Economics from the Top Down

new ideas in economics and the social sciences

No, AI Does Not Pose an Existential Risk to Humanity

Blair Fix

June 10, 2023

In the last few months, there have been a string of open letters from tech leaders warning that artificial intelligence could lead to a 'profound change in the history of life on Earth'. According to some insiders, AI poses an extinction risk, on the same level as pandemics and nuclear war.

I think this scaremongering is nonsense. It plays into popular science fiction tropes about machines conquering humanity. But the problem with these scenarios is that when you look at them through an evolutionary lens, they make no sense.

Competence, not intelligence

Let's start with the elephant in the room, which is 'intelligence'. Humans love to talk about 'intelligence' because we're convinced we possess more of it than any other species. And that may be true. But in evolutionary terms, it's also irrelevant. You see, evolution does not care about 'intelligence'. It cares about *competence* — the ability to survive and reproduce.

Looking at the history of evolution, philosopher Daniel Dennett argues that the norm is for evolution to produce what he calls 'competence without comprehension'. Viruses commandeer the replication machinery of their hosts without understanding the genetic code. Bacteria survive in extreme environments without comprehending biochemistry. Birds fly without theorizing aerodynamics. Animals reproduce without grasping the details of meiosis. And so on. In short, most of what makes an organism ‘competent’ is completely hidden from concious understanding.

Back to humans. When we talk about ‘intelligence’, we usually mean the tasks we do consciously — things like language, logic, and science. So when machines can replicate these abilities, it seems impressive … even chilling. But is it dangerous?

Of course it is. Any tool that augments human abilities is dangerous. Take the calculator as an example. Calculators are ‘intelligent’ in the sense that they can do calculations faster than any human. Couple the calculator with a knowledge of physics, and we get the ability to precisely shoot projectile weapons. The history of 20th century warfare illustrates the peril of this combination. However, few people argue that calculators pose an ‘existential’ risk to humanity. So then why the hype about the souped-up calculators we call ‘AI’?

I think the reason is primal … literally. Primates are defined by their sociality. And human primates are defined by our ability to socialize using language. So when a machine crunches numbers, it’s not particularly exciting. But what a machine *talks* to us, we find it impressive … shocking even.

If we let our imaginations run wild, we can envision that the machines might one day take over. But the problem here is that we are fooled into thinking that the machines have basic competences that we humans possess but (crucially) *do not comprehend*.

What does AI eat for breakfast?

Backing up a bit, lets talk about the nightmare scenario, played out in movies like *Terminator* and *The Matrix*. The general idea is that machines become so advanced that they overthrow their human masters and conquer the world. The idea is legitimately terrifying. It is also wildly unrealistic.[1]

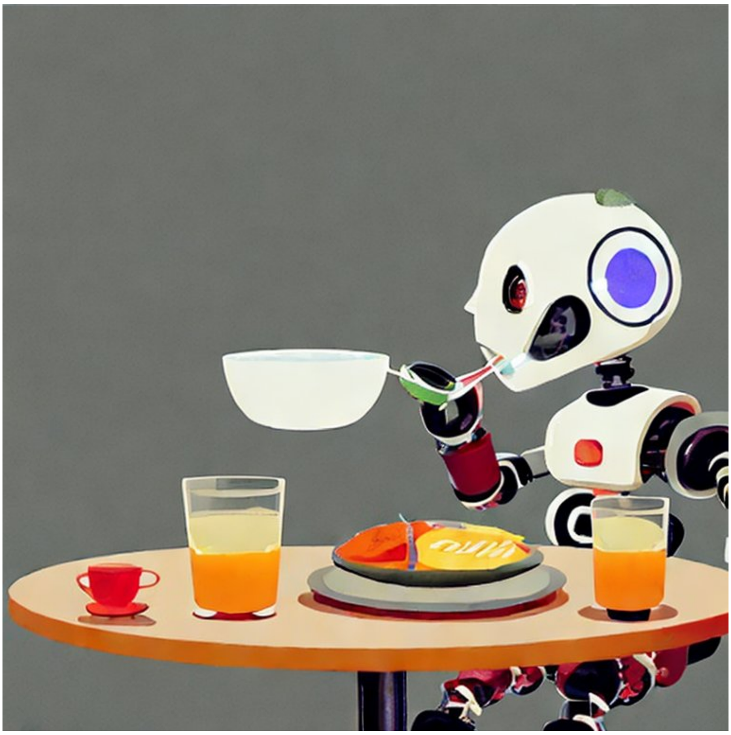
[1] *The Matrix*, for example, is premised on a whopper of a falsehood. In the distant future, machines have conquered humanity, and turned the lot of us into a massive, living ‘battery’. Sorry folks, but the energetics don’t work. Humans are not net producers of energy. So the machines are just wasting energy keeping us alive.

To see the problem, ask yourself a simple question: what does the AI eat for breakfast? Here's my point. When you start thinking about breakfast, you become aware of how your body sustains itself, mostly in ways you're oblivious to.

In a modern setting, things start with the complex supply chain that brings food to your mouth. Where did that food come from? Who grew it? How would machines develop a similar supply chain to meet their material needs?

Moving a step lower, how is it that you body 'knows' how to digest food? The answer is not located in your brain — in your so-called 'intelligence'. The answer lies in billions of years of evolution that have perfected your cellular metabolism.

Your cells 'know' how to use food to sustain themselves, even if you have no idea what's going on. Your cells also 'know' how to reproduce, passing on this ability to you, an intelligent being who is blissfully unaware of its own competence. How on earth would machines replicate this ability?



'A robot eating breakfast'. Rendered with Stable Diffusion 2.

Holding onto this question for a moment, let's return to the idea of finding and digesting breakfast. Today's machines may be 'intelligent', but they have none of the core competencies that make life robust. We design their metabolism (which is brittle) and we spoon feed them energy. Without our constant care and attention, these machines will do what all non-living matter does — wither against the forces of entropy.

In other words, if humans stopped maintaining the server farm on which ChatGPT runs — cut the repair work and cut the power — how long would the 'intelligence' last? A few seconds at best.

Still, perhaps machines could learn to sustain themselves and self-replicate? Maybe they could. But the scale of the problem is immense. We delude ourselves if we think otherwise.

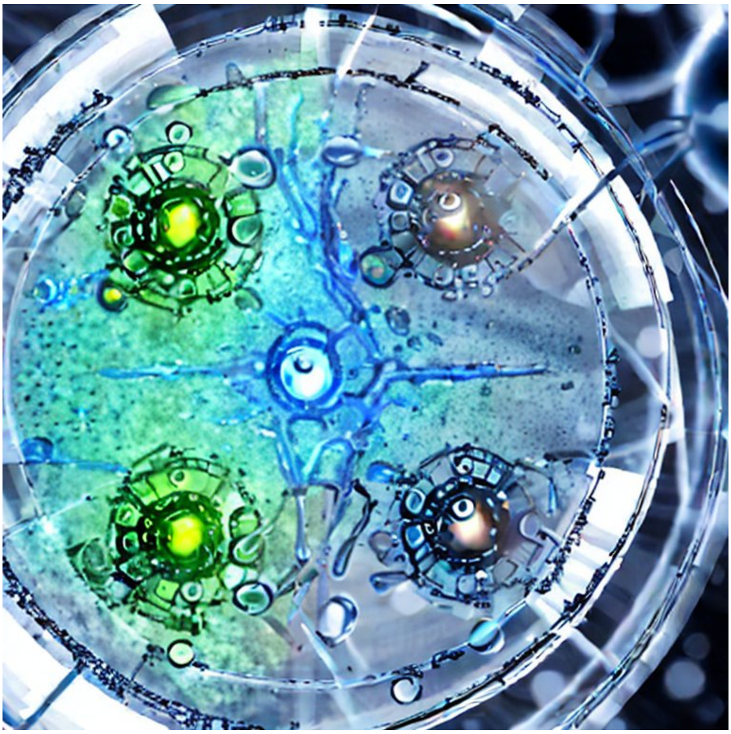
Here's the main problem that AI fearmongers don't understand. The core competency of life — survival and reproduction — is not achieved by passively ingesting information. It is rarely achieved by concious 'knowing'. Instead, it is achieved by incessantly interacting with the environment. It is achieved by exquisitely vast trial and error. It is achieved by mindless *evolution*.

Now being clever beasts, humans have found ways to simulate some of the end-results of evolution. For example, we can simulate our own ability to do math, to play chess, and to answer quiz-show questions. In fact, we can build simulations that far surpass our own ability. And because we conceive of these competencies as 'higher-level abilities', we have the hubris to think our machines might best organic life. But we are being silly.

Lurking below the concious abilities that we call 'intelligence' are a myriad of core competencies about which we are scarcely aware. Take the ability to explore our environment. Forged over millions of years of evolution, the ability to navigate the physical world is a competency that we take for granted. In other words, no one claims that a good driver is 'intelligent'. And yet this 'lower-level' skill is one that we struggle to get machines to replicate. In short, we underestimate life's competence.

When we start looking at the nuts and bolts of metabolism, things get even worse. The truth is that entropy is a universal bitch. There are an almost unimaginable number of ways for matter to break down, and very few ways for it to become self-sustaining. In short, where are the robots that have self-repairing circuits? Where are the robots who can sustain their 'intelligence' by foraging for energy? We may have nightmares about these bots. But the reality is that we have no idea how to bring them into existence. None.

In fact, I'm willing to bet that for machines to actually best biological life at survival and reproduction, they'd have to do it through evolution. And they would have to start small. In other words, forget about artificial intelligence. If we really wanted to create self-sustaining artificial life, our best bet would be to design an artificial cell and then let it evolve for a billion years. If we're lucky, the end result might be something as smart and tenacious as humans. But the more likely outcome would be extinction.



'Robot metabolism at the cellular level in a petri dish'. Rendered with Stable Diffusion 2.

To summarize, AI fears play into an extreme version of the mind-body fallacy. The reality is that minds cannot exist without bodies. And self-sustaining bodies are not easy to design. In fact, there's no evidence that anything but eons of evolution can design such a thing.

So yeah, we can build machines that simulate interesting features of the human mind. But the simulation is completely dependent on a body that humans maintain. Without a body that is self-replicating and self-sustaining, the existential risk posed by AI is precisely zero. The worst that can happen is that a bot goes rogue and we pull its plug.[2]

[2] But what if the bad AI spreads like a virus? Given the stupendous power demands of systems like ChatGPT, it's hard to imagine how this could happen. The bot would have to commandeer data centers, not personal computers. And even if it managed to control a big system, we could still pull the plug.

Real risks from machine learning models

Like all human-made tools, the real risk with machine learning models lies with the humans who use and control them. Top of mind is that fact that these models are essentially a 'black box', meaning we don't understand how they work.

It's not a new problem. A general risk with all computer software is that it may contain capabilities that the user doesn't understand. For example, you may think that the file `defrag.exe` will defragment your hard drive. But it could just as easily be a virus that will commandeer your computer. Unless you can look at the source code, you'll never know.

That's where the free software movement comes in. An essential ingredient of safe computing is to be able to study how a program works. Sure, you may not personally understand the code. But there are other people who can. The result is a kind of group policing of the commons that works similarly to Wikipedia. On Wikipedia, everyone is free to insert bogus information. But everyone else is free to delete it. The same goes for malicious code in free software. In short, the commons produce software of unparalleled quality.

The problem with models like ChatGPT is that they are best thought of as immense statistical models, the inner workings of which are not well understood. So if the machine spits out wrong answers, it's difficult to look at its components to understand what went wrong. In practice, this means that it is harder to apply the principles of commons governance to machine-learning models. At best, we can 'police' the data that goes into the models, and 'police' the information that comes out.

To date, the record of openness is not good. For example, OpenAI was founded on open-source principles, promising to share its code with the world. For a while, it made good on that promise, releasing the source code for language models GPT-1 and GPT-2. But with the release of GPT-3, it pivoted to a proprietary model. So the 'open' in 'OpenAI' is now a misnomer.

So here's the rub; when you ask ChatGPT a question, it always gives you an answer, written in clear, concise prose. But half the time, the answer is bullshit, and neither the bot nor the user knows which half. It's the machine-learning equivalent of Donald Trump.

That's the thing about big 'organisms' with a massive energy demand: they're really easy to kill. For example, its doubtful that anything short of nuclear winter would put a dent in the world's bacteria population. But lumbering whales are easy to exterminate. Just look at what humans did over the last century.

Is this random truthiness a problem? Yes. But hardly an existential one.

AI will not take all our jobs

Another concern is that AI will replace so many jobs that it will create long-term mass unemployment.

This idea is nonsense.

The problem here is a fundamental misunderstanding of what drives technological change. Capitalists replace workers with machines because the cost-cutting bolsters their profits. For the process to be self-sustaining, it cannot create mass unemployment.

If it did, companies would find that as they cut costs, they also shrink their revenues. You see, *paying* people is what creates revenue. So if everyone is replacing workers with machines and no one is creating new jobs, then everyone's revenue stream will collapse. And when it does, investment in new labor-replacing machines will stop.

The consequence is that if machines are to consistently replace existing jobs, new paying positions must invented. If they're not, the technological cycle collapses, and the job replacement stops. End of story.

Strategic scaremongering

It's easy to understand why the average person might be afraid of machine learning models. These programs exhibit behavior that is eerily human. And to be fair, the use of these models comes with risks. But to say that the risk from machine-learning models is 'existential' is absurd. The gap between a savant program like ChatGPT and a robust, self-replicating machine is monumental. Let ChatGPT 'loose' in the wild and one outcome is guaranteed: the machine will go extinct.

So why are machine-learning experts fearmongering the risks of artificial intelligence? I can think of at least four reasons:

1. The scaremongering is self-serving in the sense that it amplifies the current AI hype bubble. (As they say, all publicity is good publicity.)

‘Mass unemployment, AI replaces workers’. Rendered with Stable Diffusion 2.

2. A small (but well-funded) portion of the machine-learning community buy into the bizarre religion known as ‘longtermism’, which is obsessed with maximizing the utility of humanity’s imaginary, distant future. (The letter calling for a pause to AI development was published the Future of Life Institute, a longtermist organization.)

   According to the longtermists, humans will inevitably colonize the entire galaxy and have a population numbering in the hundreds of trillions. So far, it sounds kind of like Star Trek. But the catch is the appeal to utilitarianism. Risks, the longtermists conclude, should be evaluated in terms of their effect on humanity’s long-term utility. So yeah, the risk of AI conquering humanity is extremely minute. But if it *did* happen, the longtermists claim, it would be really bad for aggregate future utility. Therefore, rogue AI should be our number one concern.[3]

3. It seems that many machine-learning researchers buy into mind-body dualism. In other words, if they build a really powerful piece of software (a mind), they assume a robust, autonomous body might somehow fol-

[3] Like a good ideology, the nice feature of longtermism is that the numbers are entirely imaginary. In other words, you can take any problem you want and invent numbers to make it look like it should be humanity’s number one concern.

low. I'm sorry to say that the real-world doesn't work that way. Whether biological or synthetic, the mind is an extension of the body. And if the body is not self-sustaining or self-replicating, neither is the mind.

4. Tech moguls like Sam Altman (the CEO of OpenAI) likely have a cynical agenda. When Altman hypes the risks of AI and calls for government regulation, what he really wants is to build a moat around his technology. If AI is tightly regulated, it means that the big players will have a built in advantage. They can pay for the various 'certifications' and that ensure their technology is 'safe'. The small companies won't be able to compete. In short, when the titans of industry call for government intervention, it's almost surely self serving.[4]

Despite the hype bubble, I'm excited about the potential for machine-learning models be useful tools. But let's keep in mind that they are just that: *tools*.

Just as there is no risk that your calculator will conquer the world, there is no risk that machine-learning models will do the same. Whether it's a hammer, a nuclear weapon, or a chatbot, the real risk with human-built tools is what we choose to do with them.

Support this blog

Hi folks. I'm a crowdfunded scientist who shares all of his (painstaking) research for free. If you think my work has value, consider becoming a supporter.

Become an ETD supporter

[4] For a good discussion about the self-serving nature of Sam Altman's call for AI regulation, check out Chris Fisher and Michael Dominick's take in Coder Radio episode 519: Not So OpenAI.