Implementing the Transparency in Frontier AI Act (TFAIA) Report for the California Department of Technology Multistakeholder Workshops

Stakeholder Workshops convened by
Rob Reich (Stanford University) and Deirdre K. Mulligan (University of California, Berkeley)

On August 12 and 13, 2026, Rob Reich and Deirdre K. Mulligan convened workshops, hosted by Stanford's Institute for Human-Centered AI, UC Berkeley's School of Information, CITRIS and the Banatao Institute, UC Berkeley Center for Long-Term Cybersecurity, and UC Berkeley Risk & Security Lab. The purpose was to facilitate robust and diverse stakeholder input into the California Department of Technology's (CDT) first annual report under the Transparency in Frontier AI Act (TFAIA). Enacted on January 1, 2026, the TFAIA, also known as SB 53, is one of the first state regulatory frameworks for frontier AI, and it includes provisions on transparency, incident reporting, and whistleblower protections. TFAIA's various provisions turn on three key definitions of "frontier model," "frontier developer" and "large frontier developer." TFAIA leaves some flexibility in the definitions, cognizant that AI is evolving quickly.

To further the TFAIA's goals, the legislature charged CDT to conduct a yearly assessment of relevant evidence and developments and to provide a report to the legislature with recommendations about whether and how to update these key definitions to ensure that they accurately reflect technological developments, scientific literature, and widely accepted national and international standards. The TFAIA directs CDT to consider input from stakeholders, including academics, civil society, industry, the open-source community, and governmental entities in forming its recommendations.

This report contains a synthesis of the discussions over the two workshops along with concrete recommendations to CDT. The report was written by Deirdre K. Mulligan and Rob Reich and draws upon the contributions of participants in the workshop. The workshops were held under Chatham House rules, so nothing is attributed to any attendee. The report is not a collectively endorsed document. An Appendix lists the participants and their organizational affiliations. *The recommendations, while reflecting the workshop, are the authors' alone.*

TFAIA is an important stage-setting law: it builds a dialogic relationship between companies, third-party testing and evaluation organizations, and the California state government. Whatever qualms with the definitions and resulting coverage — reflected in the discussion and recommendations below — institutionalizing transparency obligations

and reporting requirements established a relationship between the companies and the California state government. In the period between the workshops and the release of this report, California has enacted two new bills that, like the TFAIA, task state agencies with developing elements of the emerging AI governance ecosystem. Robust implementation of the TFAIA and these new authorities requires new resources – infrastructure and experts. The authors encourage state leaders to prioritize building additional state capacity to support its ambitious governance efforts.

Finally, the authors believe that TFAIA embodies the longstanding view of state-level policy as a laboratory for democracy. In this respect, California's efforts have potential implications for the policymaking activities in other states and jurisdictions.

Part 1: Reflection on Definitions

Beyond the compute threshold.

The TFAIA defines "frontier model" as a foundation model that was trained using a quantity of computing power–including computing for the original training run and for any subsequent fine-tuning, reinforcement learning, or other material modifications the developer applies to a preceding foundation model–greater than 10^26 integer or floating-point operations (FLOPs).

There was widespread recognition that a FLOP threshold, initially adopted in the Biden Executive Order on AI in 2023, is an imperfect proxy for risk. For example, stakeholders explained that smaller models can pose catastrophic risks, for example those trained on biological data, so focusing on compute, or compute alone, is decidedly underinclusive. Stakeholders emphasized that raw compute at training time matters less and less given new strategies (e.g., reinforcement learning). A FLOP threshold–at least on its own–makes no sense as a determinant of risk.

What would be better? Stakeholders suggested that a more robust definition would consider a mix of compute, algorithms, and data, all of which contribute to a model's capabilities and associated risks. And ultimately, the goal must be to advance the scientific practice of AI evaluation and the ability to understand what happens in the black box of an AI model. In the absence of robust mechanistic interpretability, policymakers are left choosing among imperfect proxies for identifying risk.

Recommendation: *Extend the definition of frontier model to include both foundation models as described and models developed through distillation of such models, and add a set of alternate catastrophic risk specific metrics based on the literature and experiences to date that would capture capabilities produced by compute, data and algorithms but that fall below the 10^{26} compute threshold.*

Proposed Language:

> (i) (1) "Frontier model" means a foundation model that was trained using a quantity of computing power greater than 10^26 integer or floating-point operations ***or models trained through distillation of models using this quantity of computing power; or a model using primarily specific kinds of data, a quantity of computing power and algorithms that experts in relevant risks believe will yield models that provide expert-level assistance in the creation or release of a chemical, biological, radiological, nuclear weapon or cyberattacks.***
>
> (2) The quantity of computing power described in paragraph (1) shall include computing for the original training run and for any subsequent fine-tuning, reinforcement learning, or other material modifications the developer applies to a preceding foundation model.

Domain-specific taxonomies of harms can help solve threshold problems.

Risks and harms are domain-specific and arise from different mixes of compute, algorithms, and data. In some use cases, a small startup producing a model well-tailored to a particular domain can pose a risk of catastrophic harm. Domain-specific taxonomies of harm along with sociotechnical maps of how aspects of AI systems–compute, algorithms, data–contribute to them can support the development of more valid and robust risk thresholds.

"Hyper-focusing on model capabilities misses real harms": A singular focus on models is misguided; a sociotechnical orientation is necessary to address catastrophic harms and risks.

Many participants explained that the vast array of social and economic activity that contributes to harms cannot be captured, or even identified, in model evaluations alone. The current emphasis on technical evaluations of AI models is important but incomplete, as it fails to capture the ways other technical elements of a system, for example, the harness (the tools, memory and workflow of a given system), alters model performance

and foreseeable risk. Some participants suggested building a cybersecurity evaluation system that captures harms and risks as they arise in everyday life. They explained the dire need for methodologies that holistically assess harms and risks and the money and people required to build them. In addition to time, money and expertise, they explained that developing scenarios and other methods that validly captured the real-world risks and harms of concern to government leaders and the public was difficult. Finally, they noted that some risks, for example misalignment, can produce broader structural harms that may be poorly captured in traditional standards in areas like cybersecurity.

The basic insight is this: valid, reliable, and robust evaluations of AI models must go beyond technical assessments alone. The risk of an AI-developed novel bio-pathogen, for example, depends first on whether the use of an AI model provides an advantage over what would be possible without an AI model, and, second, whether with the AI-assisted knowledge the human is able to convert this knowledge into the manufacture of a novel pathogen in the real world. The former calls for a combination of technical and social scientific methods; the latter for social science alone. Hence the significance of a sociotechnical approach to AI evaluation.

In the absence of valid and robust standards, both process and technical evaluations are unlikely to reduce risk.

Participants cautioned against conflating current industry practices of red-teaming and benchmarking with the standards common in other technical and engineering fields. Standards and evaluations against them can serve an important role in safety engineering. But to do so, the standards must be a product of expert consensus and have valid domain-specific meaning. Today's AI evaluations lack relevant expertise, consensus, and validity. The lack of standards in the AI space is challenging; many operate with their own articulated standards for what they think the model should do. This has yielded a chaotic ecosystem. It's not clear what the standards should be if everyone has proprietary standards. The current situation is not a sound basis for safety evaluation science.

Standards are also required for auditors. In more mature areas, there are clear standards — term limits, conflict-of-interest rules, etc. — that along with expert consensus standards for conducting evaluations produce more meaningful audits. Attention in the workshops was directed at the AI Evaluator Forum and its draft standards for conducting an independent technical evaluation of an AI model or system.

Independent oversight can be an important part of a regulatory framework as long as there are clear standards for evaluation. Independent assessors can audit compliance with standards set in law, process standards such as those found in the NIST AI Risk Management Framework, and/or technical standards. In the absence of clarity about the standards however, independent oversight can create a race-to-the-bottom dynamic and paper-over and legitimate poor practices and designs.

The TFAIA's definition of catastrophic handicaps the capacity of policymakers to generate the shared knowledge base essential to ensure "frontier developers tak[e] due care in their development and deployment of frontier models proportional to the scale of the foreseeable risks."

Participants explained that the aim of TFAIA to use transparency to foster sound risk management was undermined by a narrow definition of catastrophic risk. The definition fails to encompass many of the documented harms arising today from the deployment of AI models, including deterioration of mental health, youth suicides, and most recently, evaluations of internal models that result in harms to third parties, as with the OpenAI/Hugging Face incident from early summer 2026.

Participants pointed to the broader definition of foreseeable risks found in the EU code of practice as a more robust alternative and observed several problems with the scope of catastrophic risk. First, TFAIA does not address harms to which covered models may meaningfully contribute, including mental health. Second, even with respect to the kinds of risks deemed catastrophic under the TFAIA–providing expert-level assistance in the creation or release of a chemical, biological, radiological, or nuclear weapon; conducting a cyberattack or what would be considered murder, assault, extortion, or theft if done by a human, without meaningful human oversight, intervention, or supervision; or evading the control of its frontier developer or user–the law only captures them if they arise from a single incident, were foreseeable, and materially contribute to the death or serious injury of more than 50 people or more than one billion dollars ($1,000,000,000) in damage to, or loss of, property. This requirement would fail to cover the loss-of-control risk in the OpenAI incident. Given the relatively low compliance burdens of reporting, participants stated that a broader definition of catastrophic risk (expanding the kinds of risks and lowering the magnitude of harm) would better promote the shared learning necessary for sound risk management and accountability. Importantly, greater information-sharing fostered by a reworked definition could provide the evidence to produce the standards necessary for meaningful evaluations.

Recommendation: *Change the definition of catastrophic risk to cover harms to people that arise across instances and harms to third parties in loss-of-control scenarios without necessary loss of life or property.*

Proposed Language:

> **22757.11. (c) (1)** "Catastrophic risk" means a foreseeable and material risk that a frontier developer's development, storage, use, or deployment of a foundation ***or frontier***[1] model will materially contribute to the death of, or serious injury to, more than 50 people or more than one billion dollars ($1,000,000,000) in damage to, or loss of, property arising from a single incident ***or incidents arising from a similar fact pattern*** involving a foundation ***or frontier*** model doing any of the following:

Recommendation: *Update 22757.11. (d) (1) to decouple model weight exfiltration, modification, and unauthorized access events from physical harm.*

Proposed Language:

> **22757.11. (d) (1)** Unauthorized access to, modification of, or exfiltration of, the model weights of a frontier model ~~that results in death or bodily injury~~.

The definition of large frontier developer should be broadened to include companies with other attributes likely to correlate with increased risk to the public.

Additional requirements in TFAIA placed on frontier developers were justified by the greater risks their models could pose to the public. However, an exclusive focus on the organization's income (annual gross revenue in excess of five hundred million dollars ($500,000,000)) does not capture the range of developers likely to create larger risks to the public. Revenue was viewed as a poor proxy for identifying companies on the frontier. Revenue is connected to some risks but not others. For example, for risks arising from internal deployment, revenue is a particularly poor proxy. A model could be widely deployed, with many users, but free. A company could be well-financed through investments but have little to no revenue.

Participants recommended creating a threshold or set of thresholds that would scope in small entities likely to be doing frontier development. One approach supported by

[1] See recommended change to definition of Frontier Model below.

several stakeholders was using research and development spending as an independent proxy for risk within the definition of large frontier developer. Another suggestion was adding a capability threshold for models as an independent trigger for categorization as a large model developer.

Recommendation: *Include additional metrics that indicate impact of developer.*

Proposed Language:

> (j) Large frontier developer means a frontier developer that together with its affiliates collectively had annual gross revenues in excess of five hundred million dollars ($500,000,000) in the preceding calendar year or ***who spent in excess of one hundred million dollars ($100,000,000) on research and development of frontier models***.

Administrability and verifiability should be two sides of a coin: regulated entities should be able to figure out when and how the law applies to them, and the public and regulators should be able to independently verify the decisions of regulated entities.

Administrability and verifiability travel together. Useful metrics triggering coverage should be easy for entities to assess and easy for the public and regulators to evaluate.

It's a challenge, of course, to create robust definitions that capture models and companies that simultaneously identify catastrophic risks, provide regulatory certainty for entities and verifiability for regulators. For example, a purely capabilities-focused definition might be robust. However, this requires anyone creating a model to be able to evaluate whether it achieves the relevant capabilities and is therefore considered frontier.

Participants noted that other states are going further than SB53. For example, New York's RAISE Act requires frontier developers to make supporting documentation available to the state to support their statements and decisions about compliance. Illinois's AI Safety Act, SB53, requires external auditors.

Recommendation: *Require developers to maintain and upon request provide to state agencies technical documentation developers view as relevant in the calculation of relevant thresholds under the law.*

Proposed Language: add new provision to 22757.12.

22757.12.

(g) Frontier developers and Large Frontier developers must keep technical and other documentation used to establish whether they meet thresholds established in this Act, and produce them at the request of relevant state agencies.

Part 2: Reflections on Scope and Requirements

Open-source models put pressure on the current use of compute and revenue as thresholds. Deployment of open source models leads to the question of how to preserve transparency and information as organizations are deploying systems developed by others.

Open-source AI models pose a challenge to TFAIA. The most capable open-source models are only a matter of months behind the capability frontier of closed models, and they are rising in adoption. What to do about open-source models and the risks they pose is one of the most important AI governance challenges.

With respect to TFAIA, one participant worried that, under current law, if an open model developer refuses to comply, any downstream refiners are not held liable. Another participant agreed, noting that a Chinese open-source model could have current frontier capabilities in around eight months and that addressing these concerns required support for the regulation ecosystem and that the EU AI act was a particularly instructive lesson.

Clarify what is covered by TFAIA with respect to internal deployments.

Participants were concerned that the law does not attend to the risks posed by model development and testing. The OpenAI/Hugging Face incident highlighted internal testing failures, organizational immaturity, and the absence of consequences. Several opined that the cyberbreach of Hugging Face would not count as a critical safety incident requiring reporting because it arose during an evaluation rather than a deployment.

Participants noted that even broadening the scope to include evaluations would not fully address the law's limitations because a loss-of-control incident is only considered a critical safety incident when it "caus[es] death or bodily injury." Given the legislature's belief that "timely reporting of critical safety incidents to the government is essential to

ensure that public authorities are promptly informed of ongoing and emerging risks to public safety" because it "enables the government to monitor, assess, and respond effectively in the event that advanced capabilities emerge in frontier artificial intelligence models that may pose a threat to the public" it is striking that the material harm caused by OpenAI's model's advanced capabilities would not, but for Hugging Face's robust security practices and sleuthing, have come to the government's or public's attention. As one participant bluntly stated, a developer should have to report if their model, during internal testing and evaluation or deployment, hacks a third-party system.

More broadly, participants suggested that reporting requirements should reflect the marginal utility of the information. The baseline information on risks discovered during internal deployments is very low relative to deployed models. The hockey stick curve on reported cybersecurity vulnerabilities has increased in recent weeks, with about 5X as many reported for any month in history.

Participants were concerned that OpenAI and other companies that inadequately contained and supervised models during internal evaluations would suffer no consequences under the State's flagship AI safety law.

One participant noted that companies have greater visibility into model behavior during pre-deployment evaluations and testing than during deployment where contractually their visibility may be limited (e.g., enterprise contracts with API access may exclude the model developer from logging use of their model). For this reason, incidents–or even near-incidents–arising during pre-deployment testing could be a value set of information about risks.

Recommendations: *Update incident reporting requirements to include incidents that arise during internal testing and evaluation. Decouple the loss-of-control category from physical harm. Narrow the evaluation carve-out.*

Proposed Language:

To eliminate harm and make containment failure the reportable event replace §22757.11(d)(3) with:

> ***(3) Loss of control of a frontier model, including any instance in which a frontier model, without authorization, (A) obtains access to computing resources, networks, accounts, credentials, or data outside the environment designated for its operation; (B) causes unauthorized access to, or material***

impairment of, an information system of another person; or (C) copies, transfers, or causes the persistence of its own weights or operating scaffolding outside that environment.

To narrow the internal evaluation carve-out in §22757.11(d)(4), replace "outside of the context of an evaluation designed to elicit this behavior" with:

> (4) A foundation model that uses deceptive techniques against the frontier developer to subvert the controls or monitoring of its frontier developer ***unless the behavior was both reasonably foreseeable by the design of the evaluation and confined to the evaluation environment*** ~~outside of the context of an evaluation designed to elicit this behavior and in a manner that demonstrates materially increased catastrophic risk~~.

Insert a safe harbor provision to avoid incentive misalignment.

Proposed Language:

> ***A report submitted under subsection §22757.11(d)(3) discovered during a good-faith evaluation shall not itself constitute evidence of a violation of this chapter or of a frontier developer's frontier AI framework, and shall not be admissible for that purpose.***

Finally, workshop participants only briefly discussed the technical possibilities and risks of recursive self-improvement (RSI), where a current AI model is used to train a future AI model and human oversight of model design and monitoring is thereby diminished. This technique poses a distinctive risk in that a frontier developer's own measurement and evaluation systems are outrun by automation. RSI is generally invisible to outsiders and not discussed in system cards. The dangerous dynamic is cross-industry. A frontier developer–wide compression of model development and evaluation can be driven by a rational response to the perception of RSI being deployed at another developer. To gain insight about whether such techniques are at use, we recommend transparency provisions, such as the following, for future reporting.

Proposed Language:

> ***Any frontier model developer shall report four quantifiable metrics every six months:***

1. ***Share of internal compute allocated to automated AI R&D.***
2. ***Number of training runs above a specified compute level initiated from model-generated configurations, and the number subject to human review.***
3. ***For each frontier model generation: interval since the prior generation, and duration of pre-deployment evaluation. (This yields a directly observable "evaluation time per capability increment" time series.)***
4. ***Whether any safety-relevant code, evaluation, or mitigation running in production was primarily model-generated, and what human review applied.***

Talent and human capital are essential to successful implementation of TFAIA.

The “trust but verify” approach of TFAIA has major implications for human capital. First, TFAIA assumes that covered entities will devote significant talent to the project of testing and evaluating for catastrophic risk. But participants noted that covered entities have few market or policy incentives to hire and retain staff to holistically address risks posed by models. Trust and safety teams within industry have been downsized and tend, on balance, to lack decision-making power in product development.

Second, participants noted that the government also needs appropriate experts to inform SB 53 implementation and revisions. They opined that creating pathways into public service for technically trained experts and encouraging students to consider public service careers are important components of addressing this need. One participant suggested a stimulation of the supply of talent. For example, Gov. Newsom could appear at a university gathering and give a JFK-style speech on serving the state—attaching a Teach For America-style civic glow to public interest technologists. Another noted the importance of demand and structural reform. For example, efforts were made during the Biden Administration to bring in tech and data experts to support AI deployment and AI governance. These efforts were chronicled in the report of the AI & Tech Talent Task Force, *Increasing AI Capacity Across the Federal Government: AI Talent Surge Progress and Recommendations*. The Biden administration engaged in other efforts such as public convenings, to highlight efforts across the public interest technology ecosystem and provided funding to support them.

Recommendation: *The Administration and Legislature should work together to develop pathways to bring more technical and sociotechnical experts into government.*

From the market side, insurance companies and customers will also push the companies to be more responsible.

The accountability ecosystem for frontier AI goes beyond model developers and government. It includes third-party testing and evaluation organizations. As one participant noted, it also includes insurance companies, which will be among the first to attach a price to AI risks. The National Venture Capital Association already has terms for AI-related risk. A participant experienced in industry agreed, noting that customers are also driving the ecosystem forward – demanding AI management standards such as ISO-42001 standard. The resulting legal ecosystem, participants agreed, will emerge from this dynamic between government, industry, and consumers.

Appendix 1: Attendees

- **Hosts**
  - Deirdre Mulligan (UC Berkeley)
  - Rob Reich (Stanford)
- **Virtual Workshop Participants (8/12)**
  - Mina Narayanan (Georgetown Center for Security and Emerging Technology)
  - Daniel Kroth (UC Berkeley Risk & Security)
  - Camille Crittenden (CITRIS)
  - Abigail Jacobs (University of Michigan)
  - Deepika Raman (UC Berkeley CLTC)
  - Nada Madkour (UC Berkeley CLTC)
  - Ian Reynolds (Hugging Face)
  - Ian Klaus (Carnegie Endowment for International Peace)
  - Caroline Siegel Singh (FAS)
  - Joal Stein (Collective Intelligence Project)
  - Kevin Bankston (Center for Democracy & Technology)
  - Severiano Christian (Encode)
  - Vinhcent Le (TechEquity)
  - Tamara Kneese (Partnership on AI)
- **In-Person Workshop Participants (8/13)**
  - Nathaniel Persily (Stanford Law)
  - Elana Cryst (Stanford HAI)
  - Rishi Bommasani (Stanford HAI)
  - Kathy Baxter (Salesforce)
  - Ellie Sakhaee (Google)
  - Thomas Woodside (Secure AI)
  - Skye Golann (Secure AI)
  - Deb Raji (Stanford CS)
- **State Observers**
  - Randi Michel (CA Governor's Office)
  - Vera Zakem (CDT)

- Justin D. Norman (CDT)
- Raymond Lee (CDT)
- Victor Wang (CDT)
- Katie Xin (CDT)

❖ **Research & Notetaking Support**

- Isabelle Qian (Berkeley)
- Dominic Zappia (Stanford)

Appendix 2: Workshop Agenda

Wednesday, August 12 10:00 am-12:00 pm and Thursday, August 13 13:00-15:00 pm

10:00-10:10 am Welcome and motivation: Deirdre K. Mulligan and Rob Reich

10:10-11:10 am Reflections on definitions of “Frontier model” “Frontier developer” and “Large frontier developer” in light of technological developments, scientific literature, and comparison with developments at the international, national and state level with particular attention to relative

- validity and robustness of thresholds;
- administrability for regulated entities;
- verifiability by regulators and the public; and
- information disclosure or other requirements to address risks, and support compliance, verifiability and administrability

11:10-11:40 am Reflections on scope and requirements of the Act in light of emerging information (including technological developments, scientific literature, and policy developments in other jurisdictions) about the range and severity of various risks and the circumstances contributing to either, for example:

- risks arising from internal testing and evaluation
- empirical evidence of frequency and severity of various risks
- pre deployment v. post deployment risks
- risks from open v closed models
- elements of a “frontier AI framework”

11:40-12:00 pm Closing reflections on the Act's implications for the open-source community, academic researchers, the broader public and future policy particularly in relation to seizing the technology’s benefits and managing its potential for risks