Position Paper: Model Access should be a Key Concern in AI Governance

Edward Kembery
ERA Fellow
edwardkembery@outlook.com

Ben Bucknall
Centre for the Governance of AI

Morgan Simpson
ERA

Abstract

The downstream use cases, benefits, and risks of AI systems depend significantly on the access afforded to the system, and to whom. However, the downstream implications of different access styles are not well understood, making it difficult for decision-makers to govern model access responsibly. Consequently, we spotlight Model Access Governance, an emerging field focused on helping organisations and governments make responsible, evidence-based access decisions. We outline the motivation for developing this field by highlighting the risks of misgoverning model access, the limitations of existing research on the topic, and the opportunity for impact. We then make four sets of recommendations, aimed at helping AI evaluation organisations, frontier AI companies, governments and international bodies build consensus around empirically-driven access governance.

Introduction

The rapid development of general-purpose AI has led to discussions and significant legislative efforts focused on governing the technology to minimise societal harm while promoting objectives such as economic growth and respect for diverse cultural norms [1, 2, 3, 4]. A key question has concerned which aspects of models to make available and to whom [5, 6, 7, 8]. Access provisions can vary: some model developers might offer the public a chat interface, while others may share model weights or training data. Access can also be distributed differently between different stakeholders, including internal company members, regulators, AI safety organisations, third-party researchers, and the public. These decisions form the basis of governing model access.

Poor model access governance could have serious consequences. Insufficiently cautious information release could increase risks [7, 9, 8], while overly restrictive policies could centralise power or stifle critical research [10, 11, 12]. Experts have argued that both risks may become more likely over time, and could be difficult to reverse [7, 11]. Failing to establish good governance principles early could lead to new risks, such as those from unsafeguarded dangerously powerful publicly available models, which would be more challenging to mitigate [9].

Despite this urgency, decision-makers in frontier AI companies and governments are ill-equipped to govern model access due to limited research, lack of empirical data, and inadequate conceptual clarity. Significant legislation acknowledges the need for risk-proportional model release [13], but no clear guidance exists on how deployers should manage different access styles [2, 3]. There also remains substantial uncertainty regarding the use cases afforded by different access styles, and the real-world impacts they could have on various stakeholders [14]. Resolving these uncertainties will require the work of a dedicated field aimed at gathering empirical data on these outstanding questions.

We consequently spotlight "Model Access Governance", an emerging field focused on answering the questions needed for organisations and governments to make responsible, evidence-based access decisions. As a domain of AI Governance research [15], this field would develop policies, norms, laws, and institutions to determine who accesses which features of AI models and under what conditions.

Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2024.

We argue that the risks of mismanagement are significant, current understanding of different model access styles and their implications is limited, and improving this understanding through further research could improve decision-making. We consequently make four recommendations to improve understanding around key issues in Model Access Governance, namely that: i) AI organisations extend evaluations across different access styles, ii) frontier AI companies commit to following responsible best practices, (iii) governments work to coordinate research on downstream implications, and iv) international organisations work to build consensus around empirically-driven access governance. Two appendix sections then set out six priority areas for future research and review existing work related to Model Access Governance.

1 Defining Model Access Governance

Model Access: A term describing the manner in which individuals, organisations, and systems can interact with or inspect different aspects of an AI model. This involves three components (see Fig. 1).

Model Aspects: AI *models* are comprised of components including code and weights [16]. They serve as the foundation of AI *systems*, which include infrastructure such as GPUs and user interfaces [16, 3]. Developers must decide which model components and broader information—such as training data or outputs—to share with different groups. We refer to these components as *model aspects*.

Access Styles: When a developer provides access to a model aspect with specific permissions or usage restrictions, we term this an access style. Bucknall and Trager (2023) outline key access styles, including *sampling, inspecting, fine-tuning,* and *modifying* [17]. However, not all model aspects support all access styles, and different or future models may enable new ones. Thus, existing taxonomies should be viewed as provisional.

Access Groups: An AI developer might decide to make particular aspects of a model or particular access styles available to particular groups with distinct needs, permissions, goals, or capabilities, referred to here as *access groups*. Access groups might include intracompany groups, governments and AISIs, trusted third-party auditors and industry partnerships, research communities, and the general public.

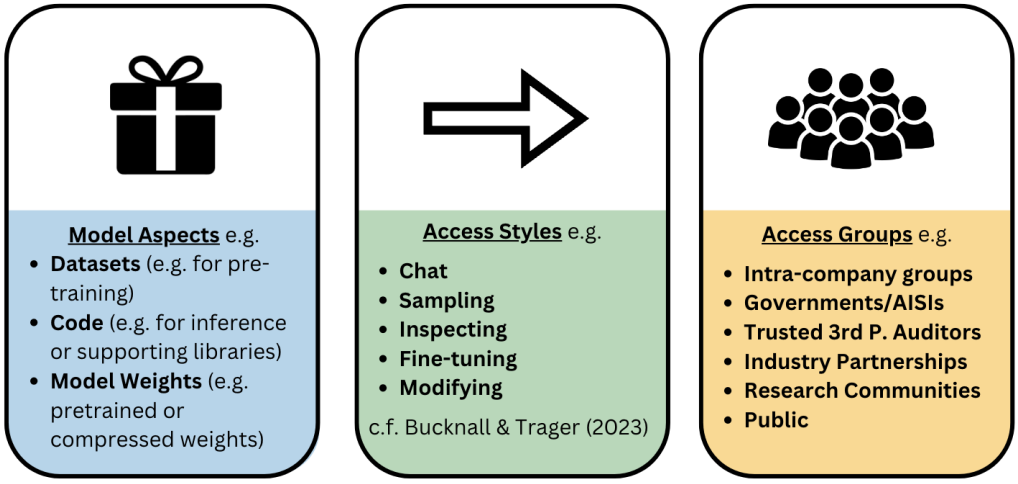

Figure 1: The Three Elements of Model Access

Model Access Governance: Governing model access refers to the process of making decisions as to which *access groups* should have what *styles of access* to which *model aspects* under what circumstances. Model Access Governance refers to an emerging research field looking to systematise and build on existing research to build evidence and develop concepts, policies, norms, laws, and institutions that help actors make these decisions better achieve diverse societal objectives.

For further information about open problems in Model Access Governance and a broader review of related work, see the Appendix.

2 The Case for Developing Model Access Governance

Our case for Model Access Governance is based on three propositions. i) Misgoverning model access could have serious consequences; ii) there is a lack of expert consensus about how to govern model access; and iii) insights and evidence provided by further targeted research can help decision-makers govern model access more effectively. This section expands on each proposition in turn.

2.1 Misgoverning model access could have serious consequences

Whilst governing model access incautiously could lead to societal harms, to do so too cautiously could create significant opportunity costs. In the first case, it is already well established that powerful AI systems might have the potential to cause significant societal harms, whether via misuse or accident [4, 18]. However, miscalibrated access governance might make these risks significantly worse, for three main reasons:

- **Incautious access policies could make models easier to jailbreak and augment**. Most leading contemporary models are straightforward to jailbreak using text-based commands [19]. However, future models might be significantly more robust to attacks afforded by some styles of access than others [20, 21]. Releasing more information about these more powerful models–for instance, internal chains of thought [20]–might make them more vulnerable to jailbreaks and enhancements, increasing the risk of more frequent and serious harm [22]. Furthermore, if future models are not robust, it might be possible to use unsecured legacy models could be used to jailbreak them [23]. In effect, this would allow malicious actors to 'bootstrap' their capabilities up to the frontier.
- **Miscalibrated access could lead to unsecured models spreading globally, with no way to retract them**. Once models are made public and downloadable, they can be freely shared worldwide, expanding the geographic risk. A poor access decision made *anywhere* creates risks *everywhere*. It also extends the time frame of the risk: a single miscalibrated decision could pose threats *indefinitely* [7, 8].
- **Miscalibrated access could reduce decision-makers' insight into evolving risks**. When models are made publicly available and downloadable, they can be shared widely and without oversight, diminishing decision-makers' understanding of how the models are used and by whom [7, 8]. This lack of insight may impair their ability to gauge the capabilities of malicious actors or the likelihood of accidents, making them less prepared to address these issues. Paradoxically, the initial "openness" might reduce overall transparency about model use, thereby increasing risks.

On the other hand, AI systems present the potential to deliver substantial benefits. Miscalibrated model access governance could fail to realise these, incurring opportunity costs in three ways:

- **Miscalibrated access might lead to underutilisation of valuable use cases**. Experts are still uncertain about the marginal utility of different access styles for various use cases and their economic value [24, 10]. However, restricting access styles for certain groups could significantly limit their ability to develop novel AI applications without substantially increasing risk. Such unnecessary limitations could result in a significant opportunity cost.
- **Miscalibrated access norms could lead to a more unequal distribution of decision-making power over advanced AI**. Sharing specific information about AI systems with certain groups might enable them to gain more benefits than others. For instance, restricting styles of access that allow groups to fine-tune models to improve their performance on low-resource languages [25] might disproportionately affect users who rely on a model's low-resource language capabilities [26]. Although the impact of access styles on AI value creation and distribution is uncertain, overly centralising decision-making and benefits could represent a significant opportunity cost and potentially lead to risks like the abuse of regulatory power.
- **Miscalibrated access could delay the development of safety research**. Research shows that safety researchers need significant access to AI systems to conduct effective studies [17, 27]. If current models are low-risk but future ones may pose greater risks, the present could be a critical period for safety research. Missing this opportunity might lead to substantial opportunity costs, resulting in increased risks from more advanced AI systems in the future.

2.2 Experts are currently uncertain about how to govern model access and unlikely to get it right by default

Experts are currently uncertain about how to govern model access [14, 28]. In particular, we note that the existing work is crucially limited by three features: lack of data, lack of adequate concepts, and limited scope.

- **Experts lack sufficient data to support empirical decision-making**. Governing model access involves understanding the net marginal benefits for different parties under certain technological assumptions and the potential downstream impacts.[1] Despite a few notable exceptions [27, 17], empirical evidence as to potential risks and benefits is generally lacking. As a result, foundational work in model access governance often relies on theoretical risks and benefits rather than concrete data. To move beyond speculation and make empirically driven decisions, experts need better, more specific evidence on the risks and benefits of various access styles [14].
- **Experts lack adequate concepts**. AI models are complex, and existing concepts often equip decision-makers with the wrong intuitions. For instance, models are referred to as 'open-source' [28, 10], even though the Open Source Initiative has yet to propose an official definition of the term [29], which may have little to do with open source software [14]. Confusingly, the term has been used to refer to many different types of access regime [6]. Papers aiming to replace 'open vs closed' with a 'spectrum of access' [30, 14] have been helpful, but risk implying that access to different components of different AI models supports downstream use cases in a way that scales linearly or generalises across domains. Nor are access styles necessarily distinct, since users may be able to extract information relating to one component of a model by accessing another [31]. Inadequate concepts serve no one: decision-makers will require appropriate concepts to navigate decisions responsibly [14, 16].
- **Existing research is limited in scope**. Significant work has begun to scope out fundamental issues in Model Access Governance such as potential risks and benefits of different access styles [7, 6, 8] (see Appendix 2 for an extended literature review). However, this research has tended to focus on public access, neglecting discussions about other stakeholders like internal actors, third party groups, AISIs, and governments. It has emphasized the importance of features like model weights, overlooking the potential for new combinations of access styles. Other important questions, like how trade-offs between risks and benefits should be navigated under uncertainty, have also been overlooked.

We think that model access governance is unlikely to go well by default for three reasons. First, companies developing may lack the incentives to govern model access responsibly, especially if responsible governance contradicts fiduciary responsibilities [32]. Second, AI regulation has proved controversial and difficult to pass [33], and lack of clarity could lead to further stagnation, or the passing of unhelpful legislation that is then difficult to retract. Finally, there is no international consensus on how to govern model access, meaning that different countries and governments are likely to take different approaches [6], undermining one another's jurisdictional integrity.

2.3 Further research will help decision-makers govern model access more effectively

Further research does not always make problems easier to solve. We believe that building Model Access Governance would support better model access governance in practice for three reasons:

- **Uncertainty is high enough for evaluations to have an impact**. When experts are uncertain, they might default to ideological positions or legacy concepts. Aspiring towards empirical evidence (for example, as to the marginal uplift from different access styles for particular tasks) might not only support accurate scenario analysis and decision-making, it would help researchers recognize shared goals.
- **The necessary institutions are already in place**. Most crucial empirical inputs for governing model access can be done by industry organisations, AISIs, governments and international bodies, many of whom are already taking modest action in this direction (see

[1] "Technological assumptions" here might include questions such as whether the model is frontier-level, or whether more powerful models are already accessible in the suggested way.

Appendix 2). These organisations are well-positioned to contribute to model access governance: for instance, by extending existing risks assessment processes across various access styles (see Recommendation 1).

- **AI research has a good track record of affecting decision-making**. AI research has repeatedly seen academic ideas like compute governance inform national legislation [34, 2]. Structured access for AI systems already a good example of an academic idea [5] that has shaped research directions and policy discourse [30, 7]. A strong research field would be well-placed to influence practical decision-making.

3 Recommendations

While governing model access effectively is challenging, it is also feasible. Research on these critical open problems could significantly improve the ability of decision-makers to make informed, responsible decisions. To support this, we make four sets of recommendations, aimed at AI evaluation organisations, frontier AI companies, governments and international bodies respectively. We divide each section into short-term, medium-term and long-term recommendations, to differentiate urgent proposals from those which might require longer timelines to achieve.

3.1 AI evaluations should be extended to different styles of access

Despite their shortcomings [35, 36], AI systems evaluations are one of the best ways to determine empirical evidence as to the capabilities of models under different scenarios [19]. Building this capacity in industry and third-party organisations like AI Safety Institutes will provide decision-makers with the empirical evidence necessary to make responsible, confident decisions.

- **Short-term (6 months):** Evaluate how different access styles impact model capabilities on harmful benchmarks, robustness to jailbreaking, and the effectiveness of post-training enhancements. Organisations should flag discrepancies and, if needed, report concerns to appropriate decision-makers.
- **Medium-term (12 months):** Assess how model properties change with user interaction. For example, compare expert prompt engineers to non-experts under various conditions. Evaluate models across diverse user cases, from low-skill actors to sophisticated ones, and under different resource levels to estimate potential harms or benefits.
- **Long-term (18+ months):** Conduct studies on long-term trends in post-training enhancements and scaffolding that might affect different styles of access. This can strengthen confidence in irreversible access decisions, such as weights release. Consider developing model interfaces which support a broader range of use cases in an reversible style, such as fine-tuning APIs.

3.2 Frontier AI companies should commit to following best practises around responsible access governance

It is crucial that empirical evidence translates to responsible decisions. Consequently, AI deployers will play a key role in ensuring model access is governed responsibly. Many industry actors already have substantial deployment policies like Responsible Scaling Policies [13], and might extend these to include 'Responsible Access Policies' [37] which stipulate how access decisions should be managed.

- **Short-term (6 months):** Frontier AI companies should expand their safety evaluations to include different styles of model access, describing any limitations and mitigation plans [37]. They should make pre-commitments to provide certain access styles only if deemed low-risk by an appropriate decision channel. They should deepen their collaboration with governments and evaluation organizations to better understand the downstream impacts of model access.
- **Medium-term (12 months):** Frontier AI companies should publish comprehensive policies around access, clearly specifying criteria for roll-forward, roll-back, and non-release decisions. Decision-making processes should be transparent to promote best practices and reduce uncertainty for groups relying on different access styles.

- **Long-term (18 months +):** Frontier AI companies should investigate, support, and consider funding research that enhances empirical understanding of model access regimes on an ongoing basis, such as AI incident reporting mechanisms [14].

3.3 Governments should build the capacity to support effective access governance

Industry actors may not have the information or the incentives to govern model access responsibly. Consequently, national governments will play a crucial role in supporting model access governance. They should support AI model evaluations, coordinate research around AI impacts, and consider regulation.

- **Short-term (6 months):** Governments should support national AI safety organisations and internal evaluation structures to evaluate risks from different styles of model access. They should consult with frontier AI companies to build an informed consensus around regulation. They should collaborate to build an international consensus (Recommendation 4).
- **Medium-term (12 months):** Governments should investigate the impact of different access styles on the AI ecosystem. They might coordinate further third party research around downstream effects like economic uplifts, and cyber-security implications.
- **Long-term (18 months +):** Governments should consider regulation to support responsible model access governance: for instance, by including a requirement that AI companies follow responsible access governance best practices in national legislation. They should consider how to best assist those affected by access governance, including victims of access retraction, or disadvantaged by access trade-offs.

3.4 International bodies should strive for a global consensus on governing model access

If access is misgoverned *anywhere*, it could create risks *everywhere*. Accordingly, international organisations like the UN and OECD may play a crucial role ensuring that jurisdictional integrity is preserved and access is managed suitably [38]. We expect that this may involve longer timeframes.

- **Short-term (12 months):** International bodies should platform discussions aiming to build consensus around empirically supported governance among member states. Questions around model access governance may be raised at conventions like AI Safety summits and through channels like the International Scientific Report on the Safety of Advanced AI.
- **Medium-term (24 months):** International bodies should platform discussions encouraging member states to adopt stipulations for requiring companies within national jurisdictions to comply with access governance best practices. These discussions should aim to identify potential empirical and value-based disagreements.
- **Long-term (36 months +):** International processes should platform discussions supporting commitments around responsible model access governance. These discussions may cover how member states can best distribute the task of evaluating evolving models, or implement verifiable agreements that could be adopted by many nations in future [39, 40].

We recognize that the future of model access governance is uncertain, and emphasize international bodies should be agile and seek to adapt their policy to reflect the changing technological landscape.

Acknowledgements

This research was supported by the ERA Fellowship. The authors would like to thank the ERA Fellowship for its financial and intellectual support, as well as Marius Hobbhahn, Guarav Sett, Lisa Soder, Jack Wadham and Tobin South for their comments and insight.

References

[1] Department for Science, Innovation and Technology, Foreign, Commonwealth & Development Office, and Prime Minister's Office, 10 Downing Street. *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*. URL: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchleydeclaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023. 2023.

[2] The White House. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. Oct. 2023.

[3] Council of European Union. *Regulation 2024/1689 of the European Parliament and of the Council*. http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1416170084502&uri=CELEX:32014R0269. June 2024.

[4] Markus Anderljung et al. *Frontier AI Regulation: Managing Emerging Risks to Public Safety*. 2023. arXiv: 2307.03718 [cs.CY]. URL: https://arxiv.org/abs/2307.03718.

[5] Toby Shevlane. "Structured access: an emerging paradigm for safe AI deployment". In: *arXiv preprint arXiv:2201.05159* (2022).

[6] Francisco Eiras et al. "Near to Mid-term Risks and Opportunities of Open Source Generative AI". In: *arXiv preprint arXiv:2404.17047* (2024).

[7] Elizabeth Seger et al. *Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives*. 2023. arXiv: 2311.09227 [cs.CY]. URL: https://arxiv.org/abs/2311.09227.

[8] Sayash Kapoor et al. *On the Societal Impact of Open Foundation Models*. 2024. arXiv: 2403.07918 [cs.CY]. URL: https://arxiv.org/abs/2403.07918.

[9] David Evan Harris. *How to Regulate Unsecured "Open-Source" AI: No Exemptions*. Tech. rep. Tech Policy Press, December 4 2023.

[10] Mark Zuckerberg. *Open Source AI Is the Path Forward | Meta*. https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/. [Accessed 27-07-2024]. 2024.

[11] Mozilla. *Joint Statement on AI Safety and Openness — open.mozilla.org*. https://open.mozilla.org/letter/. [Accessed 27-07-2024]. October 31, 2023.

[12] Shayne Longpre et al. *A Safe Harbor for AI Evaluation and Red Teaming*. 2024. arXiv: 2403.04893 [cs.AI]. URL: https://arxiv.org/abs/2403.04893.

[13] METR. *Responsible Scaling Policies (https://metr.org/blog/2023-09-26-rsp/)*. Tech. rep. METR, 26 September 2023.

[14] Jon Bateman et al. *Beyond Open vs. Closed: Emerging Consensus and Key Questions for Foundation AI Model Governance*. https://carnegieendowment.org/research/2024/07/beyond-open-vs-closed-emerging-consensus-and-key-questions-for-foundation-ai-model-governance?lang=en. [Accessed 01-08-2024]. July 2024.

[15] Allan Dafoe. "AI governance: a research agenda". In: *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK* 1442 (2018), p. 1443.

[16] Adrien Basdevant et al. *Towards a Framework for Openness in Foundation Models: Proceedings from the Columbia Convening on Openness in Artificial Intelligence*. 2024. arXiv: 2405.15802 [cs.SE]. URL: https://arxiv.org/abs/2405.15802.

[17] Benjamin S. Bucknall and Robert F. Trager. *Structured access for third-party research on frontier AI models: Investigating researchers' model access requirements*. Tech. rep. Oxford Martin School of Governance, Oct. 2023.

[18] Laura Weidinger et al. "Ethical and social risks of harm from language models". In: *arXiv preprint arXiv:2112.04359* (2021).

[19] Technical Staff. *Advanced AI evaluations at AISI: May update | AISI Work — aisi.gov.uk*. https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update. [Accessed 19-08-2024]. 20 May 2024.

[20] OpenAI. *OpenAI o1 system card*. https://openai.com/index/openai-o1-system-card/. [Accessed 14-09-2024]. 2024.

[21] Gray Swan. *About Cygnet — grayswan.ai*. https://docs.grayswan.ai/introduction. [Accessed 14-09-2024]. 2024.

[22] Xiangyu Qi et al. *Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!* 2023. arXiv: 2310.03693 [cs.CL]. URL: https://arxiv.org/abs/2310.03693.

[23] Richard Fang et al. *Teams of LLM Agents can Exploit Zero-Day Vulnerabilities*. 2024. arXiv: 2406.01637 [cs.MA]. URL: https://arxiv.org/abs/2406.01637.

[24] Daron Acemoglu. *The Simple Macroeconomics of AI*. Tech. rep. National Bureau of Economic Research, 2024.

[25] Kunat Pipatanakul et al. *Typhoon: Thai Large Language Models*. 2023. arXiv: 2312.13951 [cs.CL]. URL: https://arxiv.org/abs/2312.13951.

[26] Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. "Systematic Inequalities in Language Technology Performance across the World's Languages". In: *arXiv preprint arXiv:2110.06733* (2021).

[27] Stephen Casper et al. *Black-Box Access is Insufficient for Rigorous AI Audits*. 2024. DOI: 10.1145/3630106.3659037. arXiv: 2401.14446 [cs.CY]. URL: https://arxiv.org/abs/2401.14446.

[28] James Ball and Carl Miller. *Open Sourcing the AI Revolution: Framing the debate on open source, artificial intelligence and regulation — demos.co.uk*. https://demos.co.uk/research/open-sourcing-the-ai-revolution-framing-the-debate-on-open-source-artificial-intelligence-and-regulation/. [Accessed 12-09-2024]. 30 October 2023.

[29] Open Source Initiative. *The Open Source AI Definition – draft v. 0.0.3*. [Accessed 01-08-2024]. Oct. 2023.

[30] Irene Solaiman. *The Gradient of Generative AI Release: Methods and Considerations*. 2023. arXiv: 2302.04844 [cs.CY]. URL: https://arxiv.org/abs/2302.04844.

[31] Nicholas Carlini et al. "Stealing part of a production language model". In: *arXiv preprint arXiv:2403.06634* (2024).

[32] Helen Toner and Tasha McCauley. *AI firms mustn't govern themselves, say ex-members of OpenAI's board — economist.com*. https://www.economist.com/by-invitation/2024/05/26/ai-firms-mustnt-govern-themselves-say-ex-members-of-openais-board. [Accessed 14-09-2024]. 26 May 2024.

[33] Scott Kohler. *All eyes on Sacramento: SB 1047 and the AI Safety Debate*. https://carnegieendowment.org/posts/2024/09/california-sb1047-ai-safety-regulation?lang=en. [Accessed 14-09-2024]. 11 September 2024.

[34] Girish Sastry et al. "Computing Power and the Governance of Artificial Intelligence". In: *arXiv preprint arXiv:2402.08797* (2024).

[35] John Burden. *Evaluating AI Evaluation: Perils and Prospects*. 2024. arXiv: 2407.09221 [cs.AI]. URL: https://arxiv.org/abs/2407.09221.

[36] Gabriel Mukobi. *Reasons to Doubt the Impact of AI Risk Evaluations*. 2024. arXiv: 2408.02565 [cs.CY]. URL: https://arxiv.org/abs/2408.02565.

[37] Edward Kembery and Tom Reed. *AI Safety Frameworks Should Include Procedures for Model Access Decisions*. 2024. arXiv: 2411.10547 [cs.CY]. URL: https://arxiv.org/abs/2411.10547.

[38] Emma Klein and Stewart Patrick. *Envisioning a Global Regime Complex to Govern Artificial Intelligence — carnegieendowment.org*. https://carnegieendowment.org/research/2024/03/envisioning-a-global-regime-complex-to-govern-artificial-intelligence?lang=en. [Accessed 14-09-2024]. 21 March 2024.

[39] Seokki Cha. "Towards an international regulatory framework for AI safety: lessons from the IAEA's nuclear safety regulations". In: *Humanities and Social Sciences Communications* 11.1 (2024), pp. 1–13.

[40] Akash R Wasil et al. "Verification methods for international AI agreements". In: *arXiv preprint arXiv:2408.16074* (2024).

[41] Andrew Trask et al. *Beyond Privacy Trade-offs with Structured Transparency*. 2024. arXiv: 2012.08347 [cs.CR]. URL: https://arxiv.org/abs/2012.08347.

[42] Shayne Longpre et al. "A safe harbor for ai evaluation and red teaming". In: *arXiv preprint arXiv:2403.04893* (2024).

[43] Jakob Mökander et al. "Auditing large language models: a three-layered approach". In: *AI and Ethics* (2023), pp. 1–31.

[44] Jakob Mökander et al. "Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation". In: *Minds and Machines* 32.2 (2022), pp. 241–268.

[45] Jakob Mökander and Luciano Floridi. "Ethics-based auditing to develop trustworthy AI". In: *Minds and Machines* 31.2 (2021), pp. 323–327.

[46] Jakob Mökander and Luciano Floridi. "Operationalising AI governance through ethics-based auditing: an industry case study". In: *AI and Ethics* 3.2 (2023), pp. 451–468.

[47] Florian Tramèr et al. *Stealing Machine Learning Models via Prediction APIs*. 2016. arXiv: 1609.02943 [cs.CR]. URL: https://arxiv.org/abs/1609.02943.

[48] Matthew Jagielski et al. "High Accuracy and High Fidelity Extraction of Neural Networks". In: *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, Aug. 2020, pp. 1345–1362. ISBN: 978-1-939133-17-5. URL: https://www.usenix.org/conference/usenixsecurity20/presentation/jagielski.

[49] Smitha Milli et al. *Model Reconstruction from Model Explanations*. 2018. arXiv: 1807.05185 [stat.ML]. URL: https://arxiv.org/abs/1807.05185.

[50] Junyin Wei et al. "Leaky DNN: Stealing Deep-Learning Model Secret with GPU Context-Switching Side-Channel". In: *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)* (2020), pp. 125–137. URL: https://api.semanticscholar.org/CorpusID:220939286.

[51] Santiago Zanella-Beguelin et al. "Grey-box Extraction of Natural Language Models". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 12278–12286. URL: https://proceedings.mlr.press/v139/zanella-beguelin21a.html.

[52] Nicholas Carlini et al. *Stealing Part of a Production Language Model*. 2024. arXiv: 2403.06634 [cs.CR]. URL: https://arxiv.org/abs/2403.06634.

[53] Sella Nevo et al. "RAND Report: Securing AI Model Weights". In: (2024).

[54] Stuart Millar. *Vulnerability Detection in Open Source Software: An Introduction*. 2022. arXiv: 2203.16428 [cs.CR]. URL: https://arxiv.org/abs/2203.16428.

[55] Rishi Bommasani et al. *Considerations for Governing Open Foundation Models*. Tech. rep. Stanford HAI, 2024.

[56] Misc. *NTIA AI Open Model Weights Request for Comments*. https://www.regulations.gov/document/NTIA-2023-0009-0001/comment. [Accessed 17-09-2024]. 26 February 2024.

[57] Anca Dragan, Helen King, and Allan Dafoe. *Introducing the Frontier Safety Framework — deepmind.google*. https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/. [Accessed 17-09-2024]. 17 May 2024.

[58] OpenAI. *OpenAI Preparedness Framework*. https://cdn.openai.com/openai-preparedness-framework-beta.pdf. [Accessed 17-09-2024]. 2023.

[59] Jide Alaga, Jonas Schuett, and Markus Anderljung. *A Grading Rubric for AI Safety Frameworks*. 2024. arXiv: 2409.08751 [cs.CY]. URL: https://arxiv.org/abs/2409.08751.

[60] Anthropic. *Responsible Scaling Policy Updates — anthropic.com*. https://www.anthropic.com/rsp-updates. [Accessed 21-10-2024]. 2024.

[61] Anthropic. *Third-party testing as a key ingredient of AI policy — anthropic.com*. https://www.anthropic.com/news/third-party-testing. [Accessed 22-10-2024].

[62] Jonathan Easton. *Meta withholds advanced AI model from EU amid regulatory uncertainty — nationaltechnology.co.uk*. https://nationaltechnology.co.uk/Meta_Witholds_Advanced_AI_Model_From_EU.php. [Accessed 17-10-2024].

[63] Office of Public Affairs NTIA. *NTIA Supports Open Models to Promote AI Innovation | National Telecommunications and Information Administration — ntia.gov*. https://www.ntia.gov/press-release/2024/ntia-supports-open-models-promote-ai-innovation. [Accessed 17-09-2024]. 30 July 2024.

[64] NTIA. *Dual-Use Foundation Models with Widely Available Model Weights*. https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf. [Accessed 17-09-2024]. 30 July 2024.

A Some Open Problems in Model Access Governance

This section gestures to six categories of open problems Model Access Governance should prioritize:

4.1. Establishing Access Elements. Decision-makers lack a clear and consistent framework for describing different elements of model access. They need options and terms to describe them which are distinct, feasible, consistent, and precise. For example:

- How can access elements be defined so as to support effective decision-making?
- How can labs ensure decisions to withold access are robust and enforceable?

4.2. Evaluating Risks. The risks of providing different access styles to different groups is currently highly unclear. Decision-makers need accurate estimates of potential uplifts in misuse risks, systemic risks, and risks to privacy given different access regimes. E.g.:

- What is the net marginal uplift in risks from variously capable malicious actors given various styles of model access, under various technological assumptions?
- How might different model access regimes affect robustness, create security vulnerabilities, or compromise privacy?

4.3. Evaluating Benefits. The benefits of providing different access styles to different groups is currently highly unclear. Decision-makers need accurate estimates of potential uplifts in value to different groups, security, and scientific research given different access regimes. E.g.:

- What is the net marginal uplift in value from AI models for different use cases given various styles of model access, and how might this be distributed differently between different groups?
- How might different model access regimes improve the robustness and security of models, improve privacy, or accelerate research into AI systems?

4.4. Navigating Trade-offs. Governing model access responsibly will require value-driven decisions under significant uncertainty. Decision-makers need guidance and principles for weighing evidence, eliciting inputs from stakeholders, and aligning corporate incentives. E.g.:

- How should decision-makers govern model access given high uncertainty or risks which affect different groups asymmetrically?
- How should decisions be structured to resist perverse incentives such as (for instance) fiduciary responsibility or pressure from a 'race' scenario?

4.5. Paths to Impact. Governing model access effectively will require inter-organisational collaboration. Labs, AISIs, governments, third parties and international bodies need clear roles and processes to support best practices in model access governance. E.g.:

- How might frontier AI companies adopt processes for empirically evaluating access styles, and providing them when benefits sufficiently outweigh the risks?
- Which organisations are best placed to play which roles in model access governance, and where should capacity be strengthened to support them?

4.6 Future-Proofing. Decisions about model access may be difficult to reverse, even as the technological landscape changes substantially. Decision-makers need a clear view of trends in AI and their potential implications for model access governance. E.g.:

- How will model access governance change as models become (for instance) cheaper, more powerful, more robust, smaller, or more valuable for certain use cases?

B Related Work

Significant contributions to Model Access Governance have already been made across diverse fields. By highlighting and presenting this work as a unified field, we aim to emphasize the value of integrating and coordinating this research, and inspire further progress in this area.

Structured access, transparency, and information security. Structured transparency—a framework designed to facilitate collaboration and information sharing whilst addressing privacy concerns [41]—has been applied to model access decisions around AI systems for several years [5]. Recent work has inquired as to the information required to do valuable safety research [17] and how audits might best be structured [42, 43, 44, 45, 46].

At the same time, recent technical work has explored the extent to which these attempts to control information flows are robust: for instance, how easy it is to "steal" the weights of a production language model through interactions [47]. Early attempts to steal models with high fidelity attacks focused on specific deep neural networks with ReLU activations [48, 49], but more recent work has demonstrated that attacks can succeed against larger models [50, 51] and even parts of production models [52]. Furthermore, recent theoretical work on information security has suggested that the weights of models developed by frontier AI companies might not be resilient to extraction attacks by sophisticated and state-level actors, suggesting the need for further security in order to govern model access responsibly [53].

Open vs closed-source debate. Early work related to model access governance transferred a framing of 'openness' from earlier software research (e.g. Millar et al., 2022 [54]). These concepts and practices have not always transferred seamlessly to the governance of AI models [16], and there is as yet no official definition of 'open source AI' [29]. Some research has argued that we should move 'beyond the open vs closed debate' [14] towards 'hybrid' access regimes that involve providing access to some aspects of the model but not others [30]. Nonetheless, several papers have usefully contributed to discussions around model access governance by exploring possible marginal risks from models with widely-available weights [55] and possible alternatives to open-weight deployment [8, 7]. Others have made a vigorous case for various definitions of open-source [6, 56].

Model Deployment Governance. Frontier AI companies that have produced safety frameworks around AI systems include Anthropic [13], Google DeepMind [57] and OpenAI [58]. However, the extent to which these frameworks are comprehensive or feasible is unclear [59], and they currently do not provide explicit commitments for how model access should be governed such as criteria and standards for the styles of access that should be provided to different groups. Companies that have made explicit comments in favour of 'open' access styles include Meta [10], Mozilla and LAION [56], but these commitments may not be binding as the technology develops. Encouragingly, a recent update to Anthropic's Responsible Scaling Policy suggested that 'developing a tiered access system' was a priority [60], suggesting that developments to incorporate responsible access governance into existing safety policies may be under consideration. However, as the earlier report noted, "we are a company that primarily develops proprietary systems, and we don't have the legitimacy to make claims here about what should or shouldn't be acceptable in openly disseminated systems", suggesting the need for more substantial collaboration between frontier AI companies, governments, and third-party auditing and evaluation organisations [61].

Government Practice, Reports and Legislation. No major legislation currently distinguishes between styles of access explicitly or provides guidance for frontier AI companies making decisions about which aspects or style to provide to particular groups. However, preliminary steps have already been taken towards governing model access responsibly. In Europe, for instance, Article 53 of the EU AI Act indicates some stipulations as to the variety of information providers of general purpose AI models must make available to different parties [3]. However, there is some ambiguity as to the nature of these requirements (e.g. what a "sufficiently detailed summary about the content used for training" might look like for contemporary large language models). There is also significant ambiguity as to what information about models is to be provided to different third-parties for performing model evaluations (Art. 92). Furthermore, the definition of "general-purpose AI models with systemic risks" in Article 51 does not indicate how the nature of access provided to the model might contribute to the risk profile of the model [3]. Meta recently chose not to open-source a model in the EU, citing "regulatory uncertainty" [62].

The US Government has also taken steps towards governing model access. Section 4.6 of the Executive Order tasked the Secretary of Commerce with soliciting feedback "from the private sector, academia, civil society, and other stakeholders through a public consultation process on the potential risks, benefits, other implications, and appropriate policy and regulatory approaches related to dual-use foundation models for which the model weights are widely available" [2]. The National Telecommunications and Information Administration (NTIA) received 332 comments from their Request for Comment published in February 2024 [56]. In July they published an explicit statement in favour of open-source [63]. However, their report on the topic, 'Dual-Use Foundation Models with Widely Available Model Weights', recognises both risks and benefits, and recommends that the US government builds capacity and considers alternatives as the technology develops [64]. Beyond the US and EU, legislation is currently meagre and reference to model access governance is limited.