise from its use. Model cards can also include information about the model’s training and evaluation processes and how the model can be deployed and integrated into different applications.

**Data cards** are a dataset documentation framework aimed at increasing transparency across dataset lifecycles. They provide structured summaries of ML datasets with explanations of processes and rationale that shape the data and describe how the data may be used to train or evaluate models. At a minimum, data cards include the following: (1) upstream sources, (2) data collection and annotation methods, (3) training and evaluation methods, (4) intended use, and (5) decisions affecting model performance.

Model and data cards can be useful for various stakeholders, including developers, users, and regulators, as they provide a standardized way to communicate important information about machine learning models. By increasing transparency and accountability, model cards can help build trust in ML and ensure that models are being developed and used in a responsible and ethical manner.

To help teams navigate transparency challenges with their ML datasets, Google introduced in late 2022 the Data Cards Playbook, which applies a human-centered design approach to documentation—from planning a transparency strategy and defining the audience to writing reader-centric summaries of complex datasets—to ensure that the usability and utility of the documented datasets are well understood. Our researchers also continue to iterate on designs for new transparency artifacts to accompany our datasets and complement model and data cards, including ones that are specific to particular sectors. For example, Google’s Healthsheets are a contextualized adaptation of the original datasheet questionnaire for health-specific applications.

---

54 The Data Cards Playbook.
21. What are the obstacles to the flow of information necessary for AI accountability either within an organization or to outside examiners? What policies might ease researcher and other third-party access to inputs necessary to conduct AI audits or assessments?

Policymakers should help support AI innovation and responsible deployment by fostering information flows that are important for AI accountability and innovation, including by:

- Establishing competition safe harbors for open public-private and cross-industry collaboration on AI safety research;
- Clarifying 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; and
- Adopting a uniform national privacy law that protects personal information, while establishing a framework for AI models’ incidental use of such data on the open web for training purposes.

22. How should the accountability process address data quality and data voids of different kinds? For example, in the context of automated employment decision tools, there may be no historical data available for assessing the performance of a newly deployed, custom-built tool. For a tool deployed by other firms, there may be data a vendor has access to, but the audited firm itself lacks. In some cases, the vendor itself may have intentionally limited its own data collection and access for privacy and security purposes. How should AI accountability requirements or practices deal with these data issues? What should be the roles of government, civil society, and academia in providing useful data sets (synthetic or otherwise) to fill gaps and create equitable access to data?

Accountability processes should include procedures for identifying problematic datasets and, when identified, include methods for rectifying those deficiencies. For its part, Google implements its AI Principles even at the earliest stages of development, including in the curation of datasets and modeling techniques to eliminate bias. For example, ML can be used to generate synthetic data and adversarial datasets, which can preserve privacy, offer a source of large-scale, representative data needed for model training, and provide other benefits.

In many contexts, data quality can play a crucial role in various aspects of an AI system’s performance, including its accuracy, effectiveness, fairness, robustness, safety, and scalability. Addressing data quality challenges is a task that academia, government, and industry should collaboratively undertake. Conventional AI/ML practices that undervalue the importance of data quality can lead to a phenomenon termed by Google researchers as “data cascades,” which refers to compounding events that trigger negative downstream effects due to underlying issues with data.
In the context of consequential decisions, poor data quality and data cascades can have particularly disproportionate effects on vulnerable communities.\(^5^6\) Paying close attention to questions of representativeness, intended use case, and the sociological contexts in which applications are deployed must be considered. An initial step toward avoiding data cascades and promoting robust data quality is comprehensive documentation for datasets, particularly for those that are publicly released or provided open source for widespread use. Universally accepted and practiced standards regarding data quality and documentation practices are underdeveloped, and more work is needed to design templates for disclosure, establish and vet best practices, and address challenges with pre-existing or widely used datasets.

In addition to documentation for datasets, organizations should consider adopting a domain-based approach that subjects data used in certain contexts to elevated accountability or reporting requirements (similar to the most recent EU approach).\(^5^7\) Ideally, specific documentation and reporting requirements could be segmented and developed into usable benchmarks over time. These benchmarks could inform appropriate use case applications for specific datasets in specific deployment contexts. And over time, with collaborative effort, domain-specific best practices could emerge as reliable guidelines for assessing whether a dataset is suitable for a particular use case or not. Eventually, governments and industry can work to establish standard datasets for model testing, related benchmarks, and test beds. This type of work is already being pursued by MLCommons, and industry should continue to play a role in developing open benchmarks.

As to government’s role in providing useful datasets, there is a wealth of data collected and stored by government that could be used for cross-sector learning and innovation.\(^5^8\) Borrowing themes from open government data work to create benchmarks for dataset use or encourage the fine-tuning of datasets for model training and testing would unlock this data for public and private research to advance the public interest and promote AI accountability.

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

This response is intentionally blank.

\(^{56}\) A significant body of research suggests that it is critical to ensure algorithms and data applications do not entrench and reinforce biases and harms.

\(^{57}\) Luca Bertuzzi, *MEPs seal the deal on Artificial Intelligence Act*, EURACTIV (Apr. 27, 2023).

\(^{58}\) See, e.g., *OECD Working Papers on Public Governance*. 

26
**Barriers to Effective Accountability**

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

AI accountability measures in the private sector, such as those utilized by Google, are effective and have helped lead to fairer, safer, and more inclusive AI technologies. But certain elements exist which make AI auditing challenging.

**Lack of agreed-upon standards** – There is a lack of consensus on technical standards (or set of standards) and responsible practices for developing and deploying AI, which can make it challenging to hold companies accountable for their use of AI. More support should be provided for efforts already underway to establish internationally recognized standards for AI, including within ISO\(^69\) and IEEE;\(^60\) as well as industry-driven initiatives such as MLCommons. While these efforts can highlight key areas for attention, it is important they are able to evolve in line with the still-rapid developments in underlying AI technologies. Ultimately it is unlikely that a single set of standards will emerge to suit all circumstances: multiple families of standards are more apt, while ensuring interoperability and common definitions.

**Challenges around explainability** – Limited explainability around many AI systems can lead to distrust and hinder accountability. Tailoring explanations to be meaningful and well suited to the needs of a range of audiences is difficult and time intensive. The ability to trace back and explain outcomes from AI systems operating at scale will likely differ substantially from the more extensive probing possible during development and upfront testing. While there has been much progress in tools to support developers, such as Google’s Explainable AI tool for Cloud AI customers,\(^61\) providing real-time explanations at scale remains challenging, in part because the detail and scope of what is needed varies significantly by sector and audience, and expectations may evolve as best practices emerge.

**Limited technical expertise** – Independent audits or risk assessments require specialized technical expertise that is often lacking in many organizations.

25. **Is the lack of a general federal data protection or privacy law a barrier to effective AI accountability?**

A U.S. federal privacy law could provide a beneficial foundation for AI accountability. Such a law should establish consistent, risk-based, U.S.-wide expectations for businesses around how

---

\(^{69}\) See Response to Question 1.  
\(^{60}\) See generally [AIS Standard: IEEE portfolio of AIS technology and impact standards and standards projects](https://standards.ieee.org/standards/ais/policy).  
\(^{61}\) [Explainable AI | Google Cloud](https://cloud.google.com/explainable-ai).
personal data is treated within AI models—including topics such as transparency, data sharing, data subject rights, and privacy by design. Without a single national standard, consumers will have uneven rights and may face overly-restrictive rules that frustrate the consumer experience. In a worst-case scenario, where patchwork data regulation creates conflicting obligations or prohibits the otherwise lawful and beneficial use of data, AI systems may be less accurate or less useful for consumers. A non-sectoral, federal privacy law could preempt patchwork state laws and create a uniform framework that provides consumers with clearer rights and privacy protections across the country. Google has long advocated for such a law, and there is bipartisan and public support for doing so. The Biden Administration has likewise emphasized the issue.

26. **Is the lack of a federal law focused on AI systems a barrier to effective AI accountability?**

Existing, sectoral-based frameworks remain a prime starting point for considering how to effectively promote accountability in AI systems. These approaches build on regulatory expertise in distinct subject areas that can be adapted to address AI in specific contexts under a hub-and-spoke model. In the Response to Question 30 below, we detail several steps policymakers can take to encourage the implementation of effective AI accountability tools.

27. **What is the role of intellectual property rights, terms of service, contractual obligations, or other legal entitlements in fostering or impeding a robust AI accountability ecosystem? For example, do nondisclosure agreements or trade secret protections impede the assessment or audit of AI systems and processes? If so, what legal or policy developments are needed to ensure an effective accountability framework?**

Certain types of AI accountability mechanisms may conflict with existing legal obligations. For example, imposing mandatory disclosure requirements for datasets used to train an AI model will rarely be practicable and could risk exposing trade secrets. It may also violate contractual obligations to not retain data supplied by business clients. Providing third-party access to data could also undermine privacy, such as by requiring a central log of data to be stored or preventing data from being deleted. More fundamentally, organizations that have built products using open-source models have no reliable way to know the provenance of the data used to train the models unless the publisher of that model has chosen to release it.

On the other hand, having clear copyright flexibilities, like fair use, is essential for data training and the development of responsible AI in the context of large language models. Training a large language model is a deeply transformative exercise that uses a vast range of data for machine

---

63 See, e.g., Cameron F. Kerry, Biden State of the Union 2023: Time to restart the privacy debate, Brookings Institution (Feb. 2, 2023).
learning. Further, training on a broad and diverse range of data is important in the responsible development of AI. For example, training only on older data (such as out-of-copyright books from more than 100 years ago) could result in model outputs being skewed based on biased or inaccurate assumptions.64

28. What do AI audits and assessments cost? Which entities should be expected to bear these costs? What are the possible consequences of AI accountability requirements that might impose significant costs on regulated entities? Are there ways to reduce these costs? What are the best ways to consider costs in relation to benefits?

While it is hard to estimate the cost of AI audits and assessments in the abstract, misdesigned AI accountability obligations could materially increase costs for AI developers and deployers. Organizations need to dedicate resources to data management, auditing, and monitoring systems, which may be passed on to customers, making AI technologies less affordable. Of course, audits and assessments can also provide significant value by promoting accountability and preventing costly product revisions and security patches that can be necessary downstream of processes without such measures.

29. How does the dearth of measurable standards or benchmarks impact the uptake of audits and assessments?

Google’s approach to AI governance using its AI RAF demonstrates how organizations can proactively assess AI risks even without standardized metrics or benchmarks. Likewise, audits focused on AI governance processes can be done without standardized metrics. On the other hand, audits or certification of a particular AI system would be hard to mandate without agreed-upon standards and benchmarks.

Regulatory agencies should take into account where benchmarks exist and where there are gaps. Where gaps are identified, agencies should request support from institutions best suited to develop standards based on robust stakeholder input, such as NIST.65

AI Accountability Policies

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

The Responses to Questions 31-33 provide additional suggestions on how government should fund and incentivize AI accountability efforts. This response focuses on regulatory efforts.

64 The only works sure to be in the public domain in the United States in 2023 are from 1927 or before. Each year on January 1, one more year’s worth of copyrighted works passes into the public domain. See Jennifer Jenkins, Public Domain Day 2023; see also Amanda Levendowski, How Copyright Law Can Fix Artificial Intelligence’s Implicit Bias Problem, 93 Wash. L. Rev. 579 (2018).
65 See AI Measurement and Evaluation, NIST.
Across all categories, AI accountability policies should be aimed at promoting opportunity, encouraging responsibility and trust, and promoting U.S. and international security.

Just as there is no horizontal regulation across the fields of mathematics or biology, the focus in AI regulation should primarily be on specific applications of AI—not the science of AI itself. There is an immense diversity of AI applications across almost all sectors of society—healthcare, financial services, transportation, energy, science, retail, agriculture, logistics, manufacturing, and beyond—and their impact on people and organizations is not the same. Additionally, many “AI issues” are actually issues common to the operation of any complex software already used by retailers, banks, insurance companies, manufacturers, and others. Thus, AI regulation is likely best addressed first through sectoral approaches that leverage existing regulatory expertise in specific domains rather than one-size-fits-all horizontal approaches.

Regarding frontier models, while sector-specific approaches are most appropriate right now, some development-stage controls may be reasonable regardless of deployment.

Governments should look first to existing regulatory experts, frameworks, and instruments that may encompass AI applications. Such sectoral experts typically will be well-positioned to assess context-specific uses and effects of AI and to determine whether and how best to regulate them, although sometimes additional resources may be required, including internal technical AI expert capacity. For instance, health-focused agencies are best positioned to evaluate the use of AI in medical devices, and energy regulators are best positioned to evaluate the use of AI in energy production and distribution. It will also be useful to have consistency in oversight and the expectations for human and machine actors performing the same task unless there are justifiable grounds for difference.

U.S. regulators, with guidance from the OMB and NIST, should update existing oversight and enforcement regimes to apply to AI systems, including clarifying how existing authorities apply to the use of AI and how to demonstrate compliance of an AI system with existing regulations. For example, risk assessments based on international consensus and multi-stakeholder standards like the ISO 42001 series or the NIST AI RMF can demonstrate compliance of a medical device with FDA rules. Regulators should reference existing standards and frameworks, such as the NIST AI RMF, and articulate how these frameworks can be used together to manage risk holistically.

If action is needed, regulators should avoid duplication and speed implementation by expanding established due diligence and regulatory review processes to include AI. When an AI application is not obviously covered by existing regulations, clear guidance should be provided on the due diligence criteria companies should use in their development processes. This would enable robust upfront self-assessment and documentation of any risks and mitigation strategies, and could also include further scrutiny after launch.
Regarding AI bug bounty programs mentioned earlier, we support the development and implementation of government programs—such as sandboxes and secure partnerships—as well as incentives and safe harbors for companies to develop and test bounty programs for AI harms, including research into methods for remedies (understanding that bias bounties will require more investigation, research, and unpacking than just a quick patch). We also welcome regulatory guidance to facilitate the integration of bounty programs in the field of AI so that the hacking and research communities alike have clarity around liabilities and predictable enforcement when it comes to finding and reporting vulnerabilities or bugs.

Finally, regarding responsible frontier model development, an ecosystem will need to develop on a global scale to address new risks posed by frontier models. This will require involvement and coordination between international institutions, standards organizations, and others—bolstered by focused collaboration, well-documented incentives, robust infrastructure, and the appropriate involvement and capacity of regulatory entities.

31-33. What specific activities should government fund to advance a strong AI accountability ecosystem? What kinds of incentives should government explore to promote the use of AI accountability measures? How can government work with the private sector to incentivize the best documentation practices?

Establishing incentives, benchmarks, and standards—for individuals, within organizations and teams, and among developers—will be central to advancing AI accountability. U.S. policymakers can advance AI accountability efforts through several initiatives:

- Increase funding for responsible AI research, including for identifying and mitigating bias in AI. This can be done by increasing funding for national and multinational research centers, fully implementing the National AI Research Resource (NAIRR) strategic plan,66 and investing in AI safety and alignment research in concert with leading AI labs;

- Establish and fund public-private partnerships to build and maintain high-quality datasets for AI research, establish best practices across industry, and push for standardization on key issues—directing resources to the NSF Technology, Innovation, and Partnerships Directorate could also further these goals;

- Establish and scale programs to educate regulators and decisionmakers about AI to enable more effective policymaking and investments, including how to consider tradeoffs among different policy objectives such as those discussed in Question 3;

- Fund programs that build technical and human capacity in the AI ecosystem to enable effective risk management;

• Devote greater investment and resources for creating evolving benchmarks and generally better metrology/metrics to measure, evaluate, and monitor AI development; and

• With respect to frontier models, establish incentives and safe harbors for companies to develop and test bounty programs for AI harms and consider guidance on integrating such programs in the AI field.

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

Google welcomes the broad-based efforts already underway to establish internationally recognized standards for AI systems accountability. An interoperable (if not identical) approach to AI governance should be preferred to a fragmented approach that could slow the pace of AI development and potentially limit the availability of new products and services to consumers in certain jurisdictions. Regulators should avoid adopting measures that inhibit cross-border research or disproportionately impact AI applications created in other countries. Cross-border cooperation among regulators is also critical to helping governments jointly develop and deploy AI to address global challenges related to public health, humanitarian assistance, sustainability, and disaster response.

Policymakers should drive international policy alignment, working with allies and partners to develop common approaches that reflect democratic values. Several steps can be taken, including:

• Prioritizing U.S. government agency and expert participation in international multi-stakeholder standards processes;

• Establishing multinational AI research resources similar to the NAIRR to offer compute and data to small- and medium-size enterprises and academics in allied countries to work on AI;

• Encouraging the adoption of common approaches to AI regulation and governance, as well as a common lexicon, based on the work of the OECD;

• Utilizing trade and other tools to prevent other governments from pursuing AI-related policies that are discriminatory;

• Establishing more effective mechanisms for information and best-practice sharing with allies and partners, as well as between governments and the private sector (e.g., for identifying actors engaging in economic espionage and attacks on AI systems);
• Advocating for trusted data flows across national borders, ensuring that allies and partners do not restrict data flows between each other to ensure high-quality data is available for AI modeling; and

• Promoting copyright systems that enable appropriate and fair use of copyrighted content to enable the training of AI models on a broad and diverse range of data, while supporting workable opt-outs for websites.

Conclusion

Google and Google DeepMind welcome the opportunity to share insight based on our experience and to learn from and engage with other participants. We look forward to sharing our experience and perspectives with NTIA and other stakeholders on these important matters.