Pausing AI Developments Isn't Enough. We Need to Shut it All Down

Illustration for TIME by Lon Tweeten

IDEAS

BY ELIEZER YUDKOWSKY

MARCH 29, 2023 6:01 PM EDT

Yudkowsky is a decision theorist from the U.S. and leads research at the Machine Intelligence Research Institute. He's been working on aligning Artificial General Intelligence since 2001 and is widely regarded as a founder of the field.

An open letter published today calls for "all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4."

This 6-month moratorium would be better than no moratorium. I have respect for everyone who stepped up and signed it. It's an improvement on the margin.

I refrained from signing because I think the letter is understating the seriousness of the situation and asking for too little to solve it.

Read More: *AI Labs Urged to Pump the Brakes in Open Letter*

The key issue is not "human-competitive" intelligence (as the open letter puts it); it's what happens after AI gets to smarter-than-human intelligence. Key thresholds there may not be obvious, we definitely can't calculate in advance what happens when, and it currently seems imaginable that a research lab would cross critical lines without noticing.

Many researchers steeped in these issues, including myself, expect that the most likely result of building a superhumanly smart AI, under anything remotely like the current circumstances, is that literally everyone on Earth will die. Not as in "maybe possibly some remote chance," but as in "that is the obvious thing that would happen." It's not that you can't, in principle, survive creating something much smarter than you; it's that it would require precision and preparation and new scientific insights, and probably not having AI systems composed of giant inscrutable arrays of fractional numbers.

Without that precision and preparation, the most likely outcome is AI that does not do what we want, and does not care for us nor for sentient life in general. That kind of caring is something that *could in principle* be imbued into an AI but *we are not ready* and *do not currently know how*.

Absent that caring, we get "the AI does not love you, nor does it hate you, and you are made of atoms it can use for something else."

The likely result of humanity facing down an opposed superhuman intelligence is a total loss. Valid metaphors include "a 10-year-old trying to play chess against Stockfish 15", "the 11th century trying to fight the 21st century," and "*Australopithecus* trying to fight *Homo sapiens*".

To visualize a hostile superhuman AI, don't imagine a lifeless book-smart thinker dwelling inside the internet and sending ill-intentioned emails. Visualize an entire alien civilization, thinking at millions of times human speeds, initially confined to computers—in a world of creatures that are, from its perspective, very stupid and very slow. A sufficiently intelligent AI won't stay confined to computers for long. In today's world you can email DNA strings to laboratories that will produce proteins on demand, allowing an AI initially confined to the internet to build artificial life forms or bootstrap straight to postbiological molecular manufacturing.

If somebody builds a too-powerful AI, under present conditions, I expect that every single member of the human species and all biological life on Earth dies shortly thereafter.

There's no *proposed plan* for how we could do any such thing and survive. OpenAI's openly declared intention is to make some future AI do our AI alignment homework. Just hearing that *this is the plan* ought to be enough to get any sensible person to panic. The other leading AI lab, DeepMind, has no plan at all.

An aside: None of this danger depends on whether or not AIs are or can be conscious; it's intrinsic to the notion of powerful cognitive systems that optimize hard and calculate outputs that meet sufficiently complicated outcome criteria. With that said, I'd be remiss in my moral duties as a human if I didn't also mention that we have no idea how to determine whether AI systems are aware of themselves—since we have no idea how to decode anything that goes on in the giant inscrutable arrays—and therefore we may at some point inadvertently create digital minds which are truly conscious and ought to have rights and shouldn't be owned.

The rule that most people aware of these issues would have endorsed 50 years earlier, was that if an AI system can speak fluently and says it's self-aware and demands human rights, that ought to be a hard stop on people just casually owning that AI and using it past that point. We already blew past that old line in the sand. And that was probably *correct*; I *agree* that current AIs are probably just imitating talk of self-awareness from their training data. But I mark that, with how little insight we have into these systems' internals, we *do not actually know.*

If that's our state of ignorance for GPT-4, and GPT-5 is the same size of giant capability step as from GPT-3 to GPT-4, I think we'll no longer be able to justifiably say "probably not self-aware" if we let people make GPT-5s. It'll just be "I don't know; nobody knows." If you can't be sure whether you're creating a self-aware AI, this is alarming not just because of the moral implications of the "self-aware" part, but because being unsure means you have no idea what you are doing and that is dangerous and you should stop.

On Feb. 7, Satya Nadella, CEO of Microsoft, publicly gloated that the new Bing would make Google "come out and show that they can dance." "I want people to know that we made them dance," he said.

This is not how the CEO of Microsoft talks in a sane world. It shows an overwhelming gap between how seriously we are taking the problem, and how seriously we needed to take the problem starting 30 years ago.

We are not going to bridge that gap in six months.

It took more than 60 years between when the notion of Artificial Intelligence was first proposed and studied, and for us to reach today's capabilities. Solving *safety* of superhuman intelligence—not perfect safety, safety in the sense of "not killing literally everyone"—could very reasonably take at least half that long. And the thing about trying this with superhuman intelligence is that if you get that wrong on the first try, you do not get to learn from your mistakes, because you are dead. Humanity does not learn from the mistake and dust itself off and try again, as in other challenges we've overcome in our history, because we are all gone.

Trying to get *anything* right on the first really critical try is an extraordinary ask, in science and in engineering. We are not coming in with anything like the approach that would be required to do it successfully. If we held anything in the nascent field of Artificial General Intelligence to the lesser standards of engineering rigor that apply to a bridge meant to carry a couple of thousand cars, the entire field would be shut down tomorrow.

We are not prepared. We are not on course to be prepared in any reasonable time window. There is no plan. Progress in AI capabilities is running vastly, vastly ahead of progress in AI alignment or even progress in understanding what the hell is going on inside those systems. If we actually do this, we are all going to die.

Read More: *The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter*

Many researchers working on these systems think that we're plunging toward a catastrophe, with more of them daring to say it in private than in public; but they think that they can't unilaterally stop the forward plunge, that others will go on even if they personally quit their jobs. And so they all think they might as well keep going. This is a stupid state of affairs, and an undignified way for Earth to die, and the rest of humanity ought to step in at this point and help the industry solve its collective action problem.

Some of my friends have recently reported to me that when people outside the AI industry hear about extinction risk from Artificial General Intelligence for the first time, their reaction is "maybe we should not build AGI, then."

Hearing this gave me a tiny flash of hope, because it's a simpler, more sensible, and frankly saner reaction than I've been hearing over the last 20 years of trying to get anyone in the industry to take things seriously. Anyone talking that sanely deserves to hear how bad the situation actually is, and not be told that a six-month moratorium is going to fix it.

On March 16, my partner sent me this email. (She later gave me permission to excerpt it here.)

"Nina lost a tooth! In the usual way that children do, not out of carelessness! Seeing GPT4 blow away those standardized tests on the same day that Nina hit a childhood milestone brought an emotional surge that swept me off my feet for a minute. It's all

going too fast. I worry that sharing this will heighten your own grief, but I'd rather be known to you than for each of us to suffer alone."

When the insider conversation is about the grief of seeing your daughter lose her first tooth, and thinking she's not going to get a chance to grow up, I believe we are past the point of playing political chess about a six-month moratorium.

If there was a plan for Earth to survive, if only we passed a six-month moratorium, I would back that plan. There isn't any such plan.

Here's what would actually need to be done:

The moratorium on new large training runs needs to be indefinite and worldwide. There can be no exceptions, including for governments or militaries. If the policy starts with the U.S., then China needs to see that the U.S. is not seeking an advantage but rather trying to prevent a horrifically dangerous technology which can have no true owner and which will kill everyone in the U.S. and in China and on Earth. If I had infinite freedom to write laws, I might carve out a single exception for AIs being trained solely to solve problems in biology and biotechnology, not trained on text from the internet, and not to the level where they start talking or planning; but if that was remotely complicating the issue I would immediately jettison that proposal and say to just shut it all down.

Shut down all the large GPU clusters (the large computer farms where the most powerful AIs are refined). Shut down all the large training runs. Put a ceiling on how much computing power anyone is allowed to use in training an AI system, and move it downward over the coming years to compensate for more efficient training algorithms. No exceptions for governments and militaries. Make immediate multinational agreements to prevent the prohibited activities from moving elsewhere. Track all GPUs sold. If intelligence says that a country outside the agreement is building a GPU cluster, be less scared of a shooting conflict between nations than of the moratorium being violated; be willing to destroy a rogue datacenter by airstrike.

Frame nothing as a conflict between national interests, have it clear that anyone talking of arms races is a fool. That we all live or die as one, in this, is not a policy but a fact of nature. Make it explicit in international diplomacy that preventing AI extinction scenarios is considered a priority above preventing a full nuclear exchange, and that allied nuclear countries are willing to run some risk of nuclear exchange if that's what it takes to reduce the risk of large AI training runs.

That's the kind of policy change that would cause my partner and I to hold each other, and say to each other that a miracle happened, and now there's a chance that maybe Nina will live. The sane people hearing about this for the first time and sensibly saying "maybe we should not" deserve to hear, honestly, what it would take to have that happen. And when your policy ask is that large, the only way it goes through is if policymakers realize that if they conduct business as usual, and do what's politically easy, that means their own kids are going to die too.

Shut it all down.

We are not ready. We are not on track to be significantly readier in the foreseeable future. If we go ahead on this everyone will die, including children who did not choose this and did not do anything wrong.