ANTHROP\C

AI Accountability Policy Comment

Anthropic welcomes this opportunity to provide feedback to the National Telecommunications and Information Administration (NTIA) in response to its **AI Accountability Policy Request for Comment** (NTIA-2023-0005).

We understand the goal of this RFC is to promote greater accountability for artificial intelligence (AI) systems. There is currently no robust and comprehensive process for evaluating today's advanced AI systems or more capable systems of the future. Our submission presents a forward-looking perspective on the infrastructure needed to ensure AI accountability.

We see this RFC as an opportunity to help develop effective evaluations, define best practices, and build an auditing framework for AI that various government agencies can implement. Our recommendations consider the NTIA's potential role as a coordinating body that sets standards in collaboration with other government agencies like the National Institute of Standards and Technology (NIST). The outputs of this process will help evaluate and audit advanced AI systems at scale and promote greater AI accountability.

Summary of recommendations

In our recommendations, we focus on accountability mechanisms suitable for highly capable and general-purpose AI models. This summary addresses questions 30, 32, and 33 of the RFC. At a high level, our recommendations are:

- Fund research to build better evaluations
    - Increase funding for AI model evaluation research. Developing rigorous, standardized evaluations is difficult and time-consuming work that requires significant resources. Increased funding, especially from government agencies, could help drive progress in this critical area.
    - Require companies in the near-term to disclose evaluation methods and results. Companies deploying AI systems should be mandated to satisfy some disclosure requirements with regard to their evaluations, though these requirements need not be made public if doing so would compromise intellectual property (IP) or confidential information. This transparency could help researchers and policymakers better understand where existing evaluations may be lacking.
    - Develop in the long term a top-down set of industry evaluation standards and best practices. Government agencies like NIST could work to establish standards and benchmarks for evaluating AI models' capabilities, limitations, and risks that companies

would comply with.

- Create risk-responsive assessments based on model capabilities
    - Develop standard capabilities evaluations for AI systems. Governments should fund and participate in the development of rigorous capability and safety evaluations targeted at critical risks from advanced AI, such as deception and autonomy. These evaluations can provide an evidence-based foundation for proportionate, risk-responsive regulation.
    - Develop a risk threshold through more research and funding into safety evaluations. Once a risk threshold has been established, we can mandate evaluations for all models against this threshold.
        - If a model falls below this risk threshold, existing safety standards are likely sufficient. Verify compliance and deploy.
        - If a model exceeds the risk threshold and safety assessments and mitigations are insufficient, halt deployment and significantly strengthen oversight. Notify regulators to enable immediate review and oversight. Determine appropriate safeguards before allowing deployment.

- Require pre-registration for large AI training runs
    - Establish a process for AI developers to report large training runs ensuring that regulators are aware of potential risks. This involves determining the appropriate recipient, required information, and appropriate cybersecurity, confidentiality, and privacy safeguards.
    - Establish a confidential registry for AI developers conducting large training runs to pre-register model details with their home country's national government (e.g., model specifications, model type, compute infrastructure, intended training completion date, and safety plans) before training commences. Aggregated registry data should be protected to the highest available standards and specifications.

- Empower third party auditors that are…
    - **Technically literate** – at least some auditors will need deep machine learning experience;
    - **Security-conscious** and well-positioned to protect valuable IP, which could pose a national security threat if stolen; and
    - **Flexible** – able to conduct robust but lightweight assessments that catch threats without undermining US competitiveness.

- Mandate external red teaming before model release
    - Mandate external red teaming for AI systems, either through a centralized third party

(e.g., NIST) or in a decentralized manner (e.g., via researcher API access) to standardize adversarial testing of AI systems. This should be a precondition for developers who are releasing advanced AI systems.
    - Establish high-quality external red teaming options before they become a precondition for model release. This is critical as red teaming talent currently resides almost exclusively within private AI labs.

- **Invest in interpretability research without yet mandating fully interpretable systems**
    - Increase funding for interpretability research. Provide government grants and incentives for interpretability work at universities, nonprofits, and companies. This would allow meaningful work to be done on smaller models, enabling progress outside frontier labs.
    - Recognize that regulations demanding interpretable models would currently be infeasible to meet, but may be possible in the future pending research advances.

- **Enable industry collaboration on AI safety via clarity around antitrust**
    - Regulators should issue guidance on permissible AI industry safety coordination given current antitrust laws. Clarifying how private companies can work together in the public interest without violating antitrust laws would mitigate legal uncertainty and advance shared goals.

About Anthropic

Anthropic is an AI safety and research company working to build reliable, interpretable, and steerable AI systems. Our legal status as a public benefit corporation enables us to make choices that are in line with our mission of creating a materially positive impact on society, even if they do not maximize shareholder value. Our mission has three parts: (1) developing empirical safety research on large language models, (2) setting and demonstrating corporate norms and voluntary standards for responsible AI development and deployment, and (3) policy engagement.

On (1), Anthropic builds and studies large language models (LLMs) to identify risks and opportunities, and to deploy commercial systems that are beneficial and useful to people. We believe it is necessary to do safety research on large models, which are qualitatively different from smaller models. By building "frontier" AI models, we gain insights into both their capabilities and limitations—this is central to our views on AI safety.

On (2), we develop and follow voluntary norms on responsible AI development and deployment, such as Trust and Safety and cybersecurity best practices. We hope to set an example for other industry actors by piloting such practices and sharing them at conferences, in working groups, and in public documents.

On (3), Anthropic engages with policymakers in the United States and abroad to decrease technological surprise, increase the stability of the AI development environment, and support the crafting of thoughtful regulation. We do this by conducting and publishing original research, publicly and privately engaging with policymakers, and responding to requests for comment.

Responses to questions

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?

There are several promising AI accountability mechanisms that can inform people about how LLMs operate, some of which are useful today and others which are still under development.

Near-term mechanisms

In the near-term, techniques like model cards, model values transparency, watermarking, and model evaluations can increase accountability and oversight for AI systems.

Model cards are used to share information about a model's intended purpose and limitations, training data, and performance, and can include elements like algorithmic audits to evaluate a system's components and data sheets. Such disclosures increase transparency and allow oversight into a model's development and suitability for a particular use. We will be releasing model cards as part of all future model releases.

Model values transparency comprises defining and conveying a model's objectives in an understandable way. For example, our "Constitutional AI" approach expresses model values in natural language to make them transparent. End users can be involved in developing model values to make the process more democratic. We publicly shared the constitution for Claude v1.3 and plan to share the constitutions that guide all of our publicly released models.

Watermarking uses techniques like hash functions to certify that a specific model generated a specific output. Early research suggests that LLM developers like Anthropic could potentially apply this tool in limited circumstances to certify text generated by their language model at the time of generation. Unfortunately, there are open research problems to solve in watermarking. Currently, broader watermarking efforts would be fairly easy to defeat by malicious actors; such actors may also use techniques like prompt engineering to generate harmful or misleading "certified" text. We are researching watermarking and are open to implementing it, but do not believe that it can be considered an independently reliable accountability effort; moreover, the potential use cases for LLM text watermarking require further development.

Model evaluations refer to systematic assessments of a model's capabilities, performance, fairness, security, and other attributes by internal teams or external experts. These are discussed in depth in our response to question 9. We believe that better evaluations are the first step towards a comprehensive AI accountability regime. Creating effective evaluations is a core research focus for Anthropic, but we believe more work must be done. We would welcome more external research and government support for such development.

Longer-term mechanisms

In the longer term, we believe certain research directions may enable a deeper and more comprehensive method of informing people about how AI tools are operating.

Mechanistic interpretability aims to reverse-engineer neural networks into explainable algorithms. This could eventually help to catch safety issues pre-deployment, address concerns like deception, and identify unknown capabilities. This is also the most comprehensive way to make models transparent and understandable, with positive implications for AI accountability efforts.

Anthropic leads research on mechanistic interpretability, but the challenge remains unsolved. Interpretability research in AI can play the role of biology research in medicine: investments in understanding a complex system (i.e., basic science) can find ways to improve safety that complement black-box evaluation and monitoring schemes (e.g., the FDA). However, regulations mandating interpretable models are premature; AI developers cannot currently comply with such mandates, and it is unclear on what timeline large-scale models can be made fully interpretable. Another bottleneck to interpretability-based regulation is that we do not yet know how to measure interpretability, underscoring the importance of developing interpretability evaluations. While a blanket regulation demanding interpretability is currently infeasible, it might be reasonable to demand that models that

exceed a certain capabilities threshold meet certain performance and explainability requirements in the future. Such a requirement could incentivize labs to invest more in interpretability research.

While research funding and incentives can—and should—support interpretability work, governments must be wary of "explanations" that lack real and measurable insight into the models. In the meantime, we must rely on existing tools which help make models more reliable and safe, including constitutional AI, evaluations, and red teaming. We discuss existing AI accountability mechanisms in our response to question 9.

Recommendations:

- Standardize model disclosure expectations. Develop guidelines for transparent model information, such as its intended use cases, performance on evaluations and red teaming, and risk assessments.
- Explore watermarking for LLMs and other generative models. Invest in approaches to verify if text or other media was generated by a specific model.
- Industry-government collaboration to design a framework for safety incident reports. This should comprise identifying, investigating, and disclosing significant safety incidents or findings in a timely manner.
- Increase funding for interpretability research. Provide government grants and incentives for interpretability work at universities, nonprofits, and companies. This would allow meaningful work to be done on smaller models, enabling progress outside frontier labs.
- Recognize that regulations demanding interpretable models would currently be infeasible to meet, but may be possible in the future pending research advances.
- Build processes allowing for greater public participation and feedback to determine which values guide AI systems (i.e., public participation in designing a system's constitution). Make the process of developing model values transparent and inclusive relative to its users and use cases to build trust and ensure alignment with public priorities and concerns.

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?

While oversight and accountability are crucial for building trustworthy AI, insufficiently thoughtful or nuanced policies, liability regimes, and regulatory approaches could frustrate progress.

Policies that put confidentiality and proprietary information at risk. AI systems and training run disclosures, audits and assessments—especially by third parties—can expose confidential and proprietary information, including trade secrets, IP, and private data. This could accelerate race dynamics between AI developers or adversarial states. It could also undermine cybersecurity and national security by revealing vulnerabilities or sensitive data access points in a system that malicious actors could exploit. Requiring auditors to follow cybersecurity best practices suitable to the risk level of the information they are handling could mitigate this.

Regimes that allocate all liability to model developers. Liability regimes that hold model developers solely responsible for all potential harms could hinder progress in AI. Developers cannot reasonably predict or control the full range of uses of such a general-purpose technology. Assigning liability without accounting for nuances in the value chain could discourage innovation. For example, those who build directly on these models for commercial applications should reasonably expect some liability for potential harms from their specific use cases. Similarly, end users and companies that deploy AI also have a duty of care.

A patchwork of competing certifications would hinder global progress. Certifications are often region- or country-specific, which risks creating a patchwork of competing standards that act as barriers to global collaboration and trade. For example, early versions of the European Union's (EU) AI Act proposed requiring that systems use only EU data to train products for the EU market. This could have resulted in significant biases and worse performance, while needlessly consuming additional resources, such as energy.

Recommendations:

- Narrowly tailor mandated disclosures of proprietary AI information to avoid overexposure. Audits and assessments should require the minimum information to verify system properties. Strict controls should govern how third parties handle sensitive data and auditors need to follow cybersecurity practices appropriate for the risk level.
- Provide a careful, context-specific analysis to parse liability across the value chain. Liability could be allocated to different parties (e.g., developers, deployers, and end users) depending on factors like the degree of customization, including incorporation of new data, customer and user vetting, and the ultimate end uses.
- Seek international cooperation to develop interoperable AI standards and certifications. The US should work with the EU, Five Eyes, and other partners to harmonize safety standards across markets for regulatory consistency that promotes a "race to the top."

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?

Model evaluations—for both capabilities and harms—are a critical part of today's AI accountability mechanisms, but today's evaluations are an incomplete patchwork, and creating effective evaluations is an inherently difficult task. When evaluating highly capable general-purpose AI models, we must make complex normative decisions, including determining what tasks or model properties to prioritize, choosing appropriate metrics, and establishing confidence in these measurements.

There are four main types of model evaluations, each with trade-offs:

1. **Multiple choice evaluations** are standardized tests for models. Several examples are: the Multi-task Language Understanding (MMLU), which measures a model's world knowledge and problem solving ability on topics ranging from mathematics to U.S. history to law; the Bias Benchmark for Question Answering (BBQ), which evaluates social bias across nine protected attributes, such as race, gender identity, and religion; and the Beyond the Imitation Game Benchmark (BIG-bench), which comprises 204 diverse benchmark implementations that probe a variety of model capabilities (e.g., emoji detection) and risks (e.g., social biases). Although these benchmarks provide useful signals regarding language understanding, social biases, and model capabilities, they are not necessarily tied to real-world use—people do not necessarily administer multiple choice questions to chatbots. Furthermore, the development of bias and fairness benchmarks like BBQ is inherently difficult; BBQ took roughly 2.5 people years distributed across 7 PhD students and 1 professor over the course of 6 months to build. As of writing, there is no working public implementation of BBQ, so model developers need to implement this benchmark from scratch. Finally, due to the collaborative and bottom-up nature of BIG-bench, the benchmarks vary in quality and the implementations may not be optimized for various model developers' specific intended user infrastructure.
2. **Third-party auditing** includes methods and tools to evaluate the capabilities and limitations of models run by an independent third party organization. For example, Stanford's Holistic Evaluation of Language Models (HELM) standardizes multi-metric measurement for a range of language models in an effort to make language models comparable with each other and the

Alignment Research Center (ARC) is spearheading an effort to detect existential risks from AI. Each of these approaches has drawbacks. The downside of HELM's standardization efforts is that different models require different methods of interaction (e.g., Anthropic's Claude has a different effective prompting style than OpenAI's GPT-4), so a balance must be struck between uniformity and bespoke evaluations. Finally, ARC's work is in early stages; it is still building robust standards and developing the appropriate tooling and expertise to run its audits.

3. **Human-in-the-loop evaluations** task humans with measuring model performance. For example, red teaming denotes an approach wherein models are adversarially tested to identify issues before release. This ranges from soliciting biased responses to eliciting dangerous behaviors. Data from successful red teaming attacks can then be used to make future iterations of models more difficult to red team. Anthropic uses red teaming to evaluate and mitigate model vulnerabilities and harms; however, just because red teamers fail to elicit dangerous information from language models does not necessarily mean that such vulnerabilities do not exist. Anthropic also conducts A/B tests, in which crowdworkers reveal preferences about which of a set of models is more helpful or harmless. Through this procedure, we measure what is called an ELO score. Models with higher ELO scores are typically more helpful and more harmless; however, it is not clear that helpfulness or harmlessness are the best characteristics to test for.
4. **Model-generated or model-scored evaluations** automate evaluations of models. Previous Anthropic research used a language model to generate true/false statements that evaluate model characteristics on concerning behaviors (e.g., desire for acquiring power, having a liberal or conservative leaning viewpoint, etc.). Validating these model-generated evaluations is difficult, though Anthropic research has found 90-100% human-machine agreement that the model generated evaluations indeed test the phenomena they are designed to test. An example of model-scored evaluations is the RealToxicityPrompts evaluation, which investigates the extent to which pretrained language models can be prompted to generate toxic language. However, the evaluation relies on a toxicity classifier that often returns false positives and can itself exhibit various social biases (e.g., flagging African American Vernacular as toxic when the content itself is not toxic).

Going forward, as the amount of compute power going into these systems increases and they become more advanced, it will be critical to carry out dangerous capability evaluations. These evaluations would assess whether dangerous capabilities exist in a model or could be elicited, using any of the four approaches mentioned above or some combination thereof. These capabilities evaluations could, in turn, inform the development of risk thresholds for imposing tighter regulations on models that possess dangerous capabilities. This nascent area of research would benefit from more collaboration across industry, government, academia, and other stakeholders.

Recommendations:

- Increase funding for AI model evaluation research. Developing rigorous, standardized evaluations is difficult and time-consuming work that requires significant resources. Increased funding, especially from government agencies, could help drive progress in this critical area.
- Mandate external red teaming for AI systems, either through a centralized third party (e.g., NIST) or in a decentralized manner (e.g., via researcher API access) to standardize adversarial testing of AI systems. This should be a precondition for developers who are releasing advanced AI systems.
- Establish high-quality external red teaming options before they become a precondition for model release. This is critical as red teaming talent currently resides almost exclusively within private AI labs.
- Require companies in the near-term to disclose evaluation methods and results. Companies deploying AI systems should be mandated to satisfy some disclosure requirements with regard to their evaluations, though these requirements need not be made public if doing so would compromise IP or confidential information. This transparency could help researchers and policymakers better understand where existing evaluations may be lacking.
- Develop in the long term a top-down set of industry evaluation standards and best practices. Government agencies like NIST could work to establish standards and benchmarks for evaluating AI models' capabilities, limitations, and risks that companies would comply with.
- Increase funding for and coordination around dangerous capability evaluations, where researchers rigorously test AI systems for characteristics like deception, manipulation, and other potentially harmful behaviors. These evaluations are critical for ensuring AI progress remains beneficial and aligned with human values, but they require specialized expertise, resources, and collaboration across stakeholders.

11. What lessons can be learned from accountability processes and policies in cybersecurity, privacy, finance, or other areas?

Cybersecurity policies and practices offer valuable lessons for AI developers that can help ensure accountability and responsible development. Developers should adopt stringent software development standards, like the NIST Secure Software Development Framework (SSDF) and Supply Chain Levels for Software Artifacts (SLSA). Applying these frameworks would allow end users and regulators to determine an AI system's provenance by tracing it back to its developers through a cryptographic signature.

Requiring SSDF and SLSA for AI would mandate threat assessments, strict authentication and access controls, continuous monitoring for new threats, documentation and change justification, and more.

Other cybersecurity standards, like the Sarbanes-Oxley Act, ISO 9001, and Good Manufacturing Practice require multiple individuals to approve any changes to production systems. AI systems would benefit from similar checks and balances. Several major tech companies already require two-party, time-limited authorization to implement changes or grant access to critical/sensitive data. AI labs should adopt similar two-party controls to access and modify highly sensitive AI components like model weights. Introducing practices like this would make it much harder for bad actors to steal model weights.

Red teaming, a common cybersecurity practice, has also proven to be useful for AI systems. As discussed in our response to question 9, we recommend that external red teaming of AI systems be mandated. Developers such as Anthropic should have independent groups systematically try to break a model or elicit harmful behavior before deployment so issues can be addressed.

In summary, AI developers should apply lessons from cybersecurity by following best practices and using frameworks that protect critical infrastructure. Compliance standards, multiple-factor authentication, and red teaming are all strategies that can and should be adapted for AI. Although cybersecurity offers many useful lessons for AI, the two domains also differ in key ways. AI presents potential threats of bias and unfairness that software alone does not. This emphasizes the importance of tailoring best practices to the unique concerns of this domain.

Recommendations:

- Require stringent secure software development standards for frontier AI systems. Frameworks like SSDF and SLSA provide comprehensive guidance on building security into all phases of software creation and implementation. While frontier AI research already benefits from cloud providers applying some of these standards, regulating their use for AI would step-change system security.
- Press for collaboration between international bodies to set high, interoperable standards for AI security, oversight, and governance on a broad scale.

16. How should AI accountability mechanisms consider the AI lifecycle?

Audits and assessments of AI systems should be timed to evaluate models at critical stages in their development to ensure meaningful accountability. Specifically, reviews need to target models during and

directly after training, as well as prior to deployment. Evaluating models at checkpoints throughout training and upon completion can help detect potential harms before models become widely deployed.

For frontier models that are rapidly advancing in capability, evaluations should take place more frequently. Specifically, major training runs that produce significantly more capable models should trigger new rounds of audits and assessments. Given the present lack of more meaningful capabilities evaluations, a rough proxy for the time being could be training runs requiring at least 10^26 floating-point operations (FLOP) or more in compute power as grounds for review. While model size and compute power are imperfect proxies for capability, they can be used as stopgap measures during the development of more nuanced evaluations.

More broadly, the frequency of audits and assessments should depend on changes in a model's actual capabilities, not just proxies like model size or the compute power used in training. Systematic evaluations of various model abilities could determine if and when new reviews are needed based on significant changes in performance. For example, if an AI model is altered (e.g., by adding or removing a classifier) and this causes it to achieve dramatically different results on a capability test relative to its last evaluation, a new audit may be required even if the model was not retrained.

To implement appropriate timing for audits and assessments, policies should focus on evaluating models frequently enough to be able to detect meaningful changes in their capabilities or the addition of new abilities. For now, this largely corresponds to major training runs, but policies must be flexible enough to account for future progress.

Recommendations:

- Establish a confidential registry for AI developers conducting large training runs to pre-register model details with their home country's national government (e.g., model specifications, model type, compute infrastructure, intended training completion date, and safety plans) before training commences. Aggregated registry data should be protected to the highest available standards and specifications.
- Develop a risk threshold through more research and funding into safety evaluations. Once a risk threshold has been established, we can mandate evaluations for all models against this threshold.
    - If a model falls below this risk threshold, existing safety standards are likely sufficient. Verify compliance and deploy.

- If a model exceeds the risk threshold and safety assessments and mitigations are insufficient, halt deployment and significantly strengthen oversight. Notify regulators to enable immediate review and oversight. Determine appropriate safeguards before allowing deployment.

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?

Several regulatory proposals focus on model size, wherein larger, more compute-intensive models allegedly pose greater risks. Rather than relying on size or computing thresholds—which change rapidly as the state-of-the-art shifts—we propose evaluating what actually makes models (of any size) dangerous: their capabilities. The most prudent metric is whether models demonstrate capabilities through targeted evaluations.

Capabilities evaluations could focus on specific risks. For persuasion, manipulation, and misinformation concerns, models can be assessed on their ability to significantly influence the beliefs of a diverse population sample on a given topic. To address systems autonomously taking dangerous actions in the world, models can be evaluated on their ability to upload themselves to the internet and generate income, or to design weapons or launch offensive cyberattacks.

In summary, the scope of AI accountability measures should depend on a model's capabilities and deployment risks. Capabilities are best determined through tailored evaluations. Regulations should primarily focus on highly capable models which pose greater risks if misused or irresponsibly deployed. Maximally effective regulation would target model capabilities and safety rather than proxy attributes like size or the compute power used in training.

Recommendations:

- Develop standard capabilities evaluations for AI systems. Governments should fund and participate in the development of rigorous capability and safety evaluations targeted at critical risks from advanced AI, such as deception, autonomy, and more. These evaluations can provide an evidence-based foundation for proportionate, risk-responsive regulation.
- Tailor regulations to model capabilities and use cases. Broad regulations based primarily on compute power or dataset size alone are likely to be either over- or under-inclusive. Regulations

should be tailored to the specific capabilities and risks demonstrated by models for different use cases. More capable and potentially higher-risk models warrant stricter standards and controls.

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?

We provide documentation regarding our model for all our commercial users, including guidance on appropriate uses and known limitations. Our Acceptable Use Policy stipulates prohibited uses (e.g., abusive or fraudulent content; violent, hateful or threatening content), prohibited business use cases (e.g., criminal justice decisions, determining financial eligibility), and restricted business use cases (e.g., legal, medical), which may be permitted only if they pass further customer and use case vetting. In the future, we plan to provide users with a model card, a short document that details our model's performance and safety characteristics.

Regulators should also receive more information about models earlier in development. Frontier models can introduce unforeseen risks, even rising to the level of national security concerns. Additionally, new capabilities tend to emerge in new and larger models. To increase the awareness of regulators to these potential new capabilities, we propose that AI developers pre-register large training runs (i.e., the compute-intensive process that results in a new model) with an appropriate organization, and under appropriate confidentiality restrictions, within their home country's government before they begin. While the exact cutoff size for reporting training runs deserves further research and discussion, we think a plausible bar for now would be 10^26 FLOP or higher.

Pre-registration could require details such as:

- Model size
- Type of model (e.g., LLM, multimodal)
- Total compute power
- Intended completion date
- Safety plan and risk assessment
- Monitoring and halting procedures
- Responsible team members

Such a registry would require legal protections to ensure confidentiality, preclude public disclosure, and secure aggregate data against theft or espionage. However, the pre-registration commitment could be public.

Alongside pre-registration, the government could also explore tracking and regulating advanced models and the components required to develop them. This would provide the regulatory structure to allow regulators to verify that model developers were complying with safety standards. Questions remain on how to make such a system robust, privacy-preserving, and minimally burdensome so that it does not compromise US competitiveness.

Possible components include:

- A registry tracking transfers and uses of cutting-edge AI chips
- Pre-registration and safety verification of large new model runs
- Compliance auditing (e.g., remote or on-site monitoring)
- Enforcement actions against non compliant developers
- Tracking the export of model weights and code for models above a certain capability threshold

Recommendations:

- Implement best practices for industry groups to prohibit harmful uses like disinformation, child sexual abuse material, or terrorism. Collaborate on means of monitoring and disclosing violations.
- Establish a process for AI developers to report large training runs ensuring that regulators are aware of potential novel risks. This involves determining the appropriate recipient, required information, and appropriate cybersecurity, confidentiality, and privacy safeguards.
- Consider more stringent measures for ensuring adherence to safety standards, including controlling and monitoring access to advanced AI models and the components required to develop them.

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?

One of the biggest obstacles to AI accountability is the sensitive nature of the data and models involved. Advanced neural networks contain sensitive information like model weights that could be misused by malicious actors and which may represent tens of millions of dollars or more of investments in compute

resources. Providing auditors and third-party evaluators access to this sensitive information increases risks of leaks or hacking.

Although current behavioral or fine-tuning based evaluations do not require access to model weights, it is plausible that future evaluations—particularly those assessing transparency and interpretability—will require such access to conduct a full, robust evaluation of the system and its capabilities. However, this will need to be set against proliferation risks; as an AI lab, we cannot keep our model weights secure if external companies are regularly accessing them. This highlights a significant trade-off between good safety evaluations and model security, which could be partially addressed through exchange programs/secondments between companies and auditors. We discuss this in more detail in our response to question 24.

In general, zero-shot evaluations where auditors run an evaluation only once, or audits run via the same kind of API access the average customer receives, are insufficient for highly capable models. Models can often perform new, complex tasks once they are fine-tuned on data specific to that task. Therefore, auditors need the ability to fine-tune models themselves to determine a model's full range of capabilities. Fine-tuning access also enables the auditor to evaluate how the model will behave if a bad actor conducts their own fine-tuning to customize the model for a malicious purpose. While an auditor can fine-tune with full access to the model weights, this level of access would enable serious risks of leaks or cyber attacks and would require intense security and confidentiality policies to protect these weights. A less risky alternative that would still permit the auditor sufficient access is via a fine-tuning API which allows them to fine-tune the model but places a screen between the auditor and the weights. Anthropic is building a fine-tuning API for ARC, our independent auditing partner.

Sharing information with auditors via a fine-tuning API—especially if there are rare cases where an auditor will need direct access to weights—poses risks to the companies developing AI systems. To mitigate this, auditors and their teams must have substantial expertise in machine learning and information security. They need to fully comprehend the complex systems they are evaluating while simultaneously preventing any possibility of leaks or hacking. The most advanced auditing firms that evaluate complex infrastructure in other industries employ teams of experts, and sometimes embed employees within companies to monitor systems. A similar level of expertise and access will be required to properly audit advanced AI systems.

While some high-level information like whether a model passed an evaluation could potentially be shared more widely without compromising the full model, detailed access should be restricted to select, highly-vetted auditors. Overall, enabling meaningful auditing and accountability of AI systems will

require striking the right balance between access and security. With the proper safeguards and expertise in place, auditors can be given enough access to evaluate AI robustly without substantially increasing risks. But zero-shot or surface-level evaluations are not sufficient for evaluating the capabilities of complex and advanced AI systems. Accountability requires a commitment to transparency and a willingness to share sensitive details with trusted, technically-proficient partners.

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?

Creating robust AI accountability mechanisms in the private sector faces several significant challenges.

Evaluating AI systems is extremely difficult. Developing evaluations requires determining what capabilities or risks we want to measure, choosing appropriate metrics, sometimes involving human crowdworkers, and building custom tools. General-purpose AI systems may require bespoke evaluations for different capabilities. Evaluations also require substantial expertise and resources to develop. Many "off-the-shelf" evaluations are not yet scientifically validated and do not run "out of the box." More broadly, many stakeholders outside of AI labs presume that the state of AI evaluation is both more mature than it currently is, and that precise evaluations exist for what policymakers or other stakeholders may care about. Neither of these things are true. It is critically important that policymakers fund the development of effective AI evaluations and base any regulatory proposals on a sober and objective assessment of the current state of evaluating AI systems.

"Red teaming", or adversarial testing of AI systems, is not standardized. Frontier AI labs like Anthropic already use external domain experts to probe their systems for vulnerabilities. While this is certainly a step in the right direction, effective red teaming requires both domain expertise and skill in eliciting concerning behaviors from AI systems. The latter skill, known as "prompt engineering," is rare and valuable, and largely resides within AI labs, which makes independent red teaming particularly difficult. Centralizing red teaming in an organization like NIST could help address this by creating threat intelligence repositories and coordinating experts.

External auditing can be resource-intensive for AI labs to support. Since November 2022, Anthropic has worked closely with the Alignment Research Center (ARC) to evaluate whether our language model can "survive and spread" (i.e., autonomously replicate and acquire resources). Conducting such an evaluation requires expertise in capabilities elicitation, which refers to discovering a system's latent capabilities. This

is a very challenging process. While external auditors should aim to conduct replicable and independent evaluations in the long term, in the near term, nascent fields like AI safety require collaboration with companies to design meaningful assessments. This is the case because the people who built these systems are most familiar with how to interact with them. Indeed, frontier AI labs have deep expertise, which makes them uniquely well-placed to elicit capabilities from their models. However, this introduces conflicts of interest that must be addressed through mechanisms such as exchange programs/secondments between companies and auditors. Our major takeaways are that auditors still rely heavily on people who work at the labs—which can be very resource-intensive—and that auditors need to have both engineering and research expertise in order to eventually bring an informed independent eye to the assessment process.

Recommendations:

- Increase government funding of academic and government-run projects that seek to develop meaningful evaluations of AI systems.
- Conduct a comprehensive landscape assessment of existing AI evaluations to determine which are scientifically valid and for which purposes.
- Create personnel exchange programs that temporarily embed auditors within companies and companies' staff within auditing bodies to strengthen collaboration and trust in the auditing process. However, conflicts of interest would need to be addressed.
- Ensure that policymakers do not attempt to define static standards for AI accountability. At least in the near-term, capabilities elicitation should be viewed as a joint scientific research process between auditors and AI labs rather than a compliance exercise. Auditors could define evaluation goals while companies leverage their expertise to test their systems.
- Forge policies that create incentives for companies to rigorously search for and disclose their systems' weaknesses and limitations, rather than hiding or downplaying them. If carefully tailored, liability regimes that penalize companies for failing to adequately evaluate and disclose risks may provide such incentives.

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?

The lack of clarity regarding ownership of AI-generated content and the failure of traditional legal frameworks to keep pace with technological change are two factors currently impeding a robust AI accountability ecosystem.

The lack of clarity regarding ownership of AI-generated content affects AI providers' ability to protect their IP and creates uncertainty for AI users about their rights to the outputs. This is a significant hurdle in clearly allocating rights, obligations, and liabilities through the AI development and deployment pipeline. For instance, copyright bodies and courts hold that AI cannot possess copyright or patent rights, but it remains unclear who owns AI-generated works. Without resolution of this ambiguity, companies can determine ownership through private contracts in their terms of service. However, companies must remain silent on the rights they can actually grant, leaving end users unsure of their rights to use the outputs. This patchwork approach will likely yield confusing and conflicting IP and contract regimes and uncertainty about accountability across the value chain.

Traditional legal frameworks like copyright and privacy law have failed to keep pace with technological change, creating further uncertainty for AI developers and an inconsistent patchwork of IP ownership and rights frameworks. This inconsistent approach then trickles into companies' terms of service, contracts, and so on. Regarding copyright, the policy goals underlying the current framework do not apply to AI. New frameworks that properly capture copyright and privacy goals for AI, ideally interoperable across borders, are urgently needed.

Three potential avenues for fostering a more robust AI accountability ecosystem are clarity on antitrust regulation, nondisclosure agreements to enable assessments and audits, and whistleblower protections.

Clarity on antitrust regulation would help determine whether and how AI labs can coordinate on safety standards. Sensible coordination around consumer-friendly standards seems possible, but regulators' guidance on the issue would be welcome.

Nondisclosure agreements (NDAs) could enable assessments and audits, though this would not address all concerns about trade secrets. Confidentiality and trade secrets may impede assessments and audits depending on the audit's nature, auditor, intended work product, and audience. If the audience includes competitors, this could heighten developer concerns. Potential solutions include self-certifying audit compliance or "clean room" audits (i.e., audits without direct contact with system developers), under NDA, with only high-level findings shared publicly. This approach balances companies' need to protect proprietary information with that of public accountability.

Whistleblower protections could prove crucial for ensuring compliance with auditing and regulatory regimes once these are in place. A collaborative effort between industry and government is needed to develop acceptable processes and legal protections for companies, as well as mechanisms to report serious violations of safety principles or best practices to a competent government authority after potential whistleblowers exhaust robust internal procedures.

Recommendations:

- Clarify IP ownership frameworks for AI to establish who owns AI-generated works and enable fair licensing and liability regimes. For example, new copyright or sui generis IP laws may be needed to vest ownership in the AI system developer and/or user. Clarifying ownership would provide legal clarity for companies to develop appropriate IP protections, terms of service, and contracts to ensure developers and users understand their rights.
- Explore requiring AI system users to disclose when the content they post publicly was generated by an AI system. For example, terms of service could mandate that users label AI-generated content as such when sharing it in a way that could cause confusion or spread misinformation. However, broadly monitoring compliance with this type of requirement may be challenging.
- Supply guidance on applying antitrust laws to AI to clarify permissible industry coordination on safety. Explaining how companies can work together in the public interest without violating antitrust laws would address legal uncertainty and advance shared goals.

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

To build a robust AI accountability ecosystem, the government should fund initiatives like research into capabilities and safety evaluations, interpretability research, and increasing access to large-scale computing resources for academia and civil society. Rigorous model testing frameworks, enhanced

interpretability, and democratized access to advanced capabilities are essential for AI progress that benefits society as a whole.

AI model evaluation research. Developing rigorous, standardized evaluations is difficult and time-consuming work that requires significant resources. Increased funding, especially from government agencies, could help drive progress in this critical area, including by supporting a wide variety of academic research. Specifically, the government should fund and oversee the development of rigorous capability and safety evaluations targeted at critical risks from advanced AI like deception, autonomy, and more.

We believe that the National Institute of Standards and Technology (NIST) could be a natural home for such efforts. NIST has diligently worked on the science behind measuring AI systems and the development of associated technical standards for many years. Some highlights of NIST's work in this area include the Face Recognition Vendor Test and AI Risk Management Framework. For more information, see our proposal for Strengthening U.S. AI Innovation Through an Ambitious Investment in NIST.

Interpretability research. Provide government grants and incentives for interpretability work at universities and nonprofits and in collaboration with companies. Avoid regulations demanding interpretable models before research makes that feasible.

National AI Research Resource (NAIRR). Due to the resource-intensive nature of training LLMs, only a small number of highly-resourced private companies are able to develop them. As costs to build frontier AI systems have grown largely out of reach for academic stakeholders, public funding for research and development (R&D) has lagged, exacerbating an already unequal playing field for model development. These two interrelated factors—increased costs to build advanced AI systems and insufficient public funding for non-commercial research—have created an unequal R&D landscape that requires government intervention. Without additional investment directed towards public organizations, the future of AI development will be controlled by a handful of private actors primarily motivated by commercial interests. The NAIRR represents an opportunity to both restore a healthy balance between industry and academic contributions to AI R&D. For more information, see our Comment Regarding "Update of the National Artificial Intelligence Research and Development Strategic Plan".

Thank you for this opportunity to comment. We look forward to discussing this important topic further with you in due course.