



The AI
Whistleblower
Initiative



CARMA
CENTER FOR AI RISK
MANAGEMENT & ALIGNMENT

The AI Whistleblower Protection Act

Suggestions & Policy Analysis

Mark Gitau, Raqda Sayidali, Karl Koch, Abra Ganz

July 2026

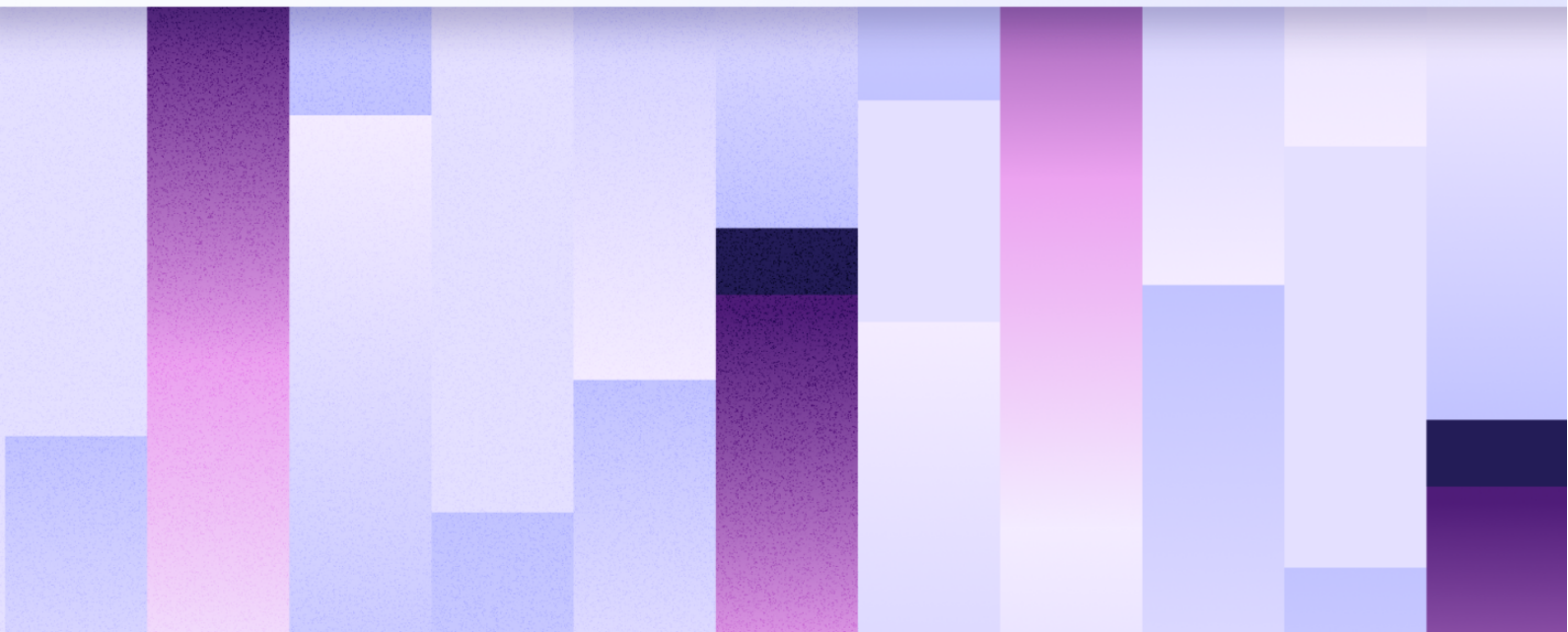


Table of Contents

Executive Summary - Assessment at a Glance	3
Introduction	5
Detailed Suggested Amendments	7
1 Material Scope	7
2 Personal Scope	9
3 Penalties	11
4 Remedies and Retaliation Protection	12
5 Reporting Channels	16
About the AI Whistleblower Initiative (AIWI)	17
About the Center for AI Risk Management & Alignment (CARMA)	17
About the Authors	18

Executive Summary - Assessment at a Glance

The AI Whistleblower Protection Act already meets most of what strong [AI whistleblower legislation requires](#). Five categories of amendments, detailed in full beginning on page 8, would close the gap between the Bill's current text and best practice. We give an overview in the table below.

These amendments draw on comparative research conducted by the Center for AI Risk Management and Alignment (CARMA) and the AI Whistleblower Initiative (AIWI), examining how whistleblowing statutes in other jurisdictions have addressed the distinctive risks and structure of the AI industry.

Category	Topic	AWPA Strengths Today	Proposed Improvement
Material Scope ("what can be reported")	Reporting on dangers to public health & safety	Protects disclosures of "a substantial and specific danger to public safety, health, or national security."	Remove the added "failure to appropriately respond" language , which creates an undefined, potentially double evidentiary burden and hands employers a built-in defence.
	Violations of the Law	Protects disclosures of any violation of Federal law, including rules and regulations, related to or committed during the development, deployment, or use of artificial intelligence.	Cover state law violations relating to AI. State AI regulation in California, New York, Colorado, and Illinois is outpacing federal law and this risks an emerging imbalance between law and what can be reported.
	Assistance and refusal protections	Protects a covered individual for testifying or assisting in any investigation or judicial or administrative action connected to a disclosure.	Add protection against retaliation for lawfully assisting another covered person's disclosure or complaint, and for refusing to obey a directive reasonably believed to require violating federal or AI-related state law.
Personal Scope ("who can report?")	'Covered individual' definition in Section 2(6)	Protects current and former employees and independent contractors.	Change 'covered individual' to 'covered person' and extend the scope to include all persons who have been granted access to a system under a written agreement.
Penalties	Civil penalties for violations	Imposes sanctions on companies indirectly, through whistleblower remedies paid out of company funds but stops short of penalising the company for the violation itself.	Add a civil monetary penalty: up to \$10,000 per violation for individuals, up to 1% of worldwide revenue per violation for organisations.

	Criminal penalties for retaliation		Apply criminal penalties when AWP is violated , building on state law precedent. ¹
Remedies	"Double backpay" remedy standard	Provides double back pay otherwise owed to the covered individual, with interest, alongside reinstatement and compensatory damages.	Explicitly include equity and equity-like protections. Case law on unvested equity is unsettled — creating uncertainty especially for senior employees.
	Causation standard for retaliation claims	Includes a best practice standard burden of proof shifting framework, requiring employees to only show that whistleblowing was a "contributing factor" to the retaliation.	Include 90-day presumption of retaliation: If an employer takes any prohibited action within 90 days of protected activity presumed (subject to rebuttable) that the action was taken because of the protected activity.
	Anti-retaliation language in Section 3(a)	Prohibits an employer from discharging, demoting, suspending, threatening, blacklisting, harassing, or in any other manner discriminating against a covered individual because of protected disclosures.	Extend protection to immigration or visa status. The current protections leave it untouched, keeping a powerful deterrent in place for visa-dependent employees.
	Confidentiality and non-disclosure agreement enforceability	Prevents an employer from enforcing an NDA against a covered individual once they've made a protected disclosure — rights under the Bill can't be waived by contract.	Prohibit employers from impeding, restricting, or discouraging protected activity, modelled on SEC Rule 21F-17, including through confidentiality or non-disparagement agreements.
	Remedies for retaliation (ex-post relief)	Provides relief once a covered individual prevails in a complaint or court action.	Add injunctive relief upon finding reasonable cause of retaliation.
Reporting Channels	Reporting channels	Protects disclosures made through external internal channels.	Add protection for public disclosure where internal or external reporting has failed, is futile, or the danger is imminent and serious.

¹ OK Stat. tit. 40, § 199 and CA Labor Code sec. 1103

Introduction

The AI Whistleblower Protection Act (introduced in the Senate as [S.1792](#) by Senator Chuck Grassley with bipartisan co-sponsorship and [in the House as H.R.3460](#)) is the first piece of US federal legislation written specifically to protect the people inside AI companies raising concerns. For the first time, frontier AI insiders would have a federal, purpose-built protection rather than a patchwork of statutes across states and designed for other industries. We believe this bill would improve the position of company insiders and want to see it enacted — ideally in a strengthened form that closes gaps to existing [best practice](#).

The AI Whistleblower Protection Act (AWPA) already performs well when measured against best practice. For example, employees are held to a reasonable-belief standard, internal reporting to supervisors with authority to act is protected alongside reporting to Congress or a regulator, and pre-dispute arbitration clauses can no longer be used to keep disclosures from surfacing. The table below sets out a complete comparison against best practice.

Best Practice	What the AWPA Does
A wide category of protected persons	Covers former employees and independent contractors; stops there
Protect disclosure of substantial danger to public health, safety, or national security	Explicitly covers “substantial and specific danger to public health and safety or national security”, although limited by the language of “appropriate response”
Reasonable belief standard	Holds employees to a “reasonable belief” standard
Internal reporting	Internal disclosure is well supported. One can report to a supervisor or another employee the whistleblower believes has “the authority to investigate, discover, or terminate the misconduct”
External reporting channels	Covers a wide range of channels such as the Attorney General, a regulatory or law enforcement agency and any committee or Member of Congress.
Assisting investigations	Protected conduct includes “initiating, testifying in, or assisting in any investigation or judicial or administrative action”, but stops short of protecting refusal to do unlawful acts
Burden of proof	Contains best practice: the complainant must show by a preponderance of evidence that the protected activity was a contributing factor to the adverse action.
Compensatory relief	Provides for reinstatement, double backpay with interest, compensatory relief including litigation costs
Statute of limitations	6 years from the retaliation, or 3 years from discovery of facts material to the retaliation claim, whichever is later capped at 10 years absolute

There are still some ways in which AWPAs falls short of best practice, potentially leaving some of the best-positioned safety witnesses without whistleblower protections. This document presents a set of recommended amendments which would help it achieve what it sets out to do: ensuring the visibility of all risks stemming from the development and deployment of AI.

This analysis is based on the [text of the bill](#) and on publicly available resources and information and represents our independent assessment.

This publication can be read alongside our companion piece, AI Whistleblowing Law: Best Practice (find the publication [here](#)), which draws on the US Office of the Whistleblower Ombuds' standards and Transparency International's International Principles for Whistleblower Legislation as well as our own additions. It sets out what is required for strong whistleblower law in general and the specifics required to create strong whistleblowing law for the AI industry.

Further reading:

- [The AI Whistleblower Protection Act](#)
- [Press Release from Sen. Grassley's Office Introducing the AI Whistleblower Protection Act Bill](#) (May 5, 2025)
- [National Whistleblower Center's Statement on The Bill](#) (May 15, 2025)
- [Coalition of Support for the Bill](#) (June 10, 2025)
- [Whistleblower Coalition Letter Supporting the AWPAs, Opposing Federal Preemption](#) (Dec 2, 2025)

Readers who wish to support the bill can contact their senators directly through the [NWC's campaign page](#).

Detailed Suggested Amendments

1 Material Scope

1.1 Remove the “failure to appropriately respond” language in Section 2(2)(B)

The concern

The language of “failure to appropriately respond” is unprecedented in (federal) whistleblowing law. We believe it will take years of case law to resolve, and it will probably weaken early claims at the moment they most need to succeed.

Suggested remedy

We recommend simplifying Section 2(2)(B) by removing the “failure to appropriately respond” wording. This emulates the Whistleblower Protection Act formulation and protects disclosures of “a substantial and specific danger to public safety, public health, or national security” arising from the development, deployment, or use of artificial intelligence, regardless of whether a company “failed to appropriately respond” to such a danger.

Reasoning

On precedent. The phrase “substantial and specific danger” mirrors the WPA (5 U.S.C. § 2302(b)(8)(A)(ii)) and has three decades of interpretive case law behind it. Adding the requirement that the employer has “failed to appropriately respond” introduces an element that the WPA does not contain. This language creates three problems. First, the whistleblower will now need to establish both the existence of the danger and the inadequacy of the employer’s response, a double evidentiary burden that no analogous statute imposes. Second, it is unclear what the threshold for an “appropriate response” is. The term is not defined in the Bill and has no settled meaning in federal whistleblower law. In fact, it will require years of case law to resolve, during which period covered individuals and their counsel will face acute uncertainty about whether their disclosures are protected. Third, an employer facing a retaliation claim can point to any response, however cosmetic or insufficient, as evidence that the statutory condition has not been met, in effect arguing that it “responded appropriately” and that the whistleblower therefore lacks protection. The Bill, in other words, gives employers a built-in defence that turns on the adequacy of the employer’s own conduct, precisely the thing the whistleblower is challenging. This addition may have been a benign drafting choice. However, given that: (i) it introduces a novel and burdensome element without clear precedent; (ii) the previous standard was demanding enough on its own; and (iii) adding a further requirement about the employer’s response weakens the protection without any clear benefit, we strongly recommend removing it.

1.2 Include state law violations within the protected class of disclosures

The concern

The definition of “AI violation” in Section 2(2)(A) is limited to federal law. This is a drafting choice that is likely to produce sharp anomalies in the coming years. State AI regulation is developing rapidly.

California (SB 53, SB 942, AB 2013), New York (the RAISE Act and associated measures), Colorado (SB 205), and Illinois all have statutes on AI disclosure, safety, or discrimination either enacted or under active consideration. Federal law has not matched the pace of state legislation, and the gap may widen before it closes. If AWPAs protect only disclosures of federal law violations, then a researcher who reports that their employer is violating California’s safety framework requirements, or New York’s impact assessment rules, is unprotected by federal whistleblower statute.

Suggested remedy

We recommend amending Section 2(2)(A) to cover violations of any federal law or any state law to the extent that the state law relates to the development, deployment, or use of artificial intelligence. Reports to state regulators or state Attorneys General should be correspondingly added to the list of covered channels in Section 3(a)(1).

Reasoning

On precedent. Whistleblower statutes that cover state law violations have precedent. The Whistleblower Protection Act (WPA) for federal employees (5 U.S.C. § 2302(b)(8)(A)(i)), which protects disclosures about “a violation of any law, rule, or regulation,” draws no federal/state distinction.² We recognize that a federal statute protecting disclosures about state law violations adds doctrinal complications. However, without this fix (which is quite narrow, as it is confined to AI-related state law), the federal statute creates a perverse situation where the federal law protects less than some state laws already do, leading to fragmented and confusing protections for covered individuals.

1.3 Protect refusal to follow presumed unlawful orders and assistance to covered persons making disclosures

The concern

The bill currently does not explicitly protect individuals assisting any other covered person in making a disclosure. Likewise, it does not protect individuals refusing to obey any directive that a covered person may believe would be a violation of the Act. Both are covered protections for federal employees under 5 U.S. Code § 2302.

Suggested remedies

Following precedent from U.S. Code, protect individuals from retaliation for:

1. lawfully assisting any other covered person in making a disclosure described in paragraph (1) or in initiating, filing, or pursuing any complaint or action under this section; or
2. refusing to obey any directive that such a covered person reasonably believes would require the covered individual to violate any Federal law, rule, or regulation, or any State law, rule, or

² Congressional Research Service, “Whistleblower Protections Under Federal Law: A Legal Overview,” R48318, updated 30 December 2024, available at, <https://www.congress.gov/crs-product/R48318>.

regulation related to the development, deployment, or operation, or use of artificial intelligence.

2 Personal Scope

2.1 Extend coverage to employees and contractors of AI evaluation and auditing organisations

The concern

Under the current definition, some of the people best positioned to detect safety failures at frontier AI developers fall outside the protected class.

We mean here researchers employed by independent evaluation organisations. Researchers at these organisations are often commissioned, through a partnership (not a paid contract) that their employer has with a frontier AI lab, to run pre-deployment evaluations on a frontier model developed by a large lab. During testing, these researchers are well positioned to observe material violations that fall within the material scope of applicable whistleblowing legislation: evidence that the lab has selectively reported evaluation results, that an evaluation was structured in a way that predictably understated capability, evidence of misrepresenting results to regulators, or deployment of a system whose evaluations flagged substantial risks.

These researchers have a contractual relationship with their organisation, not with the frontier lab. The lab is the target of the concern, but the researcher has no employer/employee or employer/contractor relationship with the lab. On a plain reading of Section 2(6), any of these researchers is not a “covered individual” vis-à-vis the lab whose conduct they want to disclose. They are an employee of their organisation, which is itself a contractor of the lab. This section thus leaves outside its reach some of the most important safety-relevant personnel.

Suggested remedies

We recommend expanding the definition of “covered individual” to include, in addition to employees and independent contractors of the employer itself, one or more of the following phrasings:

- Following language precedent in the False Claims Act, move from “individual” to “person”, following the USC definition – including organizations.
- Expand the definition of “Covered Person” to include a person who has been granted access to a non-public AI system or related safety-relevant information by an employer under a written agreement.
- Permit a covered person described above to bring a civil action in the appropriate federal district court in lieu of filing a complaint with the Labor Secretary, if that pathway is not appropriate for a covered person.

A parallel change is needed in Section 3, so that “employer” in the retaliation prohibition includes any entity that can influence (whether paid or not), based on a written contract, a covered person’s work for the developer in question, not only the covered individual’s nominal employer of record.

Reasoning

On precedent. The National Defense Authorization Act for Fiscal Year 2013³ codified what is now 41 U.S.C. § 4712 (the Enhancement of Protections for Federal Contractor Employees from Reprisal). That provision protects employees of federal contractors, subcontractors, grantees, and subgrantees who report wrongdoing, even though those employees are not direct employees of the federal government.⁴ The logic of that statute applies directly here. When an outside party is engaged to inspect or evaluate work, for private gain or for the public interest, the people doing the inspection need protection that travels with their work rather than with their formal employment relationship.

Section 806 of the Sarbanes-Oxley Act (SOX) (18 U.S.C. § 1514A), which protects employees of publicly traded companies who report shareholder fraud, was amended through the Dodd-Frank Wall Street Reform and Consumer Protection Act (Pub. L. 111-203, § 929A, 2010) to cover employees of subsidiaries and affiliates of public companies, after courts had read the original language to protect only direct employees of the issuer. The Dodd-Frank amendment was itself a response to protests by the original authors of the SOX whistleblower provision, Senators Patrick Leahy and Chuck Grassley, who objected that narrow readings by the Department of Labor had resulted in the dismissal of over 1,000 of the first 1,600 SOX whistleblower cases filed, many on the grounds that the complainant worked for a private subsidiary.⁵ The Supreme Court in *Lawson v. FMR LLC*, confirmed that SOX Section 806 extends beyond employees of the publicly traded company itself to employees of its contractors and subcontractors.⁶ The structural concern in *Lawson* is directly analogous to the evaluator problem here.

Federal whistleblower law has in past instances recognized that multi-tier contractual relationships do not eliminate the public interest in disclosure, that narrow definitions of “employee” have historically failed to protect the people best positioned to uncover concerns, and that Congress has intervened each time courts or agencies read the definitions too restrictively.

The amendment simply applies that recognition to a specific feature of the AI industry that creates the gap.

³ Pub. L. 112-239, § 828.

⁴ 41 U.S.C. § 4712(a)(1), full text available at <https://www.law.cornell.edu/uscode/text/41/4712>. See also Congressional Research Service, “Whistleblower Protections Under Federal Law: A Legal Overview,” R48318, updated 30 December 2024, available at, <https://www.congress.gov/crs-product/R48318>. The statute was later amended by Pub. L. 114-261 (December 14, 2016) to extend coverage further to personal services contractors.

⁵ Chuck Grassley, “Leahy, Grassley Press for Update on Labor Department’s Handling of Whistleblower Cases” October 7 2010, available at, <https://www.grassley.senate.gov/news/news-releases/leahy-grassley-press-update-labor-departments-handling-whistleblower-cases>.

⁶ 571 U.S. 44 (2014), at pp. 45–62, available at, <https://supreme.justia.com/cases/federal/us/571/429/>.

3 Penalties

3.1 Include Civil Penalties for Violation of the Act

The concern

The AWPB currently includes no civil penalties for violations. This matters for two reasons. Firstly, establishing a deterrent to whistleblowing is generally valuable. Secondly, not all violations covered under the AWPB are violations of other laws, where a violation could introduce a penalty dimension.

Civil penalties from state laws are also wholly insufficient deterrents given the financial resources available to companies covered under the Act.

Suggested remedy

Companies should be fined for retaliatory behaviour against whistleblowers. For these fines to be meaningful in the AI industry where revenue is often greater than \$5bn annually, fines should be formulated as a proportion of revenue. We suggest adding the following:.

(1) CIVIL PENALTY.—In addition to any relief under subsection (b), the Secretary of Labor may, after notice and an opportunity for a hearing on the record, assess a civil monetary penalty against any person that violates subsection (a). The penalty shall not exceed—

(A) in the case of an individual, \$10,000 for each violation; and

(B) in the case of an organization, a penalty of not more than 1 percent of the total worldwide annual revenue of the organization for the preceding fiscal year per violation.

3.2 Include Criminal Penalties for Retaliation

The concern

Currently, the bill includes no sanctions for retaliation against whistleblowers. Sanctions are only included indirectly via whistleblower remedies, which are paid by companies. Where responsibility is assigned to companies, especially those with access to large funds, responsibility as well as deterrence is diluted.

To combat these issues, section 1107 of the Sarbanes-Oxley Act added [18 U.S.C. § 1513\(e\)](#), making it a federal crime to retaliate against anyone who provides a law enforcement officer with truthful information about any federal offense. Penalties include up to 10 years' imprisonment for individuals. §1513 does not apply however in the majority of whistleblowing cases, but only to the narrow case of reporting violations of the law to law enforcement. Reporting a violation of the law to Congress or an Attorney General, for example, would not be covered.

Suggested Remedy

We suggest that criminal penalties apply when AWPB is violated, building on state law precedent OK Stat. tit. 40, § 199 and CA Labor Code sec. 1103.

(2) CRIMINAL PENALTY.—

(A) OFFENSE.—An individual who knowingly and willfully, and with intent to retaliate, violates subsection (a) is guilty of a misdemeanor, punishable by imprisonment for not more than one year.

(B) AGGRAVATED OFFENSE.—If a violation of subparagraph (A) involves a threat or use of force or violence, or is committed against a person who disclosed information concerning a substantial and specific danger to public health, safety or national security, the offense is punishable by imprisonment for not more than 5 years.

(C) INDIVIDUAL LIABILITY.—Subparagraphs (A) and (B) apply to any individual officer, director, manager, employee, or agent who authorizes or knowingly participates in the prohibited conduct.

(D) RULE OF CONSTRUCTION.—The penalties under this subsection are in addition to, and do not preempt or diminish, any civil remedy or administrative penalty available under this Act or any other law.

Reasoning

Criminal exposure that attaches to individuals personally, not just the company treasury, changes the calculus in a way civil damages cannot. The expectation is that the criminal provision operates primarily through its deterrent shadow and does not usually result in prosecutions. Precedent for criminalisation of retaliation against whistleblowers in general exists at a state level. For example, OK. Stat. tit. 40, § 199 and CA Labor Code sec. 1103, on which the above draft text has been modelled.

4 Remedies and Retaliation Protection

4.1 Explicitly Include Equity Protections

The concern

Compensation at frontier AI firms such as OpenAI and Anthropic is heavily weighted toward equity or [equity-like instruments](#). At Anthropic, self-reported compensation data indicates a median total pay for engineers of approximately \$557k per annum, with 44 to 57% coming from equity. OpenAI has a mixed compensation structure, in part using a traditional equity mechanism, restricted stock units (RSUs), and in part offering employees [profit participation units](#) (PPUs) rather than traditional equity. As equity vests on a schedule and can only be sold at given times, frontier labs' equity structures could result in an employee having multiples of their salary locked in equity. Employees experiencing retaliation might stand to lose the vast majority of their savings – a situation we have seen play out in practice [before](#).

Existing whistleblower statutes do not explicitly account for this. As compared to SOX' broad 'make whole' standard, Dodd-Frank, and the proposed AI Whistleblower Protection Act provide double "backpay", reinstatement, and damages, but do not explicitly address equity instruments. The case law on whether unvested equity, for the purpose of 'double backpay', is compensable under these provisions is thin and contested. Courts have [recognised that equity-based compensation](#), including stock options, may be recoverable as damages in employment disputes. On its face today, the AWPA

hence likely allows a) reclaiming of lost vested equity and b) an inclusion of unvested equity in 'double backpay' or other damages. However, courts have also characterised such damages as inherently speculative and fact-dependent, and commentators frequently note a relative lack of settled, authoritative doctrine governing the treatment of equity-based compensation in employment retaliation and wrongful termination cases. Most recently, in [*Shah v. Skillz Inc.*](#), the court held that [stock options are not "wages"](#).

This lack of explicit clarity may deter especially the most senior AI company employees, who are both most able to spot violations and most exposed to the deterrent this risk presents.

Suggested Remedy

The remedies available should be explicit in including "(B) Two times the amount of back pay, including benefits such as, but not limited to, equity or equity-like compensation, otherwise owed to such covered person, with interest.

Reasoning

The reasoning for equity protections is threefold: current remedies are inadequate; high payouts function as deterrence; protection of equity reduces the chilling effect. First, a remedy covering 20–30% of the actual economic harm is not adequate. These provisions close the gap by ensuring that the remedy reflects how compensation actually works. The aim is to make the employee whole again. Second, a \$600K civil payout does not register as a meaningful cost for a company spending \$6 billion a year on stock compensation. Equity-inclusive remedies ensure the cost of retaliation is proportionate to the harm inflicted. Third, under current law, a would-be whistleblower must gamble on whether a court will treat their unvested equity as compensable, which remains unsettled. Codifying equity protections eliminates this threshold uncertainty.

Last, an "accelerated vesting" front-pay approach would, comparable to the SEC whistleblowing bounty programme, provide a strong incentive to report violations of the law.

4.2 Include 90-day presumption of retaliation

The concern

Establishing causation in retaliation claims can be challenging – even in cases where a protected activity is temporarily close to a retaliatory action. California has addressed this issue in SB 497, which posits that any adverse employment action (such as firing, demoting, or disciplining) taken within 90 days of a protected activity should be deemed unlawful retaliation by default. This also more closely mirrors existing best practice globally, which does not require employees to provide evidence of reporting being a contributing factor for a burden of proof shift at all.

Suggested remedy

Replicating the language of SB 497, we suggest adding the following: If an employer takes any action prohibited under subsection (a) within 90 days of a covered person's protected activity under subsection (a), there shall be a rebuttable presumption that such action was taken because of the protected activity.

4.3 Introduce protection specifically for non-US citizens on visas

The concern

Many workers at frontier AI companies who might report safety violations are on nonimmigrant visas. H-1B most commonly, but also O-1, L-1, and others. Their legal status in the United States depends on continued employer sponsorship. This creates a chilling effect: an employee who knows that disclosure could cost them their job knows simultaneously that it could cost them their right to remain in the country, upend their family, and potentially void years of progress toward permanent residency. S.1792's existing anti-retaliation language addresses employment consequences but does not touch immigration status, leaving a powerful deterrent intact.

At the same time, the individuals working for these companies are valuable talent for the U.S. Labor Market and should be retained.

Suggested remedy

Add language like the following, mirroring the proposed POWER Act.

"Notwithstanding any other provision of law, withdrawal or non-renewal of an employer's petition for a covered individual's nonimmigrant classification shall constitute retaliation under Section 3(a) if taken in response to protected activity under that section. Upon a credible showing (as determined by the Secretary of Homeland Security based on documentary evidence of the disclosure and subsequent employer action) that nonimmigrant status has lapsed or is imminently at risk due to such retaliation, the Secretary of Homeland Security shall grant the covered individual employment authorization and a stay of removal for up to 2 years, renewable while any complaint under Section 3(b) remains pending."

In addition, the injunctive relief (see below) adjustment should also cover visa-related retaliation, in combination with expanding the list of prohibited conduct by

"No employer may, directly or indirectly, discharge, demote, suspend, threaten, blacklist, harass, or in any other manner discriminate against a covered individual, **including through any action affecting the covered individual's immigration status**, in the terms and conditions of employment..."

Reasoning

Protection that arrives after a DOL investigation concludes does not work when it comes to visa retaliation. Tying immigration protection to the act of filing a complaint, rather than winning it, is what neutralizes the deterrent. This structure has direct precedent in the POWER Act, which used identical logic (temporary status upon filing) and the same "notwithstanding any other provision of law" construction to bind DHS.

4.4 Prohibit acts that impede whistleblowing

The concern

The bill provides that a covered individual's rights cannot be waived and, as a result, an employer cannot enforce a confidentiality or non-disclosure agreement against a whistleblower who makes a

protected disclosure. However, this does not address the chilling effect such agreements can have before any disclosure is made. A broad confidentiality, non-disparagement, or non-disclosure provision may deter an employee from reporting safety concerns even where the provision would ultimately be unenforceable. An employee who is unsure of their rights, concerned about litigation, or unwilling to risk a dispute with their employer may choose not to report at all.⁷

Suggested remedy

Add a provision modelled on SEC Rule 21F-17, making it unlawful to impede, interfere with, or discourage a protected disclosure, including through the use of contractual restrictions.

"No employer may directly or indirectly impede, interfere with, restrict, or discourage a covered individual from engaging in protected activity under this Act, including by imposing, maintaining, enforcing, or threatening to enforce any confidentiality, non-disclosure, non-disparagement, or similar agreement that could reasonably be understood to prohibit or restrict such activity."

Reasoning

The current provision operates after a whistleblower has already come forward. It protects individuals who know their rights and are willing to assume the risks associated with disclosure. The larger group of employees remains silent because they believe, correctly or incorrectly, that a contractual restriction prevents them from reporting.

Securities law confronted this problem and responded by targeting the deterrent itself. SEC Rule 21F-17 makes it unlawful to impede communications with the Commission, regardless of whether a particular employee was actually prevented from reporting. The SEC's enforcement practice reflects a broad and robust approach to interpreting the rule. The SEC has repeatedly used Rule 21F-17 to check contractual provisions that could discourage whistleblowing. Adopting a similar approach here would ensure that the legislation addresses not only retaliation after disclosure, but also the mechanisms that prevent disclosures from occurring in the first place.

4.5 Introduce Injunctive Relief

The concern

Injunctive relief protects whistleblowers by allowing them (in cases where there is reasonable cause to believe that they have been retaliated against) to maintain their position while litigation is on-going rather than only receive ex-post relief.

Suggested remedy

We suggest adopting the text from CA Labour Code 98.7 which allows the Labor Commissioner to petition the superior court for temporary or preliminary injunctive relief, upon finding reasonable cause to believe that a person is engaging in a violation.

⁷U.S. Government Accountability Office, "Protections for Whistleblowers and Others: Selected Agency Actions Regarding Reports of Potential Wrongdoing," GAO-26-107650, March 2026, pp. 1-2, 11, and 16, available at, <https://www.gao.gov/assets/gao-26-107650.pdf>.

Reasoning

Injunctive relief protects whistleblowers by allowing them (in cases where there is reasonable cause to believe that they have been retaliated against) to maintain their position while litigation is on-going rather than only receive ex-post relief. State precedent exists in California and federal precedent for injunctive relief for employees who raise health and safety concerns is present in the Mine Act.

5 Reporting Channels

5.1 Include Conditions for public disclosure

The concern

The Bill is entirely silent on the circumstances, if any, under which a covered individual may go public with a concern, whether to the press, to the research community, or to the general public, and retain protection. While it is unprecedented in the US to allow employees to directly make a disclosure to the public, other jurisdictions protect disclosures of risk to public health and safety in cases where the risk is imminent and material. Those developing AI state that it is an unprecedented and high-risk technology — the law should reflect this. Public disclosure is an alternative to increasing the regulation on companies: instead of introducing burdensome risk management legislation, one can allow for a more nimble approach that allows companies to innovate, but then immediately respond and alert the public if risks arise.

Suggested remedy

We recommend including a provision that protects public disclosure where: (a) the covered individual has first reported through internal or external channels and no adequate action has been taken within a reasonable period; or (b) the covered individual reasonably believes that the matter involves an imminent and serious danger to public safety or health; or (c) the covered individual reasonably believes that reporting through internal or external channels would be futile or would result in retaliation or concealment of evidence. Disclosures involving classified information should not qualify for public disclosure protection.

Reasoning

The absence of any protection for public disclosure creates an all-or-nothing dynamic where the whistleblower either stays within the official channels (even when they are demonstrably ineffective) or goes public and loses all protection. The conditions we propose are designed to discourage premature or irresponsible disclosures: the default remains internal and external reporting, and public disclosure is protected only when those channels have failed or pose demonstrable risks.

About the AI Whistleblower Initiative (AIWI)

[The AI Whistleblower Initiative \(AIWI\)](#) is an independent, non-profit dedicated to strengthening the position of frontier AI insiders in raising concerns and seeing them addressed effectively — through research, policy advocacy, and access to vetted legal and financial resources. Our research strengthens internal reporting channels, supports relevant legislation, and promotes the conditions under which concerns can reach individuals best placed to act on them, with source protection at its core.

In practice, we are working on several fronts at once. We strengthen regulator reporting channels through our policy thought leadership and work, e.g. in [California](#) or [the EU](#). We aim to educate, and strengthen company reporting channels, through our [Publish Your Policies Campaign](#) and by providing input to frontier AI companies.

We help insiders access specialised, often pro bono, legal and financial support through the [AIWI Contact Hub](#) and [Defense Grants For AI Whistleblowers](#). We also offer resources on [Digital Privacy](#) and, for insiders that wish to ask questions surrounding a concern and without revealing confidential information, [Third Opinion](#), which allows anonymous consultation with jointly selected independent (technical) experts via a secure platform.

Lastly, we conduct research on barriers to AI whistleblowing, study impacts of disclosures on global safety levels, as well as impacts of disclosures.

AIWI remains responsible for any errors and welcomes constructive discussion. Contact us through our Proton Mail at hello@aiwi.org ([PGP keys](#)) or visit aiwi.org/contact for guidance on safe and secure communication methods.

About the Center for AI Risk Management & Alignment (CARMA)

The Center for AI Risk Management & Alignment (CARMA) is a research and policy think tank dedicated to more safely managing the progression and effects of rapid advances in artificial intelligence. Through rigorous analysis and strategic intervention, they work to help ensure that transformative AI technologies remain controllable, aligned with human values, trustworthy, and beneficial to society.

CARMA brings together experts in artificial intelligence, broader computer science, policy, infrastructure resilience, complex systems, mechanism design, and international technology governance to address both acute and systemic risks from increasingly powerful AI systems.

About the Authors

Mark Gitau

Mark completed this work as a Legal Research Fellow at the Centre for AI Risk Management and Alignment (CARMA), where he worked with Abra Ganz on whistleblower protections for AI safety professionals. He was previously a Research Scholar at ILINA, where he focused on the role of law and policy in strengthening model evaluations, a Researcher at the University of Cape Town African Hub on AI Safety, Peace and Security, where he conducted Africa-centric model safety evaluations, and a Fall Research Fellow at the Vista Institute for AI Policy, where he examined how evaluation evidence can establish developer fault in AI agent-caused harm.

Mark holds an undergraduate law degree from Strathmore University, where he graduated top of his year with first-class honours, and will be joining Harvard Law School in the fall for his Master of Laws degree.

Raqda Sayidali

Raqda Sayidali is a Research Scholar at the ILINA Program and a Legal Research Fellow at the Center for AI Risk Management and Alignment (CARMA). Her research focuses on legal accountability mechanisms for frontier AI. At ILINA, she works on AI liability regimes and how tort doctrine can respond to AI-caused harms, and at CARMA, on mapping AI whistleblowing channels and protections across the US, UK, and EU.

Her published research includes an analysis of *Megan Garcia v Character Technologies* for the Robotics and AI Law Society, examining how US tort doctrine might adapt to respond to AI-caused harms, and a paper with the Centre for International Governance Innovation on why Global South countries should be concerned about frontier AI.

Karl Koch

Karl Koch is the Founder and Managing Director of AIWI, leading the organisation across its research & guidance, education, and policy work.

He has co-authored [commentary on whistleblower protections in California's SB 53](#), a [resource on whistleblowing under the EU AI Act](#) with Santeri Koivula of the Future of Life Institute, and contributed subject-matter expertise to the EU AI Office during the communication and policy drafting phases of what became the [world's first AI-specialised whistleblowing channel](#). Before AIWI, he worked as a management consultant and founded a SaaS business, which was acquired in 2023.

He has appeared on [The Cognitive Revolution](#) with Nathan Labenz and the [Future of Life Institute Podcast](#), discussing legal gaps in insider protections, how to evaluate disclosure decisions, and whistleblowing as a mechanism for surfacing AI risks and shifting incentives in frontier AI development.

Abra Ganz

Abra Ganz is a senior researcher at the Center for AI Risk Management and Alignment (CARMA). She has advised on the establishment of the [EU AI Office's whistleblowing channel](#) and her work has been featured in [Transformer](#), by [LawAI](#), presented at [Harvard](#) and IASEAI, among others. Abra is now also the head of AI Policy at Pour Demain and also holds an affiliation with the Oxford Martin AI Governance Initiative.

Before working on oversight of the AI industry, Abra's previous research spans various topics and institutions: Yale's Digital Ethics Center (on infrastructure of the internet), ETH Zürich (on adversarial robustness), and MIT (on inverse reinforcement learning). Abra holds an undergraduate degree in Classics from the University of Oxford and a Master's in Logic from the Institute of Logic, Language, and Computation at the University of Amsterdam.



The AI
Whistleblower
Initiative



CARMA

CENTER FOR AI RISK
MANAGEMENT & ALIGNMENT