Re: Draft Report of the Joint California Policy Working Group on AI Frontier Models

The Center for Democracy & Technology (CDT) respectfully submits this response to the Joint California Policy Working Group on AI Frontier Models' request for feedback on its Draft Report.[1] CDT is a nonprofit 501(c)(3) organization that works to advance civil rights and civil liberties in the digital age. Among our priorities, CDT advocates for the responsible and equitable design, deployment, and use of new technologies such as artificial intelligence (AI), and promotes the adoption of robust, technically-informed solutions for the effective regulation and governance of AI systems.

We commend the Working Group on its thoughtful exploration of some of the key policy issues facing California and other states as they consider how to effectively govern advanced AI. The Working Group's draft report is clear and thorough: it draws on revealing case studies and theoretical principles to make a compelling argument for targeted interventions that would, if implemented, help mitigate the risks of AI while allowing Californians to reap its benefits.

The draft report is admirably evidence-focused. Rather than basing its recommendations on speculative or purely theoretical concerns, it recommends policies that are firmly grounded in evidence. At the same time, however, the draft report crucially emphasizes that evidence-based policy should not be purely reactive. By default, the broad, rigorous evidence that policymakers need in order to develop evidence-based AI policies is and will likely remain sorely lacking. Not only is the public's visibility into the practices of AI companies inadequate,[2] so too is the scientific and policy communities' understanding of the capabilities, behavior, risks, and real-world impacts of the AI systems those companies have developed (let alone the systems they could develop in the next few years).[3] Historical precedent — including the examples discussed in the draft report — suggests that without action from policymakers, the evidence base necessary to inform rigorous policymaking will be too slow to develop.[4] As such, we appreciate the draft report's focus on *evidence-seeking* policy: targeted measures that can proactively surface the information that regulators and the public need to effectively manage the risks and impacts of AI.

[1] Jennifer Tour Chayes, Mariano-Florentino Cuéllar, and Li Fei-Fei, "Draft Report of the Joint California Policy Working Group on Frontier AI Models," March 2025, https://www.cafrontieraigov.org/.

[2] Rishi Bommasani et al., "The Foundation Model Transparency Index," *Stanford Center for Research on Foundation Models* (May 2024), https://crfm.stanford.edu/fmti/May-2024/index.html (finding that developers satisfied 37 out of 100 transparency indicators on average in October 2023, and 58 out of 100 on average in May 2024).

[3] Percy Liang et al., "Holistic Evaluation of Language Models," *Transactions on Machine Learning Research* (October 2023), https://arxiv.org/abs/2211.09110; Amy Winecoff and Miranda Bogen, "Trustworthy AI Needs Trustworthy Measurements," *Center for Democracy & Technology* (March 2024), https://cdt.org/insights/trustworthy-ai-needs-trustworthy-measurements/.

[4] See Section 2.3 of the draft report. *See also* Stephen Casper et al., "Pitfalls of Evidence-Based AI Policy," *arXiv* (February 2025), https://arxiv.org/abs/2502.09618.

We particularly applaud the draft report's focus on transparency, which is a cornerstone of effective AI governance. For years, researchers and advocates have argued that transparency is a vital lever for managing the risks of AI and making its benefits available to a variety of stakeholders.[5] Transparency can allow researchers to rigorously study the impacts, risks, and benefits of AI, and it can provide the public with critical information about AI systems that affect their lives in important ways.[6] Transparency can also push companies to develop AI systems more responsibly and make it possible to hold those companies accountable when AI systems cause harm.[7] Moreover, as the draft report points out, transparency is a key plank of evidence-seeking policy.

For the most part, the transparency that currently exists in the AI ecosystem is due to voluntary commitments on the part of AI companies (though there are notable exceptions).[8] Yet, such commitments are not a stable foundation for managing AI risks. Different companies' transparency commitments are very different — for example, companies use different technical methods to assess the same capability or risk vector — making it difficult for researchers and policymakers to make direct comparisons between them. Moreover, as business incentives change, companies' willingness to abide by their previous commitments may change too, making their purely voluntary commitments to transparency inherently unstable.[9] Indeed, some

[5] *E.g.,* Bommasani et al., *supra* note 2.

[6] Stephen Casper et al., "Black-Box Access Is Insufficient for Rigorous AI Audits," *ACM Conference on Fairness, Accountability, and Transparency* (June 2024), https://dl.acm.org/doi/10.1145/3630106.3659037 (arguing that transparency into training and deployment information is essential for rigorous independent AI evaluation); Matt Scherer, "Colorado's AI Act is a Step in the Right Direction," *Center for Democracy & Technology* (May 2024), https://cdt.org/insights/colorados-artificial-intelligence-act-is-a-step-in-the-right-direction-it-must-be-strengthened-not-weakened/ (explaining the benefits of public transparency about AI systems used to make important decisions).

[7] *See, e.g.,* National Telecommunications and Information Administration, *AI Accountability Policy Report*, March 2024, https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report. *See also* Kevin Klyman, "How to Promote Responsible Open Foundation Models," *Stanford Institute for Human-Centered AI* (October 2023), https://hai.stanford.edu/news/how-promote-responsible-open-foundation-models.

[8] For examples of voluntary transparency-related commitments, *see, e.g.,* "Biden-Harris Administration Secures Voluntary Commitments From Leading Artificial Intelligence Companies to Manage the Risks Posed by AI," *The White House* (July 2023), https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/; "Anthropic's Responsible Scaling Policy," *Anthropic* (September 2023), https://www.anthropic.com/news/anthropics-responsible-scaling-policy/. (Examples of existing transparency requirements include the EU AI Act and Colorado's SB 24-205.)

[9] *E.g.,* Shakeel Hashim, "Google Breaks Its Promises," *Transformer* (March 2025), https://www.transformernews.ai/p/google-breaks-its-promises (describing Google's violation of its 2023 commitment to "publish reports for all new significant model public releases"). Note that as advanced AI systems become more capable, companies may face stronger incentives to renege on their earlier transparency commitments. On the one hand, more capable models may generate more revenue, so companies will pay a higher cost if complying with transparency requirements forces them to delay model development or release. On the other hand, since more capable models may have significant hazardous or dual-use capabilities, companies may see disclosing information about those capabilities as reputationally risky.

companies have recently begun to roll back the level of transparency they maintain in response to increasing competitive pressures.[10] As such, regulators — including those at the state level — have a vital role to play in ensuring sufficient transparency into AI systems and their development.[11]

I. Adding detail to transparency recommendations

While transparency in AI is vital, not all types of transparency are equally useful. Research from CDT and others has found that in order for calls for transparency to be genuinely beneficial, they must be precisely scoped and backed by a clear theory of change.[12] Similarly, they must also come with incentives that both encourage accurate, good-faith disclosures from companies and make it possible to hold those companies accountable when their behavior is unacceptable.[13] Transparency measures with these features can have immense positive effects, while vague, overbroad calls for transparency can easily lead to "transparency-washing" — token disclosures that, while imposing no real costs on developers, also fail to create real accountability or produce other positive benefits.[14] The cost of transparency-washing is high: there are real limits on the political will and regulatory capacity for addressing the risks and impacts of AI, and poorly-formulated demands for transparency can pull attention away from more careful proposals that would, if implemented, bring genuine benefits to the AI ecosystem.[15] As such, it is critical that the Working Group's final report include transparency recommendations that are specific and clearly tied to specific regulatory objectives.

While the draft report's recommendations are already admirably specific, we believe that still further detail would be both feasible and beneficial. The draft report highlights several important aspects of AI model development into which transparency is warranted, including developers' safety practices and pre-deployment testing. We agree with the draft report that "[t]ransparency into the risks associated with foundation models [and] what mitigations are implemented to address risks ... is the foundation for understanding how model developers manage risk." Indeed, information about model developers' safety practices can inform the behavior of not just

[10] Maxwell Zeff, "Google Is Shipping Gemini Models Faster Than Its AI Safety Reports," *TechCrunch* (April 2025), https://techcrunch.com/2025/04/03/google-is-shipping-gemini-models-faster-than-its-ai-safety-reports/.

[11] Scherer, *supra* note 6.

[12] Amy Winecoff and Miranda Bogen, "Improving Governance Outcomes Through AI Documentation: Bridging Theory and Practice," *Center for Democracy & Technology* (September 2024), https://cdt.org/insights/report-improving-governance-outcomes-through-ai-documentation-bridging-theory-and-practice/. Bommasani et al., *supra* note 2.

[13] National Telecommunications and Information Administration, *supra* note 7.

[14] Bommasani et al., *supra* note 2; Anna Kawakami, Daricia Wilkinson, and Alexandra Chouldechova, "Do Responsible AI Artifacts Advance Stakeholder Goals? Four Key Barriers Perceived by Legal and Civil Stakeholders," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (October 2024), https://doi.org/10.1609/aies.v7i1.31669; Monika Zalnierute, "'Transparency-Washing' in the Digital Age: A Corporate Agenda of Procedural Fetishism," *Critical Analysis of Law* (March 2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3805492.

[15] danah boyd, "Transparency != Accountability," *EU Parliament Roundtable on Algorithmic Accountability and Transparency* (November 2016), https://www.danah.org/papers/talks/2016/EUParliament.html.

regulators and researchers,[16] but also model deployers whose interests and obligations require them to be sensitive to the degree and type of risk posed by the models they rely on.[17]

However, recognizing that AI safety is a fundamentally *sociotechnical* problem,[18] we emphasize the importance of transparency into not only the technical safeguards a developer uses, but also relevant internal governance practices that the developer uses to manage risk. Without context, it may be difficult for external actors to assess the appropriateness or adequacy of a particular set of safeguards. As such, information about how the developer itself assesses the efficacy of its safeguards would be a crucial complement to technical information about those safeguards.[19] So too would information about how the developer decides that risks have been mitigated adequately enough to deploy a model, as well as how the developer intends to respond to significant risks that materialize.

II. Adding rigor to pre-deployment evaluations

The draft report also rightly calls for transparency into the pre-deployment assessments that developers use to understand their models' capabilities and risks. The public sharing of these assessments gives researchers, regulators, and the public a critical advance window into the potential impacts of AI systems before those impacts (which may be both consequential and irreversible) manifest themselves at a large scale. Moreover, the public sharing of these assessments can also push developers to behave more responsibly. Developers will likely be more reluctant to skip critical pre-deployment tests in order to rush the release of a model if they are required to publicly admit to skipping those tests. And advocates, researchers, and regulators can hold developers appropriately accountable if their pre-deployment assessments indicate an unacceptable level of risk.

However, it is essential that the information developers disclose about their pre-deployment assessments be sufficiently detailed. For instance, high-level summaries of a developer's pre-deployment tests (e.g., classifying a model's overall risk level as "low" or "medium") are of limited value on their own: they lack the detail necessary to be useful to independent researchers, as well as the nuance needed to usefully inform regulators, deployers, or the public. Moreover, non-detailed information about pre-deployment assessments often has questionable value as evidence to researchers or regulators. Because the science of AI

[16] Risto Uuk et al., "Effective Mitigations for Systemic Risks from General-Purpose AI," *arXiv* (November 2024), https://arxiv.org/abs/2412.02145.

[17] Winecoff and Bogen, *supra* note 12.

[18] Miranda Bogen and Amy Winecoff, "Applying Sociotechnical Approaches to AI Governance in Practice," *Center for Democracy & Technology* (May 2024), https://cdt.org/insights/applying-sociotechnical-approaches-to-ai-governance-in-practice/; Brian J. Chen and Jacob Metcalf, "A Sociotechnical Approach to AI Policy," *Data & Society* (May 2024), https://datasociety.net/library/a-sociotechnical-approach-to-ai-policy/. *See also* Seth Lazar and Alondra Nelson, "AI Safety on Whose Terms?" *Science* (July 2023), https://www.science.org/doi/10.1126/science.adi8982.

[19] For an example of how information about safeguards divorced from social context can be hard to interpret, one can consider much of the recent work on defending against jailbreaks in foundation models. Participants in this literature compete to create defenses that score highly on jailbreak benchmarks, but without context from developers on how they make decisions relating to the jailbreakability of their models, it is difficult to understand whether a novel method for defending against jailbreaks is adequate.

assessment is so nascent, external actors cannot reasonably assume that a developer's internal assessments are sufficiently rigorous without reviewing or probing their methodology themselves — even if the developer performed those assessments in good faith. It is all too common for such assessments to be flawed in ways that seriously undermine their conclusions, either because of issues with their implementation[20] or because the assessments themselves are premised on false assumptions.[21] For this reason, it is crucial that when developers disclose information about their pre-deployment assessments, they include enough scientific detail for independent experts to be able to analyze, critique, and replicate them.

III. Incentivizing third-party assessments

We are pleased that the draft report promotes third-party assessment as a key lever for foundation model risk management. The report correctly notes that third parties (as opposed to developers themselves and second parties who maintain contractual relationships with developers) have a crucial role in building a rigorous evidence base concerning AI risks because they are uniquely disinterested actors in the AI ecosystem.[22] The measures recommended by the report, such as safe harbor protections for independent researchers and responsible disclosure policies, are important steps towards preserving the ability of third-party researchers to rigorously assess the capabilities and risks of foundation models. However, the report leaves out a crucial class of independent assessments: pre-deployment risk assessments performed by third parties.[23] As such, the Working Group should consider how to incentivize developers to grant pre-deployment access to qualified third-party evaluators where warranted.

IV. Reporting Intervals and Scoping Metrics

The draft report does not discuss *when* developers should be required to disclose key information regarding their models, but this question is an important and challenging one. Currently, most information regarding new models is provided alongside the release of that model, in the form of a system card.[24] This approach is intuitive, but it also leaves notable gaps. Most importantly, developers' new releases increasingly take the form of relatively frequent

[20] *See, e.g.,* Epoch AI (@EpochAIResearch), *Twitter* (March 7, 2025), https://twitter.com/EpochAIResearch/status/1898149924556226908 (providing an example of an independent evaluation of OpenAI's o3-mini model on the FrontierMath benchmark that arrived at substantially different results than OpenAI's internal evaluation on the same benchmark).

[21] *See* Arvind Narayanan and Sayash Kapoor, "GPT-4 and Professional Benchmarks: The Wrong Answer to the Wrong Question," *AI Snake Oil* (March 2023), https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks (explaining that some internal assessments have limited value because they are premised on false assumptions). *See also* Anson Ho and Jean-Stanislas Denain, "The Real Reason AI Benchmarks Haven't Reflected Economic Impacts," *Epoch AI* (March 2025), https://epoch.ai/gradient-updates/the-real-reason-ai-benchmarks-havent-reflected-economic-impacts (supporting Narayanan and Kapoor's central claim while offering a different theoretical explanation).

[22] Miranda Bogen, "Assessing AI: Surveying the Spectrum of Approaches to Understanding and Auditing AI Systems," *Center for Democracy & Technology* (January 2025), https://cdt.org/insights/assessing-ai-surveying-the-spectrum-of-approaches-to-understanding-and-auditing-ai-systems/.

[23] Such as those by the U.S. AI Safety Institute, UK AI Security Institute, and METR, among others.

[24] *See, e.g.,* Anthropic, "Claude 3.7 Sonnet System Card," February 2025, https://www.anthropic.com/claude-3-7-sonnet-system-card.

incremental updates to existing foundation models, rather than infrequent major releases of entirely new models. These incremental updates are somewhat minor: individual updates may not be significant enough to warrant an accompanying system card. However, when taken together, these incremental updates can represent considerable changes in model characteristics and capabilities, and therefore considerable changes in the associated risks. For this reason, the working group should explore alternatives to tying transparency only to major model releases. One alternative may be to require developers to update their transparency reports at regular intervals (along the lines of the Foundation Model Transparency Index),[25] in addition to providing reports with major releases.

Lastly, we are optimistic about the draft report's emphasis on adverse event reporting as a key means of better understanding the impacts and risks of AI. Researcher access to AI usage data is a necessary prerequisite for a fully rigorous analysis of AI risks, and companies can share key information about how their models are used without compromising user privacy.[26] We are also encouraged that the draft report discusses how cost-based thresholds and other metrics can be used to complement compute-based thresholds. Methods for scoping AI regulation are an important emerging topic, one which deserves further investigation.

A final consideration regarding the transparency portions of the Working Group's Draft Report are the constitutional tensions inherent to government-mandated transparency. The Draft Report appropriately notes that transparency mandates may implicate concerns regarding privacy, security, intellectual property and trade secrets, innovation, and transparency washing, and in addition to these considerations, transparency mandates in the United States may also need to contend with the First Amendment.[27] The Working Group should clarify that its recommendations are consistent with the fact that requirements that an individual or entity "speak a particular message" are legally considered compelled speech, which is generally subject to the most stringent form of First Amendment review.[28] The First Amendment is not absolute, and courts routinely recognize that certain forms of transparency and compelled speech are more justifiable than others and apply less stringent forms of review to those mandates. In light of recent court cases calling into question the constitutionality of transparency provisions enacted by the State of California,[29] policymakers should be sure to formulate the transparency measures recommended by the Draft Report in a manner that can both empower meaningful transparency and ensure that those measures are protected from constitutional challenge.[30]

[25] Bommasani et al., *supra* note 2.
[26] Gabriel Nicholas, "Grounding AI Policy: Towards Researcher Access to AI Usage Data," *Center for Democracy & Technology* (August 2024), https://cdt.org/insights/grounding-ai-policy-towards-researcher-access-to-ai-usage-data/.
[27] B. Branum, First Amendment Tech Transparency Roadmap, at: https://cdt.org/wp-content/uploads/2025/02/2025-02-13-CDT-FX-Tech-Transparency-Roadmap-final.pdf.
[28] National Institute of Family & Life Advocates v. Becerra, 585 U.S. 755 (2018).
[29] See, e.g., *X Corp. v. Bonta*, 116 F.4th 888 (9th Cir. 2024).
[30] For more on tailoring transparency mandates to the appropriate level of First Amendment scrutiny, please see B. Branum, First Amendment Tech Transparency Roadmap, at: https://cdt.org/wp-content/uploads/2025/02/2025-02-13-CDT-FX-Tech-Transparency-Roadmap-final.pdf.

V. Conclusion

In conclusion, we commend the Working Group for its thoughtful and evidence-focused approach to AI governance. Our feedback has emphasized the need for more specific transparency requirements covering both technical safeguards and internal governance practices, more detailed pre-deployment assessment disclosures, and stronger incentives for third-party evaluations. We look forward to working with California policymakers to help translate the Working Group's recommendations into concrete policies that help California manage AI risks while fostering responsible innovation. CDT remains committed to supporting evidence-based and evidence-seeking policies that will build the knowledge base necessary for effective AI governance in California and beyond.

***

We appreciate the Working Group's solicitation of feedback from stakeholders and affected communities on these important matters. For additional information, or any inquiries, please contact Miranda Bogen (mbogen@cdt.org), Director of CDT's AI Governance Lab.