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;

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.

Signed by (alphabetical order):

Jacob Hilton,	formerly OpenAI
Daniel Kokotajlo,	formerly OpenAI
Ramana Kumar,	formerly Google DeepMind
Neel Nanda,	currently Google DeepMind, formerly Anthropic
William Saunders,	formerly OpenAI
Carroll Wainwright,	formerly OpenAI
Daniel Ziegler,	formerly OpenAI
Anonymous,	currently OpenAI
Anonymous,	currently OpenAI
Anonymous,	currently OpenAI
Anonymous,	currently OpenAI
Anonymous,	formerly OpenAI
Anonymous,	formerly OpenAI

Endorsed by (alphabetical order):

Yoshua Bengio
Geoffrey Hinton
Stuart Russell

References

1. OpenAI: “AGI would also come with serious risk of misuse, drastic accidents, and societal disruption … we are going to operate as if these risks are existential.” ⏎
2. Anthropic: “If we build an AI system that’s significantly more competent than human experts but it pursues goals that conflict with our best interests, the consequences could be dire … rapid AI progress would be very disruptive, changing employment, macroeconomics, and power structures … [we have already encountered] toxicity, bias, unreliability, dishonesty” ⏎
3. Google DeepMind: “it is plausible that future AI systems could conduct offensive cyber operations, deceive people through dialogue, manipulate people into carrying out harmful actions, develop weapons (e.g. biological, chemical), … due to failures of alignment, these AI models might take harmful actions even without anyone intending so.” ⏎
4. US government: “irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security.” ⏎
5. UK government: “[AI systems] could also further concentrate unaccountable power into the hands of a few, or be maliciously used to undermine societal trust, erode public safety, or threaten international security … [AI could be misused] to generate disinformation, conduct sophisticated cyberattacks or help develop chemical weapons.” ⏎
6. Bletchley Declaration (29 countries represented): “we are especially concerned by such risks in domains such as cybersecurity and biotechnology, … There is potential for serious, even catastrophic, harm” ⏎
7. Statement on AI Harms and Policy (FAccT) (over 250 signatories): “From the dangers of inaccurate or biased algorithms that deny life-saving healthcare to language models exacerbating manipulation and misinformation, …” ⏎
8. Encode Justice and the Future of Life Institute: “we find ourselves face-to-face with tangible, wide-reaching challenges from AI like algorithmic bias, disinformation, democratic erosion, and labor displacement. We simultaneously stand on the brink of even larger-scale risks from increasingly powerful systems” ⏎
9. Statement on AI Risk (CAIS) (over 1,000 signatories): “Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.” ⏎