Letter from OpenAI Whistleblowers

August 22, 2024

Dear Governor Newsom, Senate President pro Tempore McGuire, and Assembly Speaker Rivas,

OpenAI and other AI companies are racing to build Artificial General Intelligence, or "AI systems that are generally smarter than humans." It's right in OpenAI's mission statement, and the company is raising billions of dollars to achieve this goal. Along the way, they may create systems that pose a risk of critical harms to society, such as unprecedented cyberattacks or assisting in the creation of biological weapons. If they succeed entirely, artificial general intelligence will be the most powerful technology ever invented.

We joined OpenAI because we wanted to ensure the safety of the incredibly powerful AI systems the company is developing. But we resigned from OpenAI because we lost trust that it would safely, honestly, and responsibly develop its AI systems. In light of that, we are disappointed but not surprised by OpenAI's decision to lobby against SB 1047.

Developing frontier AI models without adequate safety precautions poses foreseeable risks of catastrophic harm to the public. We are not the only ones concerned about the rapid advances in AI technology: earlier this year, Science published *Managing extreme AI risks amid rapid progress*, a consensus paper from twenty-five leading scientists describing "extreme risks from upcoming, advanced AI systems." Leaders in the field, including Sam Altman, agree: he has stated that the worst case scenario for AI could be "lights out for all of us."

For there to be genuine public involvement in decisions around high risk systems, the public needs accurate information, not just voluntary disclosures from companies. As we discussed in the Right to Warn letter, whistleblowers need to feel protected if they come forward to warn officials about these risks and any other risks that future AI models may ultimately present. If an AI developer is doing something seriously unsafe, the employees at the lab will be the first to know. Silencing them puts everyone at risk.

OpenAI has not given the public reason for confidence.

- In the absence of whistleblower protections, **OpenAI demanded we sign away our rights to ever criticize the company** under threat of losing millions of dollars in vested equity when we resigned from the company.
- Despite touting "cautious" and "gradual" deployment practices, GPT-4 was deployed prematurely in India in **direct violation of OpenAI's internal safety procedures**.
- More famously, OpenAI provided technology to Bing's chatbot, which then threatened and attempted to manipulate users.
- OpenAI claimed to have "strict internal security controls" despite **a major security breach and a series of other internal security concerns**. The company also fired a colleague in part for raising concerns about security practices.
- **Prominent safety researchers have left the company,** including co-founders. The head of the team responsible for research into controlling smarter than human AI systems, said on resignation that the company was "long overdue in getting incredibly serious about the implications of AGI" and that "safety culture and processes have taken a backseat to shiny products."

While these incidents did not cause catastrophic harms, that's only because truly dangerous systems have not yet been built, not because companies have safety processes that could handle truly dangerous systems.

We believe that there should be public involvement in decisions around high risk AI systems, and SB 1047 creates a space for this to happen. It requires publishing a safety and security protocol to inform the public about safety standards. It protects whistleblowers who raise concerns to the California Attorney General if a model poses an unreasonable risk of causing or materially enabling critical harm. It provides a possibility for consequences for companies if they mislead the public and doing so leads to harm or an imminent threat to public safety. And it strikes a careful balance that protects legitimate intellectual property interests, by allowing redaction of sensitive information and only protecting disclosures to government officials.

OpenAI's complaints about SB 1047 are not constructive and don't seem in good faith.

- **Existing federal efforts and proposed legislation they reference are woefully inadequate to address these problems**. They don't protect whistleblowers, and they do nothing to prevent a company from releasing a product that would foreseeably cause catastrophic harm to the public. It is perfectly clear that they are not a substitute for SB 1047, and OpenAI knows as much.
- **We cannot wait for Congress to act** — they've explicitly said that they aren't willing to pass meaningful AI regulation. If they ever do, it can preempt CA legislation. Anthropic joins sensible observers when it worries "Congressional action will not occur in the necessary window of time."
- **SB 1047's requirements are things that AI developers—including OpenAI—have already largely agreed to in voluntary commitments to the White House and at Seoul.** The main difference is that SB 1047 would force AI developers to show the public that they are keeping these commitments, and hold them accountable if they don't.
- **The fears of a mass exodus of AI developers from the state are contrived.** OpenAI said the same thing about the EU AI Act, but it didn't happen. California is the best place in the world to do AI research. What's more, the bill's requirements would apply to anyone doing business in CA regardless of their location. It is extremely disappointing to see our former employer pursue scare tactics to derail AI safety legislation.

Sam Altman, our former boss, has repeatedly called for AI regulation. Now, when actual regulation is on the table, he opposes it. He said that "obviously [OpenAI would] comply with/aggressively support all regulation," and he testified in front of Congress calling for government intervention. Yet OpenAI opposes even the extremely light-touch requirements in SB 1047, most of which OpenAI claims to *voluntarily* commit to, raising questions about the strength of those commitments.

OpenAI's approach is in contrast to Anthropic's engagement, though we disagree with some changes that Anthropic requested. Anthropic expressed specific concerns, asked for changes, and afterwards concluded that the bill was likely beneficial on net, and that it "presents a feasible compliance burden." OpenAI has instead chosen to fearmonger and offer excuses.

We hope that the California Legislature and Governor Newsom will do the right thing and pass SB 1047 into law. With appropriate regulation, we hope OpenAI may yet live up to its mission statement of building AGI safely.

Sincerely,

William Saunders, former OpenAI Member of Technical Staff
Daniel Kokotajlo, former OpenAI Member of Policy Staff