Towards Publicly Accountable Frontier LLMs

Building an External Scrutiny Ecosystem under the ASPIRE Framework

Markus Anderljung[1,2,*], Everett Thornton Smith[1], Joe O'Brien[1,3], Lisa Soder[1], Benjamin Bucknall[1], Emma Bluemke[1], Jonas Schuett[1], Robert Trager[1,4], Lacey Strahm[5], and Rumman Chowdhury[6,7]

[1]*Centre for the Governance of AI*
[2]*Center for a New American Security*
[3]*Institute for AI Policy and Strategy*
[4]*Blavatnik School of Government, University of Oxford*
[5]*OpenMined*
[6]*Humane Intelligence*
[7]*Harvard Berkman Klein Center*
[*]*Corresponding Author: markus.anderljung@governance.ai*

Abstract

With the increasing integration of frontier large language models (LLMs) into society and the economy, decisions related to their training, deployment, and use have far-reaching implications. These decisions should not be left solely in the hands of frontier LLM developers. LLM users, civil society and policymakers need trustworthy sources of information to steer such decisions for the better. Involving outside actors in the evaluation of these systems – what we term "external scrutiny" – via red-teaming, auditing, and external researcher access, offers a solution. Though there are encouraging signs of increasing external scrutiny of frontier LLMs, its success is not assured. In this paper, we survey six requirements for effective external scrutiny of frontier AI systems and organize them under the ASPIRE framework: Access, Searching attitude, Proportionality to the risks, Independence, Resources, and Expertise. We then illustrate how external scrutiny might function throughout the AI lifecycle and offer recommendations to policymakers.

1 Risks from Frontier LLMs

The most capable large language models (frontier LLMs) [1] pose significant risks of harm, both now and in the future [2]. These risks include LLMs exacerbating discrimination [3], disinformation and election interference [4], authoritarian or corporate surveillance [3], cyberattacks [5], or the proliferation of weapons of mass destruction, especially biological weapons [6]. However, these risks are not inevitable; they are a function of decisions around model development, deployment and use.

2 AI Governance as an Information Problem

Just as risk of harm warrants governance in industries as diverse as aviation, pharmaceuticals, finance, and nuclear power, so too do the risks from articicial intelligence (AI). However, in order to make informed decisions and design well-targeted policies, stakeholders outside of AI developers need reliable information about frontier LLMs [7, 8]. As such, good governance is in part an information problem [9–11]. When users, civil society and policymakers possess reliable information, they can contribute to the responsible development, deployment, and use of frontier LLMs in a variety of ways:

- **Users.** With a firmer understanding of LLMs, users can make better choices about what systems to use and pressure AI developers to act more responsibly [12–14].
- **Civil society.** More accurate information about LLMs can allow civil society to better research and advocate for policies, standards, or other methods to reduce harm [12, 15].
- **Policymakers.** By better understanding the potential risks and impacts of AI – present and future – policymakers can legislate and regulate more effectively.

While reliable information alone is far from sufficient for good governance, it is close to a necessary condition. It is hard to imagine how a society could ensure that its pharmaceuticals, airplanes, financial system, or nuclear power plants were safe and dependable without a reliable source of information about them. The same is true for AI.

3 Sourcing Reliable Information

AI developers are the primary party responsible for ensuring their systems are safe [16], and the risk assessments they perform and other information they provide are a valuable resource for policymakers [17]. However, policy decisions must not solely rest on information provided by AI developers.

External actors – actors not employed by an AI developer – must be involved in the generation and distribution of risk assessments and other information necessary for good governance. This involvement of external actors is a process we call "external scrutiny." Such scrutiny can involve auditing [18, 12, 19], evaluations [20], red-teaming [8], or other research performed on an AI system by e.g. government agencies, academics, nonprofits, or private companies. By doing so, control over the impacts of frontier LLMs can be "democratized" [14].

We believe three types of information about frontier LLMs are particularly important for good governance [18]:

- **Model capabilities.** What is the model capable of? What tasks can it perform? This can inform appropriate use-cases, as well as identify the model's potential for misuse [20].
- **Model controllability.** Does the model reliably act in accordance with user or developer intentions (e.g. does it produce unintentional toxic content [21], or behave predictably out of distribution)? This informs the likelihood that the model will cause unintended harm.
- **Model impacts.** What effects might the model have on the world (e.g. exacerbating disinformation, or displacing workers) [3]? This can inform regulator or AI company decisions to introduce additional safeguards or even to recall a deployed model.

External scrutiny can help provide more reliable information by:

- **Verifying developer claims.** Developers might misrepresent what they know about their LLMs. In the tobacco industry, companies had data that their products were harmful, but concealed evidence of harm and made fraudulent claims about product safety for approximately fifty years [22]. AI developers will face similar incentive problems when informing stakeholders about risks from their products. External scrutiny can reduce such information asymmetries between AI developers and external actors [8, 15].
- **Uncovering new information.** Developers may fail to identify issues with their systems. They may have incentives to put blinders on and remain unaware of significant issues. This problem is exacerbated by the vast space of potential model behaviors and uses, and the fact that understanding the inner workings of LLMs is notoriously difficult; even if they earnestly tried, AI developers are unlikely to uncover all important system flaws. External scrutiny can overcome these challenges by bringing a wider set of expertise, perspectives, and motivations to bear on the problem of identifying risks and issues with LLMs.

4 Calls for External Scrutiny

In recognition of these benefits, there are many efforts to expand external scrutiny of frontier LLMs. The United Kingdom has established the Frontier AI Task Force with £100 million in funding to ensure that AI risk assessments are conducted by neutral third parties [23, 24]. In the United States,

15 leading AI companies have signed onto the White House Voluntary AI Commitments, which call for external red-teaming of AI models [25]. There are also several related proposals in the US Congress [26–28]. In the European Union, the latest parliamentary proposal on the AI Act would require involvement of "independent experts" in the design and testing of foundation models [29]. AI developers have also begun voluntarily inviting external parties to evaluate their systems [30–33].

Though developing quickly, the external scrutiny ecosystem for frontier LLMs is nascent and its effectiveness will depend on specific decisions. In the next section, we outline what some of these decisions are and offer a guide for navigating them.

5   Requirements for Effective External Scrutiny

Successfully designing and implementing external scrutiny of frontier LLMs will be a difficult task, with many potential pitfalls. If designed or implemented poorly, external scrutiny might not only fail to provide crucial information to reduce risks from AI, but could also create a false sense of security.

The collapse of Enron provides a notorious example of failed external scrutiny. Enron's external auditors failed to uncover and report ongoing financial fraud due to a lack of independence from company management, incentives to not critique Enron or its managers, formulaic auditing standards, and a lack of expertise and resources among the auditor's oversight committee [34].

In light of such potential failure modes, we propose the ASPIRE framework for ensuring effective external scrutiny, illustrated in Figure 1, comprising six requirements: Access, Searching attitude, Proportionality, Independence, Resources, and Expertise. For policy recommendations building on this framework, see Appendix I.
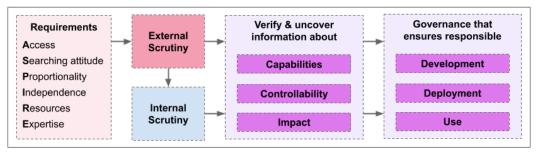

Figure 1. External Scrutiny Requirements and Goals

5.1   Access

Scrutinizers will require access to AI models and information about their development and use in order to evaluate them. Specifically, they will need access to, among other things: "base models," "model families," the components of a deployed AI system, background information on the model, third-party data on the model's impacts, and the ability to fine-tune the model. For further discussion see Appendix II.

However, providing wider access can increase the risk of unintentional and irreversible model proliferation [35]. This proliferation can limit the tools available to govern AI in a variety of ways, as well as place dangerous capabilities in the hands of malicious actors. Proliferation can occur through the model itself being leaked or stolen, or through the spread of information about it, such as its training methods, size, or capabilities, making it easier for malicious actors to create a similarly capable model [36]. Policymakers and AI developers must therefore balance providing access with ensuring information security. One such approach is "structured access", facilitating "controlled, arm's length interactions with AI systems" through cloud-based interfaces [37].

5.2   Searching Attitude

Formulaic approaches to scrutinize models (such as benchmarks) have the benefit of being easily standardized, and while many are highly informative, they are insufficient to comprehensively identify risk [38, 39]. In addition to formulaic methods, external scrutiny will require a *searching attitude*.

For many risks, like LLMs increasing bioweapons proliferation, benchmarks are still in development. Many benchmarks are rendered obsolete quickly as AI systems improve [40, 41]. Further, formulaic methods can be more easily gamed, similar to how Volkswagen designed their cars to meet emissions standards only while being tested by regulators [42]. They can also fail to accurately simulate real world conditions. As such, they might fail to elicit dangerous behavior or overlook broader societal implications [3].

Scrutinizers must therefore be actively incentivized to discover issues. In addition, they need appropriate legal protections against liability or retaliation for good faith scrutiny [43]. They must be creative and exploratory in their attempts to "break" models – to elicit some kind of undesirable behavior – and to "stretch" them – to elicit upper-bound capabilities. They must go beyond testing for known risks or capabilities; they must try to discover unknown capabilities [8]. Ultimately, external scrutinizers might be better understood as scientists rather than auditors – trying to push the frontier of understanding forward, rather than evaluating a system against best practice.

5.3 Proportionality to risk

The level of scrutiny an LLM faces should scale with the level of risk it poses. Proxies for risk include on one hand those relating to the development process, such as the model's capabilities, its novelty of design, susceptibility to accidents, and potential for misuse and on the other hand, by the LLMs deployment characteristics, including intended domain of application and the number of people likely to be affected. In aviation, for instance, alterations to an existing aircraft design face less scrutiny during certification than entirely new designs do [44].

More capable systems are more likely to possess significant dangerous capabilities, with accompanying misuse risks. Further, even when benign, higher performing systems will be used more widely and relied upon for more important decisions, increasing the stakes of accidents. As it is impossible to know the exact capabilities of a system before those capabilities are evaluated, regulators should use predictions of a model's capabilities based on previous models and be able to increase or decrease scrutiny as the model's true capabilities become apparent.

More novel systems are less well understood and therefore more likely to produce unknown risks. For instance, a system might be considered more novel if it uses more data and compute than previous systems, or a new training method. Certain safety-critical applications of AI (such as aviation, medicine, or law enforcement) will tend to be higher risk. Similarly, a model which might affect many individuals should also be considered higher risk, all else being equal.

5.4 Independence

The quality of external scrutiny is partially determined by the independence of scrutinizers. As argued by Raji et al. [12], to avoid poor incentives and guarantee sufficient independence, the AI developer must give up some control over the scrutiny process. Specifically, they must relinquish some control over decisions related to:

- **Selection and compensation.** How are scrutinizers selected to evaluate a particular model, and by whom? How are they compensated? An AI developer who controls the selection and compensation of scrutinizers can apply pressure to them and incentivize friendly treatment.

- **Scope and methods.** What kinds of questions will scrutinizers be answering? What methods will they employ to answer these questions? An AI developer that controls the scope of external scrutiny could mark important questions as off-limits.

- **Access.** What level of access is given to which scrutinizers, and who makes that decision? Since access is a key input for external scrutiny, whoever controls how access is granted holds de-facto veto power over any scrutinizing activity.

- **Post-scrutiny actions.** What is done with the results of scrutiny? Who are the results reported to and how? If an AI developer can bar a scrutinizer from informing relevant actors, then it can block information from being used effectively.

5.5 Resources

External scrutinizers need to be given the time, financial, and the computational resources necessary to carry out their task effectively.

- **Time.** Rushing scrutiny can lead to poor risk assessments. The Boeing 737 MAX crashes were caused in part by a rushed certification process attempting to keep up with competitive airplane delivery schedules [45]. GPT-4, a recent frontier LLM, received six months of pre-deployment evaluations [30]. As model risks or the complexity of evaluations increase, time for pre-deployment evaluations should increase accordingly. For reference, new pharmaceuticals or aircraft designs can receive several years of testing and evaluation before becoming commercially available [44, 46].

- **Financial resources and compute.** Scrutinizers should be compensated at rates competitive enough to attract top experts in relevant fields. Further, AI developers should make sure that scrutinzers have access to sufficient computing resources, for example by offering API credits.

5.6 Expertise

External scrutinizers must have sufficient expertise. Given LLMs' general-purpose nature and complexity [47], no single individual or team could possess all the wide-ranging expertise and perspectives needed to answer all the relevant questions [20]. For external scrutiny to be effective, it must represent and incorporate "a diversity of institutions, cultures, demographic groups, languages, and disciplines to be able to critically examine foundation models from different perspectives" [48]. Further, deep expertise will be required in a number of areas. For instance, to assess biosecurity risks, expertise countering biological weapons proliferation is likely necessary, as well an understanding of what harm different pathogens can cause. This expertise may be possessed by only a few government actors, and is therefore difficult to hire [6, 49].

6 External Scrutiny in the AI Lifecycle

Just as risks from LLMs can emerge at various stages of the AI lifecycle, external scrutiny must be applied throughout these phases. In this section, we illustrate the role external scrutiny can play across the development, pre-deployment and post-deployment stages.
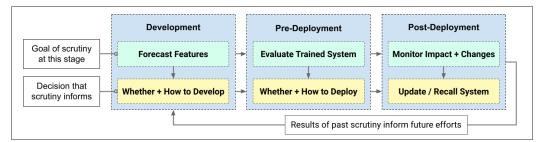

Figure 2. External Scrutiny throughout the AI Lifecycle.

6.1 Development

The results of external scrutiny should inform decisions about frontier LLM development such as whether to train the model in the first place and what information security measures might be warranted. By identifying safety issues early in the lifecycle, they may be less costly to fix [50]. Moreover, high information security standards may need to be adopted for models with significant misuse potential, as they could be stolen after development and before deployment [51].

External scrutinizers should support and assess AI developer's adherence to their public commitments. One class of such commitments are responsible scaling policies, which outline "conditions under which it would be too dangerous to continue deploying AI systems and/or scaling up AI capabilities until protective measures improve [52, 53]." External scrutinizers could predict a model's capabilities

5

and controllability by studying similar, smaller models, or by inspecting the training data [20]. During training, they should regularly evaluate the model to inform whether to continue or adjust the development process. Thereby, external scrutinizers could assess whether a model is more capable than the AI developer's safety measures can handle.

6.2 Pre-deployment

The results of external scrutiny should inform decisions on whether and how to deploy the LLM. They can inform these decisions by simulating how malicious actors could make use of the LLM, for instance by assessing the models' ability to assist in acquiring biological weapons. They can also stress-test the safeguards placed on the LLM and find areas where it could engage in harmful behavior or cause accidents. Further, they can help developers identify relevant information that should be reported to downstream users, helping them determine whether the model is suitable for a particular use-case, and preparing for ways it might cause harm [54].

6.3 Post-deployment

Once models are deployed, external scrutinizers can play an important role in assessing changes in system capabilities and controllability, as well as its downstream impacts. Historically, many model capabilities were only discovered after deployment by users, such as through prompt engineering or giving the model access to tools by e.g. combining it with other software [20]. Models which appeared harmless at the time of first deployment might prove more dangerous years later as users discover new capabilities. As such, for models deployed via API, external scrutiny should inform decisions about what updates to safeguards are needed or whether to restrict access to the model, such as by recalling it [55].

Appendix I: Policy Recommendations

Building an external scrutiny ecosystem that can contribute to public accountability in a meaningful way will be a difficult and lengthy task. However, we believe that there are concrete actions across the ASPIRE framework that policymakers can take into account now when designing new standards, regulations, or legislation.

Access

- **Mandate sufficient and secure model access** to qualified external scrutinizers and develop robust norms and criteria to provide such access, such as a scrutinizer accreditation or vetting processes. Qualification might depend on the scrutinizer's expertise, independence, and trustworthiness to not leak dangerous information.
- **Develop a research API and provide access to third-party data**. This research API could be developed by AI companies, a government body, or through a public-private partnership. It could be managed by a third party, such as the National AI Research Resource in the US.
- **Support the development of structured access tools** and privacy-enhancing technologies.

Searching attitude

- **Ensure scrutinizers are not subject to undue liability or retaliation**, e.g., by creating safe harbors. Otherwise, fear of repercussions for honest mistakes might discourage exploratory scrutiny and bias the scrutinizer towards check-list, compliance-focused approaches to avoid liability [43].
- **Promote competition among scrutinizers** for instance by engaging multiple independent teams to elicit dangerous behavior of a model, in an "adversarial audit". Ideally, compensation should be designed to reward teams that successfully expose flaws.
- **Have standards focus on identifying key objectives for evaluation,** rather than prescribing methods. For example, defining specific undesirable and desirable features of AI models, including worrisome dangerous capabilities, can help steer scrutinizer efforts. This can allow for a more flexible, future-proofed approach that adapts to evolving scrutiny practices while maintaining a targeted assessment.

Proportionality

- **Scale the intensity of mandated external scrutiny in proportion to the risks** the model poses. This could be guided by a classification scheme, analogous to biosafety levels, which necessitate varying degrees of scrutiny and regulatory oversight based on the level of risk they pose [53].

Independence

- **Develop standards of independence** for external scrutinizers that take into account how decisions on selection and compensation, scope and methods, access, and post-scrutiny actions are made.
- **Reduce the administrative burden for external scrutiny**, e.g., by promoting the development of standardized contracts to decrease red tape and ensuring companies are not able to impose unreasonable NDA requirements.
- **Establish oversight mechanisms** that can hold scrutinizers accountable, e.g., by setting and monitoring standards on scrutiny, such as rules about conflict of interest.

Resources

- **Ensure scrutiny is sufficiently resourced**, i.e., AI developers should provide enough financial and computational resources for scrutinizers, and not rush their development or deployment timelines. Current frontier LLMs should see at least six months of external scrutiny or red-teaming before deployment. As potential risks increase, scrutiny time should increase as well.

Expertise

- **Fund research programs** aimed at increasing capacity to evaluate LLMs outside AI developers. This could include training of diverse research talent, e.g., through the National AI Research Resource.
- **Provide government expertise** for scrutiny in areas where it is particularly necessary, such as in assessments relevant to national security.

Appendix II: Access Requirements

There are many different kinds of access that scrutinizers will require to evaluate frontier LLMs. For more details, see Bucknall et al [56]. Seeing as such access would risk revealing proprietary or sensitive information, it should be designed such as to minimize undesirable information or model leakage. Scrutinizers will require sufficient access to:

- **Efficient sampling.** Scrutinizers must be able to interact with the model, querying it and seeing its outputs. Scrutinizers will likely need to sample the AI model many times, so the ability to do so in an automated and systematic manner is important. Scrutinizers should have more functionality than that, as model behavior can be improved and modified by implementing different sampling algorithms. Further, scrutinizers should have access to the logits and derived probabilities of a model's output as they are crucial for calculating cross-entropy loss and perplexity – two standard measures of a model's performance.

- **Fine-tune the AI model.** Fine-tuning is a process that alters an AI model to exhibit new capabilities and tendencies. Fine-tuning is necessary to allow red-teamers to understand the full range of capabilities and behavior the model possesses.

- **The "base model".** Access to versions of the model that lack safety mitigations, like fine-tuning, is useful to understand latent capabilities within the model, and to understand what the model would be capable of if users are ever able to disable its safeguards. However, base model access can be facilitated through an API and does not require giving researchers access to "model weights".

- **Model families and previous versions.** Models often come in families that vary along one or more dimensions such as amount of data, training compute, and the type and extent of fine-tuning. Access to such model families is especially useful for studying scaling laws, which important for predicting the capabilities of future models. In addition, researchers benefit from studying older models after newer models are released, though these older models are often made unavailable.

- **Model internals.** To understand why models behave a certain way and make better predictions about their behavior, scrutinizers may need access to model internals, including e.g. model activations, attention, and embeddings. Though it should be noted that current internals-based approaches to model evaluation are in their infancy.

- **Training data**. The data used to train a model may have significant influence over its eventual behavior and failure modes. By studying the training data of a model, scrutinizers may be better able to predict if a model will exhibit dangerous capabilities. Access to the training data itself may not be necessary if scrutinizers have a way to understand the data sources, data cleaning decisions, and access to metadata about the data composition.

- **All the components of the deployed AI system.** Deployed AI systems typically combine a core model with smaller models and other software components and tools. Where such components exist, scrutinizers benefit from accessing them to understand how the AI system is likely to behave in production.

- **Third-party data on the system's impact.** To understand an LLM's impact, it will be necessary to access data from sources other than just the AI developer. For instance, to understand the impact of AI generated disinformation on social media, some kind of privacy-preserving data sharing agreement could be established with social media companies [57].

- **Model information.** Researchers will find it useful to know various kinds of meta-information about the model, including the input data, underlying training algorithm, internal testing results and amount of training compute. Developers publish such information in model or system cards [30–32, 54], though there are privacy and proliferation concerns associated with making such information publicly available.

Appendix III: Social Impacts Statement

Aim. Our paper aims to inform ongoing policy discussions on how model evaluations, auditing, red-teaming, and independent researcher access can contribute to public accountability. We hope to provide decision-makers with a framework to assess key design considerations for external scrutiny to ensure that AI policy is well-informed, capable of serving the public interest and of holding extremely powerful AI companies accountable. We particularly hope that by raising the salience of external scrutiny as a policy tool, and by articulating how it can be done effectively, audits, model evaluations, and red teaming can pave the way for broader stakeholder involvement in governing frontier LLMs.

Uncertainties. While we believe there is value to this framework as a high-level design tool, more granular, concrete decisions will need to be made to implement external scrutiny, and those granular decisions are beyond the scope of this paper. Although we would like to provide more concrete recommendations for policymakers, they crucially depend on the details of a situation and political context. Therefore, we have erred on the side of advocating for what seems clearly beneficial across many situations at the expense of being less concrete.

Limitations. It is important to acknowledge that even well-designed external scrutiny has its limitations. For instance, external scrutiny is unlikely to identify all the important risks in AI, and too much faith might be placed in the results of scrutiny, leading to a sense of false security. Further, external scrutiny will only have an impact insofar as it informs and changes important decisions, which may require e.g. regulators and the public to have more levers of influence over AI developers. As such, external scrutiny must be seen as only one tool in the AI governance toolbox. Further, we would like to underscore that our focus on the role external actors should play does not exempt AI developers from being ultimately responsible for guaranteeing their systems are safe.

References

[1] GOV.UK, "Safety and security risks of generative artificial intelligence to 2025 (annex b)," GOV.UK, Nov. 2023, available online. [Online]. Available: https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/safety-and-security-risks-of-generative-artificial-intelligence-to-2025-annex-b

[2] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, and I. Gabriel, "Taxonomy of risks posed by language models," *2022 ACM Conference on Fairness, Accountability, and Transparency*, jun 2022. [Online]. Available: https://doi.org/10.1145%2F3531146.3533088

[3] I. Solaiman, Z. Talat, W. Agnew, L. Ahmad, D. Baker, S. L. Blodgett, H. Daumé III, J. Dodge, E. Evans, S. Hooker *et al.*, "Evaluating the social impact of generative AI systems in systems and society," *arXiv preprint arXiv:2306.05949*, 2023.

[4] J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova, "Generative language models and automated influence operations: Emerging threats and potential mitigations," *arXiv preprint arXiv:2301.04246*, 2023.

[5] B. Buchanan, J. Bansemer, D. Cary, J. Lucas, and M. Musser, "Automating cyber attacks," Center for Security and Emerging Technology, Tech. Rep., November 2020. [Online]. Available: https://doi.org/10.51593/2020CA002

[6] J. B. Sandbrink, "Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools," *arXiv preprint arXiv:2306.13952*, 2023.

[7] J. Whittlestone and J. Clark, "Why and how governments should monitor ai development," accessed on November 15, 2023.

[8] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P. W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensold, C. O'Keefe, M. Koren, T. Ryffel, J. B. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askell, R. Cammarota, A. Lohn, D. Krueger, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, N. Zilberman, S. Ó. hÉigeartaigh, F. Kroeger, G. Sastry, R. Kagan, A. Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T. K. Gilbert, L. Dyer, S. Khan, Y. Bengio, and M. Anderljung, "Toward trustworthy AI development: Mechanisms for supporting verifiable claims," *arXiv preprint arxiv.org/abs/2004.07213*, Apr. 2020.

[9] C. Coglianese, R. Zeckhauser, and E. Parson, "Seeking truth for power: Informational strategy and regulatory policymaking," *Minn. L. Rev.*, vol. 89, p. 277, 2004.

[10] M. C. Stephenson, "Information acquisition and institutional design," *Harvard Law Review*, vol. 124, p. 1422, 2010.

[11] F. R. Baumgartner and B. D. Jones, *The politics of information: Problem definition and the course of public policy in America*. University of Chicago Press, 2015.

[12] I. D. Raji, P. Xu, C. Honigsberg, and D. Ho, "Outsider oversight: Designing a third party audit ecosystem for AI governance," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 557–571.

[13] Christchurch Call. Christchurch call initiative on algorithmic outcomes. Accessed on November 15, 2023. [Online]. Available: https://www.christchurchcall.com/media-and-resources/news-and-updates/christchurch-call-initiative-on-algorithmic-outcomes/

[14] E. Seger, A. Ovadya, D. Siddarth, B. Garfinkel, and A. Dafoe, "Democratising AI: Multiple meanings, goals, and methods," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 715–722.

[15] P. Cihon, J. Schuett, and S. D. Baum, "Corporate governance of artificial intelligence in the public interest," *Information*, vol. 12, no. 7, p. 275, Jul. 2021.

[16] I. D. Raji, P. Xu, C. Honigsberg, and D. Ho, "Outsider oversight: Designing a third party audit ecosystem for AI governance," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 557–571.

[17] N. Mulani and J. Whittlestone. Proposing a foundation model information sharing regime for the UK. Accessed on November 15, 2023. [Online]. Available: https://www.governance.ai/post/proposing-a-foundation-model-information-sharing-regime-for-the-uk

[18] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi, "Auditing large language models: a three-layered approach," *AI and Ethics*, pp. 1–31, 2023.

[19] E. P. Goodman and J. Trehu, "ALGORITHMIC AUDITING: CHASING AI ACCOUNTABILITY," *Santa Clara High Technology Law Journal*, vol. 39, no. 3, p. 289, 2023.

[20] T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt *et al.*, "Model evaluation for extreme risks," *arXiv preprint arXiv:2305.15324*, 2023.

[21] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 3356–3369. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.301

[22] A. B. Association *et al.*, "United states district court for the district of columbia, civil no. 95-1211 (crr); united states of america, plaintiff, v. american bar association, defendant: Report of the american bar association board of govenors," American Bar Association, 1996.

[23] Frontier AI taskforce: First progress report. Accessed on November 15, 2023. [Online]. Available: https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report/frontier-ai-taskforce-first-progress-report

[24] C. Criddle, "UK pushes for greater access to AI's inner workings to assess risks," *Financial Times*, Sep. 2023.

[25] The White House. Fact sheet: Biden-harris administration secures voluntary commitments from leading artificial intelligence companies to manage the risks posed by AI. Accessed on November 15, 2023. [Online]. Available: https://tinyurl.com/3cavwczp

[26] S. R. Blumenthal and S. J. Hawley. Bipartisan framework for U.S. AI act. USA. [Online]. Available: https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf

[27] ED Markey. Sens. markey, budd announce legislation to assess health security risks of AI. Accessed on November 15, 2023. [Online]. Available: https://www.markey.senate.gov/news/press-releases/sens-markey-budd-announce-legislation-to-assess-health-security-risks-of-ai

[28] Schumer. Schumer safe innovation framework. Accessed on November 15, 2023. [Online]. Available: https://www.democrats.senate.gov/imo/media/doc/schumer_ai_framework.pdf

[29] Article 28b: Proposal for a regulation recital 9 b. Accessed on November 15, 2023. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

[30] Open AI. GPT-4 system card. Accessed on November 15, 2023. [Online]. Available: https://cdn.openai.com/papers/gpt-4-system-card.pdf

[31] Meta. Meta and microsoft introduce the next generation of llama. Accessed on November 15, 2023. [Online]. Available: https://about.fb.com/news/2023/07/llama-2/

[32] Model card and evaluations for claude models. Accessed on 26 July 2023. [Online]. Available: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

[33] Open AI. Announcing OpenAI's bug bounty program. Accessed on November 15, 2023. [Online]. Available: https://openai.com/blog/bug-bounty-program

[34] P. M. Healy and K. G. Palepu, "The fall of enron," *Journal of economic perspectives*, vol. 17, no. 2, pp. 3–26, 2003.

[35] Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Markus Anderljung, Ben Bucknall, Alan Chan, Eoghan Sta!ord, Leonie Koessler, Aviv Ovadya, Ben Garfinkel, Emma Bluemke, Michael Aird, Patrick Levermore, Julian Hazell, Abhishek Gupta, "Open-Sourcing highly capable foundation models," Centre for the Governance of AI, Tech. Rep., Sep. 2023. [Online]. Available: https://cdn.governance.ai/Open_Sourcing_Highly_Capable_Foundation_Models_GovAI_2023.pdf

[36] M. Anderljung, J. Barnhart, J. Leung, A. Korinek, C. O'Keefe, J. Whittlestone, S. Avin, M. Brundage, J. Bullock, D. Cass-Beggs *et al.*, "Frontier AI regulation: Managing emerging risks to public safety," *arXiv preprint arXiv:2307.03718*, 2023.

[37] T. Shevlane, "Structured Access: An Emerging Paradigm for Safe AI Deployment," in *The Oxford Handbook of AI Governance*. Oxford University Press, 2022.

[38] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar *et al.*, "Holistic evaluation of language models," *arXiv preprint arXiv:2211.09110*, 2022.

[39] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[40] N. Maslej, L. Fattorini, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, H. Ngo, J. C. Niebles, V. Parli, Y. Shoham, R. Wald, J. Clark, and R. Perrault, "The AI index 2023 annual report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, Tech. Rep., April 2023.

[41] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia *et al.*, "Dynabench: Rethinking benchmarking in NLP," *arXiv preprint arXiv:2104.14337*, 2021.

[42] J. C. Jung and E. Sharon, "The Volkswagen emissions scandal and its aftermath," *Global business and organizational excellence*, vol. 38, no. 4, pp. 6–15, 2019.

[43] R. Chowdhury. Artificial intelligence: Advancing innovation towards the national interest. Accessed on November 15, 2023. [Online]. Available: https://republicans-science.house.gov/_cache/files/6/8/68b1083c-d768-4982-a8f9-74b0e771a2bc/E551A6FE9CEB156D4DE626417352ED0E.2023-06-22-dr.-chowdhury-testimony.pdf

[44] F. A. Administration. Airworthiness certification. Accessed on November 15, 2023. [Online]. Available: https://www.faa.gov/aircraft/air_cert/airworthiness_certification#:~:text=By%20comparison%2C%20the%20certification%20of,control%20system%2C%20including%20the%20MCAS%20

[45] P. DeFazio and R. Larsen, "Final committee report—the design, development & certification of the Boeing 737 MAX," *The US House Committee on Transportation and Infrastructure, Subcommittee on Aviation. Washington, DC: US House Committee on Transportation and Infrastructure*, 2020.

[46] A. Sertkaya, A. Birkenbach, A. Berlind, and J. Eyraud, "Examination of clinical trial costs and barriers for drug development," accessed on November 15, 2023. [Online]. Available: https://aspe.hhs.gov/reports/examination-clinical-trial-costs-barriers-drug-development-0

[47] S. R. Bowman, "Eight things to know about large language models," *arXiv preprint arXiv:2304.00612*, 2023.

[48] P. Liang, R. Bommasani, K. Creel, and R. Reich. The time is now to develop community norms and release foundation models. Accessed on November 15, 2023. [Online]. Available: https://hai.stanford.edu/news/time-now-develop-community-norms-release-foundation-models

[49] Anthropic, "Frontier threats red teaming for AI safety," https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety, Jul. 2023, accessed: 2023-9-30.

[50] T. Korbak, K. Shi, A. Chen, R. V. Bhalerao, C. Buckley, J. Phang, S. R. Bowman, and E. Perez, "Pretraining language models with human preferences," in *International Conference on Machine Learning*. PMLR, 2023, pp. 17 506–17 533.

[51] J. Bansemer and A. Lohn. Securing AI makes for safer AI. Accessed on November 15, 2023. [Online]. Available: https://cset.georgetown.edu/article/securing-ai-makes-for-safer-ai/

[52] Alignment Research Center, "Responsible scaling policies (RSPs)," https://evals.alignment.org/blog/2023-09-26-rsp/, Sep. 2023, accessed: 2023-10-3.

[53] Anthropic, "Anthropic's responsible scaling policy," https://www.anthropic.com/index/anthropics-responsible-scaling-policy, Sep. 2023, accessed: 2023-10-3.

[54] R. Burnell, W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein, M. Mitchell, and et al., "Rethink reporting of evaluation results in ai," *Science*, vol. 380, no. 6641, p. 136–138, 2023.

[55] O'Brien, Ee, and Williams, "Deployment corrections: An incident response framework for frontier AI models," Institute for AI Policy and Strategy, Tech. Rep., November 2023. [Online]. Available: https://www.iaps.ai/research/deployment-corrections

[56] B. Bucknall and R. Trager, "Structured access for third-party research on frontier AI models," Oxford Martin School, Tech. Rep., 2023. [Online]. Available: https://cdn.governance.ai/Structured_Access_for_Third-Party_Research.pdf

[57] A. Trask, A. Sukumar, A. Kalliokoski, B. Farkas, C. Ezenwaka, C. Popa, C. Mitchell, D. Hrebenach, G.-C. Muraru, I. Junior, I. Bejan, I. Mishra, I. Ngong, J. Bandy, J. Stahl, J. Cardonnet, K. Trask, K. Nguyen, K. Dang, K. van der Veen, K. Eng, L. Strahm, L. Ayre, M. Jay, O. Lytvyn, O. Kyemenu-Sarsah, P. Chung, P. Smith, R. S, R. Falcon, S. Gupta, S. Gabriel, T. Milea, T. Thoraldson, T. Porto, T. Cebere, Y. Gorana, and Z. Reza. How to audit an AI model owned by someone else (part 1). Accessed on November 15, 2023. [Online]. Available: https://blog.openmined.org/ai-audit-part-1/