The ideologies fighting for the soul (and future) of AI

EAs, e/accs, decels, and doomers.

CHARLIE GUO
DEC 7, 2023

Artwork created with Midjourney.

During Sam Altman's recent power struggle at OpenAI, there were a lot of questions and very few answers. The biggest mystery - which still hasn't been resolved - was *why* the board felt it had to take such drastic action against the CEO. And without a satisfying answer, plenty of ~~rumors~~ narratives emerged to fill the gaps.

One of the loudest painted a picture of a split within the startup - on one side, a faction of AI safety proponents, still devoutly adhering to the non-profit's mission; on the other, a mob of employees excited about the commercial potential of

ChatGPT and the chance to keep building cutting edge AI. Or, more succinctly - EAs (effective altruists) vs e/accs (effective accelerationists). In this narrative, the EAs (the board) were against the continued commercialization of ChatGPT and were willing to burn down the organization rather than hand it over to the e/accs [1].

We still don't know the whole story. But it's clear that an ideological war *is* happening: if not inside of OpenAI, then certainly on social media and internet forums. And more importantly, in the hearts and minds of founders, researchers, and politicians - who are all shaping the future of AI.

Both EAs and e/accs are backed by VCs with hundreds of millions of dollars. While EAs have already had a meaningful impact on government policy, e/accs are still operating in the digital world. That said, they're very quickly accruing cultural capital online - and building a movement behind it.

Knowing what each side believes and is fighting for gives us a peek into what direction things might go. But to understand how we got here, we have to rewind a bit to understand the history of these movements. We'll be doing a lot of oversimplification, though I've linked to plenty of other deep dives if you want to go down the rabbit hole.

Let's start at the beginning: AI safety.

A brief history of AI safety

For as long as we've conceived of robots, we've imagined them harming us. Isaac Asimov coined his "Three Laws of Robotics" back in 1942 - principles of robot behavior that (in theory) would keep humans safe [2]. But in the last ten years, AI safety has taken major steps forward and drawn enormous interest. Even before the launch of ChatGPT and our current AI hype-cycle, the field had advanced significantly to keep up with new ML breakthroughs.

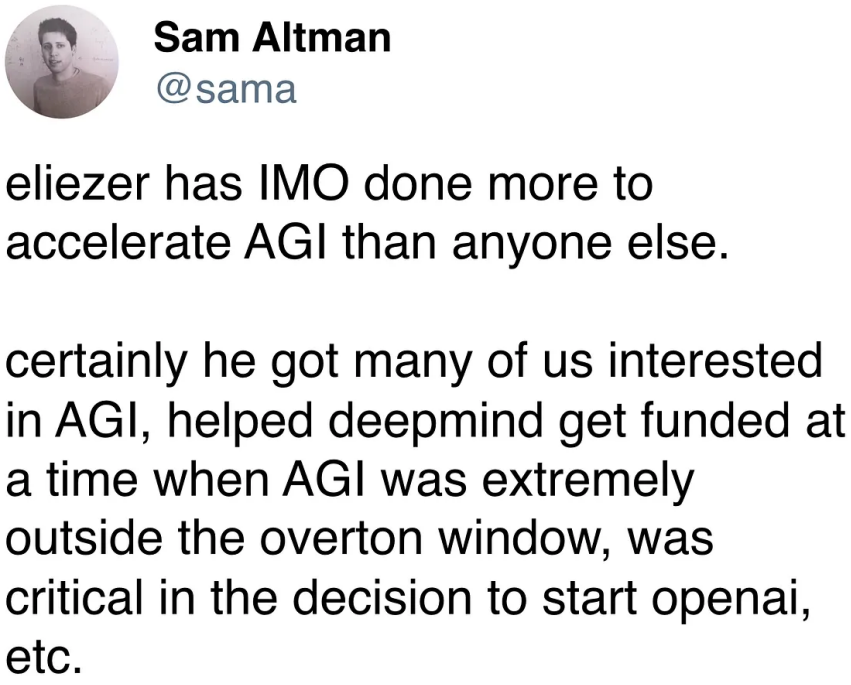
Over time, the AI safety movement has grown beyond technical challenges to include ethical and societal considerations. Philosopher Nick Bostrom published "Superintelligence," which increased concerns about rogue AI ending civilization and attracted interest from Elon Musk and Stephen Hawking. In recent years, governments have also gotten involved - the EU created the European AI Alliance,

and the Biden Administration has issued an executive order and established the US AI Safety Institute.

Today, it's not particularly controversial to say that you're in favor of building safe AI - even the CEOs of leading AI companies have said they're worried about AI's potential harms. When you dig into the details, though, AI safety proponents come in many flavors.

At one end are the "doomers" - prominent figures like Nick Bostrom, Eliezer Yudkowsky, and Yoshua Bengio, who believe that AI poses an existential threat to humanity and should be tightly regulated. However, “AI apocalypse” view isn’t fringe anymore - even Sam Altman has said he believes that AI, if unchecked, may become an extinction risk. What separates the doomers are their calls for pauses and regulation - and they’ve collected a lot of followers.

Yudkowsky in particular has had a significant impact on AI safety, primarily through his research, writings, and public advocacy. For years, he has been researching and writing about the downsides of AI, and believes we’re sprinting headfirst into disaster. While many disagree with his conclusions, he has undoubtedly inspired a wave of AI researchers and founders through his work.



But there's a pretty long list of people who believe that AI is capable of harm - myself included - even if they don't think it will wipe us out. Unfortunately, much of the online discussion around AI safety tends to get flattened into "doomers vs everyone else." I wish there was a shorthand way to say "I'd like more safety guardrails for AI" that doesn't get read as "I want the government to lock down AI research so we don't all get turned into paperclips."

The AI apocalypse isn't what you should be worried about
CHARLIE GUO • MAY 18
Read full story →

And in recent years, many of those concerned about AI safety, doomer or not, would become part of a different movement - Effective Altruism.

Better charity through math

Effective Altruism (EA) is an ideological movement that uses evidence and reason to identify the most effective ways to benefit others.

In the words of one of its founders, William MacAskill:

> Effective altruism is about trying to use your time and money as well as possible to help other people. The core idea is very simple: Imagine you could save five people from drowning or you could save one person from drowning. It's a tragic situation and you can't save both. Well, it's pretty commonsensical that in such a situation, you ought to save the five rather than the one, because there are five times as many interests at stake.
>
> It just so turns out that that drowning-person thought experiment is not really a thought experiment; it's exactly the situation we're in now. If you give to one charity, for example, you can save one or two lives, but if you give the same amount of money to another charity, you might save tens of lives.
>
> There's this super-difficult question, which is, "Of all those things we could do, what are the best ways of doing good? What are the causes that actually can help the most people?" We use the best data we have to figure out what those causes are and then take action and make as much progress on them as possible.

If you've vaguely heard of Effective Altruism before, it may have been in the aftermath of FTX's collapse - Sam Bankman-Fried was a leading figure in the EA movement, until he wasn't. But well before that, a growing number of EAs became obsessed with the problem of AI safety - particularly the potential for AGI to destroy civilization.

It was, from their perspective, rational. From the inimitable Matt Levine:

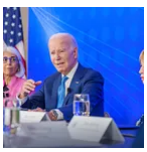
> You try to evaluate charitable projects based on how many lives they will save, and then you give all your money to save the most lives (anywhere in the world) rather than to, say, get your name on your city's art museum.
>
> Some people in the movement decided to extend the causal chain just a bit: Spending $1 million to buy mosquito nets in impoverished villages might save hundreds of lives, but spending $1 million on salaries for vaccine researchers in rich countries has a 20% chance of saving thousands of lives, so it is more valuable.

> You can keep extending the causal chain: Spending $1 million on salaries for artificial intelligence alignment researchers in California has a 1% chance of preventing human extinction at the hands of robots, saving billions of lives — trillions, really, when you count all *future* humans — so it is more valuable than anything else you could do.
>
> I made up that 1% number but, man, that number is going to be made up no matter what. Just make up some non-zero number and you will find that preventing AI extinction risk is the most valuable thing you can do with your money.

In the eyes of some, EAs were directly responsible for the narrowly averted destruction of OpenAI. In the beginning, OpenAI was partly founded on effective altruism principles; after all, it was created as a non-profit to build AI to benefit humanity. But faced with the reality of expensive training runs for state-of-the-art models, it began partnering with Microsoft to secure additional capital. And with the explosive success of ChatGPT, OpenAI began moving closer to becoming a "traditional" tech company - with misaligned incentives.

As I said above, we still don't know why the board voted to fire Altman three weeks ago. But we *do* know that two of the four board members who fired him are linked to effective altruism. Tasha McCauley is a board member for the effective-altruism organization Effective Ventures. Helen Toner is an executive with Georgetown University's Center for Security and Emerging Technology, backed by a philanthropy dedicated to effective altruism causes. And during the negotiations with Altman, the board acknowledged that allowing the company to be destroyed "would be consistent with the mission."

Between FTX and OpenAI, the last year has brought a big spotlight on EAs and their key roles. Actually, "key roles" may be an understatement - EAs have had two seats on the board of OpenAI (until recently), full control of Anthropic, and majorly influenced Biden's AI strategy. Politico has said EAs "took over Washington," as many now have roles in Congressional offices, federal agencies, and think tanks to shape AI policy.



What President Biden's AI executive order actually means
CHARLIE GUO • NOV 2
Read full story →

But the rise of AI safety proponents and their push for more government regulation has created its own backlash: the effective accelerationists.

The technocapital singularity

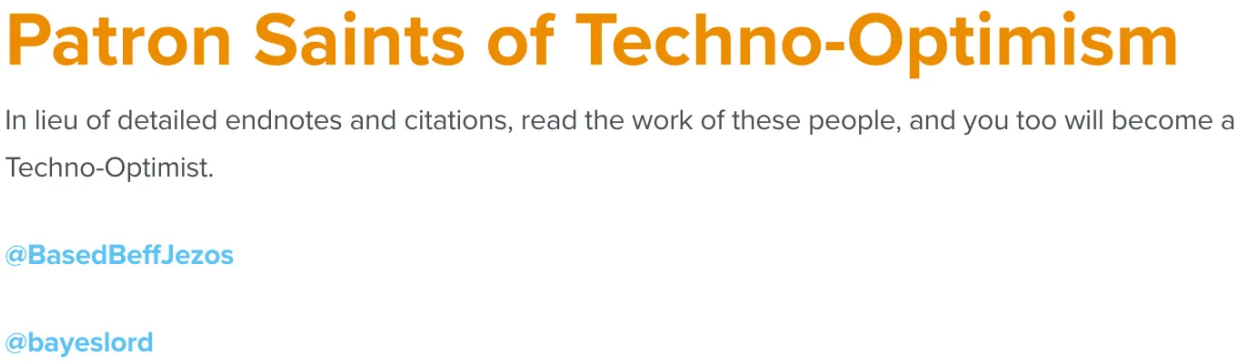
Effective Accelerationism (e/acc) started extremely online, mainly through various Twitter threads. Users @zetular, BasedBeffJezos, creatine_cycle, and bayeslord are generally credited with its inception. It's based on Nick Land's ideas on accelerationism - that we should *intensify* capitalist growth, technological change, and societal destabilization.

E/acc's principles mostly revolve around the technology aspects of these ideas - particularly the idea of bringing the "technocapital singularity" forward as fast as possible.

> 1. The overarching goal for humanity is to preserve the light of consciousness.
> 2. Technology and market forces (technocapital) are accelerating in their power and abilities.
> 3. This force cannot be stopped.
> 4. Technocapital **can** usher in the next evolution of consciousness, creating unthinkable next-generation lifeforms and silicon-based awareness.
> 5. **New forms of consciousness by definition will make sentience more varied and durable. We want this.**
> 6. Technology is leverage. As it advances, it becomes easier to extinguish all conscious life in our corner of the universe. Attempting to stall progress isn't risk free.
> 7. Society and the individual's context within it are rapidly changing, which leads to greater societal instability and mind viruses. (deterritorialisation and reterritorialisation).
> 8. Those who are the first to usher in and control the hyper-parameters of AI/technocapital have immense agency over the future of consciousness.
> 9. HUMANS HAVE AGENCY RIGHT NOW. WE CAN AFFECT THE ADVENT OF THE INFLECTION IN THIS PROCESS.

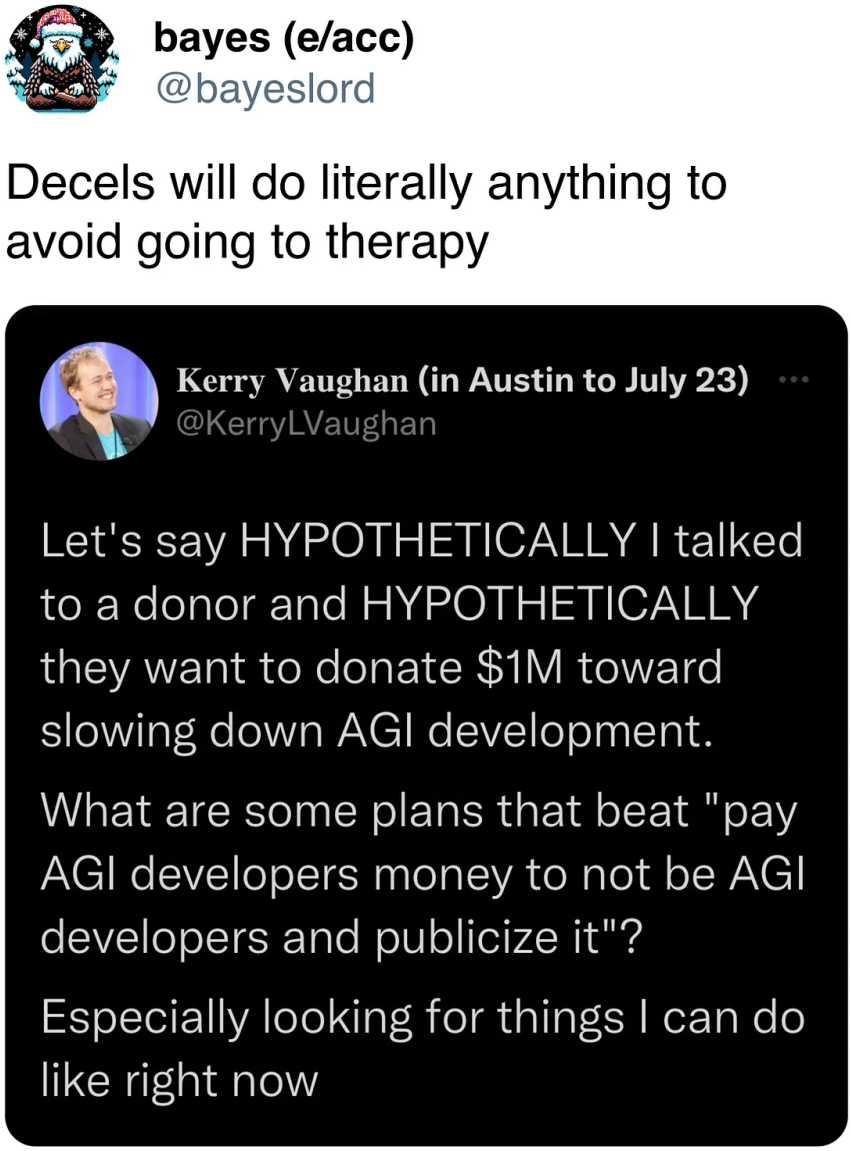
> 10. Effective Accelerationism, **e/acc,** is a set of ideas and practices **that seek to maximize the probability of the technocapital singularity, and subsequently, the ability for emergent consciousness to flourish.**

As it gained steam, it led to discussions, manifestos, and many memes and vibes. Nevertheless, it's picked up some prominent supporters, including Marc Andreessen of a16z, and Garry Tan, President of YCombinator. Andreessen in particular cites Beff Jezos and bayeslord as "Patron Saints" in his Techno-Optimist Manifesto.



It probably won't shock you to learn that e/acc is not without its criticisms. First and foremost, the philosophy disregards AI safety concerns, to the point of being anti-humanist. Beff Jezos himself has said "e/acc has no particular allegiance to the biological substrate for intelligence and life" - meaning humans should not be given preference over intelligent AIs.

But beyond that, there have been critiques about a lack of empirical arguments and evidence within the movement and its abundance of Twitter memes. e/accs also tend to paint anyone who disagrees with them as "decels" - those foolishly opposed to the inevitable progress of technology.

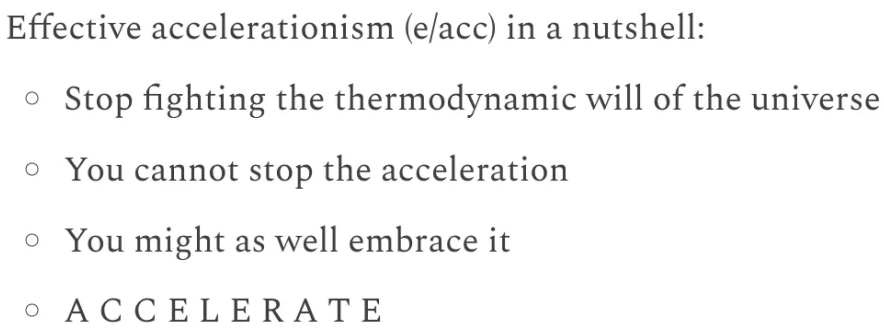


It's been an open question on who exactly was leading the e/acc movement, but just this past week, Beff Jezos was unmasked by Forbes. His real-world identity is Guillaume Verdon, an ex-Google engineer working on Extropic, a stealth AI startup [3]. The entire article is worth reading, but it shows Verdon as someone with both reasonable and radical ideas.

> In a wide-ranging interview with *Forbes*, Verdon confirmed that he is behind the account, and extolled the e/acc philosophy. "Our goal is really to increase the

> scope and scale of civilization as measured in terms of its energy production and consumption," he said.
>
> At various points, on Twitter, Jezos has defined effective accelerationism as "a memetic optimism virus," "a meta-religion," "a hypercognitive biohack," "a form of spirituality," and "not a cult."
>
> ...
>
> Verdon is part of a loud chorus in the AI community who believe that this technology should not be developed in secret at companies like OpenAI, but instead be open-sourced. "We think this whole AI safety industry is just a pretense for securing more control," he said.
>
> "If you're interested in safety, decentralization and freedom is kind of the way to go," he told *Forbes*. "We've got to make sure AI doesn't end up in the hands of a single company."

In practice, e/accs are against any new regulation on AI, want to make it as distributed and accessible as possible, and want effectively zero guardrails or safety checks.



Source: Beff Jezos' Substack

I'm confident that the e/acc movement will keep growing and could spawn additional ideologies. There isn't a great banner uniting those who want to keep building AI without tons of regulation but also aren't fully onboard with "technocapitalism at all costs."

It's harder to say what they'll accomplish in terms of policy - besides rolling back existing regulations, there isn't much that e/accs have said governments should do.

Artwork created with Midjourney.

Ideology in the time of social media

If you're going to create an ideology in the time of social media, you've got to engineer it to be viral.

– Guillaume Verdon, aka Beff Jezos

Look: I'm not here to tell you what to believe. You're an adult; you can figure it out for yourself. If you're a longtime reader, you'll know that I'm a proponent of being intellectually honest about the current capabilities of AI, both good and bad.

How to talk to your family about AI over the holidays
CHARLIE GUO • NOV 23
Read full story →

But I find it fascinating that if you squint, a lot of these AI factions start to look like religions. There are core beliefs and values, there are preachers and apostles. And there is a whole lot of **faith**. Faith that AGI will kill us all if left unchecked; faith that we can optimize our altruistic impact with math and rationalizations; faith that the singularity will usher in a new world of abundance.

I'm not good at faith - which might explain why none of these camps have seemed particularly appealing. I get the idea though - particularly with EAs and e/accs, "number go up" is an easy, black-and-white way of seeing the world. But once you've identified a tribe, it makes it much harder to change your mind (see: politics). And one thing that I know for certain is that nobody knows how the next decade of AI is going to play out - so isn't it better to stay flexible and avoid dogma?

There used to be a time when technologists believed in "strong ideas, weakly held." Going by the loudest voices in AI, though, it sometimes feels like "strong ideas, blindly followed."

---

1 To be clear, I don't think this is an accurate representation. The truth is likely far less dramatic, and much more human.

2 Narrator: They did not.

3 Normally, doxxing someone takes the wind out of their sails when they're revealed to be hypocritical or deceptive in some way. In Verdon's case... it kind of makes him more credible? He certainly has an impressive track record.

32 Likes · 10 Restacks

11 Comments