INSTITUTE
FOR LAW & AI

Draft Report of the Joint California Policy Working Group on AI Frontier Models – Scoping and Definitions Comments

Mackenzie Arnold, Sarah Bernardo | Institute for Law & AI | April 8, 2025

These comments on the Draft Report of the Joint California Policy Working Group on AI Frontier Models were submitted to the Working Group as feedback on April 8, 2025. The opinions expressed in these comments are those of the authors and do not reflect the views of the Institute for Law & AI.

Commendations

1. **The Report correctly identifies that AI models and their risks vary significantly and thus merit different policies with different inclusion criteria.** Not all AI policies are made alike. Those that target algorithmic discrimination, for example, concern a meaningfully different subset of systems, actors, and tradeoffs than a policy that targets cybersecurity threats. What's more, the market forces affecting these different policies vary considerably. For example, one might be far more concerned about limiting innovation in a policy context where many small startups are attempting to integrate AI into novel, high-liability-risk contexts (e.g., healthcare) and less concerned in contexts that involve a few large actors receiving large, stable investments, where the rate of tort litigation is much lower absent grievous harms (e.g., frontier model development). That's all to say: It makes sense to foreground the need to scope AI policies according to the unique issue at hand.

2. **We agree that at least some policies should squarely address foundation models as a distinct category.** Foundation models, in particular those that present the most advanced or novel capabilities in critical domains, present unique challenges that merit separate treatment. These differences emerge from the unique characteristics of the models themselves, not their creators (who vary considerably) or their users. And the potential benefits and risks that foundation models present cut across clean sectoral categories.

3. **We agree that thresholds are a useful and necessary tool for tailoring laws and regulations (even if they are imperfect).** Thresholds are easy targets for criticism. After all, there is something inherently arbitrary about setting a speed limit at 65 miles per hour rather than 66. Characteristics are more often continuous than binary, so typically there isn't a clear category shift after you cross over some talismanic number. But this issue isn't unique to AI policy, and in every other context, government goes on nonetheless. As the Report notes, policy should be proportional in its effects and appropriately narrow in its application. Thresholds help make that possible.

4. **The Report correctly acknowledges the need to update thresholds and definitional criteria over time.** We agree that specific threshold values and related definitional criteria will likely need to be updated to keep up with technological advances. Discrete, quantitative thresholds are particularly at risk of becoming obsolete. For instance, thresholds based on training compute may

become obsolete due to a variety of AI developments, including improvements in compute and algorithmic efficiency, techniques such as distillation, and/or the growing impact of inference scaling. Given the competing truths that setting some threshold is necessary and that any threshold will inevitably become obsolete, ensuring that definitions can be quickly, regularly, and easily updated should be a core design consideration.

5. **We agree that, at present, compute thresholds (combined with other metrics and/or thresholds) are preferable to developer-level thresholds.** Ultimately, the goal of a threshold is to set a clear, measurable, and verifiable bar that correlates with the risk or benefit the policy attempts to address. In this case, a compute threshold best satisfies those criteria—even if it is imperfect. For more discussion, see Training Compute Thresholds: Features and Functions in AI Regulation and The Role of Compute Thresholds for AI Governance.

Recommendations

1. **The Report should further emphasize the centrality of updating thresholds and definitional criteria.** Updating is perhaps *the* most important element of an AI policy. Without it, the entire law may in short time cease to cover the conduct or systems policymakers aimed to target. We should expect this to happen by default. The error may be one of overinclusion—for example, large systems may present few or manageable risks even after a compute threshold is crossed. After some time, we will be confident that these systems do not merit special government attention and will want to remove obligations that attach to them. The error may be one of underinclusion—for example, improvements in compute or algorithmic efficiency, techniques such as distillation, and/or the growing impact of inference scaling may mean that models below the threshold merit inclusion. The error may be in both directions—a truly unfortunate, but entirely plausible, result. Either way, updating will be necessary for policy to remain effective.

   We raise this point because without key champions, updating mechanisms will likely be left out of California AI legislation—leading to predictable policy failures. While updating has been incorporated into many laws and regulations, it was notably absent from the final draft of SB 1047 (save for an adjustment for inflation). A similar result cannot befall future bills if they are to remain effective long-term. A clear statement by the authors of the Report would go a long way toward making updating feasible in future legislation.

   Recommendation: The Report should clearly state that updating is necessary for effective AI policy and explain why policy is likely to become ineffective if updating is not included. It should further point to best practices (discussed below) to address common concerns about updating.

2. **The Report should highlight key barriers to effective updating and tools to manage those barriers.** Three major barriers stand in the way of effective updating. First is the concern that updating may lead to large or unpredictable changes, creating uncertainty or surprise and making it more difficult for companies to engage in long-term planning or fulfill their compliance obligations. Second, some (understandably) worry that overly broad grants of discretion to agencies to update the scope of regulation will lead to future overreach, extending powers to contexts far beyond what was originally intended by legislators. Third, state agencies may lack sufficient capacity or knowledge to effectively update definitions.

   The good news: These concerns can be addressed. Establishing predictable periodic reviews, requiring specific procedures for updates, and ensuring consistent timelines can limit

uncertainty. Designating a competent updater and supplying them with the resources, data, and expert consultation they need can address concerns about agency competency. And constraining the option space of future updates can limit both surprise and the risk of overreach. When legislators are worried about agency overreach, their concern is typically that the law will be altered to extend to an unexpected context far beyond what the original drafters intended—for example, using a law focused on extreme risks to regulate mundane online chatbots or in a way that increases the number of regulated models by several orders of magnitude. To combat this worry, legislators can include a purpose clause that directly states the intended scope of the law and the boundaries of future updates. For example, a purpose clause could specify that future updates extend "only to those models that represent the most advanced models to date in at least one domain or materially and substantially increase the risk of harm X." Purpose clauses can also come in the imperative or negative. For example, "in updating the definition in Section X, Regulator Y should aim to adjust the scope of coverage to exclude models that Regulator Y confidently believes pose little or no material risk to public health and safety."

Recommendation: The Report should highlight the need to address the risks of uncertainty, agency overreach, and insufficient agency capacity when updating the scope of legislation. It should further highlight useful techniques to manage these issues, namely, (a) including purpose clauses or limitations in the relevant definitions, (b) specifying the data, criteria, and public input to be considered in updating definitions, (c) establishing periodic reviews with predictable frequencies, specific procedures, and consistent timelines, (d) designating a competent updater that has adequate access to expertise in making their determinations, (e) ensuring sufficient capacity to carry out periodic reviews and quickly make updates outside of such reviews when necessary, and (f) providing adequate notice and opportunity for input.

3. **The Report should highlight other tools beyond thresholds to narrow the scope of regulations and laws—namely, carve-outs, tiered requirements, multiple definitions, and exemption processes.** Thresholds are not the only option for narrowing the scope of a law or regulation, and highlighting other options increases the odds that a consensus will emerge. Too often, debates around the scope of AI policy get caught on whether a certain threshold is overly burdensome for a particular class of actor. But adjusting the threshold itself is often not the most effective way to limit these spillover effects. The tools below are strong complements to the recommendations currently made in the Report.

   By carve-outs, we mean a full statutory exclusion from coverage (at least for purposes of these comments). Common carve-outs to consider include:

   - Small businesses
   - Startups in particularly fragile funding ecosystems, onerous regulatory environments, or high-upside sectors that merit regulatory favoritism on innovation grounds
   - Open-source model developers or hosts with the caveats noted below
   - Providers of high-volume, low-cost services that could not feasibly exist with additional regulatory costs due to their volume or margins (e.g., some chat bots)
   - Social service providers or governments who provide a socially valuable service at low or no cost, especially where we expect that these actors may under-adopt useful technology due to other frictions

   This is not to say that these categories should always be exempt, but rather that making explicit carve-outs for these categories will often ease tensions over specific thresholds. In particular, it is worth noting that while current open-source systems are clearly net-positive according to any

reasonable cost-benefit calculus, future advances could plausibly merit some regulatory oversight. For this reason, any carve-out for open-source systems should be capable of being updated if and when that balance changes, perhaps with a heightened evidentiary burden for beginning to include such systems. For example, open-source systems might be generally exempt, but a restriction may be imposed upon a showing that the open-source systems materially increase marginal risk in a specific category, that other less onerous restrictions do not adequately limit this risk, and that the restriction is narrowly tailored.

Related, but less binary, is the use of tiered requirements that impose only a subset of requirements or weaker requirements on these favored models or entities, such as, requiring certain reporting requirements of smaller entities while not requiring them to perform the same evaluations. For this reason, more legislation should likely include multiple or separate definitions of covered models to enable a more nimble, select-only-those-that-apply approach to requirements.

Another option is to create exemption processes whereby entities can be relieved of their obligations if certain criteria are met. For example, a model might be exempt from certain requirements if it has not, after months of deployment, materially contributed to a specific risk category or if the model has fallen out of use. Unlike the former two options, these exemption processes can be tailored to case-by-case fact patterns and occur long after the legislative or regulatory process. They may also better handle harder-to-pin-down factors like whether a model creates exceptional risk. These exemption processes can vary in a few key respects, namely:

- Evidentiary: Presumptive or requiring a showing of evidence
- Decision maker: Self-attested, certified by a third party, or approved by a regulator
- Duration: Permanent or temporary
- Rigidity: Formulaic or factor-based with flexible considerations
- Speed: Automatic or requiring action or review

Recommendation: The Report already mentions that exempting small businesses from regulations will sometimes be desirable. It should build on this suggestion by emphasizing the utility of carve-outs, tiered requirements, multiple definitions, and exemption processes (in addition to thresholds) to further refine the category of regulated models. It should also outline some of the common carve-out categories (noting the value of maintaining option value by ensuring that carve-outs for open-source systems are revised and updated if the cost-benefit balance changes in the future) as well as key considerations in creating exemption processes.

4. **We recommend that the Report elaborate on the approach of combining different types of thresholds by discussing the complementary pairing of compute and capabilities thresholds.** It is important to provide additional detail about other metrics that could be combined with compute thresholds because this approach is promising and one of the most actionable items in the Report. We recommend capabilities thresholds as a complement to compute thresholds in order to leverage the advantages of compute that make it an excellent initial filter, while making up for its limitations with evaluations of capabilities, which are better proxies for risk and more future-proof. Other metrics could also be paired with compute thresholds in order to more closely track the desired policy outcome, such as risk thresholds or impact-level properties; however, they have practical issues, as discussed in the Report.

   Recommendation: The Report should expand on its suggestion that compute thresholds be combined with other metrics and thresholds by noting that capabilities evaluations may be a particularly promising complement to compute thresholds, as they more closely correspond to risk and are more adaptable to future developments and deployment in different contexts. Other

metrics could also be paired with compute thresholds in order to more closely track the desired policy outcome, such as risk evaluations or impact-level properties.

5. **The Report should note additional definitional considerations in the list in Section 5.1—namely, risk-tracking, resilience to circumvention, clarity, and flexibility.** The Report correctly highlights three considerations that influence threshold design: determination time, measurability, and external verifiability.

   Recommendation: We recommend that the Report note four additional definitional considerations, namely:

   - Risk-Tracking: How closely is the proxy correlated with the risks a policy looks to manage? Currently, compute correlates strongly with advanced capabilities. While there are some exceptions amongst specialized models, bigger is generally better. This remains true even after meaningful gains in inference scaling; it is true both that more inference compute leads to better results and that for any fixed amount of inference compute, a model with more training compute tends to perform better. Generally, the most compute-intensive models are the most likely to be deployed widely in new contexts and the most likely to exhibit emergent capabilities that pose unique risks. Compute is less correlated with risk than more direct measures like capabilities or risk itself, but both of these proxies are harder to measure and define.
   - Resilience to Circumvention: How difficult is it to game the proxy or evade its application? Thresholds that are more difficult to circumvent are more effective, while easily circumvented thresholds risk becoming useless once a few actors demonstrate the ease of circumvention. Training compute is a difficult proxy to circumvent. While a threshold that focuses solely on training compute could miss models that rely heavily on inference, training compute is still a significant contributor to the capabilities of a model. Derivative models and distillations pose a meaningful obstacle here, as policymakers must decide what and how to cover models with similar performance but different compute inputs. Generally speaking, requirements that lead to paperwork redundancies for similar models can likely be collapsed so that only one model is governed, while rules that relate to preventing or governing specific uses or risks may need to extend to derivatives and distillations to avoid becoming ineffective.
   - Clarity: How certainly can a regulated party predict that they will be affected by regulation? And how quickly and clearly can regulators clarify ambiguities through interpretations and guidances? Compute thresholds are clear relative to more subjective alternatives. While there are some open questions regarding who measures and how to measure compute, order-of-magnitude differences in compute usage will typically allow actors to know whether they fall in or out of scope of a regulation.
   - Flexibility: Will the proxy remain accurate over time—because it remains the same, naturally adjusts, or allows for easy updating? Compute is less naturally adaptable than risk-based or capabilities-based thresholds.

For more discussion, see Training Compute Thresholds: Features and Functions in AI Regulation and The Role of Compute Thresholds for AI Governance.