AI and Ethics
https://doi.org/10.1007/s43681-023-00336-y

COMMENTARY

Resolving the battle of short- vs. long-term AI risks

Henrik Skaug Sætra[1] · John Danaher[2]

Received: 28 June 2023 / Accepted: 17 August 2023

Abstract

AI poses both short- and long-term risks, but the AI ethics and regulatory communities are struggling to agree on how to think two thoughts at the same time. While disagreements over the exact probabilities and impacts of risks will remain, fostering a more productive dialogue will be important. This entails, for example, distinguishing between evaluations of particular risks and the politics of risk. Without proper discussions of AI risk, it will be difficult to properly manage them, and we could end up in a situation where neither short- nor long-term risks are managed and mitigated.

Keywords AI · Existential risk · Risk management · Risk analysis · Risk assessment

1 Introduction

As AI development increasingly falls under regulatory scrutiny, debates about how to deal with AI risks intensify. One example is the debate following the brief statement from the Center for AI Safety (CAIS) aimed at making it easier to 'voice concerns about some of advanced AI's most severe risks'. The full statement reads:

> Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war [1].

Other examples include the open letter from the Future of Life Institute (FLI) in March 2023 that warned, amongst other things, of 'ever more powerful digital minds that no one—not even their creators—can understand, predict, or reliably control' [2].

These calls for precaution and regulatory intervention to mitigate existential risks have been met with considerable scepticism in some circles. The main issues raised by critics are that (a) the risks are speculative and uncertain, (b) these warnings divert attention from real short-term risks and harms, (c) these statements and letters are in reality strategic manipulation aimed at avoiding regulation, (d) they prevent us from exploiting the positive potential of AI, and (e) the signatories are just fuelling counterproductive AI hype.

Previous research has suggested that there is common ground to be found between these competing perspectives [3]. While this strategy has some merit, we suggest that the debates about AI risk could benefit from adopting established risk analysis and management practices. This allows for systematic and transparent assessment of risk, despite genuine disagreement, and could help engender a more fruitful dialogue between opposed groups.

2 Controversial long-term risks

Existential risks (or x-risks) stemming from AI have long been debated in scholarly circles, and depicted in science fiction books and movies. While superintelligent AI systems could easily be imagined to aid human and (post)human development, what scares many of the originators of these systems, such as George Hinton, is the potential for such systems to take control and act in a manner contrary to human survival and flourishing.

Some members of the AI ethics community are, however, outraged by the attention devoted to existential AI risk. The list of experts attacking the FLI open letter and the CAIS statement is very long, and includes, for example, Timnit Gebru, Safiya Umoja Noble, Emily Bender, Meredith Whittaker, and Deb Raji and many more [4]. Their reactions are too numerous to recount in full, but here we provide some select examples.

✉ Henrik Skaug Sætra
Henrik.satra@hiof.no

[1] Faculty of Computer Science, Engineering and Economics, Østfold University College, Remmen, 1757 Halden, Norway

[2] School of Law, NUI Galway, Galway, Ireland

Published online: 04 September 2023

Ryan Calo responds to the CAIS statement by saying that 'if AI threatens humanity, it's by accelerating trends of wealth and income inequality, lack of integrity in information' and the exploitation of natural resources.[1] We should instead focus our time and attention on issues of privacy, bias, and environmental and social impacts, he says, as these things are 'are actually happening'. This is echoed by Mark Riedl, who argues that focusing on existential threats implies that 'other harms are not happening or are not of consequence' [5]. The problem, he says, is that research funding and attention is limited, which seems to suggest that we must choose our worries wisely.

Joanna Bryson states that the CAIS statement is 'openly regulatory interference'.[2] She calls existential risk a 'fantasy' that distracts from 'real issues', and argues that 'the elite'—referring to those signing the statement of the FLI open letter—seeks to 'build regulatory institutions to consolidate' the current status quo. She also suggested that the statement is really about 'slowing/misdirecting/perverting' the European Union's coming AI Act.[3] Such concerns were recently mirrored in a Nature editorial highlighting the danger that discussing x-risk entails overlooking immediate concerns and preventing us from living 'in harmony with the technology' [6].

3 Understanding risk

These reactions might seem strange to anyone used to corporate and social risk management. In this world, assessing and simultaneously dealing with very different types of risks, even if they are unlikely, is the name of the game. This requires some finesse, and different risks must be dealt with differently. Crucially, however, achieving this requires a systematic approach and a shared vocabulary.

AI risks might seem both novel and recent, but concerns about emerging technologies go way back [7], and systematic approaches for assessing and managing risk can be found as far back as 3200 B.C. [8]. The notion of the risk society—often linked to Ulrich Beck and Anthony Giddens—provides some cues for identifying key sources of difficulties in achieving fruitful discussions of risk in the AI ethics field. Mainstream approaches to risk are usually referred to as risk management, and while most corporations already have some sort of risk management system in place, it is not often used or referred to by AI ethicists debating, for example, extinction risks from AI.

The Institute of Risk Management (IRM) defines risk as 'the combination of the probability of an event and its consequence', where consequences can be positive (opportunity risk) and negative [9]. In risk management, risk matrices are often used to get an overview over how to prioritize risk responses by their likelihood and impact. Risk assessment necessarily precedes such matrices, and the following aspects are considered for each potential risk:

- magnitude of the event (harm or benefit) should the risk materialize;
- size of the impact that the event would have on the organization;
- likelihood of the risk materializing at or above the benchmark;
- scope for further improvement in control [9].

Examples of potentially relevant AI risks are tentatively placed in the risk matrix shown in Fig. 1. For example, runaway AI resulting in human extinction could be categorized as low in likelihood, and very high in impact if it should occur. Bias and discrimination resulting from the use of data-based AI systems could be argued to have a lower potential overall impact (relative to extinction risk), but the likelihood of occurrence is extremely high. Likewise, the environmental impacts of training and running AI systems in vast data centers are high-probability events, but could be argued to have a medium/low impact if compared to high emitting activities such as construction. Proper assessment of the risks shown in the figure requires extensive analysis, of course, and the end placement and evaluation of likelihood and impact could be very different from the hypothetical illustration we have here provided.

How does this help to address the challenges related to fostering fruitful debates about AI risks and regulation? The example of the CAIS statement illustrates some challenges related to debates about existential risk. Such risks are problematic in terms of traditional risk management, as 'identifying, evaluating, and managing such existential threats is often extremely difficult and ultimately may be uncontrollable' [10].

Nevertheless, all major organizations that occupy themselves with risk analysis include various low likelihood and high impact risks. While existential AI risk has not yet entered most of the risk and trend reports, the US National Intelligence Council, for example, in their Global Trends Report 2040, from 2021, includes 'runaway AI' as one existential risk, alongside 'engineered pandemics, nanotechnology weapons, or nuclear war' [11].

Non-AI-related existential risks have been discussed for a long time. In a 1989 article on risk management and

[1] https://twitter.com/rcalo/status/1663614234369097728?s=20.

[2] https://www.linkedin.com/posts/bryson_openai-reveals-the-anti-regulatory-intent-activity-7069579849245347840-4Ts_?utm_source=share&utm_medium=member_desktop.

[3] https://twitter.com/j2bryson/status/1663794967804882944?s=20.

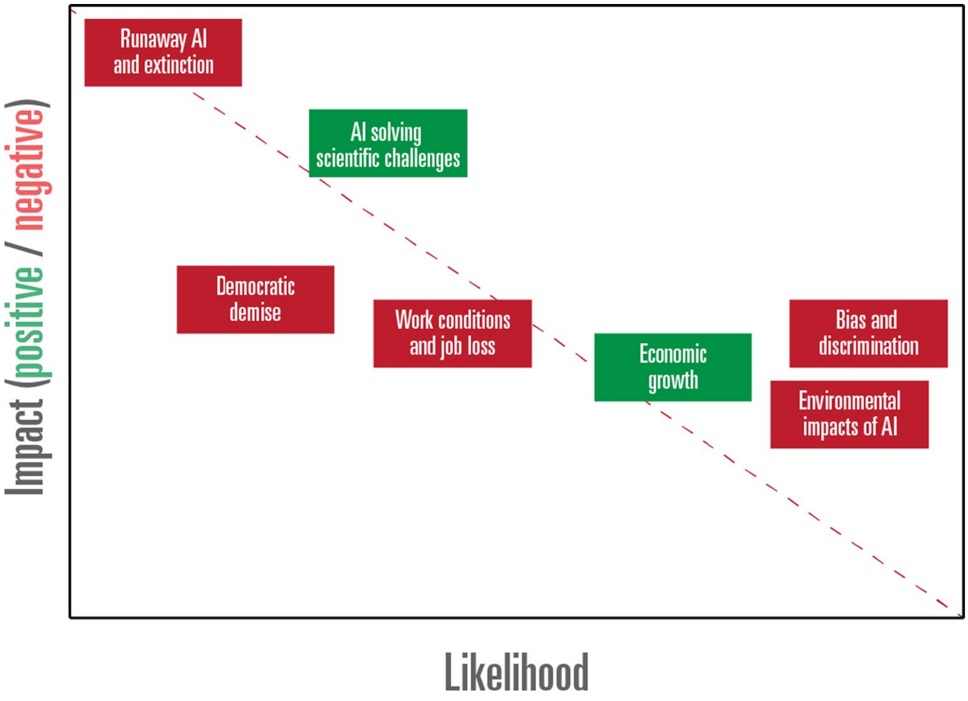

Fig. 1 Risk matrix on a societal level, categorizing risks by impact and likelihood

existential risk, it was pointed out that there are two cognitive traps into which we may fall when assessing such risks: the 'non-perception' of real existential risks; and 'the conjuring up of imaginary risks' [12]. CAIS believe that non-perception is a problem, while its critics fear the latter.

From a risk management perspective, we must avoid both these traps, and this can arguably be achieved through probability-based risk management. If we overlook unlikely but real risks, this could be catastrophic. Certain long-term and potentially unlikely risks must be seriously considered to properly prepare for the future and mitigate the risks. Runaway AI is one such example, and preparation for extraterrestrial contact [13] and discussions about potential robot rights [14] are others. Gordon [15] discusses the latter, and highlights how 'uncertain but not impossible' risks cannot be neglected. On the other hand, if we exaggerate the risk, the cost of mitigation could outweigh the real risk, and we could lose out on many potential benefits to be had from appropriate use of AI.

However, taking such risks seriously does not mean we should divert attention from more immediate risks. Two of the IRM's principles of risk management—proportionality and alignment—provide us with a framework for understanding how to deal with different types of risks simultaneously. The proportionality principle highlights that risk management activities must reflect the level of risk, while the alignment principle stresses the need to see risks as a whole, so that various risk management responses must be aligned [9]. For example, dealing with AI bias and discrimination must be done while simultaneously taking proportionate actions to mitigate long-term risks.

An analogy might help us illustrate this point. An energy- and carbon-intensive organization producing chemicals might face a range of short-term hazards and risks, such as environmental spills, while also facing long-term risks that fundamentally challenge its basic operating model and future existence. In a strategy meeting, these two concerns could be voiced by two different executives. A risk management-trained C-suite would recognize the merit of both and proceed to develop risk mitigation strategies for short-term environmental harms while also preparing long-term changes of their operating model.

A similar situation arises for AI systems developed, for example, for use in political contexts [16, 17]. Short-term and certain risks related to bias and discrimination in public services are clear and obvious short-term risks with high probabilities. Gradually losing control over water and electrical infrastructure if these become increasingly reliant on AI is somewhat longer term risk, while the 'runaway AI' system intentionally or negligently actively destroying humanity through, for example, water distribution systems and the control of chemicals in the water, would be a long-term and low probability risk. The coming European 'AI act' is 'risk-based,' and it presumes that the gravity of risk is linked to the area of deployment. The benefits of this approach can be seen in this example, as AI used to control critical infrastructure, for example, requires extra care. However, the area of deployment-specific approach is not sufficient for understanding all AI risk, and we agree with others that this is not a sufficiently sophisticated approach [18].

On the company level, taking strong action in response to all risks is not the only possible outcome. The four standard hazard risk management actions—to tolerate, treat, or transfer the risk, or to terminate the activities causing the risk [9]—demonstrate how the analysis and recognition of a risk can be coupled with non-action once all risks are analysed. The company could, for example, acknowledge the CAIS warning but tolerate the risk for now. They could also terminate certain actions, as called for in the FLI letter requesting a (temporary) moratorium. Treating the risk means making changes to mitigate the risk, and transferring implies that they try to make it someone else's problem—a nice option for corporations—not for society.

However, private sector self-regulation is not the only option, and it has historically been less prominent than the three other social level strategies for social risk mitigation: insurance, law, and government intervention [8]. Insurance entails acceptance of a certain degree of risk, and the marketization of harms and damages. Whenever risk is seen as unavoidable, or its effects outweigh the negative implications, individuals can pool their resources and eliminate individual level risk through markets. Law and regulation could entail introducing liability for risky behaviour and systems and introduces both means of compensation for those harmed and deterrence. Direct government intervention goes further and opens for bans and mandated changes in behaviour. Avoiding the latter two actions is at times seen as the goal of the big tech companies concerned, and explains why some refer to industry talk of extinction risk as regulatory interference or sabotage. This assumes that the goal of the companies is not primarily to reduce risk, but to establish an impression of private sector self-regulation. To avoid regulation and interference, companies must both show that they are responsible and avoid overly risky action that would prompt intervention.

4 Can established risk management approaches reconcile warring camps?

If we now return to the AI ethicists quarrelling over which AI risks matter and why, the ongoing debates suggest that they could be sorted into two groups, with some pointing to the divide between what is labelled the 'AI safety' group and the 'AI ethics' group [19]. Those concerned with artificial general intelligence, extinction risk, alignment, and (very) long-term risks are often placed in the former category, while those working on, for example, mitigating bias and discrimination in AI systems are placed in the latter. The first group is largely open to the idea of quantitatively and objectively assessing all risks; the second more often view the conflict through political and ideological lenses not amenable to the language of traditional risk management. This divide might, but need not, also follow various disciplinary fields and backgrounds, as some have extensive training in computer science but not philosophy and economics, for example, while others might have an opposite balance of backgrounds [15].

However, this dichotomy oversimplifies the issue, and obscures the fact that different AI ethicists might in fact be discussing completely different aspect of AI risk, and the split into two camps is arbitrary and unnecessary. One particularly important point is that the very idea of quantitatively and objectively assessing various risks can be seen as a political undertaking [7]. Some see risk as socially constructed and reject the idea of 'real' and objective risk [7]. By choosing to focus on certain risks, we privilege certain perspectives and positions, and some might also see it as validating and opening for futures that they do not want. AI risk is seen as a part of politics and power relations. When arguments about risk are seen as a deeply political, it makes sense to focus on various individual's motivations and the consequences of just talking about a particular risk, rather than agreeing to engage in quantification and traditional risk assessments.

Despite these concerns, we argue that both groups could find the risk management framework suggested in this article useful for engaging with their opposition. The group that balks at talk of x-risk could, for example, argue that those calling for a prioritisation of existential risk are overweighting these risks and underweighting the more pressing ones. This is problematic if scientific, public, and/or regulatory attention is a scarce good, leading to the conclusion that it might be morally right to suppress certain approaches [20]. This is a view that can be meaningfully engaged with by proponents of existential risk, and the disagreement could be subject to a deliberative resolution. For example, it might be concluded that extinction risks will be acknowledged but tolerated, or that regulatory attention is not so scarce, or that there are sufficient/insufficient resources available to manage the different sets of risks. The two sides might even agree to strategically align for the sake of passing systematic regulation that addresses both sets of risks, during a political window of opportunity.

However, some might still think that the risk management approach concedes too much. For them, talk of 'existential risk' is a red flag, because its purveyors are thought to harbour a more sinister agenda. Emily Bender, for instance, has argued that x-riskers are 'not natural allies' of those concerned with real, short-term risks, because they are powerful 'johnny-come-latelys' that engage in 'ridiculous distraction tactics', overlooking the longstanding work of AI critics.[4] Timnit Gebru[5] and Émile Torres [21, 22] go so far as to link the x-risk fixation and 'longtermism' to an ideology grounded in eugenics. While we would not claim to speak for them, we suspect that sitting down and agreeing upon a risk matrix is unlikely to be seen as a viable option for those adopting such a view, though there is also the possibility (perhaps slim) of forming a politically convenient détente between such critics, as suggested by Stix and Maas, to pursue regulatory intervention in a few areas of overlapping concern (e.g., algorithm audits, bans on military use of AI) [24]. This also highlights the need to distinguish between the idea of actual and perceived risk, and research has shown that perceived risks varies by a range of variables, such as gender, ethnicity, age, education, experience, etc. [7]. Nevertheless, it would be helpful for all parties if we managed to acknowledge the differences that stem from a) the objective evaluation of isolated risks and b) the political implications of even debating risks as isolated and objectively evaluable.

If we assume that both sides discussed in this article are motivated by a desire to avoid human and societal harm from AI, they could well benefit from real engagement with each other's positions. Without such engagement, we run the risk of 'organized irresponsibility, unaccountability and uninsurability' [23], which Beck feared would ensue if we got into a situation of 'a collective avoidance of responsibility for risk management' [7].

Author contributions HSS and JD contributed equally to this research. All authors read and approved the final manuscript.

Funding Open access funding provided by Ostfold University College. The authors declare that no funds, grants, or other supports were received during the preparation of this manuscript.

Data availability Not applicable.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical approval Not applicable

Consent to participate Not applicable.

Consent for publication Not applicable.

[4] https://twitter.com/emilymbender/status/1665054683311017984.

[5] https://twitter.com/timnitGebru/status/1665449574159339526.

References

1. Center for AI Safety. Statement on AI Risk: AI experts and public figures express their concern about AI risk. https://www.safe.ai/statement-on-ai-risk. Accessed 30 May 2023
2. Future of Life Institute. Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/. Accessed 31 May 2023
3. Cave, S., ÓhÉigeartaigh, S.S.: Bridging near-and long-term concerns about AI. Nat Mach Intell **1**(1), 5–6 (2019). https://doi.org/10.1038/s42256-018-0003-2
4. Wong, M. (ed): AI doomerism is a decoy. In: The Atlantic. (2023)
5. Goldman, S. (ed): AI experts challenge 'doomer' narrative, including 'extinction risk' claims. In: VentureBeat. (2023)
6. The editorial board.: Stop talking about tomorrow's AI doomsday when AI poses risks today. Nature **618**, 885–886 (2023)
7. Ekberg, M.: The parameters of the risk society: a review and exploration. Curr. Sociol. **55**(3), 343–366 (2007)
8. Covello, V.T., Mumpower, J.: Risk analysis and risk management: an historical perspective. Risk Anal. **5**(2), 103–120 (1985). https://doi.org/10.1111/j.1539-6924.1985.tb00159.x
9. Hopkin, P., Thompson, C.: Fundamentals of risk management: understanding, evaluating and implementing effective risk management, 5th edn. Kogan Page Publishers, London (2021)
10. Glendon, A.I., Clarke, S., McKenna, E.: Human safety and risk management. CRC Press, Florida (2016)
11. National Intelligence Council, Global Trends 2040: A More Contested World 2021.: https://www.dni.gov/index.php/gt2040-home. Accessed 15 Apr 2023
12. Best, C.F.: Risk takes on an existential nature. Risk Manage. **36**(2), 52–53 (1989)
13. Neal, M.: Preparing for extraterrestrial contact. Risk Manage. **16**, 63–87 (2014)
14. Gunkel, D.J.: Robot rights. MIT Press, London (2018)
15. Gordon, J.-S.: The impact of artificial intelligence on human rights legislation: a plea for an AI convention. Palgrave Macmillan, Cham (2023)
16. Sadowski, J., Selinger, E.: Creating a taxonomic tool for technocracy and applying it to silicon valley. Technol. Soc. **38**, 161–168 (2014)
17. Sætra, H.S.: A shallow defence of a technocracy of artificial intelligence: Examining the political harms of algorithmic governance in the domain of government. Technol. Soc. **62**, 101283 (2020)
18. Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., Floridi, L.: How to evaluate the risks of artificial intelligence: a proportionality-based, risk model for the AI Act. SSRN Electron J (2023). https://doi.org/10.2139/ssrn.4464783
19. Piper, K. (ed): There are two factions working to prevent AI dangers. Here's why they're deeply divided. In: Vox (2022)

20. Sætra, H.S., Fosch-Villaronga, E.: Research in AI has implications for society: how do we respond? Morals & Mach **1**(1), 60–73 (2021)
21. Torres, É.P. (ed): Eugenics in the Twenty-First Century: New Names, Old Ideas. In: Truthdig (2023)
22. S. Linton. Tech Elite's AI Ideologies Have Racist Foundations, Say AI Ethicists. People of Color in Tech. https://peopleofcolorintech.com/articles/timnit-gebru-and-emile-torres-call-out-racist-roots-of-the-tech-elites-ai-ideologies/. Accessed 9 Aug 2023
23. Beck, U.: Ch 1 Politics of risk society. In: Franklin, J. (ed.) Politics of risk society, pp. 9–22. Polity Press, Cambridge (1998)
24. Stix, C., Maas, M.M.: Bridging the gap: the case for an 'incompletely theorized agreement'on AI policy. AI and Ethics **1**(3), 261–271 (2021). https://doi.org/10.1007/s43681-020-00037-w

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.