On the troubled relation between AI ethics and AI safety

Olle Häggström

June 27, 2024

1. Introduction

In November 2022, just a couple of weeks before OpenAI became a household name by releasing ChatGPT, world-leading quantum computing expert Scott Aaronson gave an influential talk about AI safety at the University of Texas at Austin. He was new to that field, having just begun his one-year leave from UT Austin to work for OpenAI on problems in AI safety. In the talk, Aaronson confessed that, only months earlier, he had believed that "AI ethics" and "AI safety" were more or less synonymous terms. He went on to clarify that he had now learned that the two communities "despise each other. It's like the People's Front of Judea versus the Judean People's Front from Monty Python" (Aaronson, 2022). While his quip is overly pointed, there is still, sadly, more truth to it than one might wish for.

This essay will examine the relationship between the fields of AI ethics and AI safety. While their boundaries are somewhat loose, I will start by roughly indicating the scope of the two fields (Section 2). Next, I outline of the main source of conflict: a perceived competition between them for attention and resources (Section 3), followed by an identification and survey of the core issue for judging the merits of the common claim among AI ethicists that AI safety is a useless distraction (Section 4). I then offer some psychological speculation that may be of relevance to the conflict (Section 5), and finally a broadening of the discussion to encompass a third pole – AI accelerationism – without which the tension between AI ethics and AI safety cannot be fully understood (Section 6).

2. What are AI ethics and AI safety?

Both AI ethics and AI safety are concerned with societal consequences of AI and how to ensure positive outcomes; hence the talk about near-synonymity in connection with the Aaronson quote I began with. The difference between the fields is mostly one of emphasis. Work in AI safety focuses mainly on what happens once AI attains capabilities sufficiently broad and powerful to rival humanity in terms of who is in control. It also addresses how to avoid a situation where such an AI with goals and incentives misaligned with core human values goes on to take over the world and possibly exterminate us. This kind of extreme risk to humanity is often labelled existential, so a short way to say what mainly characterizes work on AI safety is to point to its focus on AI-related existential risk.

In contrast, work in AI ethics tends to focus on more down-to-Earth risks and concerns emanating from present-day AI technology. These include, e.g., AI bias and its impact on social justice, misinformation based on deepfakes and related threats to democracy, intellectual property issues, privacy concerns, and the energy consumption and carbon footprint from the training and use of AI systems.

Distinguishing between AI ethics and AI safety is sometimes convenient, and it does point to an existing division of labor in the research community, although Cave and ÓhÉigeartaigh (2019) and many others have stressed grey areas and continuities between the fields. In recent years, such continuities have become particularly conspicuous in the study of large language models; see, e.g., OpenAI (2023b) and Anwar et al. (2024).

A common way to phrase the distinction between AI ethics and AI safety is in terms of near-term vs long-term issues; see, e.g., Cave and ÓhÉigeartaigh (2019), Stix and Maas (2021), and Saetra and Danaher (2023). I feel, however, that this terminology has become increasingly misleading due to shortened estimates of timelines until the civilizationally critical AI breakthrough. In other words, these estimates denote the time we can expect until the crucial fork-in-the-road when AI becomes so capable that we need to get AI safety right in order to reap the enormous benefits of AI rather than be destroyed by it. As late as in 2020 and even 2021 such timelines tended to be measured in decades. See Cotra (2020) for the most ambitious and respected analysis of AI timelines from that epoch, arriving at highly spread-out a posteriori distribution of the time at which we achieve such transformative AI, with a median estimate three decades away from 2020. The rapid advancement of generative AI since then has led to drastically shortened timelines that now often end up in single-digit years rather than decades. Furthermore, the closer we get to the epicenter of AI in San Francisco and Silicon Valley, the more these timelines seem to shorten; see, e.g., Altman et al. (2023) and Henshall (2024), and the more ambitious and detailed recent report by Aschenbrenner (2024). To speak about an event that many leading experts think may happen this decade as "long-term" invites confusion, and I therefore propose that the terminology is abandoned.

Before moving on, I should state for transparency where I locate myself with respect to the fields of AI ethics and AI safety. My main academic background is in mathematics, but over the last 10-15 years I have become increasingly interested in AI and its societal implications, and while the distinction between AI ethics and AI safety didn't come clearly into focus until more recently, I nevertheless aspired throughout this period to work in both fields. With that said, my actual engagement in AI ethics has mostly amounted to teaching, committee work and so on, whereas my true heart and whatever intellectual contributions I can lay claim to reside mostly on the AI safety side of the spectrum.

3. The main conflict: distracting attention

Prima facie, it is strange for antagonism to arise between AI ethics and AI safety, given their shared goal of ensuring that AI benefits society and human welfare. The most salient feature of the conflict is the (real or just perceived) competition for attention and resources. Those favoring AI ethics over AI safety may for instance complain that the latter "sucks all the oxygen out of the room", leading to a neglect of the more mundane concerns of AI ethics (Stilgoe, 2024). In the same spirit, the title of a recent Nature editorial calls to "stop talking about tomorrow's AI doomsday when AI poses risks today" (Anonymous, 2023). A surge of such complaints followed the two much-publicized open letters on AI existential risk and AI safety in the spring of 2023 (Bengio et al., 2023a; Hinton et al., 2023); see Grunewald (2023) for a list of further examples. Sometimes, the complaints are accompanied by conspiracy theories suggesting that AI safety concerns are raised with the express purpose of distracting from the real issues (Bryson, 2023; Wong, 2023).

Corresponding calls for AI ethics people to lower their profile so as not distract from more important issues in AI safety are harder to come by, beyond the occasional snide remark from the leading AI safety pioneer (Yudkowsky, 2020), and comparisons like the following by Geoffrey Hinton:

> I believe that the possibility that digital intelligence will become much smarter than humans and will replace us as the apex intelligence is a more serious threat to humanity than bias and discrimination, even though bias and discrimination are happening now and need to be confronted urgently. (Quoted in O'Neil, 2023)

When I speak to AI safetyists, they pretty much never express annoyance over the actual work that AI ethicists engage in. Whatever frustration there is tends to be about the AI ethicists' insistence that AI safety talk is pernicious and ought to stop. When I started planning this essay, I expected to describe a more-or-less symmetric situation where AI ethicists and AI safetyists alike accuse each other of stealing all the oxygen and distracting from the truly important issues. But on closer inspection, the literature seems to indicate that the situation is less symmetric than suggested in the Aaronson quote I began with in Section 1. Calls for the other side to back away from the limelight come overwhelmingly from the AI ethics camp.

Before concluding that the tension between the two communities are primarily the fault of AI ethicists, we should ask whether the crowding-out phenomenon they complain about is real. Does increased attention to AI safety lead to a depletion of attention and resources for AI ethics?

In principle, the effect could go either way. Imagine two adjacent fields A and B, and denote their union by C (so that here A could be AI safety, B could be AI ethics, and C could be described as systematic thinking about the societal and other consequences of AI and how to ensure positive outcomes). If there is a fixed pool of attention and resources for C, then any increase in attention for A leads to a corresponding decrease in the attention for B, and likewise for resources. But such a zero-sum model need not be accurate. On the contrary, it might be that an increase in attention and resources for A leads to an increased awareness of the importance of field C as a whole, causing an increase in attention and resources for C. Some of this is likely to spill over to B, and if enough of it does, the net result might well be positive for B.

It is far from obvious which way the effect on B should be expected to go, and at the end of the day this is an empirical question. Also in the special case of AI safety vs AI ethics, the question remains unanswered, but there are a couple of preliminary studies. Grunewald (2023) looks at a variety of measures of and proxies for attention and resources to AI ethics, and what happened to these in the wake of the upsurge of media attention on AI existential risk and AI safety in the spring of 2023, following the release of GPT-4 and the two aforementioned open letters. Although there are plenty of confounders preventing any strong conclusions about causal effects, the results point quite clearly towards AI ethics having had a substantial net increase of attention and resources during the rest of that year. Friedrich (2023) takes an experimental approach to the same question and finds that prior exposure to AI safety concerns causes test subjects to show greater support for addressing problems in the realm of AI ethics. None of this evidence is conclusive, but it does underline that the oft-expressed certitude that media attention on AI safety hurts the field of AI ethics is premature.

4. A core issue: is AI existential risk real?

Beyond the possible crowding-out effect discussed in Section 3, another perhaps even more crucial issue in evaluating calls by AI ethicists to limit AI safety discussions is whether AI existential risk is real. AI ethicist Meredith Whittaker dismisses it as "a fantastical, adrenalizing ghost story" (Wong, 2023). If true, then there is indeed a strong case for redirecting resources to less ghostly issues. Conversely, if AI existential risk is real, addressing it is critical, making AI safety essential and efforts to silence it misguided.

So, is AI existential risk real? A growing body of AI safety literature argues that it is. While I can't cover it all here, I can highlight a few key ideas and references. The outstanding modern classic on AI existential risk is Nick Bostrom's (2014) *Superintelligence*, which beautifully outlines many of the fundamental ideas. However, it predates the ongoing deep learning and generative AI revolutions. To a lesser extent this is also true for Stuart Russell's (2019) excellent *Human Compatible*. A newcomer to the field would do well to begin with one of these two books, complemented with one or two more recent survey papers such as Ngo et al. (2022), Burden et al. (2023) and Hendrycks et al. (2023).

The basic case for the reality of AI existential risk can be outlined as follows.

> (1) **AI capabilities.**
>
> > (1a) It is possible to build a machine which, when it comes to cognitive capabilities relevant to the ability to take over the world, is vastly superior to a human.
> >
> > (1b) This is achievable in the not-too-distant future.
>
> (2) **AI motivations.** Unless we make sure that such an AI's goals are highly aligned with human values, or whatever values it is we want it to pursue to benefit humans, this AI may develop very different goals, and go on to prioritize those over human welfare.

If we accept these premises, it pretty much follows that if we continue to build ever more capable AIs without solving the AI alignment problem (more on which later in this section) indicated in (2), we risk creating superintelligent AIs with goals alien to ours, leading to potential conflict and possibly even the end of humanity. The classic semi-concrete example of how this could play out is the so-called paperclip apocalypse (outlined, e.g., in Bostrom, 2014; Häggström, 2021a; and Burden et al., 2023), but for readers who find themselves unable to take such a cartoonish scenario seriously I recommend the Production Web scenario described in some detail by Critch and Russell (2023).

A common (but naïve) intuition, fueled by various science fiction tropes, is that large language models are irrelevant to catastrophic AI risk because an AI with power-grabbing ambitions would need to influence the physical would rather than just produce text. And since robotics still lags far behind the generative AI frontier, there is no need to worry. What this overlooks is that a rogue AI without access to advanced robotics has a powerful alternative: humans. The key cognitive skill for the AI to access this resource is deception and social manipulation. This is why OpenAI includes persuasion as one of the four main categories in its program for monitoring and evaluating dangerous AI capabilities (OpenAI, 2023b). In fact, already GPT-4

shows signs of such competence, such as in much-discussed examples involving CAPTCHA (OpenAI, 2023a) and insider trading (Scheurer et al., 2023). While these are isolated and relatively harmless instances, the concern is that a further scaled-up large language model might cross some critical threshold where it learns to deceive more broadly and systematically towards some hidden goal.

We need, of course, to evaluate claims (1a), (1b) and (2). Regarding (1a), we know for a fact the biological evolution was able to create human intelligence, and even if we ignored the slowness and the various glaring deficiencies of the human mind, it would be bizarre to think that this evolutionary process has managed to find a global maximum in the construction of intelligent brains. Whatever evolution managed to do with its crude local search in the biological fitness landscape, we should be able to replicate, especially as we have evolutionary methods at our disposal (see, e.g., Lehman et al., 2020), and even more so as we have the extra advantage of being able to look to the human brain as a kind of template. If we accept a naturalistic worldview (without which all bets are off) and that intelligence is fundamentally an information processing phenomenon, we come very close to being able to conclude (1a). A possible attempt to nevertheless resist this conclusion is to insist that intelligence is exclusively biological. This claim strikes me as extremely speculative, and it becomes increasingly untenable the more intelligent behavior we see from GPT-4 and other AIs. Arkoudas (2023) flatly denies this last kind of observation by pointing to a long list of reasoning mistakes performed by GPT-4, as if humans never made such mistakes; see Häggström (2023a) for a more detailed response.

Moving on to (1b), we are faced with the much more refined issue of how long we can expect continued AI development to take until it reaches levels where AI becomes capable of challenging humans for world dominance. This is of course a very delicate question, and until recently the best approach was to compare exponential technology trends to biological benchmarks. The report by Cotra (2020) mentioned in Section 2 offers the most ambitious Bayesian attempt to handle the many uncertainties such as the computational efficiency of contemporary AIs compared to the human brain, the expected rate of continued algorithmic progress, and so on. She lands in an a posteriori distribution for the time of the crucial breakthrough which is highly spread out but has most of its mass around mid-century. However, in just the few years that has passed since Cotra's report, it has become increasingly evident that we need replace or at least complement this approach by closely examining the capabilities of current cutting-edge AIs and comparing them to humans. The high-dimensional nature of intelligence makes this comparison a very delicate matter: I am smarter than GPT-4 in some respects, it is smarter than me in others, and how to relevantly weigh the various dimensions against each other is far from obvious. Still, this shift in perspective is at the heart of the drastically shortened timeline estimates which predict extreme events within the next ten years, and which are increasingly becoming the consensus view at and around the leading AI labs; see, e.g., Aschenbrenner (2024).

One intelligence dimension that seems especially crucial is the ability to carry out AI research. Once we get AIs that match or surpass human capabilities in this respect we enter a new regime of AI development that may well become orders of magnitude faster than what we are now seeing. The classic theoretical treatment of the dynamics of the self-improvement feedback loop that would arise in this new regime is Yudkowsky (2013), but see Davidson (2023) for a more refined treatments that take into account the particulars of current trendlines in deep

learning and large language models, and Aschenbrenner (2024) for some staggering conclusions about short timelines.

Of course, uncertainties abound regarding nearly all I've said so far in this section, but it would be epistemically reckless to take these uncertainties as an excuse to assume that across-the-board superhumanly capable AIs are at least decades away or even impossible. We therefore urgently need to think about how to handle a world with such AI and whether we might unintentionally give those AIs goals and values that clash with ours, as in item (2) above, with potentially catastrophic consequences.

The best theoretical approach available today for answering this question is the so-called Omohundro-Bostrom framework for AI motivations (Omohundro, 2008; Bostrom, 2014; Häggström, 2019). The framework distinguishes between instrumental and final goals: an AI's final goal is what it values as an end-in-itself and optimizes for, while the instrumental goals are steps towards achieving the final goal. Of the framework's two cornerstones, the first is the orthogonality thesis, which states that almost any final goal is compatible with arbitrarily high levels of intelligence. This is part of what the paperclip thought experiment is meant to illustrate: there is nothing inherently unintelligent about wanting to maximize paperclip production. Together with so-called fragility of human values (meaning that even a minor perturbation from them can have disastrous consequences) and the fact that only a miniscule fraction of the space of possible values prioritize human welfare at all, this spells trouble in case we don't take extreme care in making sure that the AI has such human-prioritizing values (Yudkowsky, 2009; Bostrom, 2014; Carlsmith, 2024).

The framework's second cornerstone is the instrumental convergence thesis, which posits that certain instrumental goals are universal. Any sufficiently intelligent AI can be expected to adopt them in order to promote its final goal, pretty much regardless of what this final goal is. One such instrumental goal is self-preservation: the AI understands that its continued existence will make it better positioned to promote its final goal than otherwise, so it will resist being turned off. The logic is similar behind other universal instrumental goals such as self-improvement, resource acquisition and goal integrity, the last one meaning that the AI wants to preserve its final goal (Bostrom, 2014; Burden et al., 2023). Especially scary is the instrumental goal of discretion, highlighted by Bostrom (2014) as the "treacherous turn", which amounts to the AI hiding its true values and/or capabilities as long as it finds a discrepancy between its own values and those of humans, biding its time until one day it decides it is strong enough to come out and overpower us. Today this is starting to look like an increasingly difficult obstacle to evaluating the safety of frontier AI models (Cohen et al., 2024; Henshall, 2024).

Sharing a world with a superhumanly intelligent agent engaged in resource acquisition and in deceiving us about its true intentions spells disaster, and this is why AI alignment, meant to ensure that the first superintelligent AIs have goal and values aligned with ours, is considered so important. Now, the Omohundro-Bostrom framework is not written in stone the way a mathematical theorem is, but it has withstood scrutiny for more than a decade. For instance, in Häggström (2021b) I show how the orthogonality thesis survives the critique from Müller and Cannon (2021). They argue that a sufficiently intelligent AI should be able to figure out that killing humans for the sake of paperclip maximization is wrong. However, this critique conflates facts and values, and overlooks that while killing humans may seem wrong from our value-laden perspective, there is no guarantee that a superintelligent paperclip-maximizer will automatically share this value.

All in all, while the claim (2) that superintelligent AI may attain humanly alien and dangerous goals is not literally proven, the literature on AI existential risk offers enough support of it that we would be foolish to ignore it. Many proposals have been made for how to approach the AI alignment problem – see, e.g., Häggström (2021a), Ngo et al. (2022) and Burden et al. (2023) for surveys – and leading AI labs like OpenAI, Anthropic and Google/DeepMind are interested in solving it, but the fact of the matter is that no convincing plan has yet been presented. This has contributed, along with the shortened timelines, to a growing feeling in large parts of the AI safety community that focusing on solving AI alignment is no longer enough, and that we may also need to pull the brakes on AI capabilities progress; see, e.g., Grace (2022) and Bengio et al. (2023a).

5. Some psychological speculations

In view of the arguments in Section 4 that AI existential risk is real, it seems reasonable to expect that AI ethicists, such as those referenced in Section 3, who oppose concerns about AI existential risk and AI safety, should explain why these arguments are flawed and why in their view existential risk from AI is negligible or nonexistent. However, they rarely provide such explanations, and when they do, their counterarguments tend to disintegrate upon closer inspection. For a typical example, see my panel debate in October 2023 with well-known AI ethicist Virginia Dignum (Häggström, 2023b).

This pattern suggests that strong arguments against the reality of AI existential risk are lacking among these AI ethicists, because if they did have such arguments, it would be puzzling why they so consistently present weak ones or none at all. And yet they persist in dismissing AI existential risk, so we have reason to look for psychological explanations for why they do so. This issue is likely complex and not reducible to a single explanation, but in what follows I offer three possible contributing factors, acknowledging that this is the most speculative part of my essay.

In a much-discussed panel debate with Yoshua Bengio, Yann LeCun, and Max Tegmark in June 2023, AI ethicist Melanie Mitchell dismissed AI existential risk as "just science fiction" (Bengio et al., 2023b). This echoes a pattern observed among AI ethicists who often start their talks by derisively declaring science fiction-like risk scenarios to be out of scope for their talk, sometimes referencing the Terminator movie franchise. This routine may have begun as an attempt to appear rational and grounded, with no particular intention to make substantial claims about existential risk. Nonetheless, with enough repetition, the habit may inadvertently lead speakers to believe that AI safety is silly. Psychology has much to say about this kind of phenomenon; see, e.g., Hutson (2017).

(It is worth noting here that likening AI existential risk scenarios to science fiction is not a good argument for dismissing the former. A person in 1920 would surely have thought of speculations about a moon landing in 1969 as outlandish science fiction, and a person in 1980 being told about our lives today with Internet and smartphones would have reacted similarly, as would a person in 2017 who was told about the capabilities of GPT-4. We already live in a science fiction world, so clearly it's a mistake to think that adding one or two astonishingly powerful technologies to a scenario automatically makes it implausible.)

A second candidate for a psychological mechanism behind the dismissive view of AI existential risk among AI ethicists relates to their ideological leanings. While presumably there are AI

ethicists on all sides of the political spectrum, many of them – including Timnit Gebru and Émile Torres whose work I will discuss in the next section – align with perspectives associated with identity politics and the social justice movement. These perspectives tend to interpret all societal issues through the lens of conflicts between privileged and marginalized groups (Mounk, 2023). Given such an inclination, it can be difficult to make sense of a catastrophic risk that strikes against all humans equally by simply killing everyone, creating a kind of cognitive dissonance that is perhaps easiest to overcome by accepting even weak of evidence against such risk as conclusive. Relatedly, Critch (2023) reports on a small but nonzero percentage of AI professionals who hold the view that "the world is very unfair, but if everyone dies from AI it will be more fair", and that "dying together is less upsetting than dying alone".

The third mechanism I'd like to suggest involves ideological connections on the AI safety side of the debate. Many AI safety researchers have a background in the effective altruism movement and its offshoot, longtermism, which emphasizes the moral importance of our actions' long-term impacts. Longtermism argues that the potential number of future human lives is vast, and that consequently a major determinant of the value of our actions today is their potential to improve these future lives (MacAskill, 2022). Mitigating existential risks, including those posed by AI, is central to this ideology, as ensuring humanity's survival allows these future lives to exist.

The overlap between the AI safety and longtermism communities has led many AI safety proponents to use longtermist arguments in support of it. This includes myself (see, e.g., Häggström, 2021a), and while I do still think the core longtermist argument for AI safety remains valid, I am no longer as enthusiastic about using it in public debate. If we want to convince AI ethicists and others to take AI existential risk seriously, appeals to whatever concerns they may or may not have for hypothetical future generations thousands or millions of years down the road seems unnecessarily roundabout, compared to simply stressing the fact that all of us, and our loved ones, risk being murdered by AI in a matter of years. This roundabout might even confuse them into thinking that the matter lacks urgency. A de-emphasis of the longtermist perspective on the part of AI safetyists might therefore help bridge the AI safety vs AI ethics chasm. Shulman and Thornley (2023) offer a similar pragmatic perspective.

6. AI ethics, AI safety and AI accelerationism: an unhappy triangle

The AI debate as a whole is a complicated landscape with many different and sometimes quite nuanced positions, but one camp that I need to bring up here in relation to AI ethics and AI safety is the one we may call AI accelerationism. It features thinkers such as Marc Andreessen, Tyler Cowen and Yann LeCun, and centers on the idea that the greatest AI-related threat to future human flourishing is neither the catastrophic AI takeover that AI safetyists point to, nor any of the more mundane risks that AI ethicists are concerned with; instead, their fear is that the path forward to AI-assisted human flourishing might be slowed down or even halted by excessive safety concerns and overregulation (Andreessen, 2023; Cowen, 2023; Levy, 2023). These AI accelerationists therefore tend to advocate a laissez-faire approach where AI companies are free to race ahead, unhindered by cumbersome safety standards.

If we were to simplify AI discourse by arranging the three positions of AI ethics, AI safety and AI accelerationism in a 2-against-1 setup along a single axis, what is the most natural such setup? In fact, all three of the possible 2-against-1 alliances have been suggested in various contexts.

For instance, in the aforementioned and much publicized Munk Debate, AI accelerationist Yann LeCun and AI ethicist Melanie Mitchell were teamed up (somewhat awkwardly) against AI safetyists Yoshua Bengio and Max Tegmark on the issue of whether AI existential risk is real (Bengio et al, 2023b).

Another 2-against-1 setup is the one I personally strongly favor, namely to see AI ethics and AI safety as natural allies against AI accelerationism. The former two camps emphasize the need to consider various AI risks, and to proceed with caution, thus finding a common enemy in the AI accelerationists who try to downplay all such concerns.

The third possibility for a 2-against-1 setup is to pair up AI safety and AI accelerationism in a joint battle against AI ethics. As someone who sees AI accelerationism as the natural antagonist to the AI safety community's focus on risk mitigation and caution, I found this setup to be patently absurd until recently. Reading up on AI ethicist Émile Torres' latest arguments for it allowed me to understand them a bit better, however. In a recent podcast conversation (Hastings-Woodhouse and Torres, 2024), Torres stresses how both the AI safety and AI accelerationist communities have roots in the so-called TESCREAL bundle (Transhumanism, Extropianism, Singularitarianism, Cosmism, Rationalism, Effective Altruism and Longtermism) – an acronym invented by Torres for a collection of interrelated futuristic and techno-optimistic movements that developed during the last few decades, and whose history is explored at some length in Gebru and Torres (2024). The genealogy is said to imply that even those AI safetyists who are now calling for slowdowns in AI progress are of the same ilk as AI accelerationists and thus cannot be trusted. As soon as they feel that AI alignment has been adequately solved, they will (or so the story goes) join hands with AI accelerationists in the race towards superintelligence. This cannot possibly bring anything good, according to Torres, who sees superintelligence as an inherently Western and colonial (and therefore flawed) project.

Looking around at the AI landscape today, some signs in support of Torres' view of its sociology can indeed be found, such as in how the heads of the three leading AI labs – OpenAI, Google/DeepMind and Anthropic – all pay at least lip service to AI safety ideals (such as by signing the open letter by Hinton et al, 2023), yet race forward in accelerationist manner. The same Janus-facedness can be found in the arguments of Aschenbrenner (2024). In fact, talk about slowing down AI development used to be more or less taboo among AI safetyists, for fear of coming across as Luddite. That situation remained until the quite recent thaw indicated by texts such as those by Alexander (2022a) and Grace (2022).

Still, the various deceptive ways in which Gebru and Torres (2024) attempt to escalate the AI ethics vs AI safety conflict strike me as uncalled for, such as their malicious cut of the last 14 words in the Hinton quote in Section 3 above. There's a revealing passage towards the end of the conversation with Hastings-Woodhouse, where Torres is asked about the way forward for humanity, and replies that "the hour is late, I don't know if there's much to do at this point" other than "pointing the finger at who the responsible parties are", namely capitalism and the TESCREAL movement (Hastings-Woodhouse and Torres, 2024). Such finger-pointing is very much the modus operandi of Gebru and Torres (2024), who spend page after page on grievance archaeology, digging up connections between TESCREAL and discredited ideas such as Christian eschatology, eugenics and racism.

But defeatism will help nobody. I favor a more constructive approach. To spell it out, let me end this essay where I started, going back again to late 2022 and the time of the release of ChatGPT. Just two weeks after the release, on December 12, tech blogger Scott Alexander commented on

the jailbreaking practices that had already become popular. Users had quickly learned how to circumvent safety measures intended to prevent ChatGPT from giving instructions on how to cook methamphetamine or from expressing racist slurs. Alexander concluded his text with a call for AI ethicists and AI safetyists to stop quarreling and to join forces:

> The people who want less racist AI now, and the people who want to not be killed by murderbots in twenty years, need to get on the same side right away. The problem isn't that we have so many great AI alignment solutions that we should squabble over who gets to implement theirs first. The problem is that *the world's leading AI companies do not know how to control their AIs*. Until we solve this, nobody is getting what they want. (Alexander, 2022b)

Apart from the implied somewhat lax timeline, I wholeheartedly agreed with Alexander at the time, and I still do. AI ethicists and AI safetyists – unite!

Acknowledgement. I am grateful to Björn Bengtsson, Emma Jonson, James Miller and Stefan Schubert whose comments on an earlier draft helped improve the manuscript significantly.

References

Aaronson, S. (2022) My AI safety lecture for UT Effective Altruism, *Shtetl-Optimized*, November 29, https://scottaaronson.blog/?p=6823

Alexander, S. (2022a) Why not slow AI progress?, *Astral Codex Ten*, August 8, https://www.astralcodexten.com/p/why-not-slow-ai-progress

Alexander, S. (2022b) Perhaps It is a bad thing that the world's leading AI companies cannot control their AIs, *Astral Codex Ten*, December 12, https://www.astralcodexten.com/p/perhaps-it-is-a-bad-thing-that-the

Altman, S., Brockman, G. and Sutskever, I. (2023) Governance of superintelligence, OpenAI, https://openai.com/index/governance-of-superintelligence/#GregBrockman

Andreessen, M. (2023) Why AI will save the world, *a16z*, June 6, https://a16z.com/ai-will-save-the-world/

Anonymous (2023) Stop talking about tomorrow's AI doomsday when AI poses risks today, *Nature* **618**, 885-886.

Anwar, U. and 37 coauthors (2024) Foundational challenges in assuring alignment and safety of large language models, https://arxiv.org/abs/2404.09932

Arkoudas, K. (2023) GPT-4 can't reason, https://arxiv.org/abs/2308.03762

Aschenbrenner, L. (2024) *Situational Awareness: The Decade Ahead*, https://situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf

Bengio, Y. and 33,000+ co-signatories (2023a) Pause giant AI experiments: An open letter, Future of Life Institute, March 22, https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Bengio, Y., LeCun, Y., Mitchell, M. and Tegmark, M. (2023b) Artificial Intelligence debate, *Munk Debates*, June 22, https://haggstrom.blogspot.com/2023/10/debating-ai-takeover-with-virginia.html

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford.

Bryson, J. (2023) The letter was published to coincide with..., *Twitter*, May 31, https://x.com/j2bryson/status/1663794967804882944

Burden, J., Clark, S. and Whittlestone, J. (2023) From Turing's Speculations to an Academic Discipline: A History of AI Existential Safety, in *The Era of Global Risk: An Introduction to Existential Risk Studies* (eds S. Beard, M. Rees, C. Richards and C. Rios Rojas), Open Book Publishers, Cambridge, UK, pp 201-236, https://www.openbookpublishers.com/books/10.11647/obp.0336/chapters/10.11647/obp.0336.09

Carlsmith, J. (2024) An even deeper atheism, https://joecarlsmith.com/2024/01/11/an-even-deeper-atheism

Cave, S. and ÓhÉigeartaigh, S. (2019) Bridging near- and long-term concerns about AI, *Nature Machine Intelligence* **1**, 5-6.

Cohen, M., Kolt, N., Bengio, Y., Hadfield, G. and Russell, S. (2024) Regulating advanced artificial agents, *Science* **384**, 36-38.

Cotra, A. (2020) Forecasting TAI with biological anchors, https://drive.google.com/drive/u/0/folders/15ArhEPZSTYU8f012bs6ehPS6-xmhtBPP

Cowen, T. (2023) Existential risk, AI, and the inevitable turn in human history, *Marginal Revolution*, March 27, https://marginalrevolution.com/marginalrevolution/2023/03/existential-risk-and-the-turn-in-human-history.html

Critch, A. (2023) From my recollection, >5% of AI professionals..., *Twitter*, July 23, https://x.com/AndrewCritchPhD/status/1683216490517135361

Critch, A. and Russell, S. (2023) TASRA: a taxonomy and analysis of societal-scale risks from AI, https://arxiv.org/abs/2306.06924

Davidson, T. (2023) What a compute-centric framework says about takeoff speeds, Open Philantropy, https://www.openphilanthropy.org/research/what-a-compute-centric-framework-says-about-takeoff-speeds/

Friedrich, D. (2023) Are AI safety and AI ethics memetic rivals?, https://osf.io/preprints/psyarxiv/3rpwt

Gebru, T. and Torres, É. (2024) The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday* ***29***(4).

Grace, K. (2022) Let's think about slowing down AI, *AI Alignment Forum*, December 22, https://www.alignmentforum.org/posts/uFNgRumrDTpBfQGrs/let-s-think-about-slowing-down-ai

Grunewald, E. (2023) Attention on AI X-Risk Likely Hasn't Distracted from Current Harms from AI, *LessWrong*, December 21, https://www.lesswrong.com/posts/5rexNxtZgkEQBi3Sd/attention-on-ai-x-risk-likely-hasn-t-distracted-from-current

Häggström, O. (2019) Challenges to the Omohundro-Bostrom framework for AI motivations, *Foresight* **21**, 153-166.

Häggström, O. (2021a) *Tänkande maskiner: Den artificiella intelligensens genombrott*, Fri Tanke, Stockholm.

Häggström, O. (2021b) AI, orthogonality and the Müller-Cannon instrumental vs general intelligence distinction, https://arxiv.org/abs/2109.07911

Häggström, O. (2023a) Are large language models intelligent? Are humans? *Computer Science and Mathematics Forum* **8**(1), 68.

Häggström, O. (2023b) Debating AI takeover with Virginia Dignum, *Häggström hävdar*, October 8, https://haggstrom.blogspot.com/2023/10/debating-ai-takeover-with-virginia.html

Hastings-Woodhouse, S. and Torres, E. (2024) Émile P. Torres and I discuss where we agree and disagree on AI safety, Consistently Candid, February 20, https://podcasts.apple.com/nz/podcast/4-%C3%A9mile-p-torres-and-i-discuss-where-we-agree/id1732326178

Hendrycks, D., Mazeika, M. and Woodside, T. (2023) An overview of catastrophic AI risks, https://arxiv.org/abs/2306.12001

Henshall, W. (2023) When might AI outsmart us? It depends who you ask, *Time Magazine*, January 19.

Henshall, W. (2024) Nobody knows how to safety-test AI, *Time Magazine*, March 21.

Hinton, G. and 500+ co-signatories (2023), Statement on AI Risk, Center for AI Safety, May 29, https://www.safe.ai/work/statement-on-ai-risk

Hutson, M. (2017) Living a lie: We deceive ourselves to better deceive others, *Scientific American*, April 4.

Lehman, J. and 52 coauthors (2020) The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities, *Artificial Life* 26, 274-306.

Levy, S. (2023) How not to be stupid about AI, with Yann LeCun, *Wired*, December 22, https://www.wired.com/story/artificial-intelligence-meta-yann-lecun-interview/

MacAskill, W. (2022) *What We Owe the Future*, Basic Books, New York.

Mounk, Y. (2023) *The Identity Trap: A Story of Ideas and Power in Our Time*, Penguin Random House, New York.

Müller, V. and Cannon, M. (2021) Existential risk from AI and orthogonality: can we have it both ways?, *Ratio*, https://doi.org/10.1111/rati.12320

Ngo, R., Chan, L. and Mindermann, S. (2022) The alignment problem from a deep learning perspective, https://arxiv.org/abs/2209.00626

Omohundro, S. (2008) The basic AI drives, *Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (ed. P. Wang, B. Goertzel and S. Franklin), IOS, Amsterdam, 483-492.

O'Neil, L. (2023) These women tried to warn us about AI, *Rolling Stone*, August 12, https://www.rollingstone.com/culture/culture-features/women-warnings-ai-danger-risk-before-chatgpt-1234804367/

OpenAI (2023a) GPT-4 technical report, March 4, https://arxiv.org/abs/2303.08774

OpenAI (2023b) Preparedness framework (beta), December 18, https://cdn.openai.com/openai-preparedness-framework-beta.pdf

Russell, S. (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking, New York.

Sætra, H.S. and Danaher, J. (2023) Resolving the battle of short- vs. long-term AI risks, *AI and Ethics,* https://doi.org/10.1007/s43681-023-00336-y

Scheurer, J., Balesni, M. and Hobbhahn, M. (2023) Large language models can strategically deceive their users when put under pressure, https://arxiv.org/abs/2311.07590

Shulman, C. and Thornley, E. (2023) How much should governments pay to prevent catastrophes? Longtermism's limited role, https://philpapers.org/archive/SHUHMS.pdf

Stilgoe, J. (2024) Technological risks are not the end of the world, *Science* **384**(6693).

Stix, C. and Maas, M. (2021) Bridging the gap: the case for an 'Incompletely Theorized Agreement' on AI policy, *AI and Ethics* **1**, 261-271.

Wong, M. (2023) AI doomerism is a decoy, *The Atlantic*, June 2.

Yudkowsky, E. (2009) Value is fragile, *LessWrong*, January 29, https://www.lesswrong.com/posts/GNnHHmm8EzePmKzPk/value-is-fragile

Yudkowsky, E. (2013) Intelligence explosion microeconomics, https://intelligence.org/files/IEM.pdf

Yudkowsky, E. (2020) Is somebody in "AI ethics" Terribly Concerned about…, *Twitter*, December 28, https://x.com/ESYudkowsky/status/1343624937030918145