https://jacobin.com/2024/01/can-humanity-survive-ai

Can Humanity Survive AI?

By Garrison Lovely 22-01-2024

With the development of artificial intelligence racing forward at warp speed, some of the richest men in the world may be deciding the fate of humanity right now.

Google cofounder Larry Page thinks superintelligent AI is "just the next step in evolution." In fact, Page, who's worth about $120 billion, has reportedly argued that efforts to prevent AI-driven extinction and protect human consciousness are "speciesist" and "sentimental nonsense."

In July, former Google DeepMind senior scientist Richard Sutton — one of the pioneers of reinforcement learning, a major subfield of AI — said that the technology "could displace us from existence," and that "we should not resist succession." In a 2015 talk, Sutton said, suppose "everything fails" and AI "kill us all"; he asked, "Is it so bad that humans are not the final form of intelligent life in the universe?"

"Biological extinction, that's not the point," Sutton, sixty-six, told me. "The light of humanity and our understanding, our intelligence — our consciousness, if you will — can go on without meat humans."

Yoshua Bengio, fifty-nine, is the second-most cited living scientist, noted for his foundational work on deep learning. Responding to Page and Sutton, Bengio told me, "What they want, I think it's playing dice with humanity's future. I personally think this should be criminalized." A bit surprised, I asked what exactly he wanted outlawed, and he said efforts to build "AI systems that could overpower us and have their own self-interest by design." In May, Bengio began writing and speaking about how advanced AI systems might go rogue and pose an extinction risk to humanity.

Bengio posits that future, genuinely human-level AI systems could improve their own capabilities, functionally creating a new, more intelligent species. Humanity has driven hundreds of other species extinct, largely by accident. He fears that we could be next — and he isn't alone.

Bengio shared the 2018 Turing Award, computing's Nobel Prize, with fellow deep learning pioneers Yann LeCun and Geoffrey Hinton. Hinton, the most cited living scientist, made waves in May when he resigned from his senior role at Google to more freely sound off about the possibility that future AI systems could wipe out humanity. Hinton and Bengio are the two most prominent AI researchers to join the "x-risk" community. Sometimes referred to as AI safety advocates or doomers, this loose-knit group worries that AI poses an existential risk to humanity.

In the same month that Hinton resigned from Google, hundreds of AI researchers and notable figures signed an open letter stating, "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." Hinton and Bengio were the lead signatories, followed by OpenAI CEO Sam Altman and the heads of other top AI labs.

Hinton and Bengio were also the first authors of an October position paper warning about the risk of "an irreversible loss of human control over autonomous AI systems," joined by famous academics like Nobel laureate Daniel Kahneman and Sapiens author Yuval Noah Harari.

LeCun, who runs AI at Meta, agrees that human-level AI is coming but said in a public debate against Bengio on AI extinction, "If it's dangerous, we won't build it."

Deep learning powers the most advanced AI systems in the world, from DeepMind's protein-folding model to large language models (LLMs) like OpenAI's ChatGPT. No one really understands how deep learning systems work, but their performance has continued to improve nonetheless. These systems aren't designed to function according to a set of well-understood principles but are instead "trained" to analyze patterns in large datasets, with complex behavior — like language understanding — emerging as a consequence. AI developer Connor Leahy told me, "It's more like we're poking something in a Petri dish" than writing a piece of code. The October position paper warns that "no one currently knows how to reliably align AI behavior with complex values."

In spite of all this uncertainty, AI companies see themselves as being in a race to make these systems as powerful as they can — without a workable plan to understand how the things they're creating actually function, all while cutting corners on safety to win more market share. Artificial general intelligence (AGI) is the holy grail that leading AI labs are explicitly working toward. AGI is often defined as a system that is at least as good as humans at almost any intellectual task. It's also the thing that Bengio and Hinton believe could lead to the end of humanity.

Bizarrely, many of the people actively advancing AI capabilities think there's a significant chance that doing so will ultimately cause the apocalypse. A 2022 survey of machine learning researchers found that nearly half of them thought there was at least a 10 percent chance advanced AI could lead to "human extinction or similarly permanent and severe disempowerment" of humanity. Just months before he cofounded OpenAI, Altman said, "AI will probably most likely lead to the end of the world, but in the meantime, there'll be great companies."

Public opinion on AI has soured, particularly in the year since ChatGPT was released. In all but one 2023 survey, more Americans than not have thought that AI could pose an existential threat to humanity. In the rare instances when pollsters asked people if they wanted human-level or beyond AI, strong majorities in the United States and the UK said they didn't.

So far, when socialists weigh in on AI, it's usually to highlight AI-powered discrimination or to warn about the potentially negative impact of automation in a world of weak unions and powerful capitalists. But the Left has been conspicuously quiet about Hinton and Bengio's nightmare scenario — that advanced AI could kill us all.

Worrying Capabilities

Illustration by Ricardo Santos

While much of the attention from the x-risk community focuses on the idea that humanity could eventually lose control of AI, many are also worried about less capable systems empowering bad actors on very short timelines.

Thankfully, it's hard to make a bioweapon. But that might change soon.

Anthropic, a leading AI lab founded by safety-forward ex-OpenAI staff, recently worked with biosecurity experts to see how much an LLM could help an aspiring bioterrorist. Testifying before a Senate subcommittee in July, Anthropic CEO Dario Amodei reported that certain steps in bioweapons production can't be found in textbooks or search engines, but that "today's AI tools can fill in some of these steps, albeit incompletely," and that "a straightforward extrapolation of today's systems to those we expect to see in two to three years suggests a substantial risk that AI systems will be able to fill in all the missing pieces."

In October, New Scientist reported that Ukraine made the first battlefield use of lethal autonomous weapons (LAWs) — literally killer robots. The United States, China, and Israel are developing their own LAWs. Russia has joined the United States and Israel in opposing new international law on LAWs.

However, the more expansive idea that AI poses an existential risk has many critics, and the roiling AI discourse is hard to parse: equally credentialed people make opposite claims about whether AI x-risk is real, and venture capitalists are signing open letters with progressive AI ethicists. And while the x-risk idea seems to be gaining ground the fastest, a major publication runs an essay seemingly every week arguing that x-risk distracts from existing harms. Meanwhile, orders of magnitude more money and people are quietly dedicated to making AI systems more powerful than to making them safer or less biased.

Some fear not the "sci-fi" scenario where AI models get so capable they wrest control from our feeble grasp, but instead that we will entrust biased, brittle, and confabulating systems with too much responsibility, opening a more pedestrian Pandora's box full of awful but familiar problems that scale with the algorithms causing them. This community of researchers and advocates — often labeled "AI ethics" — tends to focus on the immediate harms being wrought by AI, exploring solutions involving model accountability, algorithmic transparency, and machine learning fairness.

I spoke with some of the most prominent voices from the AI ethics community, like computer scientists Joy Buolamwini, thirty-three, and Inioluwa Deborah Raji, twenty-seven. Each has conducted pathbreaking research into existing harms caused by discriminatory and flawed AI models whose impacts, in their view, are obscured one day and overhyped the next. Like that of many AI ethics researchers, their work blends science and activism.

Those I spoke to within the AI ethics world largely expressed a view that, rather than facing fundamentally new challenges like the prospect of complete technological unemployment or extinction, the future of AI looks more like intensified racial

discrimination in incarceration and loan decisions, the Amazon warehouse-ification of workplaces, attacks on the working poor, and a further entrenched and enriched techno-elite.

Illustration by Ricardo Santos

A frequent argument from this crowd is that the extinction narrative overhypes the capabilities of Big Tech's products and dangerously "distracts" from AI's immediate harms. At best, they say, entertaining the x-risk idea is a waste of time and money. At worst, it leads to disastrous policy ideas.

But many of the x-risk believers highlighted that the positions "AI causes harm now" and "AI could end the world" are not mutually exclusive. Some researchers have tried explicitly to bridge the divide between those focused on existing harms and those focused on extinction, highlighting potential shared policy goals. AI professor Sam Bowman, another person whose name is on the extinction letter, has done research to reveal and reduce algorithmic bias and reviews submissions to the main AI ethics conference. Simultaneously, Bowman has called for more researchers to work on AI safety and wrote of the "dangers of underclaiming" the abilities of LLMs.

The x-risk community commonly invokes climate advocacy as an analogy, asking whether the focus on reducing the long-term harms of climate change dangerously distracts from the near-term harms from air pollution and oil spills.

But by their own admission, not everyone from the x-risk side has been as diplomatic. In an August 2022 thread of spicy AI policy takes, Anthropic cofounder Jack Clark tweeted that "Some people who work on long-term/AGI-style policy tend to ignore, minimize, or just not consider the immediate problems of AI deployment/harms."

"AI Will Save the World"

A third camp worries that when it comes to AI, we're not actually moving fast enough. Prominent capitalists like billionaire Marc Andreessen agree with safety folks that AGI is possible but argue that, rather than killing us all, it will usher in an indefinite golden age of radical abundance and borderline magical technologies. This group, largely coming from Silicon Valley and commonly referred to as AI boosters, tends to worry far more that regulatory overreaction to AI will smother a transformative, world-saving technology in its crib, dooming humanity to economic stagnation.

Some techno-optimists envision an AI-powered utopia that makes Karl Marx seem unimaginative. The Guardian recently released a mini-documentary featuring interviews from 2016 through 2019 with OpenAI's chief scientist, Ilya Sutskever, who boldly pronounces: "AI will solve all the problems that we have today. It will solve employment, it will solve disease, it will solve poverty. But it will also create new problems."

Andreessen is with Sutskever — right up until the "but." In June, Andreessen published an essay called "Why AI Will Save the World," where he explains how AI will make "everything we care about better," as long as we don't regulate it to death. He

followed it up in October with his "Techno-Optimist Manifesto," which, in addition to praising a founder of Italian fascism, named as enemies of progress ideas like "existential risk," "sustainability," "trust and safety," and "tech ethics." Andreessen does not mince words, writing, "We believe any deceleration of AI will cost lives. Deaths that were preventable by the AI that was prevented from existing a form of murder."

Andreessen, along with "pharma bro" Martin Shkreli, is perhaps the most famous proponent of "effective accelerationism," also called "e/acc," a mostly online network that mixes cultish scientism, hypercapitalism, and the naturalistic fallacy. E/acc, which went viral this summer, builds on reactionary writer Nick Land's theory of accelerationism, which argues that we need to intensify capitalism to propel ourselves into a posthuman, AI-powered future. E/acc takes this idea and adds a layer of physics and memes, mainstreaming it for a certain subset of Silicon Valley elites. It was formed in reaction to calls from "decels" to slow down AI, which have come significantly from the effective altruism (EA) community, from which e/acc takes its name.

AI booster Richard Sutton — the scientist ready to say his goodbyes to "meat humans" — is now working at Keen AGI, a new start-up from John Carmack, the legendary programmer behind the 1990s video game Doom. The company mission, according to Carmack: "AGI or bust, by way of Mad Science!"

Capitalism Makes It Worse

In February, Sam Altman tweeted that Eliezer Yudkowsky might eventually "deserve the Nobel Peace Prize." Why? Because Altman thought the autodidactic researcher and Harry Potter fan-fiction author had done "more to accelerate AGI than anyone else." Altman cited how Yudkowsky helped DeepMind secure pivotal early-stage funding from Peter Thiel as well as Yudkowsky's "critical" role "in the decision to start OpenAI."

Yudkowsky was an accelerationist before the term was even coined. At the age of seventeen — fed up with dictatorships, world hunger, and even death itself — he published a manifesto demanding the creation of a digital superintelligence to "solve" all of humanity's problems. Over the next decade of his life, his "technophilia" turned to phobia, and in 2008 he wrote about his conversion story, admitting that "to say, I almost destroyed the world!, would have been too prideful."

Yudkowsky is now famous for popularizing the idea that AGI could kill everyone, and he has become the doomiest of the AI doomers. A generation of techies grew up reading Yudkowsky's blog posts, but more of them (perhaps most consequentially, Altman) internalized his arguments that AGI would be the most important thing ever than his beliefs about how hard it would be to get it not to kill us.

During our conversation, Yudkowsky compared AI to a machine that "prints gold," right up until it "ignite the atmosphere."

And whether or not it will ignite the atmosphere, that machine is printing gold faster than ever. The "generative AI" boom is making some people very, very rich. Since 2019, Microsoft has invested a cumulative $13 billion into OpenAI. Buoyed by the wild success of ChatGPT, Microsoft gained nearly $1 trillion in value in the year following

the product's release. Today the nearly fifty-year-old corporation is worth more than Google and Meta combined.

Profit-maximizing actors will continue barreling forward, externalizing risks the rest of us never agreed to bear, in the pursuit of riches — or simply the glory of creating digital superintelligence, which Sutton tweeted "will be the greatest intellectual achievement of all time … whose significance is beyond humanity, beyond life, beyond good and bad." Market pressures will likely push companies to transfer more and more power and autonomy to AI systems as they improve.

One Google AI researcher wrote to me, "I think big corps are in such a rush to win market share that safety is seen as a kind of silly distraction." Bengio told me he sees "a dangerous race between companies" that could get even worse.

Panicking in response to the OpenAI-powered Bing search engine, Google declared a "code red," "recalibrate" their risk appetite, and rushed to release Bard, their LLM, over staff opposition. In internal discussions, employees called Bard "a pathological liar" and "cringe-worthy." Google published it anyway.

Dan Hendrycks, the director of the Center for AI Safety, said that "cutting corners on safety . . . is largely what AI development is driven by. . . . I don't think, actually, in the presence of these intense competitive pressures, that intentions particularly matter." Ironically, Hendrycks is also the safety adviser to xAI, Elon Musk's latest venture.

The three leading AI labs all began as independent, mission-driven organizations, but they are now either full subsidiaries of tech behemoths (Google DeepMind) or have taken on so many billions of dollars in investment from trillion-dollar companies that their altruistic missions may get subsumed by the endless quest for shareholder value (Anthropic has taken up to $6 billion from Google and Amazon combined, and Microsoft's $13 billion bought them 49 percent of OpenAI's for-profit arm). The New York Times recently reported that DeepMind's founders became "increasingly worried about what Google would do with their inventions. In 2017, they tried to break away from the company. Google responded by increasing the salaries and stock award packages of the DeepMind founders and their staff. They stayed put."

One developer at a leading lab wrote to me in October that, since the leadership of these labs typically truly believes AI will obviate the need for money, profit-seeking is "largely instrumental" for fundraising purposes. But "then the investors (whether it's a VC firm or Microsoft) exert pressure for profit-seeking."

Between 2020 and 2022, more than $600 billion in corporate investment flowed into the industry, and a single 2021 AI conference hosted nearly thirty thousand researchers. At the same time, a September 2022 estimate found only four hundred full-time AI safety researchers, and the primary AI ethics conference had fewer than nine hundred attendees in 2023.

The way software "ate the world," we should expect AI to exhibit a similar winner-takes-all dynamic that will lead to even greater concentrations of wealth and power. Altman has predicted that the "cost of intelligence" will drop to near zero as a result of AI, and in 2021 he wrote that "even more power will shift from labor to capital." He

continued, “If public policy doesn’t adapt accordingly, most people will end up worse off than they are today.” Also in his “spicy take” thread, Jack Clark wrote, “economy-of-scale capitalism is, by nature, anti-democratic, and capex-intensive AI is therefore anti-democratic.”

Markus Anderljung is the policy chief at GovAI, a leading AI safety think tank, and the first author on an influential white paper focused on regulating “frontier AI.” He wrote to me and said, “If you’re worried about capitalism in its current form, you should be even more worried about a world where huge parts of the economy are run by AI systems explicitly trained to maximize profit.”

Sam Altman, circa June 2021, agreed, telling Ezra Klein about the founding philosophy of OpenAI: “One of the incentives that we were very nervous about was the incentive for unlimited profit, where more is always better. . . . And I think with these very powerful general purpose AI systems, in particular, you do not want an incentive to maximize profit indefinitely.”

In a stunning move that has become widely seen as the biggest flash point in the AI safety debate so far, Open-AI’s nonprofit board fired CEO Sam Altman on November 17, 2023, the Friday before Thanksgiving. The board, per OpenAI’s unusual charter, has a fiduciary duty to “humanity,” rather than to investors or employees. As justification, the board vaguely cited Altman’s lack of candor but then ironically largely kept quiet about its decision.

Around 3 a.m. the following Monday, Microsoft announced that Altman would be spinning up an advanced research lab with positions for every OpenAI employee, the vast majority of whom signed a letter threatening to take Microsoft’s offer if Altman wasn’t reinstated. (While he appears to be a popular CEO, it’s worth noting that the firing disrupted a planned sale of OpenAI’s employee-owned stock at a company valuation of $86 billion.) Just after 1 a.m. on Wednesday, OpenAI announced Altman’s return as CEO and two new board members: the former Twitter board chair, and former Treasury secretary Larry Summers.

Within less than a week, OpenAI executives and Altman had collaborated with Microsoft and the company’s staff to engineer his successful return and the removal of most of the board members behind his firing. Microsoft’s first preference was having Altman back as CEO. The unexpected ouster initially sent the legacy tech giant’s stock plunging 5 percent ($140 billion), and the announcement of Altman’s reinstatement took it to an all-time high. Loath to be “blindsided” again, Microsoft is now taking a nonvoting seat on the nonprofit board.

Immediately after Altman’s firing, X exploded, and a narrative largely fueled by online rumors and anonymously sourced articles emerged that safety-focused effective altruists on the board had fired Altman over his aggressive commercialization of OpenAI’s models at the expense of safety. Capturing the tenor of the overwhelming e/acc response, then pseudonymous founder @BasedBeffJezos posted, “EAs are basically terrorists. Destroying 80B of value overnight is an act of terrorism.”

The picture that emerged from subsequent reporting was that a fundamental mistrust of Altman, not immediate concerns about AI safety, drove the board’s choice. The Wall

Street Journal found that "there wasn't one incident that led to their decision to eject Altman, but a consistent, slow erosion of trust over time that made them increasingly uneasy."

Weeks before the firing, Altman reportedly) used dishonest tactics to try to remove board member Helen Toner over an academic paper she coauthored that he felt was critical of OpenAI's commitment to AI safety. In the paper, Toner, an EA-aligned AI governance researcher, lauded Anthropic for avoiding "the kind of frantic corner-cutting that the release of ChatGPT appeared to spur."

The New Yorker reported that "some of the board's six members found Altman manipulative and conniving." Days after the firing, a DeepMind AI safety researcher who used to work for OpenAI wrote that Altman "lied to me on various occasions" and "was deceptive, manipulative, and worse to others," an assessment echoed by recent reporting in Time.

This wasn't Altman's first time being fired. In 2019, Y Combinator founder Paul Graham removed Altman from the incubator's helm over concerns that he was prioritizing his own interests over those of the organization. Graham has previously said, "Sam is extremely good at becoming powerful."

OpenAI's strange governance model was established specifically to prevent the corrupting influence of profit-seeking, but as the Atlantic rightly proclaimed, "the money always wins." And more money than ever is going into advancing AI capabilities.

Full Speed Ahead

Recent AI progress has been driven by the culmination of many decades-long trends: increases in the amount of computing power (referred to as "compute") and data used to train AI models, which themselves have been amplified by significant improvements in algorithmic efficiency. Since 2010, the amount of compute used to train AI models has increased roughly *one-hundred-millionfold*. Most of the advances we're seeing now are the product of what was at the time a much smaller and poorer field.

And while the last year has certainly contained more than its fair share of AI hype, the confluence of these three trends has led to quantifiable results. The time it takes AI systems to achieve human-level performance on many benchmark tasks has shortened dramatically in the last decade.

It's possible, of course, that AI capability gains will hit a wall. Researchers may run out of good data to use. Moore's law — the observation that the number of transistors on a microchip doubles every two years — will eventually become history. Political events could disrupt manufacturing and supply chains, driving up compute costs. And scaling up systems may no longer lead to better performance.

But the reality is that no one knows the true limits of existing approaches. A clip of a January 2022 Yann LeCun interview resurfaced on Twitter this year. LeCun said, "I don't think we can train a machine to be intelligent purely from text, because I think the amount of information about the world that's contained in text is tiny compared to what we need to know." To illustrate his point, he gave an example: "I take an object, I

put it on the table, and I push the table. It's completely obvious to you that the object would be pushed with the table." However, with "a text-based model, if you train a machine, as powerful as it could be, your 'GPT-5000' . . . it's never gonna learn about this."

But if you give ChatGPT-3.5 that example, it instantly spits out the correct answer.

In an interview published four days before his firing, Altman said, "Until we go train that model , it's like a fun guessing game for us. We're trying to get better at it, because I think it's important from a safety perspective to predict the capabilities. But I can't tell you here's exactly what it's going to do that GPT-4 didn't."

History is littered with bad predictions about the pace of innovation. A New York Times editorial claimed it might take "one million to ten million years" to develop a flying machine — sixty-nine days before the Wright Brothers first flew. In 1933, Ernest Rutherford, the "father of nuclear physics," confidently dismissed the possibility of a neutron-induced chain reaction, inspiring physicist Leo Szilard to hypothesize a working solution *the very next day* — a solution that ended up being foundational to the creation of the atomic bomb.

One conclusion that seems hard to avoid is that, recently, the people who are best at building AI systems believe AGI is both possible and imminent. Perhaps the two leading AI labs, OpenAI and DeepMind, have been working toward AGI since their inception, starting when admitting you believed it was possible anytime soon could get you laughed out of the room. (Ilya Sutskever led a chant of "Feel the AGI" at OpenAI's 2022 holiday party.)

Perfect Workers

Employers are already using AI to surveil, control, and exploit workers. But the real dream is to cut humans out of the loop. After all, as Marx wrote, "The machine is a means for producing surplus-value."

Open Philanthropy (OP) AI risk researcher Ajeya Cotra wrote to me that "the logical end point of a maximally efficient capitalist or market economy" wouldn't involve humans because "humans are just very inefficient creatures for making money." We value all these "commercially unproductive" emotions, she writes, "so if we end up having a good time and liking the outcome, it'll be because we started off with the power and shaped the system to be accommodating to human values."

OP is an EA-inspired foundation financed by Facebook cofounder Dustin Moskovitz. It's the leading funder of AI safety organizations, many of which are mentioned in this article. OP also granted $30 million to OpenAI to support AI safety work two years before the lab spun up a for-profit arm in 2019. I previously received a onetime grant to support publishing work at New York Focus, an investigative news nonprofit covering New York politics, from EA Funds, which itself receives funding from OP. After I first encountered EA in 2017, I began donating 10 to 20 percent of my income to global health and anti–factory farming nonprofits, volunteered as a local group organizer, and worked at an adjacent global poverty nonprofit. EA was one of the earliest communities to seriously engage with AI existential risk, but I looked at the AI folks

with some wariness, given the uncertainty of the problem and the immense, avoidable suffering happening now.

A compliant AGI would be the worker capitalists can only dream of: tireless, motivated, and unburdened by the need for bathroom breaks. Managers from Frederick Taylor to Jeff Bezos resent the various ways in which humans aren't optimized for output — and, therefore, their employer's bottom line. Even before the days of Taylor's scientific management, industrial capitalism has sought to make workers more like the machines they work alongside and are increasingly replaced by. As The Communist Manifesto presciently observed, capitalists' extensive use of machinery turns a worker into "an appendage of the machine."

But according to the AI safety community, the very same inhuman capabilities that would make Bezos salivate also make AGI a mortal danger to humans.

Explosion: The Extinction Case

The common x-risk argument goes: once AI systems reach a certain threshold, they'll be able to recursively self-improve, kicking off an "intelligence explosion." If a new AI system becomes smart — or just scaled up — enough, it will be able to permanently disempower humanity.

The October "Managing AI Risks" paper states:

There is no fundamental reason why AI progress would slow or halt when it reaches human-level abilities. . . . Compared to humans, AI systems can act faster, absorb more knowledge, and communicate at a far higher bandwidth. Additionally, they can be scaled to use immense computational resources and can be replicated by the millions.

These features have already enabled superhuman abilities: LLMs can "read" much of the internet in months, and DeepMind's AlphaFold can perform years of human lab work in a few days.

Here's a stylized version of the idea of "population" growth spurring an intelligence explosion: if AI systems rival human scientists at research and development, the systems will quickly proliferate, leading to the equivalent of an enormous number of new, highly productive workers entering the economy. Put another way, if GPT-7 can perform most of the tasks of a human worker and it only costs a few bucks to put the trained model to work on a day's worth of tasks, each instance of the model would be wildly profitable, kicking off a positive feedback loop. This could lead to a virtual "population" of billions or more digital workers, each worth much more than the cost of the energy it takes to run them. Sutskever thinks it's likely that "the entire surface of the earth will be covered with solar panels and data centers."

These digital workers might be able to improve on our AI designs and bootstrap their way to creating "superintelligent" systems, whose abilities Alan Turing speculated in 1951 would soon "outstrip our feeble powers." And, as some AI safety proponents argue, an individual AI model doesn't have to be superintelligent to pose an existential threat; there might just need to be enough copies of it. Many of my sources likened corporations to superintelligences, whose capabilities clearly exceed those of their constituent members.

"Just unplug it," goes the common objection. But once an AI model is powerful enough to threaten humanity, it will probably be the most valuable thing in existence. You might have an easier time "unplugging" the New York Stock Exchange or Amazon Web Services.

A lazy superintelligence may not pose much of a risk, and skeptics like Allen Institute for AI CEO Oren Etzioni, complexity professor Melanie Mitchell, and AI Now Institute managing director Sarah Myers West all told me they haven't seen convincing evidence that AI systems are becoming more autonomous. Anthropic's Dario Amodei seems to agree that current systems don't exhibit a concerning level of agency. However, a completely passive but sufficiently powerful system wielded by a bad actor is enough to worry people like Bengio.

Further, academics and industrialists alike are increasing efforts to make AI models more autonomous. Days prior to his firing, Altman told the Financial Times: "We will make these agents more and more powerful . . . and the actions will get more and more complex from here. . . . The amount of business value that will come from being able to do that in every category, I think, is pretty good."

What's Behind the Hype?

The fear that keeps many x-risk people up at night is not that an advanced AI would "wake up," "turn evil," and decide to kill everyone out of malice, but rather that it comes to see us as an obstacle to whatever goals it does have. In his final book, Brief Answers to the Big Questions, Stephen Hawking articulated this, saying, "You're probably not an evil ant-hater who steps on ants out of malice, but if you're in charge of a hydroelectric green-energy project and there's an anthill in the region to be flooded, too bad for the ants."

Unexpected and undesirable behaviors can result from simple goals, whether it's profit or an AI's reward function. In a "free" market, profit-seeking leads to monopolies, multi-level marketing schemes, poisoned air and rivers, and innumerable other harms.

There are abundant examples of AI systems exhibiting surprising and unwanted behaviors. A program meant to eliminate sorting errors in a list deleted the list entirely. One researcher was surprised to find an AI model "playing dead" to avoid being identified on safety tests.

Yet others see a Big Tech conspiracy looming behind these concerns. Some people focused on immediate harms from AI argue that the industry is actively promoting the idea that their products might end the world, like Myers West of the AI Now Institute, who says she "see the narratives around so-called existential risk as really a play to take all the air out of the room, in order to ensure that there's not meaningful movement in the present moment." Strangely enough, Yann LeCun and Baidu AI chief scientist Andrew Ng purport to agree.

When I put the idea to x-risk believers, they often responded with a mixture of confusion and exasperation. OP's Ajeya Cotra wrote back: "I wish it were less of an industry-associated thing to be concerned about x-risk, because I think it's just really fundamentally, on the merits, a very anti-industry belief to have. . . . If the companies

are building things that are going to kill us all, that's really bad, and they should be restricted very stringently by the law."

GovAI's Markus Anderljung called fears of regulatory capture "a natural reaction for folks to have," but he emphasized that his preferred policies may well harm the industry's biggest players.

One understandable source of suspicion is that Sam Altman is now one of the people most associated with the existential risk idea, but his company has done more than any other to advance the frontier of general-purpose AI.

Additionally, as OpenAI got closer to profitability and Altman got closer to power, the CEO changed his public tune. In a January 2023 Q and A, when asked about his worst-case scenario for AI, he replied, "Lights out for all of us." But while answering a similar question under oath before senators in May, Altman doesn't mention extinction. And, in perhaps his last interview before his firing, Altman said, "I actually don't think we're all going to go extinct. I think it's going to be great. I think we're heading towards the best world ever."

Altman implored Congress in May to regulate the AI industry, but a November investigation found that OpenAI's quasi-parent company Microsoft was influential in the ultimately unsuccessful lobbying to exclude "foundation models" like ChatGPT from regulation by the forthcoming EU AI Act. And Altman did plenty of his own lobbying in the EU, even threatening to pull out of the region if regulations became too onerous (threats he quickly walked back). Speaking on a CEO panel in San Francisco days before his ouster, Altman said that "current models are fine. We don't need heavy regulation here. Probably not even for the next couple of generations."

President Joe Biden's recent "sweeping" executive order on AI seems to agree: its safety test information sharing requirements only affect models larger than any that have likely been trained so far. Myers West called these kinds of "scale thresholds" a "massive carveout." Anderljung wrote to me that regulation should scale with a system's capabilities and usage, and said that he "would like some regulation of today's most capable and widely used models," but he thinks it will "be a lot more politically viable to impose requirements on systems that are yet to be developed."

Inioluwa Deborah Raji ventured that if the tech giants "know that they have to be the bad guy in some dimension . . . they would prefer for it to be abstract and long-term in timeline." This sounds far more plausible to me than the idea that Big Tech actually wants to promote the idea that their products have a decent chance of *literally killing everyone*.

Nearly seven hundred people signed the extinction letter, the majority of them academics. Only one of them runs a publicly traded company: OP funder Moskovitz, who is also cofounder and CEO of Asana, a productivity app. There were zero employees from Amazon, Apple, IBM, or any leading AI hardware firms. No Meta executives signed.

If the heads of the Big Tech firms wanted to amplify the extinction narrative, why haven't they added their names to the list?

Why Build the "Doom Machine?"

If AI actually does save the world, whoever created it may hope to be lauded like a modern Julius Caesar. And even if it doesn't, whoever first builds "the last invention that man need ever make" will not have to worry about being forgotten by history — unless, of course, history ends abruptly after their invention.

Connor Leahy thinks that, on our current path, the end of history will shortly follow the advent of AGI. With his flowing hair and unkempt goatee, he would probably look at home wearing a sandwich board reading "The end is nigh" — though that hasn't prevented him from being invited to address the British House of Lords or CNN. The twenty-eight-year-old CEO of Conjecture and cofounder of EleutherAI, an influential open-source collective, told me that a lot of the motivation to build AI boils down to: "'Oh, you're building the ultimate doom machine that makes you billions of dollars and also king-emperor of earth or kills everybody?' Yeah, that's like the masculine dream. You're like, 'Fuck yeah. I am the doom king.'" He continues, "Like, I get it. This is very much in the Silicon Valley aesthetic."

Leahy also conveyed some-thing that won't surprise people who have spent significant time in the Bay Area or certain corners of the internet:

There are actual, completely unaccountable, unelected, techno-utopian businesspeople and technologists, living mostly in San Francisco, who are willing to risk the lives of you, your children, your grandchildren, and all of future humanity just because they might have a chance to live forever.

In March, the MIT Technology Review reported that Altman "says he's emptied his bank account to fund two . . . goals: limitless energy and extended life span."

Given all this, you might expect the ethics community to see the safety community as a natural ally in a common struggle to reign in unaccountable tech elites who are unilaterally building risky and harmful products. And, as we saw earlier, many safety advocates have made overtures to the AI ethicists. It's also rare for people from the x-risk community to publicly attack AI ethics (while the reverse is . . . not true), but the reality is that safety proponents have sometimes been hard to stomach.

AI ethicists, like the people they advocate for, often report feeling marginalized and cut off from real power, fighting an uphill battle with tech companies who see them as a way to cover their asses rather than as a true priority. Lending credence to this feeling is the gutting of AI ethics teams at many Big Tech companies in recent years (or days). And, in a number of cases, these companies have retaliated against ethics-oriented whistleblowers and labor organizers.

This doesn't necessarily imply that these companies are instead seriously prioritizing x-risk. Google DeepMind's ethics board, which included Larry Page and prominent existential risk researcher Toby Ord, had its first meeting in 2015, but it never had a second one. One Google AI researcher wrote to me that they "don't talk about long-term risk . . . in the office," continuing, "Google is more focused on building the tech and on safety in the sense of legality and offensiveness."

Software engineer Timnit Gebru co-led Google's ethical AI team until she was forced out of the company in late 2020 following a dispute over a draft paper — now one of the most famous machine learning publications ever. In the "stochastic parrots" paper, Gebru and her coauthors argue that LLMs damage the environment, amplify social biases, and use statistics to "haphazardly" stitch together language "without any reference to meaning."

Gebru, who is no fan of the AI safety community, has called for enhanced whistleblower protections for AI researchers, which are also one of the main recommendations made in GovAI's white paper. Since Gebru was pushed out of Google, nearly 2,700 staffers have signed a solidaristic letter, but then Googler Geoff Hinton was not one of them. When asked on CNN why he didn't support a fellow whistleblower, Hinton replied that Gebru's critiques of AI "were rather different concerns from mine" that "aren't as existentially serious as the idea of these things getting more intelligent than us and taking over."

Raji told me that "a lot of cause for frustration and animosity" between the ethics and safety camps is that "one side has just way more money and power than the other side," which "allows them to push their agenda way more directly."

According to one estimate, the amount of money moving into AI safety start-ups and nonprofits in 2022 quadrupled since 2020, reaching $144 million. It's difficult to find an equivalent figure for the AI ethics community. However, civil society from either camp is dwarfed by industry spending. In just the first quarter of 2023, OpenSecrets reported roughly $94 million was spent on AI lobbying in the United States. LobbyControl estimated tech firms spent €113 million this year lobbying the EU, and we'll recall that hundreds of billions of dollars are being invested in the AI industry as we speak.

One thing that may drive the animosity even more than any perceived difference in power and money is the trend line. Following widely praised books like 2016's Weapons of Math Destruction, by data scientist Cathy O'Neil, and bombshell discoveries of algorithmic bias, like the 2018 "Gender Shades" paper by Buolamwini and Gebru, the AI ethics perspective had captured the public's attention and support.

In 2014, the AI x-risk cause had its own surprise bestseller, philosopher Nick Bostrom's Superintelligence, which argued that beyond-human AI could lead to extinction and earned praise from figures like Elon Musk and Bill Gates. But Yudkowsky told me that, pre-ChatGPT, outside of certain Silicon Valley circles, seriously entertaining the book's thesis would make people look at you funny. Early AI safety proponents like Yudkowsky have occupied the strange position of maintaining close ties to wealth and power through Bay Area techies while remaining marginalized in the wider discourse.

In the post-ChatGPT world, Turing recipients and Nobel laureates are coming out of the AI safety closet and embracing arguments popularized by Yudkowsky, whose best-known publication is a piece of Harry Potter fan fiction totaling more than 660,000 words.

Perhaps the most shocking portent of this new world was broadcast in November, when the hosts of a New York Times tech podcast, Hard Fork, asked the Federal Trade Commission chair: "What is your p(doom), Lina Khan? What is your probability that AI

will kill us all?" EA water cooler talk has gone mainstream. (Khan said she's "an optimist" and gave a "low" estimate of *15 percent*.)

It would be easy to observe all the open letters and media cycles and think that the majority of AI researchers are mobilizing against existential risk. But when I asked Bengio about how x-risk is perceived today in the machine learning community, he said, "Oh, it's changed a lot. It used to be, like, 0.1 percent of people paid attention to the question. And maybe now it's 5 percent."

Probabilities

Like many others concerned about AI x-risk, the renowned philosopher of mind David Chalmers made a probabilistic argument during our conversation: "This is not a situation where you have to be 100 percent certain that we'll have human-level AI to worry about it. If it's 5 percent, that's something we have to worry about."

This kind of statistical thinking is popular in the EA community and is a large part of what led its members to focus on AI in the first place. If you defer to expert arguments, you could end up more confused. But if you try to average the expert concern from the handful of surveys, you might end up thinking there's at least a few-percent chance that AI extinction could happen, which could be enough to make it the most important thing in the world. And if you put any value on all the future generations that could exist, human extinction is categorically worse than survivable catastrophes.

However, in the AI debate, allegations of arrogance abound. Skeptics like Melanie Mitchell and Oren Etzioni told me there wasn't evidence to support the x-risk case, while believers like Bengio and Leahy point to surprising capability gains and ask: What if progress doesn't stop? An academic AI researcher friend has likened the advent of AGI to throwing global economics and politics into a blender.

Even if, for some reason, AGI can only match and not exceed human intelligence, the prospect of sharing the earth with an almost arbitrarily large number of human-level digital agents is terrifying, especially when they'll probably be trying to make someone money.

There are far too many policy ideas about how to reduce existential risk from AI to properly discuss here. But one of the clearer messages coming from the AI safety community is that we should "slow down." Advocates for such a deceleration hope it would give policymakers and broader society a chance to catch up and actively decide how a potentially transformative technology is developed and deployed.

International Cooperation

One of the most common responses to any effort to regulate AI is the "but China!" objection. Altman, for example, told a Senate committee in May that "we want America to lead" and acknowledged that a peril of slowing down is that "China or somebody else makes faster progress."

Anderljung wrote to me that this "isn't a strong enough reason not to regulate AI."

In a June Foreign Affairs article, Helen Toner and two political scientists reported that the Chinese AI researchers they interviewed thought Chinese LLMs are at least two to three years behind the American state-of-the-art models. Further, the authors argue

that since Chinese AI advances "rely a great deal on reproducing and tweaking research published abroad," a unilateral slowdown "would likely decelerate" Chinese progress as well. China has also moved faster than any other major country to meaningfully regulate AI, as Anthropic policy chief Jack Clark has observed.

Yudkowsky says, "It's not actually in China's interest to commit suicide along with the rest of humanity."

If advanced AI really threatens the whole world, domestic regulation alone won't cut it. But robust national restrictions could credibly signal to other countries how seriously you take the risks. Prominent AI ethicist Rumman Chowdhury has called for global oversight. Bengio says we "have to do both."

Yudkowsky, unsurprisingly, has taken a maximalist position, telling me that "the correct direction looks more like putting all of the AI hardware into a limited number of data centers under international supervision by bodies with a symmetric treaty whereby nobody — including the militaries, governments, China, or the CIA — can do any of the really awful things, including building superintelligences."

In a controversial Time op-ed from March, Yudkowsky argued to "shut it all down" by establishing an international moratorium on "new large training runs" backed by the threat of military force. Given Yudkowsky's strong beliefs that advanced AI would be much more dangerous than any nuclear or biological weapon, this radical stance follows naturally.

All twenty-eight countries at the recent AI Safety Summit, including the United States and China, signed the Bletchley Declaration, which recognized existing harms from AI and the fact that "substantial risks may arise from potential intentional misuse or unintended issues of control relating to alignment with human intent."

At the summit, the hosting British government commissioned Bengio to lead production of the first "State of the Science" report on the "capabilities and risks of frontier AI," in a significant step toward a permanent expert body like the Intergovernmental Panel on Climate Change.

Cooperation between the United States and China will be imperative for meaningful international coordination on AI development. And when it comes to AI, the two countries aren't exactly on the best terms. With the 2022 CHIPS Act export controls, the United States tried to kneecap China's AI capabilities, something an industry analyst would have previously considered an "act of war." As Jacobin reported in May, some x-risk-oriented policy researchers likely played a role in passing the onerous controls. In October, the United States tightened CHIPS Act restrictions to close loopholes.

However, in an encouraging sign, Biden and Xi Jinping discussed AI safety and a ban on AI in lethal weapons systems in November. A White House press release stated, "The leaders affirmed the need to address the risks of advanced AI systems and improve AI safety through U.S.-China government talks."

Lethal autonomous weapons are also an area of relative agreement in the AI debates. In her new book Unmasking AI: My Mission to Protect What Is Human in a World of

Machines, Joy Buolamwini advocates for the Stop Killer Robots campaign, echoing a longtime concern of many AI safety proponents. The Future of Life Institute, an x-risk organization, assembled ideological opponents to sign a 2016 open letter calling for a ban on offensive LAWs, including Bengio, Hinton, Sutton, Etzioni, LeCun, Musk, Hawking, and Noam Chomsky.

A Seat at the Table

After years of inaction, the world's governments are finally turning their attention to AI. But by not seriously engaging with what future systems could do, socialists are ceding their seat at the table.

In no small part because of the types of people who became attracted to AI, many of the earliest serious adopters of the x-risk idea decided to either engage in extremely theoretical research on how to control advanced AI or started AI companies. But for a different type of person, the response to believing that AI could end the world is to try to *get people to stop building it*.

Boosters keep saying that AI development is inevitable — and if enough people believe it, it becomes true. But "there is nothing about artificial intelligence that is inevitable," writes the AI Now Institute. Managing director Myers West echoed this, mentioning that facial recognition technology looked inevitable in 2018 but has since been banned in many places. And as x-risk researcher Katja Grace points out, we shouldn't feel the need to build every technology simply because we can.

Additionally, many policymakers are looking at recent AI advances and *freaking out*. Senator Mitt Romney is "more terrified about AI" than optimistic, and his colleague Chris Murphy says, "The consequences of so many human functions being outsourced to AI is potentially disastrous." Congresspeople Ted Lieu and Mike Johnson are literally "freaked out" by AI. If certain techies are the only people willing to acknowledge that AI capabilities have dramatically improved and could pose a species-level threat in the future, that's who policymakers will disproportionately listen to. In May, professor and AI ethicist Kristian Lum tweeted: "There's one existential risk I'm certain LLMs pose and that's to the credibility of the field of FAccT / Ethical AI if we keep pushing the snake oil narrative about them."

Even if the idea of AI-driven extinction strikes you as more fi than sci, there could still be enormous impact in influencing how a transformative technology is developed and what values it represents. Assuming we can get a hypothetical AGI to do what we want raises perhaps the most important question humanity will ever face: What should we *want* it to want?

When I asked Chalmers about this, he said, "At some point we recapitulate all the questions of political philosophy: What kind of society do we actually want and actually value?"

One way to think about the advent of human-level AI is that it would be like creating a new country's constitution (Anthropic's "constitutional AI" takes this idea literally, and the company recently experimented with incorporating democratic input into its model's foundational document). Governments are complex systems that wield

enormous power. The foundation upon which they're established can influence the lives of millions now and in the future. Americans live under the yoke of dead men who were so afraid of the public, they built antidemocratic measures that continue to plague our political system more than two centuries later.

AI may be more revolutionary than any past innovation. It's also a uniquely normative technology, given how much we build it to reflect our preferences. As Jack Clark recently mused to Vox, "It's a real weird thing that this is not a government project." Chalmers said to me, "Once we suddenly have the tech companies trying to build these goals into AI systems, we have to really trust the tech companies to get these very deep social and political questions right. I'm not sure I do." He emphasized, "You're not just in technical reflection on this but in social and political reflection."

False Choices

We may not need to wait to find superintelligent systems that don't prioritize humanity. Superhuman agents ruthlessly optimize for a reward at the expense of anything else we might care about. The more capable the agent and the more ruthless the optimizer, the more extreme the results.

Sound familiar? If so, you're not alone. The AI Objectives Institute (AOI) looks at both capitalism and AI as examples of misaligned optimizers. Cofounded by former public radio show host Brittney Gallagher and "privacy hero" Peter Eckersley shortly before his unexpected death, the research lab examines the space between annihilation and utopia, "a continuation of existing trends of concentration of power in fewer hands — super-charged by advancing AI — rather than a sharp break with the present." AOI president Deger Turan told me, "Existential risk is failure to coordinate in the face of a risk." He says that "we need to create bridges between" AI safety and AI ethics.

One of the more influential ideas in x-risk circles is the unilateralist's curse, a term for situations in which a lone actor can ruin things for the whole group. For example, if a group of biologists discovers a way to make a disease more deadly, it only takes one to publish it. Over the last few decades, many people have become convinced that AI could wipe out humanity, but only the most ambitious and risk-tolerant of them have started the companies that are now advancing the frontier of AI capabilities, or, as Sam Altman recently put it, pushing the "veil of ignorance back." As the CEO alludes, we have no way of truly knowing what lies beyond the technological limit.

Some of us fully understand the risks but plow forward anyway. With the help of top scientists, ExxonMobil had discovered conclusively by 1977 that their product caused global warming. They then lied to the public about it, all while building their oil platforms higher.

The idea that burning carbon could warm the climate was first hypothesized in the late nineteenth century, but the scientific consensus on climate change took nearly one hundred years to form. The idea that we could permanently lose control to machines is older than digital computing, but it remains far from a scientific consensus. And if recent AI progress continues at pace, we may not have decades to form a consensus before meaningfully acting.

The debate playing out in the public square may lead you to believe that we have to choose between addressing AI's immediate harms and its inherently speculative existential risks. And there are certainly trade-offs that require careful consideration.

But when you look at the material forces at play, a different picture emerges: in one corner are trillion-dollar companies trying to make AI models more powerful and profitable; in another, you find civil society groups trying to make AI reflect values that routinely clash with profit maximization.

In short, it's capitalism versus humanity.