The AI Whistleblower Initiative

CARMA
CENTER FOR AI RISK MANAGEMENT & ALIGNMENT

AI Whistleblowing Law: A Best Practice Guide

Abra Ganz, Karl Koch

Version 1, July 2026

Content

Introduction

Whistleblower protections are a critical mechanism for surfacing risks and have existed across many US states and the federal level for decades – albeit with strongly varying degrees of effectiveness.

Whistleblower protections enable those best able to spot risks to raise them when their company has not addressed them without having to fear, or at least reduce the impacts of company retaliation. They are extremely effective: in other industries, they surface 40% of fraud, as compared to only 16.5% surfaced by internal audits.[1] They are also particularly important in the AI industry where the risks are novel and legislation has not caught up. The AI industry also has a stark information asymmetry between regulatory oversight and industry insiders: AI develops rapidly, with technology and culture being both rarely understood by those outside of the companies who created them[2]. Thus employees are significantly better able to detect risks than hypothetical regulators.

Yet existing whistleblower statutes were designed for industries where risks are familiar, partially mismatching the idiosyncrasies of frontier AI development. For example, whistleblower protections often rely on disclosures of violations of the law — which it is not possible to make when no law exists to violate, with only a few states including the better "substantial and specific danger to public health and safety" standard. Equity-heavy compensation structures are standard across frontier AI labs, but not explicitly protected in legislation (although precedent covers equity make-whole provisions in some cases – consultations with specialized attorneys are critical for those affected[3]). Further, some of the people best positioned to surface safety concerns often fall outside the scope of existing protections by virtue of acting in a voluntary capacity. Safety researchers and red-teamers surface risks as a routine part of their work, but duty speech (speech which is part of your job) is not consistently protected across states, leaving employers free to argue that these disclosures are not protected acts. Non-US-citizen employees, who make up a large proportion of the AI-industry workforce, face risks of visa-sponsorship retaliation when they make a protected disclosure, with statutory injunctive relief inconsistently available to prevent harm from materializing even where retaliation is found to be unlawful.

Legislation can close these gaps. This guide sets out what strong AI whistleblower law looks like, covering seven areas: the scope of protected disclosures, who is protected, what conduct is prohibited, available disclosure channels, remedies, employee notice requirements, and agency requirements.

These best practices draw on the US Office of the Whistleblower Ombuds' *Best Practice Whistleblower Law Standards* and Transparency International's guide, *International Principles for Whistleblower Legislation*, as well as our own analysis of previous AI whistleblowing legislation such as SB-53 and AWPA. We note that while whistleblower protections are extremely important in the AI context, it is equally important that there exists an agency with the power to act on them.

The recommendations below can be applied either to a specific AI bill with whistleblower protections or as amendments to existing general-purpose whistleblowing legislation. AI-specific recommendations are preceded by [AI], while other recommendations are general whistleblowing law best practice.

---

[1] Katyal, S. (2018). Private accountability in the age of artificial intelligence. *UCLA Law Review, 66*, 54–141.
[2] Green-Lowe, J., Fehrenbach, F., & Reddish, M. (2025). Silicon sentinels: Using whistleblower protections to manage information asymmetry and AI risk. Liberty University Law Review, 19(4), Article 5.
[3] Visit the AIWI Contact Hub for a list of pro-bono whistleblower support counsel handling AI cases

Recommendations

1. Scope of Protected Disclosures

[AI] Include Dangers to Public Safety, Health, National Security independent of pathway. AI whistleblower laws should specifically protect disclosures related to catastrophic risks. Ideally, this would be by including disclosures of 'dangers to public health and safety' as in New York's state whistleblower protection laws. A lower bar would be to use language similar to SB-53 where a catastrophic risk is defined as a foreseeable material risk that a model could materially contribute to the death of or serious injury to more than 50 people or more than one billion dollars in damages. While California's SB-53 includes specific risk pathways in their definition of Catastrophic/Critical Risks, and hence restricts material scope on this basis. Best practice would be to leave this open, allowing for all possible risks.

Include Violations of State and Federal Law. Whistleblower laws should protect reporting of reasonably believed (see below) violations of state and federal law. In a specific AI whistleblower law, it is best practice to re-state that whistleblowers are protected for reporting a violation of the law. While this protection may already exist in law for employees under pre-existing state whistleblowing law, if the scope of protected individuals in the new law is wider than in the basic whistleblower law (for example, by including evals orgs), then stating that violations of the law are included ensures that more people can report them.

Reasonable Belief Standard. Protection should be granted for disclosures that the whistleblower reasonably believes to be true. That is, where a reasonable individual with a similar level of expertise in equivalent circumstances would believe the same.
Protect disclosures regardless of motive, prior disclosure, and form. Protection should apply regardless of the whistleblower's motive; whether the information was previously disclosed; the amount of time that has passed since the alleged wrongdoing; and whether the disclosure was made orally or in writing. Employers frequently invoke prior disclosure and mixed-motive arguments to defeat otherwise valid claims; the statute should explicitly foreclose these arguments.

Protect "duty speech" disclosures. Disclosures made as part of an employee's normal work duties — so-called "duty speech" — should be protected without heightened scrutiny. In AI labs, safety researchers, red-teamers, and evaluators surface concerns as routine parts of their jobs. Without explicit duty-speech protection, employers may argue those individuals were "merely doing their job" and deny them whistleblower protection.

Protect Assisting Investigations and Refusing to Break the Law. Whistleblowing laws usually protect testifying, assisting in lawful investigations, and exercising appeal/complaint/grievance rights. In the AI context, this is especially important to protect evals orgs which might be called upon to assist an investigation. Best practice is also to protect refusal to follow an order that would require violating law, rule, or regulation.

2. Protected Individuals

Protections must extend to all workers involved with AI development. Basic protections include all employees. Best practice would extend whistleblower protections to:

- **[AI]** Third parties that provide services for assessment, management, or addressing critical risk, or employees of such organisations
- Board members, former employees, applicants, consultants, and contractors
- Those who assist or attempt to assist a whistleblower
- Those who are perceived to be, or to become, whistleblowers
- Close relatives and co-workers, where adverse action against them would reasonably dissuade the covered individual from engaging in protected activity
- Use 'persons' instead of 'individuals' (following USC definition and as precedented in False Claims Act)

3. Protected Behaviour

Prohibited Conduct. AI Companies should not retaliate against covered persons for disclosing protected information (as defined in section 1). Retaliation includes all forms of disadvantage or discrimination at the workplace, including dismissal, probation, job sanctions, punitive transfers, harassment, reduced duties or hours, withholding of promotions or training, loss of status and benefits, and threats of such actions. AI companies should neither adopt any policies violating covered persons rights to make protected disclosures.

Make Gag Orders Invalid. Whistleblower rights must override confidentiality/nondisclosure agreements. Companies cannot enter into contracts that prevent covered persons from making disclosures protected under existing whistleblower statutes. Companies must be prohibited from requiring a covered person to engage in arbitration, mediation, or any other alternative dispute resolution process before seeking relief under the law.

Burden of Proof. Best practice is a two-step standard. First, the whistleblower must demonstrate by a preponderance of the evidence that their whistleblowing was a contributing factor in the retaliation they faced. Second, the employer has the opportunity to rebut by 'clear and convincing evidence' that it would have taken the same action for independent, legitimate reasons regardless of the whistleblowing.

90-day presumption of retaliation: Following SB497, 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.

Civil Penalties. 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. Ideally, civil penalties are sent to the harmed individual (as in California), not the state.

Criminal Penalties and Individual Liability for Retaliation. Civil remedies alone are unlikely to deter large AI Companies, for whom such payouts can be budgeted as a cost of doing business. Best practice is to provide for criminal penalties for retaliation against whistleblowers (as in 18 U.S.C. § 1513(e)) which apply to the individual officers, directors, and managers who retaliated.

Additional provisions:

1. Extend immunity from criminal or civil liability with respect to protected disclosures.
2. Include a minimum 3-year statute of limitations on claims of whistleblower retaliation.
3. State that these protections are in addition to existing legal protections, not a replacement.

4. Disclosure Channels

Internal Channels

Protect internal reporting. Employees must be protected from reporting to persons with authority over the employee and/or to persons whom a whistleblower believes in good faith to be able to rectify an issue (i.e. employees should be able to express concern to their manager(s) without fear of reprisals).

Provide an anonymous and confidential internal channel. Companies must provide a reasonable, anonymous and confidential internal process through which covered employees may disclose information. Additionally, they should include the following features: (i) independence from executive management and legal team, best practice is governance by the board; (ii) good faith requirement, as noted in section 1; (iii) monthly updates to the disclosing person regarding the investigation status and actions taken; (iv) quarterly sharing of disclosures and responses with officers and directors. Ideally, there is also the option to use the internal channel to report directly to the board.

External Reporting to Authorities

Covered persons must be able to disclose information confidentially and anonymously to the Attorney General, relevant state or federal agencies, authorities such as law enforcement and regulators, as well as members of Congress.

Public Disclosure

While public disclosures are protected in international best practices in cases of urgent or grave public danger, or persistently unaddressed wrongdoing, this is unprecedented in the US, so we would not recommend it.

5. Remedies

[AI] Equity Protection. Compensation at frontier AI firms is heavily weighted toward equity or equity-like instruments (RSUs, stock options, profit participation units). Self-reported data indicates median total pay for engineers at Anthropic is approximately $557k per annum, with 44–57% from equity. Existing whistleblower statutes do not explicitly account for this, and, while case law makes it likely that lost equity can be recovered or potentially included in (doubled) backpay, introducing more clarity here is critical given the importance of the deterrent at hand. Best practice in this context requires explicitly including benefits, including but not limited to equity or equity-like compensation, in compensatory relief remedies.

[AI] Visa Protection. Many of the workers in the tech industry in the US do not have US citizenship, which means that if they are fired due to making a whistleblower disclosure their visa will be revoked. This may make it difficult for the whistleblower to file a whistleblower protection claim and defend their claim in court. As such, to ensure that individuals are not deterred from blowing the whistle, they should be granted temporary lawful presence if there exists reasonable cause to believe that the disclosed behaviour occurred (the same standard that is often applied for injunctive relief).

Injunctive Relief. Covered persons may petition for temporary or preliminary injunctive relief, which should be granted upon showing reasonable cause to believe a violation occurred. Such relief cannot be stayed pending appeal.

Compensatory Relief. Relief should cover all direct, indirect, and future consequences of reprisals, including back pay, reasonable attorney fees (one-directional, i.e. only claimable by an employee, not an employer, in case of a rejected retaliation claim, as in California), compensation for lost earnings and status, and compensation for pain and suffering. Ideally, following precedent in Dodd-Frank, back pay should be doubled.

6. Employee Notice Requirements

AI Companies must provide clear notice to all covered employees of their rights and responsibilities, either through workplace postings or annual written notices that must be received and acknowledged.

7. Agency Requirements

[AI] Technical AI Expertise. Agencies receiving AI-related disclosures (e.g. AGs' offices, OES, federal regulators) must have, or have ready access to, technical AI expertise sufficient to evaluate disclosures. Without this, capacity to respond may be weak, undermining the effectiveness of the legislation.

Confidentiality and Anonymity. There must be systems in place to ensure that any information reported to recipient authorities is kept confidential as well as the identity of the whistleblower. There must also be a way to disclose information anonymously.

Transparency. Best practice requires whistleblower complaints authorities collect and regularly publish anonymised data on the number of cases received, their outcomes, and time to process. Ideally, data would also be collected on whistleblowers' compensation and recoveries.
Timeliness of Investigation. Investigations into alleged misconduct should be overseen by independent fact finders with clear deadlines, opportunities for whistleblower input, and public reporting requirements.

Acknowledgment and investigation timelines. Investigating bodies should acknowledge receipt of a disclosure within a defined period (the EU Whistleblowing Directive requires 7 days) and provide substantive feedback to the disclosing individual within a defined period (the EU Directive requires 3 months). Without mandatory timelines, potential whistleblowers are prone to believe that no action will be taken which is the primary deterrent to whistleblowing.

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 welcome 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

Abra Ganz

Abra Ganz is the Head of AI Policy at Pour Demain, where her focus is improving oversight of AI systems. 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. She holds affiliations with the Center for AI Risk Management and Alignment and 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.

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 to whistleblowing under the EU AI Act with Santeri Koivula of the Future of Life Institute, and contributing 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.

The AI
Whistleblower
Initiative

CARMA

CENTER FOR AI RISK
MANAGEMENT & ALIGNMENT