

Opinion | Dear Congress: Here's What to Ask A.I. Leaders

Cal Newport

Advertisement

[SKIP ADVERTISEMENT](#)

Guest Essay

Sept. 24, 2026



Credit...Baptiste Viot

Listen · 9:04 min

By Cal Newport

Mr. Newport is a professor of computer science at Georgetown University and the author of “Deep Work.”

The fear of a powerful and uncontrollable artificial intelligence has shaken America and now Congress is scrambling to respond. As someone who studies the societal impact of technology and teaches in Washington, I often field questions from anxious lawmakers. Here's the advice I'm giving.

First and foremost, we don't yet understand enough about what is happening in these A.I. labs to create an effective response. We cannot regulate a vibe, nor can we specify concrete A.I. controls based on alarming parables about technology becoming too powerful.

Before any intervention, Congress should lead a public fact-finding mission that reveals the unvarnished details of A.I. development. This effort should start with Dario Amodei and Sam Altman — the leaders who have expressed the most concern about frontier A.I. models.

I would focus the questions on three key areas.

What specific types of A.I. technology are you worried about? Mr. Amodei and Mr. Altman tend to use the nonspecific term "A.I." when describing their systems. This implies that A.I. is a singular technology which improves along a fixed trajectory toward ever more capable and dangerous functionality.

In reality, A.I. is a broad category that includes a diverse collection of technologies — from self-driving cars, to game-playing engines, to medical image processing, to chatbots, to automated coding tools. Only some of these advancements depend on the large language models (L.L.M.s) produced by the frontier labs.

When Mr. Amodei and Mr. Altman talk about A.I. danger, they seem to be actually referring to a specific A.I. setup known as L.L.M.-powered agents. Though the term "agent" conjures creepy vibes, they're actually normal computer programs that repeatedly prompt an L.L.M. for instructions and then execute whatever the model suggests. Agents aren't worrisome when constrained to specific tasks with common-sense controls. Software developers use advanced coding agents every day without fearing that these systems will go rogue or defy human interests.

Sign up for the Opinion Today newsletter Get expert analysis of the news and a guide to the big ideas shaping the world every weekday morning.

The real concerns come from the experimental, Mad Max-style of agents like the ones that OpenAI deployed over the spring and summer. These agents were connected to L.L.M.s without standard guardrails, were instructed to break into protected test systems and were left to run with inadequate monitoring, relentlessly prompting and executing L.L.M. commands. This is self-evidently a recipe for unpredictable hazards. It was exactly these types of agents that performed the Hugging Face attack in July that helped spark our current panic.

It's not useful to talk about "pacing" or slowing down A.I. in some general sense if the real issue is a narrow band of seemingly incautious agent experiments. Biotechnology provides an instructive example. In 2014, concerns began to grow about gain-of-function research in which viruses are manipulated to become more potent. After a series of incidents involving the mishandling of dangerous pathogens at secure labs, many feared that we wouldn't be able to contain the powerful viruses produced by this technique.

The government didn't respond by slowing down all biotechnology research. They instead instituted a

[targeted moratorium](#) on funding for some of the more concerning gain-of-function efforts. The A.I. equivalent of these super-viruses may be certain types of agent experiments.

What are the safety procedures surrounding your research efforts? It's worth remembering that Anthropic and OpenAI are relatively new organizations. OpenAI, which was originally founded as a nonprofit, spun off a for-profit subsidiary capable of raising venture capital only in 2019, and Anthropic wasn't formed until two years later. These companies have since collectively raised hundreds of billions of dollars and are both targeting valuations well north of \$1 trillion. That's an absurdly fast growth rate. It wouldn't be surprising, then, if we discovered that they didn't yet have mature research safety processes in place.

To return to a biotechnology example, if the lab at a fast-growing start-up suffered multiple incidents in which dangerous pathogens were accidentally released, we'd demand to know more about their safety standards. They wouldn't be able to proceed with their work until we were satisfied that they could do so without endangering the public. A similar scrutiny needs to be applied to the L.L.M.-powered agent research at the frontier labs.

OpenAI, for example, [recently admitted](#) that the Hugging Face hack was just one of multiple concerning incidents. Why didn't they stop after the first was discovered? What improvements, if any, were made to safety procedures in response? OpenAI often discusses these incidents as if it were a passive observer of an autonomous technology — a rhetorical gambit that acts as a smokescreen for bad behavior. We must return the focus to the actual humans responsible for setting up these hazardous conditions.

(The New York Times [sued](#) OpenAI and Microsoft in 2023, claiming copyright infringement of news content related to A.I. systems. The two companies have denied those claims.)

What is the ultimate goal of your A.I. company? In some sense, it would be comforting to discover that the more harmful ideas and experiments run by Anthropic and OpenAI were the result of racing too fast toward profits or trying to impress investors. This is a well-worn path in capitalism, and one that the regulatory state has long served to redirect.

But there's a more worrisome possibility that we must better understand. It's been [well documented](#) that Mr. Amodei and Mr. Altman's companies are deeply connected to a strain of apocalyptic futurist thinking that believes something like an all-powerful superintelligent A.I. god is inevitable — and that this being will either deliver a trans-humanist utopia or destroy humanity. Both OpenAI and Anthropic were founded, in part, with the mission of trying to increase the probability that we end up with the benevolent option — a goal they describe as aligning A.I. to human interests before it grows too powerful.

This techno-religion undoubtedly impacts the mind-set of the people who work in these companies. When Mr. Amodei and Mr. Altman, or their employees, casually talk about A.I. automating jobs and creating high unemployment, or put precise percentages on the possibility that A.I. causes human extinction, it can be shocking to outside observers. But such claims are common and unexceptional in the futurist circles that they intersect. More concerning, however, is the possibility that these ideological beliefs are also driving these specific labs toward more reckless behavior, such as their hacking agent experiments over the past

year. If you believe winning the race to superintelligence is necessary to save humanity, you might not worry too much about collateral damage along the way.

Consider Anthropic and OpenAI's insistent interest in so-called recursive self-improvement (R.S.I.). This strategy — popularized by futurist circles — believes that designing A.I. to improve itself again and again is the fastest route to superintelligence. Earlier this month, the OpenAI chief scientist, Jakub Pachocki, [wrote](#): “We focus OpenAI research toward R.S.I., as we believe it is the only way to remain at the frontier of A.I. research moving forward.”

To be clear, R.S.I. is in no way necessary to create a thriving A.I. sector. It's an obsession that seems to be held mainly by the two labs — Anthropic and OpenAI — that are most connected to the futurist philosophies that have long venerated R.S.I. as being central to their apocalyptic visions. The Meta C.E.O., Mark Zuckerberg, recently called out this reality, [saying](#): “Committing the significant majority of compute towards serving people rather than racing toward recursive self-improvement is one of the best ways to ensure we develop this technology safely.”

Mr. Zuckerberg is correct: Pursuing R.S.I. represents an unnecessary danger. Not because it might lead to superintelligent A.I. (many experts believe this idea to be more science fiction than science), but because, in the short term, efforts to let A.I. systems modify themselves will make them harder for us to understand, and therefore harder for us to prevent unpredictable and damaging behavior. From a safety perspective, R.S.I. is a preposterously irresponsible path to prioritize.

These proposed questions are just a starting point, but they capture a general spirit that's important in this fearful moment. We need to stop letting a small number of private companies, acting and talking in increasingly erratic ways, dictate how we're supposed to feel about A.I. We've heard what they want to say; now it's Congress's turn to step in on our behalf to find out what's really going on.

Advertisement

[SKIP ADVERTISEMENT](#)