Fathom's Response to the Joint California Policy Working Group on AI Frontier Models

In September 2024, California Governor Gavin Newsom requested that Dr. Fei-Fei Li, Dr. Mariano-Florentino Cuéllar, and Dr. Jennifer Tour Chayes prepare a report shedding light on the path forward for AI governance in California. Drawing extensively on historical case studies from the consumer products sector, energy industry, and the evolution and governance of the internet, the recently released draft report offers a wealth of valuable insight to guide the development, evaluation, and governance of frontier AI. Below is Fathom's feedback to the Joint California Policy Working Group on AI Frontier Models.

---

1. The structure of the Draft Report includes sections about Context, Transparency and Third-Party Risk Assessment, Adverse Event Reporting, and Scoping. At a high level, what might you find valuable? What types of questions are you most interested in, and how might you use the report in your work?

Fathom is investigating private governance solutions to align commercial incentives with public welfare. In particular, we are developing potential California legislation to create voluntary certification bodies to define and enforce heightened standards of care for the AI industry. Fathom expects to draw heavily on the report's sections on transparency as it refines its private governance model, having already found inspiration in the work of many esteemed scholars, including Professor Gillian Hadfield's and Jack Clark's *Regulatory Markets: The Future of AI Governance* paper. Fathom has already sponsored a bill of this nature in the California legislature.

Fathom would be particularly interested to see the following in the final report:

- Mechanisms for incentivizing developer transparency;
- Recommendations as to which agencies or types of entity should manage, receive, and oversee disclosures, including those from third-party auditors.

2. From your perspective and experience, what key factors do you see affecting California's path forward in AI governance? Please feel free to provide specific feedback referring to the sections of the draft report.

Fathom agrees with the Joint California Policy Working Group on the need for a "thriving policy environment that encourages creativity, attracts top talent, and creates thoughtful synergies across industry, academia, civil society, and policy" (p.14) It also affirms the need for proper safeguards to mitigate the possibility of harms from AI. To achieve this balance, California should prioritize AI governance solutions with six key features:

1. **Subject Matter Expertise.** The draft report indicates a need to minimize the "non-expert application of laws that will inevitably happen if a growing number of frontier AI-related cases appear in front of the courts." (p.14) Fathom believes that AI is uniquely poorly suited to governance by non-experts given its incredible complexity. The California Legislature should seek out governance solutions that utilize subject matter experts to develop standards and best practices.
2. **National, If Not Federal.** The absence of a unified set of standards at the national level threatens to impede innovation by increasing regulatory burden and uncertainty for developers. Federal regulation does not appear forthcoming, but the report is quick to recognize the "critical role states are increasingly playing in building AI policy, with national ripple effects." (p.7) The California Legislature should prioritize governance solutions that create pathways to a unified, national standards system through state-based legislation.
3. **Scalability.** Just as California should spearhead national standards through its state legislation, it should seek out governance models that are interoperable and capable of scaling internationally in recognition of the fact that AI presents global challenges in need of global solutions.
4. **Adaptability.** Fathom's research reveals a strong concern across senior stakeholders that any near-term governance may overindex on current-day model capabilities, quickly becoming outdated as the frontier moves on, and potentially entrenching long-term challenges as with the initial protocols and security frameworks of the internet. Fathom urges California to pursue "governance approaches [that] capitalize on early policy windows when harms can be minimized while simultaneously allowing for adaptation to technical realities as they evolve." (p.12)
5. **Proportionality.** Fathom agrees that "well-designed regulation is proportionate" (p.31) and should meet AI companies where they are in the entrepreneurial cycle to ensure continued innovation.

6. **Clarity for Integration.** Regulated industries have emphatically communicated the need for clear rules of the road regarding liability to de-risk and encourage the widespread adoption of AI technologies.

---

3. Numerous frontier AI governance-focused groups have been working on frameworks, guidance, and reports aiming to leverage scientific research. For what topics are you observing challenges in reaching scientific consensus? Do you have recommendations to bridge gaps?

Fathom's view is that gaps will continue to emerge as capabilities progress, and that effective governance approaches should acknowledge and accommodate this perennial process rather than seek to "solve" it, continually adapting in response to the bleeding-edge of scientific consensus.

---

4. What could be done individually or collectively to leverage frontier AI for Californians' benefit?

Fathom urges the California Legislature to investigate further "policy initiatives that need not explicitly regulate the industry, but instead reconstitute market incentives for companies to internalize societal externalities." (p.7) The Legislature should take inspiration from this country's rich history of successful public-private partnerships and consider multistakeholder regulatory organizations (MRO), through which developers would adopt voluntary standards in exchange for economic and legal incentives. MROs can accommodate expert input, national standards, scalability, adaptability, proportionality, and clarity of integration, and are therefore uniquely well-suited to AI governance.

MROs would be authorized by the state to create and administer opt-in certification processes for AI companies. MROs, comprised of independent subject matter experts, civil society voices, and industry representatives, would certify that AI developers meet rigorous technical and operational standards, validating that benchmarks, evaluation methods, and governance practices have been deployed to manage safety and security risks. Certified developers would earn an affirmative defense against future negligence claims resulting in personal injury or property damage, reducing their exposure to lawsuits and enabling them to drive at the frontier with confidence.

1. The report highlights that "without proactive regulatory frameworks, litigation—an inherently reactive and piecemeal process—[becomes] the default mechanism for

addressing novel technological challenges," (p.13) resulting in the non-expert application of laws. An MRO could pre-empt such an outcome by having subject matter experts proactively define the standard of care up-front.

2. Tort law exists in every state as part of common law. Authorizing MROs to define and certify for the standard of care would help create national standards in two ways. First, defining the standard of care in California would shape the standard of care nationally, driving a national system of standards in the process. Second, other states could decide to point to the same MROs for certification in their state, further cementing these national standards.
3. International governments may opt to follow California's lead and authorize MROs themselves. This would enable U.S. MROs to seek authorization abroad and drive a globally unified system of standards, de-risking global expansion for U.S. companies. MROs may also collaborate and coordinate with international governments, sharing critical insights and promoting the standardisation of best practices to identify, monitor, and mitigate risks of harm from AI platforms.
4. MROs' autonomy affords them the flexibility to evolve standards and best practices as model capabilities evolve. California could incentivize this by requiring the revocation of an MRO's state license should the licensing authority find their methods obsolete for ensuring acceptable levels of risk, or should a certified model cause significant harm.
5. MROs are well-positioned to ensure proportionate governance given their balanced makeup, with independent subject matter experts, civil society voices, and industry representatives shaping standards, and their ability to develop customized compliance pathways for companies at different stages of growth. Moreover, should the set of standards required for certification prove burdensome, developers retain the right to opt-out of the MRO in question and, should they wish, opt-in with a competitor. MROs are thus incentivized to identify pro-innovation pathways lest they be undercut by competing MROs.
6. Specialized MROs could certify fine-tuned models for consumer-facing applications, offering regulated industries the opportunity to clarify and manage liability.

Fathom believes that this private governance approach would succeed in acknowledging "industry expertise while establishing robust mechanisms to independently verify safety claims and risk assessments," (p.10) aligning commercial incentives with public welfare. Similarly, this is an approach that would secure for California a policy environment that promotes innovation, ensuring that the technologies that emerge from California can continue to shape the world for the better.