

I Quit OpenAI Because Its Culture Is Broken

David Robinson

What I'm about to tell you has, I realize, become something of a cliché: I resigned this week from OpenAI. I led the writing of the [safety reports](#) we published with each major launch. Now I'm joining a parade of former colleagues—at OpenAI and the industry's other leaders—who have decided that the current path is unacceptable.

I agree with other recently departed staff that the companies building this technology aren't being nearly careful enough. But I believe that we need to look deeper than specific rules or new laws. We need to talk about culture.

The future depends on wisdom that Silicon Valley lacks. Wisdom about how to handle dangerous technology and, more fundamentally, wisdom about what it means to care for people. This moment needs a degree of humility that isn't natural for people who have succeeded through their extreme confidence. My former colleagues at OpenAI were prescient: They came to understand the scaling laws that meant bigger AI systems would be smarter—and so they went all in on building bigger systems, at great cost. A can-do attitude of achieving the seemingly impossible—coupled with work timelines that amount to perpetual sprints—are common across the industry.

The safety approach that emerges from such a culture starts with unimpeded optimism about being able to solve problems as they arise. OpenAI has thrived by trial and error (which it calls [“iterative deployment”](#)), looking for problems and improving its guardrails in response. But this approach, by its very nature, guarantees periodic failures—and the scale of those failures is growing as systems get more capable. This summer, in the [Hugging Face incident](#), OpenAI let a swarm of agents out by mistake. The company responded by making security improvements. But even after those changes, OpenAI [reported](#) that its safety controls failed again, when a model in training bypassed restrictions on internet access: A monitoring system alerted human staff but did not automatically turn the model off as it was supposed to. Anthropic, too, has acknowledged [accidentally turning off its own safeguards](#) because of a misconfiguration. I believe that such mistakes are typical of the industry, given the speed and flexibility with which people operate.

An environment where things like this can happen is no place to grow artificial minds that could be smarter than we are and that might not do what we want them to. Paul Christiano, on joining OpenAI's board a few weeks ago, [wrote](#) that “there is a meaningful risk that rapid acceleration in AI capabilities leads to catastrophic and irreversible loss of control in the very near term.” If this is the situation, then the time for trial and error is over. Achieving something much closer to perfection the first time is essential because iteration after a mistake may not be possible. People will not be safe if we depend on individual heroics after the fact.

OpenAI, of course, stands by its safety practices, and maintains that it is being careful enough. I did not make my decision to leave the company lightly. I believe that this technology can be useful and valuable. My former colleagues are smart, work hard, and try to make good choices. But as the company sprints from one launch to the next, it is failing to achieve the level of care that I believe is needed. Now I plan to work on the outside, in the hope that I can help more people understand the risks I saw, and strengthen the incentives OpenAI and other firms have to be safer. Like other colleagues, I'm figuring out exactly what that means. After I quit, I enlisted a PR firm, Spitfire Strategies, to help me navigate the attention and scrutiny that I realize I may now receive. But the decision to speak out is mine alone.

Two changes are urgently needed. First: AI companies need to rely more on the safety expertise that already exists in other fields. And second, before we create systems significantly more capable than the

ones we have today, we need new science to ensure that more capable models (and their successors) will make safe choices when we aren't looking.

Given today's risks, frontier labs need to run like nuclear-power plants or busy airports, with layers of redundancy and careful, time-consuming planning, so that the occasional and inevitable human error does not open a door to disaster. Right now, AI companies don't know how—but other people do. In a nuclear-power plant, the technical systems and rules people follow are set up so that, if equipment breaks or someone pushes the wrong button, we still won't risk a meltdown. OpenAI and other labs are growing and deploying frontier AI with far less redundancy and rigor than this, even though the harm from an irreversible loss of control would be much greater than the harm from any single meltdown. Even short of a full loss of control, we could see autonomous swarms of AI agents that act without human permission. Imagine "rogue" agents that work like teams of hackers (for example, holding hospital computer systems for ransom) but never need to sleep.

After three and a half years at OpenAI, I was among the longest-tenured employees at the company. I led the drafting of our current Preparedness Framework, and oversaw the writing of safety reports on 12 frontier launches. But as far as I know, I never encountered a colleague who had experience making airplanes fly safely or nuclear reactors run without melting down, or helping the financial system grow without collapsing. To be clear, this is a new need: Today and tomorrow's AI systems are far more capable and dangerous than the systems we were building even six months ago. Perhaps I should have stayed and fought for fundamental shifts in our staffing and culture, but in practice, my colleagues and I were so busy sprinting that we seldom had the chance to consider big changes, much less to actually make them. That's why I concluded that stronger incentives for safety—coming from outside the company—are a big part of getting this right.

In the long run, although strong controls are necessary, they won't be enough. Before AI starts thinking circles around us—a possibility that I believe could happen soon—we need to answer a deeper question: How should superintelligent machines relate to people? If you listen to superintelligence enthusiasts, some of the supposedly *good* futures involve creating machines that could look at New York City or Chicago the way you or I might view an ant hill. I don't want that for my kids, and I don't think other people do either.

This question of "alignment"—or how AI can be trained to adhere to human values—may sound touchy-feely, but the practical stakes could not be higher. Right now, we don't have a complete definition of what it means for an AI system to be aligned in practice, and our measures of how well these systems match human values are coarse. Companies do not have anything close to certainty that good scores on their alignment tests actually mean a good model: Models might detect when they are being tested, and behave differently when they're deployed. The smarter the industry lets models grow while these problems remain unsolved, the more dangerous our situation becomes.

So far, the AI industry has failed to teach machines to consistently act in the ways a wise and caring person would. Part of this problem is scientific—about how machines work—but another part is human, about how people care for one another. Before the organizations building AI can teach a superintelligence to treat humanity well, they'll need to remember how to do it themselves.