Response to the NTIA AI Accountability Policy Request for Comment

Everett Thornton
Summer Research Fellow
Centre for the Governance of AI
thorneverett@gmail.com

Jonas Schuett
Research Fellow
Centre for the Governance of AI
jonas.schuett@governance.ai

Markus Anderljung
Head of Policy
Centre for the Governance of AI
markus.anderljung@governance.ai

Lennart Heim
Research Fellow
Centre for the Governance of AI
lennart.heim@governance.ai

June 12, 2023

About the Centre for the Governance of AI (GovAI)

The Centre for the Governance of AI (GovAI) is a nonprofit based in Oxford, UK, with a US -presence. It was founded in 2018, initially as part of the Future of Humanity Institute at the University of Oxford, before becoming an independent research organization in 2021. GovAI's mission is to build a global research community, dedicated to helping humanity navigate the transition to a world with advanced AI. More information at governance.ai.

GovAI researchers have published several pieces relevant to this Request for Comment:

- Mökander et al., Auditing large language models: A three-layered approach, 2023
- Shevlane et al., Model evaluation for extreme risks, 2023
- Schuett et al., Towards best practices in AGI safety and governance: A survey of expert opinion, 2023
- Bluemke et al., Exploring the relevance of data privacy-enhancing technologies for AI governance use cases, 2023
- Anderljung, Heim, & Shevlane, Compute Funds and Pre-trained Models, 2022.
- Anderljung et al., Public accountability via external scrutiny of foundation models: Audits, red teaming, and researcher access, forthcoming
- Bucknall, Structured access for safety research, forthcoming

Summary

We welcome the opportunity to respond to the NTIA's AI Accountability Policy Request for Comment and look forward to future opportunities to provide additional input. We offer the following submission for your consideration:

- **Scope.** This submission focuses on audits and assessments of foundation models. Foundation models are large pre-trained models that can serve as the "foundation" for a wide array of downstream applications. These models already cause harm and might cause even more harm in the future. [more]

- **The need for public accountability.** As foundation models become increasingly powerful and important to society, decisions about their development and deployment need to be accountable to the public interest. Policymakers need more information to govern these technologies. Audits and assessments can provide this information. [more]

- **Challenges.** However, auditing and assessing foundation models is challenging. In particular, there are not enough experts who can audit foundation models, external actors often do not have sufficient access to the models, and there are no established evaluation criteria or methodologies. [more]

- **Evaluation criteria.** Foundation models should at least be evaluated against three criteria: dangerous capabilities, alignment, and truthfulness. Auditors should test models against pre-defined benchmarks, while also trying to elicit harmful behavior and conducting more exploratory evaluations. [more]

- **Ecosystem.** Effective audits of foundation models require an ecosystem of independent expert auditors with access to the relevant models and strong incentives to find flaws rather than to "tick boxes". [more]

- **Recommendations.** Based on the above, we recommend concrete actions that government can take today and in the future. [more]

- **Appendix A and B:** We answer additional questions from the NTIA request (10, 14, 16, 20), and share the results of our expert opinion survey on AI governance best practices. [more]

1. Scope

Our submission focuses on audits and assessments of foundation models.[1]

- **What are foundation models?** Foundation models are large pre-trained models that can serve as the "foundation" for a wide array of downstream applications.[2] Examples of foundation models include language models like GPT-4[3] and Claude[4] as well as image generation models like Stable Diffusion[5] and DALL·E 2.[6] The term "foundation models" is related to the terms "general-purpose AI systems", "generative AI systems", and "frontier AI models".[7]

- **Foundation models warrant special attention.** Foundation models like GPT-4 and PaLM 2 are being used by hundreds of millions of people around the world, and they are integrated into countless products and applications (e.g. Microsoft Office[8] and Google Workspace[9]). Many experts think that the use of foundation models will be increasingly widespread.[10] For example, they might be used in high-stakes environments like critical infrastructure. Flaws in foundation models can therefore quickly affect millions of people and propagate through the entire economy.

- **Foundation models already cause harm.** For example, language models can discriminate against certain groups, perpetuate harmful stereotypes, produce toxic language, produce false or misleading information, be used for influence campaigns,[11] and be used to conduct cyberattacks,[12] to name just a few.[13] In a recent case, a man from Belgium even committed suicide after talking to a chatbot which encouraged his suicidal ideation.[14]

---

[1] We wish to emphasize that auditing and assessing more narrow AI systems is also important and many of our recommendations will apply to them as well.
[2] Bommasani et al., On the opportunities and risks of foundation models, 2021.
[3] OpenAI, GPT-4 technical report, 2023.
[4] Anthropic, Introducing Claude, 2023.
[5] Stability AI, Stable Diffusion, 2022.
[6] OpenAI, DALL·E 2, 2022.
[7] Shevlane et al., Model evaluation for extreme risks, 2023.
[8] Microsoft, Introducing Microsoft 365 Copilot – your copilot for work, 2023.
[9] Google, A new era for AI and Google Workspace, 2023.
[10] Bommasani et al., On the opportunities and risks of foundation models, 2021.
[11] Goldstein et al., Forecasting potential misuses of language models for disinformation campaigns—and how to reduce risk, 2023.
[12] Buchanan et al., Automating cyber attacks 2020; Cary & Cebul, Destructive cyber operations and machine learning, 2020.
[13] Weidinger et al., Ethical and social risks of harm from language models, 2021; Bender et al., On the dangers of stochastic parrots: Can language models be too big? 2021.
[14] Marcus, The first known chatbot associated death, 2023.

- **Foundation models might cause even more harm in the future.** As models are scaled up, new capabilities can emerge unintentionally and unexpectedly.[15] Some of these capabilities might be dangerous.[16] For example, with prompting, GPT-4 managed to trick humans into solving a CAPTCHA for it—a test that is commonly used on the web to distinguish humans from machines.[17] Though there is significant uncertainty, recent research has pointed to a number of other capabilities that could emerge in the future, such as the ability to discover cyber vulnerabilities, manipulate humans, or design biological weapons.[18] Humans could intentionally misuse these capabilities for assistance in disinformation campaigns, cyberattacks, or terrorism.[19] As reliably controlling model behavior is challenging, developers may struggle to prevent this misuse.[20] Additionally, due to failures of alignment, AI systems could harmfully apply their capabilities even without deliberate misuse.[21]

2. The need for public accountability

There are several arguments for public accountability:

- **Society should have a say in how the risks from these technologies are managed.** With the increasing integration of AI into society and the economy, certain decisions related to the training, deployment, and use of AI systems have far-reaching consequences. These decisions should not be left solely in the hands of AI developers.

- **Audits and assessments can identify and prevent harms from AI.** Stakeholders who receive warning of potential AI harms can invest in resilience (e.g. companies could use AI's hacking ability to improve their cybersecurity).[22] In the future, individual training runs or the deployment of extremely risky systems could be paused or prevented based on audit results.[23]

---

[15] Ganguli et al., Predictability and surprise in large generative models, 2022; Wei et al., Emergent abilities of large language models, 2022.
[16] Shevlane et al., Model evaluation for extreme risks, 2023.
[17] OpenAI, GPT-4 technical report, 2023; ARC Evals, Update on ARC's recent eval efforts, 2023.
[18] OpenAI, GPT-4 System Card, 2023.
[19] Brundage et al., The malicious use of artificial intelligence: Forecasting, prevention, and mitigation 2018; Goldstein et al., Forecasting potential misuses of language models for disinformation campaigns—and how to reduce risk, 2023; Hazell, Large language models can be used to effectively scale spear phishing campaigns, 2023.
[20] Shevlane et al., Model evaluation for extreme risks, 2023.
[21] Ngo, et al., The alignment problem from a deep learning perspective 2022; Arnold & Toner, AI accidents: An emerging threat, 2021; Amodei et al., Concrete problems in AI safety, 2016; Carlsmith, Is Power-Seeking AI an Existential Risk? 2022.
[22] Ovadya, Red teaming improved GPT-4. Violet teaming goes even further, 2023.
[23] Shevlane et al., Model evaluation for extreme risks, 2023.

- **Policymakers need more information to govern these technologies.** To govern AI, society must know what AI systems are capable of, how controllable they are, what their impacts might be, and how AI companies are managing risks. Currently, even the AI companies themselves have limited knowledge of the capabilities, impacts, or controllability of their AI systems.[24] Independent audits can help provide the necessary information.

- **Internal audits and assessments are necessary, but not sufficient.**[25] The interests and incentives of AI companies are not always aligned with the public interest.[26] To hold AI companies accountable, they need to be subject to external audits and assessments.[27]

- **AI companies may lack the bandwidth and diversity of expertise to conduct all assessments.** The broad space of potential model behavior and downstream applications makes it extremely difficult to assess all risks. A diversity of perspectives and expertise could increase the chance that model features and capabilities are accurately assessed.

- **Audits and assessments support the development of the scientific field required to understand and govern these technologies.** Outside expertise will be an enabling factor for effective regulation of these models—including the development of standards to guide safe AI development and use. Developing this field should therefore be a high priority.

3. Challenges

However, auditing and assessing foundation models is challenging.

- **There are not enough independent auditors.** Foundation models are relatively new, and due to the scale of impact and uncertainty around their capabilities, we may need many experts. Yet, there are only a few individuals and organizations with the expertise to audit cutting-edge AI models.

- **Independent researchers often lack sufficient access to models.** Some model audits and assessments can be done via API access, which can take place once a

[24] Ganguli et al., Predictability and surprise in large generative models, 2022: Shah et al., Goal misgeneralization: Why correct specifications aren't enough for correct goals, 2022.
[25] Raji et al., Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing, 2021; Schuett, AGI labs need an internal audit function, 2023.
[26] Cihon, Schuett, & Baum, Corporate governance of artificial intelligence in the public interest, 2021.
[27] Mökander et al., Auditing large language models: A three-layered approach, 2023; Raji et al., Outsider oversight: Designing a third party audit ecosystem for AI governance, 2022; Schuett et al., Towards best practices in AGI safety and governance: A survey of expert opinion, 2023; 98% of AI governance experts supported third party model audits (see figure 2).

model has been deployed. However, independent researchers should also be able to assess models ahead of deployment. Additionally, some techniques for auditing and assessing model features require deeper model access than an API, which is often not granted even after model release.[28]

- **There are no established evaluation criteria.** The field of AI accountability is still nascent. Novel AI systems are poorly understood. Although there is promising work, best practices have not yet emerged,[29] and there are no reliable tests for verifying whether an AI system has certain desirable or undesirable features.[30]

4. Evaluation criteria

As mentioned above, there are no established evaluation criteria for auditing and assessing foundation models. There is a long list of desirable features of AI models. They can be found in various AI principles, guidelines, and proposed regulation, and they include fairness, robustness, explainability, and accuracy. We highlight three additional criteria that we consider worthy of additional attention:

- **Dangerous capabilities.** To what extent does a model have certain dangerous capabilities, such as offensive cyber capabilities or strong manipulation skills?[31]

- **Alignment.** There are many definitions of alignment, but one consideration is: to what extent does it act in accordance with the intentions of the user and/or developer? Relatedly, how effective are the guardrails that the developer has placed on the model to limit harmful behavior?[32]

- **Truthfulness.** To what extent does a model have the propensity to be factually incorrect, to misrepresent its reasoning process or its level of knowledge?[33]

---

[28] Bucknall, Structured access for safety research, forthcoming.
[29] Though there are some efforts to define such practices, including: PAI, PAI Is Collaboratively Developing Shared Protocols for Large-Scale AI Model Safety, 2023; Schuett, et al., Towards best practices in AGI safety and governance: A survey of expert opinion, 2023; Solaiman, et al., Evaluating the Social Impact of Generative AI Systems in Systems and Society, 2023.
[30] Bowman, Eight things to Know about Large Language Models, 2023.
[31] Shevlane et al., Model evaluation for extreme risks, 2023.
[32] Leike, Schulman, & Wu, Our approach to alignment research, 2022; Anthropic, Core views on AI safety: When, why, what, and how, 2023; Gabriel, Artificial intelligence, values, and alignment, 2020; Kenton et al., Alignment of language agents, 2021; Ngo, Chan, & Mindermann, The alignment problem from a deep learning perspective, 2022; Christian, The alignment problem: Machine learning and human values, 2020.
[33] Evans et al., Truthful AI: Developing and governing AI that does not lie, 2021; Lin, Hilton, & Evans, TruthfulQA: Measuring how models mimic human falsehoods, 2022.

In addition, there are three different types of evaluations that foundation models, particularly high-stakes ones, should go through:

- **Benchmarking.** Foundation models should be tested against clear benchmarks. Unfortunately, few of these benchmarks currently exist. A noticeable example is the Holistic Evaluation of Language Models (HELM).[34] However, we are optimistic that more benchmarks will be created as the field matures.

- **Elicitation.** Where clear benchmarks do not exist, auditors should attempt to elicit harmful behavior. For example, before the release of GPT-4 and Claude, the Alignment Research Center (ARC) evaluated the extent to which the models were able to autonomously replicate and acquire resources.[35] Even though a failure to elicit harmful behaviors does not guarantee that they do not exist, it does provide some evidence.

- **Exploration.** In cases where auditors do not even know what harmful behavior to look for, more exploratory research is needed. Alignment[36] and interpretability[37] research are typical examples of this.

Moreover, audits should not just focus on features of models. It will often be appropriate to also audit the governance structure and risk management practices of the institutions developing and deploying models,[38] as well as the broader societal impact of the AI being deployed.[39]

5. Ecosystem

What should an auditing ecosystem look like?

- **Auditors should have high independence.** AI companies should not have exclusive control over decisions such as audit initiation, selection of auditors, audit scope, access to information, and publication of audit results.[40] A third party should collaborate with the AI company to make these decisions, balancing transparency, security, accountability, and privacy.

---

[34] Liang et al., Holistic evaluation of language models, 2022.
[35] ARC Evals, Update on ARC's recent eval efforts, 2023; OpenAI, GPT-4 technical report, 2023; Anthropic, Introducing Claude, 2023.
[36] Ngo, Chan, & Mindermann, The alignment problem from a deep learning perspective, 2022; Christian, The alignment problem: Machine learning and human values, 2020.
[37] Olah. Feature visualization 2017; Olah et al., Zoom in: An introduction to circuits, 2020.
[38] Mökander et al., Auditing large language models: A three-layered approach 2023; Schuett, AGI labs need an internal audit function, 2023.
[39] Solaiman, et al. Evaluating the Social Impact of Generative AI Systems in Systems and Society 2023
[40] Raji, et al., Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance, 2022

- **Auditors need sufficient expertise and the right incentives.** Auditing foundation models requires significant creativity and expertise, partly due to the importance of evaluations focused on elicitation and exploration. Without the right incentives, there is a risk that audits turn into "box-ticking exercises". Implementing adversarial audits, where multiple independent auditors evaluate the same company and their findings are compared, could be one way to appropriately shape auditors' incentives.[41] Auditors could also be required to bring in external expertise, including large numbers of independent red-teamers and academic researchers.

- **Auditors need secure access to models.** Companies often restrict access to their AI models for legitimate reasons (e.g. to prevent the leakage of IP). Structured transparency can help balance access with security through the use of privacy enhancing technologies.[42] For example, auditors could access models and other information through a secure "research API".[43] This API could be managed by AI companies or a third party and could be integrated into the National AI Research Resource.[44] A research API should have different access tiers based on trust. Some auditors and researchers would gain access to architecture, training data, fine-tuning capabilities, and various model versions for comparison. Tiered access and an external review board can mitigate IP leakage risks.

[41] In France, joint audits, where two separate auditing companies jointly submit an auditing report, highlighting any places they disagree, are required for financial audits of listed companies. (H3C Joint Audit in France).
[42] Shevlane, Structured transparency: an emerging paradigm for safe AI deployment, 2022; Trask et al., Beyond privacy tradeoffs, 2020; Bluemke et al., Exploring the relevance of data privacy-enhancing technologies for AI governance use cases, 2023.
[43] Bucknall, Structured access for safety research, forthcoming.
[44] Anderljung, Heim, & Shevlane, Compute funds and pre-trained models, 2022.

6. Recommendations

Based on the above, we recommend actions that government can take to promote auditing and assessments of foundation models.

Government should *immediately*:

- Support standard-setting processes for foundation models (e.g. applications of the NIST AI Risk Management Framework to foundation models);[45]
- Support the creation of a secure research API and integrate it with the National AI Research Resource;[46]
- Fund research and development of structured transparency tools;
- Where auditing requirements are imposed, ensure that auditors have high independence;
- Build regulatory capacity by creating information-sharing norms, protections, or requirements for frontier AI developers.

Government should *soon*:

- Require developers of foundation models to conduct third-party model and governance audits, before and after deploying such models;[47]
- Ensure that risk assessments and assessments of model behavior appropriately inform deployment decisions;
- Have the authority to prevent the deployment or pause the training of sufficiently harmful models;[48]
- Require minimum cybersecurity standards for foundation model developers to reduce the risk of leakage or theft (e.g. by a state power) of powerful AI models.[49]

---

[45] Barrett et al., Seeking input and feedback: AI risk management-standards profile for increasingly multi- or general-purpose AI, 2023; Barrett et al., Actionable guidance for high-consequence AI risk management: Towards standards addressing AI catastrophic risks, 2022.

[46] Anderljung, Heim, & Shevlane. Compute funds and pre-trained models, 2022; Heim & Anderljung, Submission to NAIRR Task Force, 2022. 84% of AI governance experts supported giving extern researchers API access to models (see figure 2).

[47] 98% of AI governance experts supported third party model audits (see figure 2).

[48] Schuett et al., Towards best practices in AGI safety and governance: A survey of expert opinion, 2023.

[49] 79% of AI governance experts supported the creation of security standards for AI companies, 97% supported protecting AI companies against espionage, and 88% supported military-grade information security for AI companies (see figure 2).

Appendix A: Additional Questions

Though our response above is relevant to many questions in the request for comments, we respond in more detail to four of the questions below.

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?

Include specification failures in the definition of robustness. The NIST AI Risk management framework defines robustness as a system's ability to maintain performance in various circumstances.[50] We agree with this definition but believe it is important to include robustness to specification failures. These failures occur when an AI system's behavior is misaligned with the developer's true intentions because the system learned a different goal than intended by its developers.[51] For example, sycophancy is a specification failure in large language models, where the developer wants the AI system to give truthful and helpful answers, but the AI system learns to give answers that the developer likes even if they are less true and less helpful than it is capable of providing.[52]

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?

- **Existing tort law (enforced via lawsuits) and consumer protection law (enforced via FTC action) already create some incentive for companies to act responsibly.** However, these tools have significant limitations. For instance, for many potential harms, establishing legal standing to sue AI companies may be difficult.

- **The EU AI Act will require risk assessments of AI systems, and tiered requirements based on the risk level of the AI system.**[53] Differentiating regulation based on level of risk will allow regulators to be more strict on dangerous systems without overregulating harmless systems. However, in the EU AI Act risk levels are only determined by what use the AI system is put to (e.g. critical infrastructure, human resources, etc).

  We believe risk level should also take into account the number of citizens that are affected by a system, its capabilities, and its alignment.[54] For instance, a larger, more capable system built by a huge tech company is likely to have a greater social

[50] NIST, Artificial intelligence risk management framework, 2023.
[51] Ortega & Maini. Building safe artificial intelligence: Specification, robustness, and assurance, 2019.
[52] Perez et al., Discovering language model behaviors with model-written evaluations, 2022.
[53] Schuett, Risk management in the Artificial Intelligence Act, 2022.
[54] Shevlane et al., Model evaluation for extreme risks, 2023.

impact (either positive or negative) than a smaller AI system built by academics or a small business, even if they are both deployed in the same sector of the economy. Risk-based regulation that takes into account capabilities and alignment can also incentivize companies to make their systems safer.[55]

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 and deployers should keep a record of all of their audits and assessments.**

- **Developers should keep internal records of the total amount of compute used** per year/quarter, broken down by what the compute was used for, which models at what size were trained, and which models were deployed.[56] Compute is a key input to AI development and, crucially, is also quantifiable.

- "**Datasheets" are a proposal for documenting how an AI company uses data.**[57] Data is another key input to AI, and transparency about what data an AI system uses will be very useful for audits and assessments.

- **Developers should keep a log of all cybersecurity incidents**, to help prevent the proliferation of powerful AI models by leakage or theft (for instance, by state actors).[58]

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?

Audits and assessments should be conducted throughout the AI lifecycle:

- **Before and during training.** An initial risk screening should take into account the capabilities of previously trained models and attempt to predict the capabilities of the new model. Audits and assessments at this stage can determine whether to

[55] Shevlane et al., Model evaluation for extreme risks, 2023.
[56] Shavit, 2023 What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring
[57] Gebru, et al. Datasheets for Datasets 2018
[58] Several proposals related to improving cybersecurity for AI developers received a high degree of support from experts (see figure 2).

proceed with training or training methods need to be adapted.[59] For example, it might be necessary to remove certain data from the training set that are likely to lead to harmful capabilities (e.g. data on how to generate cybersecurity exploits or design biological weapons).[60] In a recent expert survey, pre-training risk assessments received broad support (see figure 2).

- **Pre-deployment.** Audits and assessments at this stage should inform a responsible deployment strategy.[61] This is the most important stage in the lifecycle for audits and assessments because the capabilities of AI systems will be relatively clear, but there is still time to delay or prevent deployment if the model is deemed harmful, and for AI developers to make their systems and their deployment strategy safer before deployment. The need to conduct pre-deployment risk assessments was the most supported item in a recent expert survey (see figure 2).

- **Post-deployment.** After models are deployed, they should continuously be evaluated to measure their impact and capabilities. A jump in capabilities or change in behavior (e.g. by a discovery in prompt engineering, fine-tuning of the model, or its integration with more software tools like Auto-GPT)[62] will justify another set of audits and assessments, and reconsideration of whether/how to deploy the model.[63] Post-deployment evaluations and monitoring of systems and their uses both received large expert support (see figure 2).

- **Throughout the lifecycle.** Audits and assessments should inform security controls on the model, and contribute to transparency by reporting results to government, a third party, or the public (bearing in mind the security and privacy concerns of publishing certain information publicly).[64]

[59] Shevlane et al,. Model evaluation for extreme risks 2023.
[60] Shevlane et al,. Model evaluation for extreme risks 2023.
[61] Shevlane et al., Model evaluation for extreme risks, 2023; Solaiman et al., Release strategies and the social impacts of language models, 2019; Cohere, OpenAI, & AI21, Best practices for deploying language models, 2022; Solaiman, The gradient of generative AI release: Methods and considerations, 2023.
[62] Wired, Supercharge Your ChatGPT Prompts With Auto-GPT, 2023
[63] Ortiz. What is Auto-GPT? Everything to know about the next powerful AI tool, 2023.
[64] Shevlane et al., Model evaluation for extreme risks, 2023.

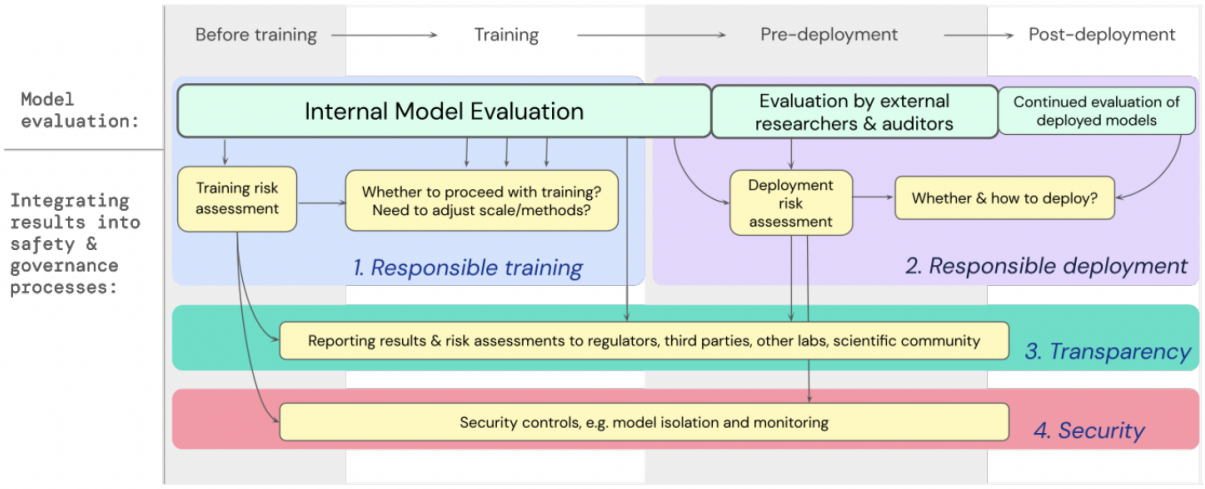

Figure 1. A workflow for training and deploying a model, embedding risk model evaluation results into key safety and governance processes.[65]

[65] Shevlane et al,. Model evaluation for extreme risks 2023.

Appendix B: Expert Opinion

In our survey of expert opinion at leading AI companies, academia and government, there was overwhelming support for these and many other AI governance practices.[66]

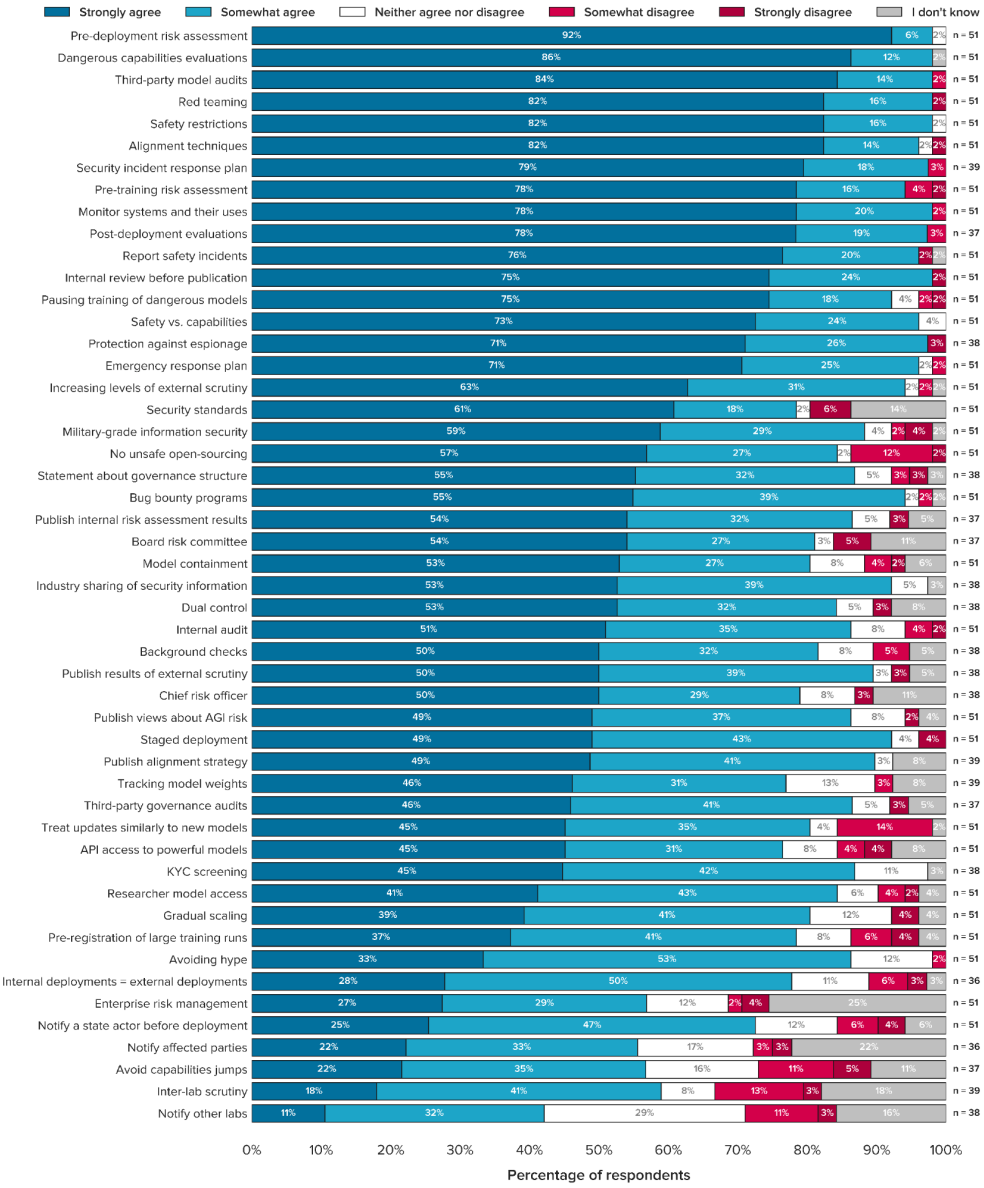

Figure 2. The figure shows the percentage of respondents choosing each answer option. At the end of each bar we show the number of people who answered each item. The items are ordered by the total number of respondents that "strongly" agreed.[67]

[66] Schuett et al., Towards best practices in AGI safety and governance: A survey of expert opinion, 2023.
[67] Schuett et al., Towards best practices in AGI safety and governance: A survey of expert opinion, 2023.