INSTITUTE
FOR LAW & AI

Draft Report of the Joint California Policy Working Group on AI Frontier Models - Whistleblower Protections Comments

Charlie Bullock, Mackenzie Arnold | Institute for Law & AI | 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.

We applaud the Working Group's decision to include a section on whistleblower protections. Whistleblower protections are light-touch, innovation-friendly interventions that protect employees who act in good faith, enable effective law enforcement, and facilitate government access to vital information about risks. Below, we make a few recommendations for changes that would help the Report more accurately describe the current state of whistleblower protections and more effectively inform California policy going forward.

1. Whistleblowers should be protected for disclosing risks to public safety even if no company policy is violated

The Draft Report correctly identifies the importance of protecting whistleblowers who disclose risks to public safety that don't involve violations of existing law. However, the Draft Report seems to suggest that this protection should be limited to circumstances where risky conduct by a company "violate[s] company policies." This would be a highly unusual limitation, and we strongly advise against including language that could be interpreted to recommend it. A whistleblower law that only applied to disclosures relating to violations of company policies would perversely discourage companies from adopting strong internal policies (such as responsible scaling policies). This would blunt the effectiveness of whistleblower protections and perhaps lead to companies engaging in riskier conduct overall.

To avoid that undesirable result, existing whistleblower laws that protect disclosures regarding risks in the absence of direct law-breaking focus on the seriousness and likelihood of the risk rather than on whether a company policy has been violated. See, for example: 5 U.S.C. § 2302(b)(8) (whistleblower must "reasonably believe" that their disclosure is evidence of a "substantial and specific danger to public health or safety"); 49 U.S.C. § 20109 (whistleblower must "report[], in good faith, a hazardous safety or security condition"); 740 ILCS 174/15 (Illinois) (whistleblower must have a "good faith belief" that disclosure relates to activity that "poses a substantial and specific danger to employees, public health, or safety."). Many items of proposed AI whistleblower legislation in various states also recognize the importance of protecting this kind of reporting. See, for example: California SB 53 (2025–2026) (protecting disclosures by AI employees related to "critical risks"); Illinois HB 3506 (2025–2026) (similar); Colorado HB25-1212 (protecting disclosures by AI employees who have "reasonable cause to believe" the disclosure relates to activities that "pose a substantial risk to public safety or security, even if the developer is not out of compliance with any law").

We recommend that the report align its recommendation with these more common, existing whistleblower protections, by (a) either omitting the language regarding violations of internal company policy or qualifying it to clarify that the Report is not recommending that such violations be used as a requirement for whistleblower protections to apply; and (b) explicitly referencing common language used to describe the type of disclosures that are protected even in the absence of lawbreaking.

- **Suggested language:** "However, some actions that clearly pose serious risks to public safety may not violate any existing laws. Therefore, policymakers may consider protections that cover a broader range of activities, which may draw upon notions of 'good faith' reporting on risks found in other domains such as cybersecurity. One possible approach is to follow the example of the federal Whistleblower Protection Act and protect disclosures made by a person who 'reasonably believes' that the disclosure relates to a 'substantial and specific danger to public health or safety.'"

2. The report's overview of existing law should discuss California's existing protections

The report's overview of existing whistleblower protections makes no mention of California's whistleblower protection law, California Labor Code § 1102.5. That law protects both public and private employees in California from retaliation for reporting violations of any state, federal, or local law or regulation to a government agency or internally within a company. It also prohibits employers from adopting any internal policies to prevent employees from whistleblowing.

This is critical context for understanding the current state of California whistleblower protections and the gaps that remain. The fact that § 1102.5 already exists and applies to California employees of AI companies means that additional laws specifically protecting AI employees from retaliation for reporting law violations would likely be redundant unless they added something new—e.g., protection for good faith disclosures relating to "substantial and specific dangers to public health or safety."

This information could be inserted into the subsection on "applicability of existing whistleblower protections."

- **Suggested language:** "Under existing California law, both public and private sector employees in California are protected from retaliation for reporting violations of any state, federal, or local law or regulation to a government or law enforcement agency or internally within their company [*reference*]."

3. The report should highlight the importance of establishing a reporting process

Protecting good-faith whistleblowers from retaliation is only one lever to ensure that governments and the public are adequately informed of risks. Perhaps even more important is ensuring that the government of California appropriately handles that information once it is received. One promising way to facilitate the secure handling of sensitive disclosures is to create a designated government hotline or office for AI whistleblower disclosures.

This approach benefits all stakeholders:

- Companies know that any sensitive business information disclosed to the government will be handled securely and appropriately and that the risk of valuable trade secrets being leaked to competitors will be minimized;

- Whistleblowers receive greater assurance that the information they bring forward will actually be put to good use (justifying the reputational and personal risk they take on);and
- The government of California becomes more capable of acting on the information it receives, responding to risks in a timely manner, updating its decision-making in light of new evidence, sharing information with key partners, and enforcing the law.

The report already touches briefly on the desirability of "ensuring clarity on the process for whistleblowers to safely report information," but a more specific and detailed recommendation would make this section of the Report more actionable. Precisely because of our uncertainty about the risks posed by future AI systems, there is great option value in building the government's capacity to quickly, competently, and securely react to new information received through whistleblowing. By default, we might expect that no clear chain of command will exist for processing this new information, sharing it securely with key decision makers, or operationalizing it to improve decision making. This increases coordination costs and may ultimately result in critical information being underutilized or ignored.

- **Suggested language**: "Ensuring clarity on the process for whistleblowers to safely report information can jointly advance accountability and manage countervailing interests, such as the disclosure of trade secrets or the misuse of information to compromise safety and security. One promising way to facilitate secure disclosures is to establish a secure government-run hotline or office for receiving AI whistleblower disclosures and to establish procedures for receiving, processing, sharing, and acting upon disclosures. Establishing such procedures may also increase government agencies' ability to quickly and competently process important information and respond to emerging issues."