

The Case for Boycotting Generative AI

I believe the time has come

Gary Marcus — Marcus on AI — Sep 09, 2026

Source: <https://garymarcus.substack.com/p/the-case-for-boycotting-generative>

It has become clear by now that the frontier labs are knowingly and by their own admission doing something irreversible and incredibly dangerous without public consent.

They are already causing harm, and things could conceivably get much worse.

They have no serious plan for mitigating the risks of generative AI.

The government, too tight with industry, doesn't either.

The time for a public boycott has come.

When an Anthropic employee, Evan Hubinger, publicly writes this (with over 28 million views), it is time to take action:

Evan Hubinger (@EvanHub), X, Sep 8, 2026:

"Jacob is correct here—we really do earnestly believe AI could kill all humans! I personally think it is >10% within the next decade. I believe Anthropic is trying its best, but we do not yet have a plan to solve alignment for superintelligence and are not clearly on track to."

Quoted: Jacob Coxon (@hilbertspaess): "The people building AI earnestly believe that it could kill us all by the end of the decade. This is not a marketing stunt. If anything, many executives and senior researchers will couch their phrasing in the press to sound sensible - but I hear the same people express fear."

21.8M views — 3.13K replies · 6.11K reposts · 38.5K likes.

A few days earlier OpenAI's chief scientist Jakub Pachocki wrote something equally candid and distressing about the lack of solution to the "alignment problem."

Some of this may just be marketing, and they may exaggerate the risks, but there are real risks, and there is no plan.

Mind you, I happen not to worry too much about extinction. And I don't think you should be either. As I wrote last year in the Times Literary Supplement, humans are geographically diverse and generically diverse, and too resilient. If machines tried to take over we would fight back; we might take casualties but we would not be annihilated altogether.

But the risk of *catastrophe* is real, e.g., from AI-generated pathogens, from wars started or escalated by AI-generated disinformation, from hacks that destroy critical infrastructure, and so on.

Nothing I have seen gives any indication that any of that is under control.

Put differently, my $p(\text{doom})$ is low; my $p(\text{dystopia})$ is high.

And that is, for me, enough to be deeply worried, especially when coupled with a lack of a plan.

As an another important aside for those who like to misinterpret my words: in my view the issue here is not that current or near-future machines are somehow "too smart"; it's that they are too unreliable, and hence not worthy of our trust. And worse, we are rapidly giving such machines too much power to affect the world. We need to pause until reliability and trustworthiness can be ensured.

§

I see no reason for a permanent ban on AI, nor for a permanent ban on superintelligence; I don't actually think AI is going to wipe humans from the earth. And I still retain a little of the ten-year-old boy in me who believes that AI could be a force for good. If the frontier labs get their house in order, we can lift the ban.

But the time has come to pause building (or at least deploying) what I will call *incorrigible AI*.

By this incorrigible AI I mean AI that we cannot trust, or, to use popular jargon, "align" to human values.

Until such time as we have better answers to that both technically and politically, we should stop building such things, and stop giving them access to the internet.

§

On the technical side, we need AI that can follow instructions reliably and consistently. Classical, deterministic AI does that, and does not get us into trouble; no search engine or recommendation engine or route recommendation engine (to take three classic and widely used forms of AI) has ever on its own unexpectedly hacked a website or broken out of a sandbox.

The dangerous kind of AI that we are facing now is the kind of AI that is built on a core of probabilistic large language models that are not deterministic. They are often spectacularly correct, but – at other times, unpredictably – bafflingly wrong, and never trustworthy. That was true in 2022 when ChatGPT came out, and it is still true now.

Four years ago I started likening such stuff to bulls in a china shop ("powerful, but reckless and difficult to control").

Today they remain reckless and difficult to control. Worse we have given them access to the internet; immense harm may come from that.

We don't have to build AI the way we are now; there is no fundamental law that says we need to use LLMs. But using them as the core technology introduces easily foreseeable risks.

We either need to find a way to make them more reliable – but by the industry's own admission they don't really have one – or we need to go back to square one, and make determinism the central mission.

It was a mistake to put all our eggs in an erratic basket.

§

On the political side, I have frankly lost faith in our government. Three-and-half years ago when I testified in the US Senate, I was optimistic. The Senators really seemed to want to do something. But nothing substantial has been put to a vote.

The White House meanwhile has a policy that is entirely opaque; they seem technically naive, having allowed GPT-6 Astra, which reduces monitorability (hence likely increases risk) without public comment. They seem too close with OpenAI to take serious action. That may burn them in the end, if anything really bad happens, but for now they seem willing to risk that potential heat and unable to take a real stand.

Replacing the word existential with catastrophic, this would be a good idea. So would this. I sure hope that happens.

But because it probably won't, I still lean towards a boycott as a last but possibly necessary resort.

§

The counterargument to a boycott I expect is "what about China?"

Alyssa Farah Griffin (@Alyssafarah), X, Sep 9, 2026:

"The AI warnings coming out in the last 24 hours are deadly serious. But any efforts to slow the pace of development & implement guardrails have to be global. It's meaningless for the US to self-regulate when China continues advancing."

71.5K views — 12 reposts · 120 likes.

The answer to this is obvious.

It is time to make a deal with China.

§

Derek Thompson also just raised a very good point in response to the above:

"If the frontier labs feel obligated to build something they think is dangerous because China is going to build it anyway, we'd better be really sure that China is going to build it anyway. Like, really, really, really sure."

Are we? Are we actually sure? The CCP [China's Communist Party] wants to build an out-of-control, recursively self-improving model bc its neurotically, control-obsessed government thinks this is a policy worth pursuing? We're 100% sure about that?"

I am pretty sure they (the Chinese Communist Party) does not in fact want that.

Which is part of why we might actually be able to strike a deal. (I will write more about that soon.)

§

While the government dithers, the rest of us need to take action.

One simple action could avert potential catastrophe. If we stop using generative AI tools like chatbots and image generators we send a message to the frontier labs, "we have had enough."

And we would threaten their economics. Every member of those companies is making a Faustian bargain, seeking untold riches balanced against the chance of great harm.

The only way to motivate them to stop what they are doing until they can find better solutions – and to motivate them to find better solutions instead of pushing ones that they know full well they cannot control – is to hit them in the wallet.

If usage falls, the IPOs won't fly. Nothing could be more incentivizing to them. In a text to me this morning, Cal Newport argued, "If you really pushed them on why they insist on building this particularly unstable type of AI system they would probably finally admit it's because of their deterministic belief in a world made whole (or destroyed) by AI. There really is no commercial justification. It's ideological..... There's no need to be building those recklessly. There's not a commercial case for them. You could stop doing that tomorrow ..." Even if you thought there was a commercial case (as some have tried to argue) for focusing on LLMs rather than other more stable technology, it's hard to see how it would justify catastrophic risks.

The AI companies have also not made the case cogently that pursuing AGI (or whatever it is they are doing) reduces the risks they see in any substantive way.

Taking a supposed 10% chance at catastrophe with no clear path to a tremendously positive outcome to offset it is insane.

Unless someone comes up with a better way to address that insanity, and I am all ears, I urge a boycott on generative AI until the companies do something tangible, besides expressing concern, to reduce the risk.

P.S. You can watch an interesting conversation from a few days ago with PauseAI's Holly Elmore here, discussing among other things when one might lift a pause if one were imposed. At the time, this was recorded, I was still resistant to an overall pause, and focused primarily on OpenAI's misconduct, but seeing an Anthropic employee's somewhat cavalier attitude towards significant risk has led me to endorse something broader.