How philosophers are giving ideological cover to big AI and what to do about it

Alice Crary, New School for Social Research/Oxford

For a public ESDIT (Ethics of Socially Disruptive Technologies) event, "Doing Good in the Age of Big Tech and Declining Democracy," together with Ingrid Robeyns, TU Delft Spui 5, April 17, 2026, 13h-16h45, with a talk at 13h10.

Thanks to Janna van Grunsven, Sabine Roeser, Stella Pekiaridi, the other organizers, those who contributed to creating this physical and intellectual space, and also Ingrid Robeyns. I am very pleased to join this conversation about such important topics. I want to get started right away and explain the angle on them I'll be taking.

1. It will not be news to anyone here that the world is home to intense and widespread misery and injustice. Daily life is, for most, shaped by effects of wars, displacements, extreme poverty and related suffering, and resurgent racism, xenophobia, misogyny, transphobia, and ableism, all exacerbated by slow and quick on-set disasters of a heating climate and the crossing of 6 other of Earth's 9 planetary boundaries. So, we confront difficult questions about how to do good. Should we go for social work or community service, try individual action, or participate in social movements? Philanthropy is one way to support different forms of engagement. There are do's and don'ts to philanthropy, and I am going to talk about some pitfalls, with particular reference to big AI.

2. Philanthropy can be mere *reputation laundering*, as when gas companies trumpet efforts to parter with conservationists or when industrial meat-producers hold forth about animal welfare. These are the things called green-washing, pink-washing, blue-washing, and humane-washing, and they have long histories. They aren't as easy to avoid as it may seem. Even honest attempts to do good can be problematic. Consider the ubiquitous philanthropic traditions of "impact investing" and "Effective Altruism" (or EA). These tap into two powerful cultural forces. They rely on assumptions of standard, **neo-classically-based economics**, treating the economy as free from ethical and ecological moorings and suppressing questions about whether the current socio-economic system's internal tendencies are harmful. They combine these assumptions with **consequentialist moral theory**, taking acting rightly to be securing the biggest return of quantifiable value. On these grounds, they recommend using *strategic investment as the model for all laudable social giving*. However straightforward this seems, it is ill-advised. Instead of addressing violent and socially disruptive chapters in the history of the world's wealth distribution, it encourages those with means to go for calculable impacts. This prioritizes working within the system over liberating social transformation, and it wrongly encourages the inordinately affluent to see themselves as the greatest altruists in human history.

3. Today, we confront philanthropy as reputation-laundering on an unprecedented scale in reference to big AI companies' efforts to build machines with human intelligence in all or most domains or artificial general intelligence, AGI. The main ideological instrument is an EA-related philanthropic tradition called **longtermism**, which represents AGI as humanity's most important philanthropic project and encourages us to downplay big AI's harms and environmental damage. So, Elon Musk says of his businesses: "They are philanthropy." Yet few know how longtermism emerged or how toxic it is.

4. Here's a version of longtermism's story. Although longtermism originates in Silicon Valley, the tradition is described as a future-oriented, late-arriving form of EA. EA was named in 2011 by Oxford philosophers Toby Ord and William MacAskill. It aims to make charitable work and gifts maximally effective. EA's originators are depicted as impressed by a 1972 argument of Peter Singer's about practical implications of applying utilitarian principles to global poverty. This omits EA's roots in AI circles.

During the early 2000s, co-founder of the Berkeley-based Machine Intelligence Research Institute Eliezer Yudkowsky started two blogs devoted to *rationalism*. Rationalism aims to "improv[e] human reasoning and decision-making" so that there are smart people to create and control intelligent machines. Some in Yudkowsky's online community wanted to use rationalist ideas to do good, and Yudkowsky once posted about "effective altruism." That was in 2007, four years before Ord and MacAskill named their movement and two years before they founded Giving What We Can and practiced EA *avant la lettre*. Ord, who by 2009 was active on one of Yudkowsky's blogs, would have been aware of routes to EA via rationalist notions and of rationality training's aspiration to control AI systems. By 2006, Ord was collaborating with philosopher Nick Bostrom, who in 2005 had founded Oxford's Future of Humanity Institute (FHI), an institute partly concerned with AGI's risks. EA wasn't AI-oriented, but it stemmed from discussions about AGI-related topics that would preoccupy longtermists, and its fundraising-virtuosity derived from its AI ties. Longtermism comes before not after EA.

Skip to 2017, when MacAskill coined the monicker 'longtermism' for an EA-related view championed at Bostrom's FHI, Ord's home institution. This view, which Ord wrote a book about and MacAskill championed in his best-selling 2022 *What We Owe the Future*, is that humankind is at a stage when we could annihilate ourselves or proceed to a radiant future, and that we should prioritize neutralizing threats to this future. A radiant future is one in which vast numbers of humans' digital descendants live for billions of years by colonizing the galaxies. Longtermists hold that anthropogenic threats are likelier than natural ones and that the biggest is machines with above-human-level intelligence whose values are unaligned with 'ours'. Ord and MacAskill claim to break from standard EA and develop longtermism via a future-indexed strain of utilitarianism. They don't say that the views that inspired longtermism circulated in Silicon Valley decades earlier.

Their mentor Bostrom provided the first formulation of longtermism. As a graduate student, he joined an email list managed by "Extropians," a group of modern transhumanists. Modern transhumanists hold that genetic engineering, AI, and molecular nanotechnology will enable us to transcend the human condition, augment our intelligence, and overcome disease and aging. In 1998, Bostrom co-founded the World Transhumanist Association and started reformulating transhumanist ideas for analytic philosophers. In 2002, he coined the now ubiquitous term *existential risk* for threats to the "potential of humankind to develop" into transhumanists' envisioned posthumanity. A 2003 article, "Astronomical Waste," presents a utilitarian case for holding that population expansion through space colonization is so valuable that reducing risks to it should be our top moral priority. In 2014, Bostrom argued, in *Superintelligence*, that non-aligned

superintelligent machines are the biggest existential risk. He was then an establishment figure who had converted disreputable Silicon Valley transhumanism into mainstream analytic philosophy. Ord and MacAskill went farther with high-profile books on longtermism that omit all mention of transhumanism.

5. Small surprise that tech leaders received longtermism admiringly. Bill Gates declared he would "highly recommend" *Superintelligence*; Musk and OpenAI's Sam Altman enthused similarly. In 2022, Musk linked to Bostrom's "Astronomical Waste," writing it was "likely the most important paper ever written," and reposted copy for MacAskill's *What We Owe the Future*, declaring it "a close match for my philosophy."

This adulation is for researchers at institutes AI billionaires fund. Here's just one example: Skype co-founder Jaan Tallin helped found Cambridge University's Center for the Study of Existential Risk and the longtermist Future of Life Institute, the latter receiving $14 million from Musk and getting most of its funding from cryptocurrency Ethereum co-creator Vitalik Buterin, who gave it $650 million in 2021. The funding of such think tanks leverages intellectual credibility and influence at global seats of power. Still, the AI-longtermism relationship rarely gets attention.

6. AI leaders are caught up with transhumanism, longtermism's antecedent. For background, note that building "safe" AGI is the framework for AI companies' jostling in the LLM-era. DeepMind in 2010 was the first company explicitly dedicated to building AGI. When, in 2015, Musk, Altman, and others founded the non-profit OpenAI, they championed AGI for all humanity. In 2018, Musk left OpenAI, while Altman started a for-profit OpenAI-division to ensure AGI "benefits all humanity." When, in 2020, Dario Amodei and others co-founded Anthropic, they sought a better approach. In 2023, Musk founded xAI "to build a good AGI," and, in 2024, Meta's Mark Zuckerberg joined in.

Preoccupation with AGI makes sense for modern transhumanism. The tradition projects a future in which technology makes humans super-intelligent, super-long-lived beings who enjoy super well-being, and it treats AGI as utopia's gateway. AGI supposedly leads to computers with above-human-level intelligence, triggering an exponentially increasing cycle of technological progress. This "Singularity" is humanity's route to techno-glory. A consolidation of money and power is required, and Silicon Valley transhumanism is staunchly libertarian.

Transhumanist themes circulated in the 1990s in AI forums, like the Extropian email list. One transhumanist on this list was engineer and author Ray Kurzweil. Another was Yudkowsky. In 2006, Kurzweil, Yudkowsky, and tech-right leader Peter Thiel started the Singularity Summit, a transhumanism-themed conference. Thiel rejects the label "transhumanism" but supports many transhumanism-dominated organizations. He also funds "network states." Similar things are true of Altman and Musk.

Until recently, most AI leaders sounded transhumanist themes about AGI's importance while debating the "existential risk" that an AGI "unaligned" with human values would end humanity. AI leaders were largely "doomers," holding care needs to be taken to avoid lethal AGI and grappling with the "alignment problem." There is a doomer-extreme represented by Yudkowsky, who urges the US to halt AI research and

destroy "rogue datacenter[s] by airstrike," even at risk of triggering a "full nuclear exchange." Opposing doomers are "accelerationists," wholly positive about technologies like AI. Trump's second term has seen a shift from doomerism to accelerationism.

The rallying-cry of doomers is *AI safety*, addressing only the "existential risk" of a theoretical rogue AGI. This excludes *AI ethics*, addressing harms of things like lack of data privacy, algorithmic biases, fake news, hate speech, worker exploitation, extractivist injuries, and devastating energy use. Doomers often combine concern with AI safety with efforts to block non-"safety"-related regulation, since it seems to them racial justice and related issues aren't "as existentially serious as…these [machines] getting more intelligent than us and taking over." Though doomers tout safety-concerns, they also take building AGI to outweigh justice and environmental issues. This is the stance longtermists defend on moral grounds, suggesting that supporting the AGI race counts as philanthropic.

7. But Longtermism is no good as a moral theory. Longtermists offer consequentialist arguments that equate value with wellbeing and so count as utilitarian doctrines. They presuppose that value is recognizable from an abstract view-of-the-universe and treat wellbeing as a metric to compare value *anywhere*, including across time to the future.

Philosophers situate longtermism in the field of population ethics, which deals with future-facing applications of utilitarianism. Longtermists index moral assessment to total aggregate wellbeing, despite recognizing that such *total utilitarianism*, entails the so-called "Repugnant Conclusion" viz, that compared with a population with good quality of life, a much larger population leading barely tolerable lives is morally preferable. Think of *Matrix*-situations with humans harvested for bio-power. Some see these posthuman outcomes as important enough to justify mass near-term suffering and mortality. Longtermist moral logic here echoes murderous autocrats and sci-fi villains, like the Marvel Universe's Thanos, who happily cause mass death to bring on utopian futures.

This moral logic is longtermists' gift to AI leaders. It seems to justify thinking AGI matters morally more than harms for which it may be responsible. But this is disastrous moral reasoning. Like other consequentialism-based theories, longtermism resembles forms of welfarism that presuppose existing social arrangements. Consequentialists' abstract methods block recognition of the injustices liberating movements protest and trace to injurious social structures. Many such injustices are invisible apart from the history of the targeted structures and so inaccessible to those reliant on consequentialists' abstract methods. **The failure of longtermist ethics extends to AI leaders' reliance on it. Projections of the total digital wellbeing in an AGI-enabled techno-future don't justify neglecting harms AI is causing now.**

A case for doubting that AI companies are hastening toward AGI represents an additional critique. The rationale for investing hope in LLMs turns on thinking these "pure language" models are candidates for natural language understanding. LLMs give the impression of intentional speech through sensitivity to the relative frequency of expressions in their databases. They have weaknesses in accuracy and causal reasoning. But even supposing accuracy-gains, there's a problem with the idea that successful models will self-improve and trigger explosive technological growth. This would require

agency, which involves the ability to step back from impulses to believe or do things, and ask whether one should believe or do them. It's unclear why we should expect LLMs to become "agentic." Complex systems can have novel emergent features, but, if beliefs about AGI's imminence refer to such features, they are more religious than scientific.

There is no excuse, of the sort AI leaders and longtermists intimate, for overlooking harms of the self-enriching efforts of a small group of Global North-based men to build ever larger LLMs. Supporting these efforts is not philanthropy.

8. You are undoubtedly familiar with some of the harms of LLMs. These harms fall within AI ethics, which I have already introduced.

Since our topic today is democratic decline, I want to sound one theme from AI ethics about how generative AI systems weaken democracy. These tools sap institutional knowledge because they encourage "cognitive offloading and skill atrophy," displace decision-making, and isolate humans by eliminating occasions for interaction. Consider the Trump Administration's "Department of Government Efficiency" or DOGE, intended to modernize "[f]ederal technology and software to maximize governmental efficiency and productivity." Abstracting from ethical breaches in its deployment, DOGE introduced systems that displace expertise and eliminate roles for decision-making, disrespecting the balance between efficiency and other civic ideals. There are similarly corrosive deployments of AI-technologies into health systems, schools, universities, and legal systems. This shrinks the space in which we see each other acting, register difference, and see the need to change. *It shrinks the space for democratic life.*

9. In closing, two strategies for resisting longtermism and related ideologies. (i) First, longtermism's story is about how money gives immensely damaging ideas a foothold with the public. But money, or worldly appeal, doesn't reliably indicate good ideas. That's obvious, but it needs to be repeated. Often. One strategy of resistance, then, is to support measures to keep private money from distorting public discourse, which means keeping it out of journalism, politics, and education. This is difficult to do given the big tech-accelerated mass inequality in wealth in the US and elsewhere.

(ii) A second strategy is to protect universities for their role as sites for studying society in ways not limited to the abstract and quantitative methods favored by longtermists. Many researchers in the humanities and social sciences, the areas most targeted by neoliberal austerity policies, cultivate and defend social understanding that essentially depends on historical, cultural, and other perspectives; that is not reducible to the narrowly factual or merely technical; and that illustrates how the world can shift with new perspectives, opening potentially liberating possibilities. Protecting this research space is key for resisting democratic decline. The plurality characteristic of democracy requires a standing willingness to respond to perspectival provocations. That is why any innovations, AI-related or otherwise, that overreach in relieving us of the need to confront the dissonance of each other's viewpoints drain democracy's lifeblood. The most vulnerable parts of universities are prime sites for investigating AI's limits and preserving democratic conversation. We should rally to protect them. Thank you.