Written Testimony of

William Saunders
Former Member of Technical Staff, OpenAI
San Francisco, California

Presented before the U.S. Senate Committee on the Judiciary
Subcommittee on Privacy, Technology, & the Law
For a Hearing on Oversight of AI: Insiders' Perspectives

September 17, 2024

Oral Testimony

Mr. Chairman, Ranking Member Hawley, and distinguished Members, thank you for the opportunity to address the Committee.

For three years, I worked as a Member of Technical Staff at OpenAI. Companies like OpenAI, are working towards building Artificial General Intelligence — A G I. They are raising billions of dollars towards this goal. OpenAI defines AGI as "highly autonomous systems that outperform humans at most economically valuable work." This means AI systems that could act on their own over long periods of time and do most jobs that humans can do.

AI companies are making rapid progress towards building AGI. A few days before this hearing, OpenAI announced a new AI system GPT-o1 that passed significant milestones including one that was personally significant for me. When I was in high school, I spent years training for a prestigious international computer science competition. OpenAI's new system leaps from failing to qualify to winning a gold medal, doing better than me in an area relevant to my own job. There are still significant gaps to close but I believe it is plausible that an AGI system could be built in as little as three years.

AGI would cause significant changes to society, including radical changes to the economy and employment. AGI could also cause the risk of catastrophic harm via systems autonomously conducting cyberattacks, or assisting in the creation of novel biological weapons. OpenAI's new AI system is the first system to show steps towards biological weapons risk, as it is capable of helping experts in planning to reproduce a known biological threat. Without rigorous testing, developers might miss this kind of dangerous capability. While OpenAI has pioneered aspects of this testing, they have also repeatedly prioritized deployment over rigor. I believe there is a real risk they will miss important dangerous capabilities in future AI systems.

AGI will also be a valuable target for theft, including by foreign adversaries of the United States. While OpenAI publicly claims to take security seriously, their internal security was not prioritized. When I was at OpenAI, there were long periods of time where there were vulnerabilities that would have allowed me or hundreds of other engineers at the company to bypass access controls and steal the company's most advanced AI systems including GPT-4.

No one knows how to ensure that AGI systems will be safe and controlled. Current AI systems are trained by human supervisors giving them a reward when they appear to be doing the right thing. We will need new approaches when handling systems that can find novel ways to manipulate their supervisors, or hide misbehavior until deployed. The Superalignment team at OpenAI was tasked with developing these approaches, but ultimately, we had to figure it out as we went along, a terrifying prospect when catastrophic harm is possible. Today, that team no longer exists; its leaders and many key researchers resigned after struggling to get the resources they needed to be successful.

OpenAI will say that they are improving. I and other employees who resigned doubt they will be ready in time. This is true not just with OpenAI; the incentives to prioritize rapid development apply to the entire industry. This is why a policy response is needed.

My fellow witnesses and I may have different specific concerns with the AI industry, but I believe we can find common ground in addressing them.

If you want insiders to communicate about problems within AI companies, you need to make such communication safe and easy: That means a clear point of contact, and legal protections for whistleblowing employees.

Regulation must also prioritize requirements for third-party testing, both before and after deployment. Results from these tests must be shared. Creating an independent oversight organization and mandated transparency requirements, as in Senator Blumenthal and Senator Hawley's proposed framework, would be important steps towards these goals.

I resigned from OpenAI because I lost faith that by themselves they will make responsible decisions about AGI. If any organization builds technology that imposes significant risks on everyone, the public and the scientific community must be involved in deciding how to avoid or minimize those risks. That was true before AI. It needs to be true today with AI.

Thank you for your work on these issues, and I look forward to your questions.

Appendix 1: A Right to Warn about Advanced Artificial Intelligence

The following is the text of a letter I and other OpenAI employees signed at righttowarn.ai

---

We are current and former employees at frontier AI companies, and we believe in the potential of AI technology to deliver unprecedented benefits to humanity.

We also understand the serious risks posed by these technologies. These risks range from the further entrenchment of existing inequalities, to manipulation and misinformation, to the loss of control of autonomous AI systems potentially resulting in human extinction. AI companies themselves have acknowledged these risks [1, 2, 3], as have governments across the world [4, 5, 6] and other AI experts [7, 8, 9].

We are hopeful that these risks can be adequately mitigated with sufficient guidance from the scientific community, policymakers, and the public. However, AI companies have strong financial incentives to avoid effective oversight, and we do not believe bespoke structures of corporate governance are sufficient to change this.

AI companies possess substantial non-public information about the capabilities and limitations of their systems, the adequacy of their protective measures, and the risk levels of different kinds of harm. However, they currently have only weak obligations to share some of this information with governments, and none with civil society. We do not think they can all be relied upon to share it voluntarily.

So long as there is no effective government oversight of these corporations, current and former employees are among the few people who can hold them accountable to the public. Yet broad confidentiality agreements block us from voicing our concerns, except to the very companies that may be failing to address these issues. Ordinary whistleblower protections are insufficient because they focus on illegal activity, whereas many of the risks we are concerned about are not yet regulated. Some of us reasonably fear various forms of retaliation, given the history of such cases across the industry. We are not the first to encounter or speak about these issues.

We therefore call upon advanced AI companies to commit to these principles:

1. That the company will not enter into or enforce any agreement that prohibits “disparagement” or criticism of the company for risk-related concerns, nor retaliate for risk-related criticism by hindering any vested economic benefit;

2. That the company will facilitate a verifiably anonymous process for current and former employees to raise risk-related concerns to the company's board, to regulators, and to an appropriate independent organization with relevant expertise;
3. That the company will support a culture of open criticism and allow its current and former employees to raise risk-related concerns about its technologies to the public, to the company's board, to regulators, or to an appropriate independent organization with relevant expertise, so long as trade secrets and other intellectual property interests are appropriately protected;
4. That the company will not retaliate against current and former employees who publicly share risk-related confidential information after other processes have failed. We accept that any effort to report risk-related concerns should avoid releasing confidential information unnecessarily. Therefore, once an adequate process for anonymously raising concerns to the company's board, to regulators, and to an appropriate independent organization with relevant expertise exists, we accept that concerns should be raised through such a process initially. However, as long as such a process does not exist, current and former employees should retain their freedom to report their concerns to the public.