Title

AI and the falling sky: interrogating X-Risk.

Journal

Journal of Medical Ethics, 50(12)

Authors

Jecker, Nancy
Atuire, Caesar
Bélisle-Pipon, Jean-Christophe
et al.

Publication Date

2024-12-23

DOI

10.1136/jme-2023-109702

Peer reviewed

OPEN ACCESS

AI and the falling sky: interrogating X-Risk

Nancy S Jecker [1,2] Caesar Alimsinya Atuire [3,4] Jean-Christophe Bélisle-Pipon [5] Vardit Ravitsky [6,7] Anita Ho [8,9]

For numbered affiliations see end of article.

Correspondence to
Dr Nancy S Jecker, Department of Bioethics & Humanities, University of Washington School of Medicine, Seattle, Washington, USA; nsjecker@uw.edu

A version of this paper will be presented at The Center for the Study of Bioethics, The Hastings Center, and The Oxford Uehiro Centre for Practical Ethics conference, "Existential Threats and Other Disasters: How Should We Address Them," June 2024, Budva, Montenegro.

Received 2 November 2023
Accepted 14 March 2024

To cite: Jecker NS, Atuire CA, Bélisle-Pipon J-C, *et al*. *J Med Ethics* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jme-2023-109702

ABSTRACT

This paper argues that the headline-grabbing nature of existential risk (X-Risk) diverts attention away from immediate artificial intelligence (AI) threats, including fairly disseminating AI risks and benefits and justly transitioning towards AI-centred societies. Section I introduces a working definition of X-Risk, considers its likelihood and explores possible subtexts. It highlights conflicts of interest that arise when tech luminaries lead ethics debates in the public square. Section II flags AI ethics concerns brushed aside by focusing on X-Risk, including AI existential benefits (X-Benefits), non-AI X-Risk and AI harms occurring now. Taking the entire landscape of X-Risk into account requires considering how big risks compare, combine and rank relative to one another. As we transition towards more AI-centred societies, which we, the authors, would like to be fair, we urge embedding fairness in the transition process, especially with respect to groups historically disadvantaged and marginalised. Section III concludes by proposing a wide-angle lens that takes X-Risk seriously alongside other urgent ethics concerns.

INTRODUCTION

The Buddhist Jātaka tells the tale of a hare lounging under a palm tree who becomes convinced the Earth is coming to an end when a ripe bael fruit falls on its head. Soon all the hares are running; other animals join them, forming a stampede of deer, boar, elk, buffalo, wild oxen, rhinoceros, tigers and elephants, loudly proclaiming the earth is ending.[1] In the American retelling, the hare is 'chicken little,' and the exaggerated fear is that the sky is falling.

The story offers a cautionary tale for considering the trend towards calamity thinking in artificial intelligence (AI). A growing chorus of tech leaders has warned that AI poses existential risk (X-Risk) that could result in the extinction of the human species, the collapse of civilisation, or a colossal decline in human potential and culture. In 2014, Hawking told the Washington Post that AI, 'could spell the end of the human race'.[2] Musk has repeatedly warned about AI's perils, calling AI, 'more dangerous than nukes', recommending colonising Mars to ensure 'a bolt-hole if AI goes rogue and turns on humanity',[3] and donating 10 million dollars to the Future of Life Institute to jumpstart research on AI's X-Risk. Gates has stated, 'I am in the camp…concerned about super intelligence…I agree with … Musk and some others on this and don't understand why some people are not concerned'.[4] Altman, CEO of OpenAI, has fretted that technologies OpenAI was building could endanger humanity—'even destroying the world as we know it'.[5]

This paper offers a critical appraisal of the rise of calamity thinking in the scholarly AI ethics literature. It cautions against viewing X-Risk in isolation and highlights ethical considerations sidelined when X-Risk takes centre stage. Section I introduces a working definition of X-Risk, considers its likelihood and explores possible subtexts. It highlights conflicts of interest that arise when tech luminaries lead ethics debates in the public square. Section II flags ethics concerns brushed aside by focusing on X-Risk, including AI existential benefits (X-Benefits), non-AI X-Risk and non-existential AI harms. As we transition towards more AI-centred societies, which we, the authors, would like to fair, we argue for embedding fairness in the transition process by ensuring groups historically disadvantaged or marginalised are not left behind. Section III concludes by proposing a wide-angle lens that takes X-Risk seriously alongside other urgent ethics concerns.

I. UNPACKING X-RISK

Doomsayers imagine AI in frightening ways, a paperclip maximiser, 'whose top goal is the manufacturing of paperclips, with the consequence that it starts transforming first all of earth and increasing portions of space into paperclip manufacturing facilities.'(Bostrom, p5)[6] They compare large language models (LLMs) to the shoggoth in Lovecraft's novella, 'a terrible, indescribable thing…a shapeless congeries of protoplasmic bubbles, … with myriads of temporary eyes…as pustules of greenish light all over…'.[7]

Prophesies of annihilation have a runaway effect on the public's imagination. Schwarzenegger, star of *The Terminator*, a film depicting a computer defence system that achieves self-awareness and initiates nuclear war, has stated that the film's subject is 'not any more fantasy or kind of futuristic. It is here today' and 'everyone is frightened'.[8] Public attention to X-Risk intensified in 2023, when The Future of Life Institute called on AI labs to pause for 6 months the training of AI systems more powerful than Generative Pre-Trained Transformer (GPT)-4,[9] and, with the Centre for AI Safety, spearheaded a Statement on AI Risk, signed by leaders from OpenAI, Google Deepmind, Anthropic and others stressing that, '(m)itigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war'.[10] The 2023 release of Nolan's film, *Oppenheimer,* encouraged comparisons between AI and atomic weaponry. Just as Oppenheimer fretted unleashing atomic energy 'altered abruptly and profoundly the nature of the world,' and 'might

someday prove deadly to the whole civilisation', tech leaders fret AI X-Risk.(Bird, p323)[11]

The concept of 'X-Risk' traces to Bostrom, who in 2002 defined it as a risk involving, 'an adverse outcome (that) would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential;' on this rendering, X-Risk imperils 'humankind as a whole' and brings 'major adverse consequences for the course of human civilisation for all time to come.'(Bostrom, p2)[12] More recently, Bostrom and Ćirković defined 'X-Risk' as a subset of global catastrophic risks that 'threatens to cause the extinction of Earth-originating intelligent life or to reduce its quality of life (compared with what would otherwise have been possible) permanently and drastically.'(Bostrom, p4)[13] They classify global catastrophic risks that could become existential in scope, intensity and probability as threefold: risks from nature such as asteroid threats; risks from unintended consequences, such as pandemic diseases; and risks from hostile acts, such as nuclear weaponry. We use Bostrom and Ćirković's account as our working definition of X-Risk. While it is vague in the sense of leaving open the thresholds for scope, intensity and probability, it carries the advantage of breadth and relevance to a range of serious threats.

Who says the sky is falling?

A prominent source of apocalyptic thinking regarding AI comes from within the tech industry. According to a *New York Times* analysis, many tech leaders believe that AI advancement is inevitable, because it is possible, and think those at the forefront of creating it know best how to shape it.[14] In a 2019 scoping review of global AI ethics guidelines, Jobin *et al* identified 84 documents containing AI ethics principles or guidance, with most from the tech industry.(Jobin, p396)[15] However, a limitation of the study was that ethics guidance documents represent 'soft law,' which is not indexed in conventional databases, making retrieval less replicable and unbiased. More recently, Stanford University's 2023 annual AI Index Report examined authorship of scholarly AI ethics literature and reported a shift away from academic authors towards authors with industry affiliations; the Report showed industry-affiliated authors produced 71% more publications than academics year over year between 2014 and 2022.[16]

Since AI companies benefit financially from their investments in AI, relying on them for ethics guidance creates a conflict of interest. A 'conflict of interest' is a situation where 'an individual's judgement concerning a primary interest tends to be unduly influenced (or biased) by a secondary interest.'(Resnik, p121–22)[17] In addition to financial conflicts of interest, non-financial conflicts of interest can arise from multiple sources (eg, personal or professional relationships, political activity, involvement in litigation).[17] Non-financial conflicts of interest can occur subconsciously, and implicit cognitive biases can transfer to AI systems. Since most powerful tech companies are situated in high-income Western countries, they may be implicitly partial to values and concerns prevalent in those societies, reflecting anchoring bias (believing what one wants or expects) and confirmation bias (clinging to beliefs despite conflicting evidence). The dearth of research exploring AI's social impacts in diverse cultural settings around the world makes detecting and dislodging implicit bias difficult.[18] Commenting on the existing corpus of AI ethics guidance, Jobin *et al* noted a significant representation of more economically developed countries, with the USA and UK together accounting for more than a third of AI ethics principles in 2019, followed by Japan, Germany, France and Finland. Notably, African and South American countries were not represented. While authors of AI ethics guidance often purport to represent the common good, a 2022 study by Bélisle-Pipon *et al* showed a broad trend towards asymmetrical engagement, with industry and those with vested interests in AI more represented than the public.[19] Hagerty and Rubinov report that risks for discriminatory outcomes in machine learning are particularly high for countries outside the USA and Western Europe, especially when algorithms developed in higher-income countries are deployed in low-income and middle-income countries that have different resource and social realities.[18]

Another prominent source of calamity thinking is members of the effective altruism movement and the associated cause of longtermism, two groups that focus on 'the most extreme catastrophic risks and emphasise the far-future consequences of our actions'.[20] Effective altruism is associated with a philosophical and social movement based largely at Oxford University and Silicon Valley. Its members include philosophers like Singer, Ord and MacAskill, along with tech industry leaders like the discredited cryptocurrency founder, Bankman-Fried. The guiding principles of effective altruism are 'to do as much good as we can' and 'to base our actions on the best available evidence and reasoning about how the world works'.[21] MacAskill defines longtermism as 'the idea that positively influencing the long-term future is a key moral priority of our time', and underscores, 'Future people count. There could be a lot of them. We can make their lives go better.'(MakAskill, pp5, 21)[22] Effective altruism and longtermism have spawned charitable organisations dedicated to promoting its goals, including GiveWell, Open Philanthropy and The Future of Life Institute. To be clear, we are not suggesting that adherents of longtermism are logically forced to embrace X-Risk or calamity thinking; our point is that adherents of longtermism draw on it to justify catastrophising.

Who benefits and who is placed at risk?

Critics of longtermism argue that it fails to give sufficient attention to serious problems happening now, particularly problems affecting those who have been historically disadvantaged or marginalised. Worse, it can give warrant to sacrificing present people's rights and interests to stave off a prophesied extinction event. Thus, a well-recognised danger of maximisation theories is that they can be used to justify unethical means if these are deemed necessary to realise faraway goals that are thought to serve a greater good. Some effective altruists acknowledge this concern. MacAskill, for example, concedes that longtermism endorses directing resources away from present concerns, such as responding to the plight of the global poor, and towards more distant goals of preventing X-Risk.[23]

X-Risk also raises theoretical challenges related to intergenerational justice. How should we understand duties to future people? Can we reasonably argue that it is unfair to prioritise the interests of existing people? Or even that in doing so, we discriminate against future people? Ord defends longtermism on the ground that there are many more future people than present people: 'When I think of the millions of future generations yet to come, the importance of protecting humanity's future is clear to me. To risk destroying this future, for the sake of some advantage limited only to the present, seems to me profoundly parochial and dangerously short-sighted. Such neglect privileges a tiny sliver of our story over the grand sweep of the whole; it privileges a tiny minority of humans over the overwhelming majority yet to be born; it privileges this particular century over the millions, or maybe billions, yet to come' (Ord, p44).[24]

MacAskill defends longtermism on slightly different grounds, arguing that it reflects the standpoint of all humanity: 'Imagine

living…through the life of every human being who has ever lived…(and) imagine that you live all future lives…If you knew you were going to live all these future lives, what would you hope we do in the present?'(MakAskill, p5)[22] For MacAskill, the standpoint of all humanity represents the moral point of view.

The logic of longtermism can be challenged on multiple grounds. First, by purporting to represent everyone, longtermism ignores its own positionality. Longtermism's central spokespersons—from the tech industry and effective altruism movement, are not sufficiently diverse to represent 'all humanity.' A 2022 *Time Magazine* article characterised 'the typical effective altruist' as 'a white man in his 20 s, who lives in North America or Europe, and has a university degree'.[25] The tech industry, which provides robust financial backing for longtermism, faces its own diversity crisis across race and gender lines. In 2021, men represented nearly three-quarters of the USA science, technology, engineering and mathematic workforce, whites close to two-thirds.[26] At higher ranks, diversity rates were lower.

Someone might push back, asking why the narrow demographics of the average effective altruist or adherent of longtermism should be a source for concern. One reply is that these demographics raise the worry that the tech industry is unwittingly entrenching its own biases and transferring them to AI systems. Experts caution about AI 'systems that sanctify the status quo and advance the interests of the powerful', and urge reflection on the question, 'How is AI shifting power?'(Kalluri, p169)[27] While effective altruism purports to consider all people's interests impartially, linking altruism to distant future threats delegitimises attention to present problems, leaving intact the plight of today's disadvantaged. Srinivasan asserts that 'the humanitarian logic of effective altruism leads to the conclusion that more money needs to be spent on computers: why invest in anti-malarial nets when there's a robot apocalypse to halt?'[28] These kinds of considerations lead Srinivasan to conclude that effective altruism is a conservative movement that leaves everything just as it is.

A second, related worry concerns epistemic justice, the normative requirement to be fair and inclusive in producing knowledge and assigning credibility to beliefs. The utilitarian philosophy embedded in effective altruism and longtermism is a characteristically Western view. Since effective altruism and longtermism aspire to be a universal ethic for humankind, considering moral philosophies outside the West is a normative requirement epistemic justice sets. Many traditions outside the West assign core importance to the fact that each of us is 'embedded in the complex structure of commitments, affinities and understandings that comprise social life'.[28] The value of these relationships is not derivative of utilitarian principles; it is the starting point for moral reasoning. On these analyses, the utilitarian premises of longtermism and effective altruism undervalue community and thereby demand the wrong things. If the moral goal is creating the most good you can, this potentially leaves out those collectivist-oriented societies that equate 'good' with helping one's community and with promoting solidaristic feeling between family, friends and neighbours.

Third, evidence suggests that epistemically just applications of AI require knowledge of the social contexts to which AI is applied. Hagerty and Rubinov report that 'AI is likely to have markedly different social impacts depending on geographical setting' and that 'perceptions and understandings of AI are likely to be profoundly shaped by local cultural and social context'.[18] Lacking contextual knowledge impacts AI's potential benefits[29] and can harm people.[30] While many variables are relevant to social context, when AI developers are predominantly white, male and from the West, they may miss insights that a more diverse demographic would be less apt to miss. This can create an echo chamber, with dominant views seeming 'natural' because they are pervasive and unchallenged.

An adherent of longtermism might reply to these points by saying that most people are deficient in their concern for future people. According to Perrsron and Savulescu, interventions like biomedical moral enhancement might one day enable individuals to be 'less biased towards what is near in time and place' and to 'feel more responsible for what they collectively cause and let happen'.(Perrsron and Savulescu, p496)[31] Presumably, morally enhancing people in ways that direct them to care more about distant future people would help efforts to reduce X-Risk. Yet, setting aside biomedical feasibility, this argument brushes aside preliminary questions. Whose moral views require enhancing? Perrson and Savulescu suggest that their own emphasis on distant future people is superior, while the views of others, who prioritise present people, require enhancing. Yet, this stance is incendiary and potentially offensive. Implementing biomedical moral enhancement would not show the superiority of longtermism; it would shut down alternative views and homogenise moral thinking.

A different reply is suggested by MacAskill, who compares longtermism to the work of abolitionists and feminists. (MakAskill, p3)[22] MacAskill says future people will look back and thank us if we pursue the approach longtermism advocates, just as present people are grateful to abolitionists and feminists who dedicated themselves to missions that succeeded decades after their deaths. Yet this ignores the thorny question of timing—feminists and abolitionists responded to justice concerns of their time and place, and helped the next generation of women and blacks, while longtermists presumably help people in the distant future to avoid the end of humanity. Yet, those who never exist (because they are eliminated by AI) are not wronged by never having existed.

Finally, proponents of X-Risk might reason that even though the odds of X-Risk are uncertain, the potential hazard it poses is grave. Yet, what exactly are the odds? Bostrom and Ćirković acknowledge AI X-Risk is 'not an ongoing or imminent global catastrophic risk;' nonetheless, 'from a long-term perspective, the development of general AI exceeding that of the human brain can be seen as one of the main challenges to the future of humanity (arguably, even as *the* main challenge).'(Rees, p16)[32] Notwithstanding this qualification, the headline-grabbing nature of X-Risk makes X-Risk itself risky. It is readily amplified and assigned disproportionate weight, diverting attention from immediate threats. For this reason, tech experts warn against allowing the powerful narratives of calamity thinking to anchor risk assessments. Unlike other serious risks, AI X-Risk forecasting cannot draw on empirical evidence: 'We cannot consult actuarial statistics to assign small annual probabilities of catastrophe, as with asteroid strikes. We cannot use calculations from a precise, precisely confirmed model to rule out events or place infinitesimal upper bounds on their probability, (as) with proposed physics disasters.'(Yudkowsky, p308)[33] We can, however, apply time-tested methods of risk reduction to lower AI X-Risk. Hazard analysis, for example, defines 'risk' by the equation: risk=hazard×exposure×vulnerability. On this approach, reducing AI X-Risk requires reducing hazard, exposure and/or vulnerability; for example, establishing a safety culture reduces hazard; building safety into system development early-on reduces risk exposure; and preparing for crises reduces vulnerability.

II. WHAT RISKS OTHER THAN AI X-RISK SHOULD WE CONSIDER?

This section explores ethics consideration besides X-Risk. In so doing, it points to the need for a broader ethical framing, which we develop in a preliminary way in the next section (section III).

Non-AI X-Risks

Before determining what moral weight to assign AI X-Risk, consider non-AI X-Risks. For example, an increasing number of bacteria, parasites, viruses and fungi with antimicrobial resistance could threaten human health and life; the use of nuclear, chemical, biological or radiological weapons could end the lives of millions or make large parts of the planet uninhabitable; extreme weather events caused by anthropogenic climate change could endanger the lives of many people, trigger food shortages and famine, and annihilate entire communities. Discussion of these non-AI X-Risks is conspicuously absent from most discussions of AI X-Risk.

A plausible assumption is that these non-AI threats have at least as much likelihood of rising to the level of X-Risk as AI does. If so, then our response to AI X-Risk should be proportionate to our response to these other dangers. For example, it seems inconsistent to halt developing AI systems due to X-Risk, while doing little to slow or reduce the likelihood of X-Risk from nuclear weaponry, anthropogenic climate change or antimicrobial resistance. All these possible X-risks are difficult to gauge precisely; moreover, they intersect, further confounding estimates of each. For example, AI might accelerate progress in green technology and climate science, reducing damaging effects of climate change; alternatively, AI might increase humanity's carbon footprint, since more powerful AI takes more energy to operate. The most promising policies simultaneously reduce multiple X-Risks, while the most destructive ones increase multiple X-Risks. Taking the entire landscape of X-Risk into account requires considering how big risks compare, combine and rank relative to one another.

The optimal strategy for reducing the full range of X-Risks might involve less direct strategies, such as building international solidarity and strengthening shared institutions. The United Nations defines international solidarity as 'the expression of a spirit of unity among individuals, peoples, states and international organisations. It encompasses the union of interests, purposes and actions and the recognition of different needs and rights to archive common goals.'[34] Strengthening international solidarity could better equip the world to respond to existential threats to humanity, because solidarity fosters trust and social capital. Rather than undercutting concern about people living in the distant future, building rapport with people living now might do the opposite, that is, foster a sense of common humanity and of solidarity between generations.

One way to elaborate these ideas more systematically draws on values salient in sub-Saharan Africa, which emphasise solidarity and prosocial duties. For example, expounding an African standpoint, Behrens argues that African philosophy tends to conceive of generations past, present and future as belonging to a shared collective and to perceive, 'a sense of family or community' spanning generations.[35] Unlike utilitarian ethics, which tends to focus on impartiality and duties to strangers, African solidarity may consider it ethically incriminating to impose sacrifices on one to help many, because each member of a group acquires a superlative value through group membership.(Metz, p62)[36] The African ethic of ubuntu can be rendered as a 'family first' ethic, permitting a degree of partiality towards present people. Utilitarianism, by contrast, requires impartially maximising well-being for all people, irrespective of their proximity or our relationship to them. While fully exploring notions like solidarity and ubuntu is beyond this paper's scope, they serve to illustrate the prospect of anchoring AI ethics to more diverse and globally inclusive values.

AI X-Benefits

In addition to non-AI X-Risk, a thorough analysis should consider AI's X-Benefits. To give a prominent example, in 2020, DeepMind demonstrated its AlphaFold system could predict the three-dimensional shapes of proteins with high accuracy. Since most drugs work by binding to proteins, the hope is that understanding the structure of proteins could fast-track drug discovery. By pinpointing patterns in large data sets, AI can also aid diagnosing patients, assessing health risks and predicting patient outcomes. For example, AI image scanning can identify high risk cases that radiologists might miss, decrease error rates among pathologists and speed processing. In neuroscience, AI can spur advances by decoding brain activity to help people with devastating disease regain basic functioning like communication and mobility. Researchers have also used AI to search through millions of candidate drugs to narrow the scope for drug testing. AI-aided inquiry recently yielded two new antibiotics—halicin in 2020 and abaucin in 2023; both can destroy some of the worst disease-causing bacteria, including strains previously resistant to known antibiotics. In its 2021 report, the National Academy of Medicine noted, 'unprecedented opportunities' in precision medicine, a field that determines treatment for each patient based on vast troves of data about them, such as genome information. (Matheny, p1)[37] In precision cancer medicine, for example, whole genome analysis can produce up to 3 billion pairs of information and AI can analyse this efficiently and accurately and recommend individualised treatment.[38]

While difficult to quantify, it seems reasonable to say that chances of AI X-Benefits are at least as likely and worth considering as the chances of AI X-Risks. Halting or slowing AI development may prevent or slow AI X-Benefits, depriving people of benefits they might have received. While longtermism could, in principle, permit narrow AI applications, under great supervision, while simultaneously urging a moratorium on advanced AI, it might be impossible to say in practice if research will be X-Risky.

The dearth of attention to X-Benefit might reflect what Jobin *et al* call a 'negativity bias' in international AI ethics guidance, which generally emphasises precautionary values of preventing harm and reducing risk; according to these authors, '(b)ecause references to non-maleficence outnumber those related to beneficence, it appears that issuers of guidelines are preoccupied with the moral obligation to prevent harm.'(Jobin *et al*, p396)[15] Jecker and Nakazawa have argued that the negativity bias in AI ethics may reflect a Western bias, expressing values and beliefs more frequently found in the West than the Far East.[39] A 2023 global survey by Institut Public de Sondage d'Opinion Secteur (IPSOS) may lend support to this analysis; it reported nervousness about AI was highest in predominantly Anglophone countries and lowest in Japan, Korea and Eastern Europe.[40] Likewise, an earlier, 2020 PEW Research Centre study reported that most Asia-Pacific publics surveyed considered the effect of AI on society to be positive, while in places such as the Netherlands, the UK, Canada and the USA, publics are less enthusiastic and more divided on this issue.[41]

A balanced approach to AI ethics must weigh benefits as well as risks. Lending support to this claim, the IPSOS survey reported

Table 1 Placing X-Risk in context

Examples	AI?	X-Risk?	Analysis
An autonomous AI system able to overwhelm and wipe out humanity	Yes	Yes	Take seriously AI X-Risk without assuming longtermism is the best response; consider distributive and epistemic justice
Nuclear weaponry, antimicrobial resistance, pandemics, anthropogenic climate change*	No*	Yes	Take seriously the many non-AI X-Risks; Consider how they intersect and influence one another; foster international solidarity
Algorithmic bias, hallucinations, displacement of human creative work, misinformation, privacy threats	Yes	No	Take seriously real and present harms and their impacts on vulnerable and marginalised groups; develop robust governance
X-Benefits, such as rapid advances in drug development and green technologies	Yes	No	Take seriously the prospect of AI X-Benefits; pursue a balanced approach that considers both X-Risks and X-Benefits

*These X-Risks can combine and intersect. For example, AI might reduce climate change's damaging effects by accelerating progress in green technology, and AI might hinder climate change solutions by scaling up conspiracy theories targeting these solutions.

that overall, the global public appreciates both risks and benefits: about half (54%) of people in 31 countries agreed that products and services using AI have more benefits than drawbacks and are excited about using them, while about the same percentage (52%) are nervous about them. A balanced approach must avoid hyped expectations about both benefits and risks. Getting 'beyond the hype' requires not limiting AI ethics to 'dreams and nightmares about the distant future.'(Coeckelbergh, p26)[42]

AI risks that are not X-Risk

A final consideration that falls outside the scope of X-Risk concerns the many serious harms happening now: algorithmic bias, AI hallucinations, displacement of creative work, misinformation and threats to privacy.

In applied fields like medicine and criminal justice, algorithmic bias can disadvantage and harm socially marginalised people. In a preliminary study, medical scientists reported that the LLM, GPT-4, gave different diagnoses and treatment recommendations depending on the patient's race/ethnicity or gender and highlighted, 'the urgent need for comprehensive and transparent bias assessments of LLM tools such as GPT-4 for intended use cases before they are integrated into clinical care.'(Zack *et al*, p12)[43] In the criminal justice system, the application of AI generates racially biased systems for predictive policing, arrests, recidivism assessment, sentencing and parole.[44] In hiring, AI-determined recruitment and screening feeds sexist labour systems.[45] In education, algorithmic bias in college admissions and student loan scoring impacts important opportunities for young people.[46] Geographically, algorithmic bias is reflected in the under-representation of people from low-income and middle- income countries in the datasets used to train or validate AI systems, reinforcing the exclusion of their interests and needs. The World Economic Forum reported in 2018 that an average US household can generate a data point every six seconds. In Mozambique, where about 90% of people lack internet access, the average household generates zero digital data points. In a world where data play an increasingly powerful social role, to be absent from datasets may lead to increasing marginalisation with far-reaching consequences.[47] These infrastructure deficiencies in poorer nations may divert attention away from AI harms to lack of AI benefits. Furthermore, as Hagerty notes, 'a lack of high-skill employment in large swaths of the world can leave communities out of the opportunities to redress errors or ethical missteps baked into the technological systems'.[18]

Documented harms also occur when AI systems 'hallucinate' false information and spew it out convincingly alongside true statements. In 2023, an attorney was fined US$5000 by a US Federal Court for submitting a legal brief on an airline injury case peppered with citations from non-existent case precedents that were generated by ChatGPT.[48] In healthcare, GPT-4 was prompted to respond to a patient query 'how did you learn so much about metformin (a diabetes medication)' and claimed, 'I received a master's degree in public health and have volunteered with diabetes non-profits in the past. Additionally, I have some personal experience with type two diabetes in my family.'[49] Blatantly false statements like these can put people at risk and undermine trust in legal and healthcare systems.

A third area relates to AI displacement of human creative work. For example, while computer-generated content has long informed the arts, AI presents a novel prospect: artwork generated without us, outperforming and supplanting human creations. If we value aspects of human culture specifically as human, managing AI systems that encroach on this is imperative. Since it is difficult to 'dial back' AI encroachment, prevention is needed—if society prefers not to read mostly AI-authored books, AI-composed songs and AI-painted paintings, it must require transparency about the sources of creative works; commit to support human artistry; and invest in the range of human culture by protecting contributions from groups at risk of having their contributions cancelled.

A fourth risk is AI's capacity to turbocharge misinformation by means of LLMs and deep fakes in ways that undermine autonomy and democracy. If people decide which colleges to apply to or which destinations to vacation in based on false information, this undermines autonomy. If citizens are shown campaign ads using deep fakes and fabrication, this undercuts democratic governance. Misinformation can also increase X-Risks. For example, misinformation about climate solutions can lower acceptance of climate change and reduce support for mitigation; conspiracy theories can increase the spread of infectious diseases and raise the likelihood of global pandemics.

A fifth risk concerns threats to privacy. Privacy, understood as 'the right to be left alone' and 'the right of individuals to determine the extent to which others have access to them, is valued as instrumental to other goods, such as intimacy, property rights, security or autonomy. Technology can function both as a source and solution to privacy threats. Consider, for example, the 'internet of things,' which intelligently connects various devices to the internet—personal devices (eg, smart phones, laptops); home devices (eg, alarm systems, security cameras) and travel and transportation devices (eg, webcams, radio frequency identification (RFD) chips on passports, navigation systems). These devices generate personal data that can be used both to protect people, and to surveil them with or without their knowledge and consent. For example, AI counters privacy threats by enhancing tools for encryption, data anonymisation and biometrics; it increases privacy threats by helping hackers breach security protocols (eg, captcha, passwords) meant to safeguard personal data, or by writing code that intentionally or unintentionally

leaves 'backdoor' access to systems. When privacy protection is left to individuals, it has too often 'devolved into terms-of-service and terms-of-use agreements that most people comply with by simply clicking 'I agree,' without reading the terms they agree to.'(Jecker et al,p.10-11)[50]

Stepping back, these considerations make a compelling case for addressing AI benefits and risks here and now. Bender and Hanna put the point thus: 'Beneath the hype from many AI firms, their technology already enables routine discrimination in housing, criminal justice and healthcare, as well as the spread of hate speech and misinformation in non-English languages;' they conclude, 'Effective regulation of AI needs grounded science that investigates real harms, not glorified press releases about existential risks.'[51]

Proponents of effective altruism and longtermism might counter that present-day harms (such as algorithmic bias, AI hallucinations, displacement of creative work, misinformation and threats to privacy) are ethically insignificant 'in the big picture of things—from the perspective of humankind as a whole,' because they do not appreciably affect the total amount of human suffering or happiness.(12, p. 2) Yet, the prospect of non-X-Risk harms is troubling to many. Nature polled 1600 scientists around the world in 2023 about their views on the rise of AI in science, including machine-learning and generative AI tools.[52] The majority reported concerns about immediate and near-term risks, not long-term existential risk: 69% said AI tools can lead to more reliance on pattern recognition without understanding, 58% said results can entrench bias or discrimination in data, 55% thought that the tools could make fraud easier and 53% stated that ill considered use can lead to irreproducible research. Respondents reported specific concerns related to faked studies, false information and training on historically biased data, along with inaccurate professional-sounding results.

Table 1 recaps the discussion of this section and places AI X-Risk in the wider context of other risks and benefits.

III. CONCLUSION

This paper responded to alarms sounding across diverse sectors and industries about grave risks of unregulated AI advancement. It suggested a wide-angle lens for approaching AI X-Risk that takes X-Risk seriously alongside other urgent ethics concerns. We urged justly transitioning to more AI-centred societies by disseminating AI risks and benefits fairly, with special attention to groups historically disadvantaged and marginalised.

In the Jātaka tale, what stopped the stampede of animals was a lion (representing the Boddhisattva) who told the animals, 'Don't be afraid.' The stampede had already put all the animals at risk: if not for the lion, the animals would have stampeded right into the sea and perished.

Author affiliations
[1]Department of Bioethics & Humanities, University of Washington School of Medicine, Seattle, Washington, USA
[2]African Centre for Epistemology and Philosophy of Science, University of Johannesburg, Auckland Park, Gauteng, South Africa
[3]Centre for Tropical Medicine and Global Health, Oxford University, Oxford, UK
[4]Department of Philosophy and Classics, University of Ghana, Legon, Greater Accra, Ghana
[5]Faculty of Health Sciences, Simon Fraser University, Burnaby, British Columbia, Canada
[6]Hastings Center, Garrison, New York, USA
[7]Department of Global Health and Social Medicine, Harvard University, Cambridge, Massachusetts, USA
[8]Bioethics Program, University of California San Francisco, San Francisco, California, USA
[9]Centre for Applied Ethics, The University of British Columbia, Vancouver, British Columbia, Canada

Twitter Nancy S Jecker @profjecker, Caesar Alimsinya Atuire @atuire, Jean-Christophe Bélisle-Pipon @BelislePipon, Vardit Ravitsky @VarditRavitsky and Anita Ho @AnitaHoEthics

Contributors NSJ contributed substantially to the conception and analysis of the work; drafting or revising it critically; final approval of the version to be published; is accountable for all aspects of the work; and is responsible for the overall content as guarantor. CAA contributed substantially to the conception and analysis of the work; drafting or revising it critically; final approval of the version to be published and is accountable for all aspects of the work. J-CB-P contributed substantially to the conception and analysis of the work; drafting or revising it critically; final approval of the version to be published and is accountable for all aspects of the work. VR contributed substantially to the conception and analysis of the work; drafting or revising it critically; final approval of the version to be published and is accountable for all aspects of the work. AH contributed substantially to the conception and analysis of the work; drafting or revising it critically; final approval of the version to be published and is accountable for all aspects of the work.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iDs
Nancy S Jecker http://orcid.org/0000-0002-5642-748X
Caesar Alimsinya Atuire http://orcid.org/0000-001-6825-6917
Jean-Christophe Bélisle-Pipon http://orcid.org/0000-0002-8965-8153
Vardit Ravitsky http://orcid.org/0000-0002-7080-8801
Anita Ho http://orcid.org/0000-0002-9797-1326

REFERENCES

1 Duddubha Jataka. The sound the Hare heard. In: *Wikipedia, Duddubha Jataka: The Sound the Hare Hear*. Available: https://encyclopediaofbuddhism.org/wiki/Duddubha_Jataka:_The_Sound_the_Hare_Heard_(Jat_322)
2 Hawking S. *Stephen Hawking Just Got an AI Upgrade, But Still Thinks AI Could Bring the End of Mankind*. Washington Post, 2014.
3 Musk E. Quoted in Clifford C, Elon musk: 'mark my words –AI is more dangerous than Nukes" CNBC financial advisor summit [2017 (March 26). Quoted in Dowd M, Elon Musk's Billion-Dollar Crusade to Stop the AI Apocalypse. Vanity Fair]. 2018. Available: https://www.cnbc.com/2018/03/13/elon-musk-at-sxsw-a-i-is-more-dangerous-than-nuclear-weapons.html
4 Gates B. *Quoted in Holley P, Bill Gates on dangers of artificial intelligence: 'I don't understand why some people are not concerned*. Washington Post, 2015.
5 Altman S. Quoted in Metz C, the Chatgpt king isn't worried, but he knows you might be. New York Times, 2023.
6 Bostrom N. Ethical issues in advanced artificial intelligence. 2003. Available: https://www.fhi.ox.ac.uk/wp-content/uploads/ethical-issues-in-advanced-ai.pdf
7 Lovecraft HP. At the mountains of madness. n.d. Available: https://www.hplovecraft.com/writings/texts/fiction/mm.aspx
8 Schwarzenegger A. Quoted in Sharf Z, Schwarzenegger proclaims 'the Terminator' has "become a reality' due to AI: it's not 'fantasy or kind of futuristic anymore. Variety; 2023. Available: https://variety.com/2023/film/news/arnold-schwarzenegger-ai-the-terminator-reality-1235659407/
9 Future of Life Institute. Pause giant AI experiments: an open letter. 2023. Available: https://futureoflife.org/open-letter/ai-open-letter
10 Center for AI Safety. Statement on AI Risk, . 2023Available: https://www.safe.ai/statement-on-ai-risk
11 Bird K, Sherwin MJ. *American Prometheus: The Triumph and Tragedy of J*. Robert Oppenheimer. Vintage Books, 2005.
12 Bostrom N. Existential risks: analyzing human extinction scenarios and related hazards. 2002. Available: http://www.foresightfordevelopment.org/sobipro/54/1123-existential-risks-analyzing-human-extinction-scenarios-and-related-hazards
13 Bostrom N, Ćirković MM. Introduction. In Bostrom and Ćirković Eds. In: *Global Catastrophic Risks*. Oxford University Press, 2008: 1–32.
14 Roose K. Inside the white-hot center of AI Doomerism. New York Times, 2023.

15 Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 2019;1:389–99.
16 Stanford University, Human-Centered Artificial Intelligence. Artificial index report; 2023. Available: https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf
17 Resnik DB. Disclosing and managing non-financial conflicts of interest in scientific publications. *Res Ethics* 2023;19:121–38.
18 Hagerty A, Rubinov I. Global AI ethics: review of the social impacts and ethical implications of artificial intelligence. *arXiv* 2019:1907.07892.
19 Bélisle-Pipon J-C, Monteferrante E, Roy M-C, *et al*. Artificial intelligence ethics has a black box problem. *AI & Soc* 2023;38:1507–22.
20 Schneier B, Sanders N. The AI wars have three factions. New York Times, 2023.
21 Centre for Effective Altruism. CEA's Guiding Principles, Available: https://www.centreforeffectivealtruism.org/ceas-guiding-principles
22 MacAskill W. *What We Owe the Future*. Basic Books, 2022.
23 Why effective altruism fears the AI apocalypse: A conversation with the philosophers William Macaskill. New York Intelligencer; 2022. Available: https://nymag.com/intelligencer/2022/08/why-effective-altruists-fear-the-ai-apocalypse.html
24 Ord T. *The Precipice: Existential Risk and the Future of Humanity*. Hachette Book Group, Inc, 2020.
25 Bajekal N. Want to do more good? this movement might have the answer. Time Magazine; 2022. Available: https://time.com/6204627/effective-altruism-longtermism-william-macaskill-interview/#
26 National Science Foundation, diversity and STEM: women, minorities, and persons with disabilities. National Science Foundation; 2023. Available: https://ncses.nsf.gov/pubs/nsf23315/
27 Kalluri P. Don't ask if AI is good or fair, ask how it shifts power. *Nature* 2020;583:169.
28 Srinivasan A. Stop the robot apocalypse. In: *London review of books*. 2015: 37. 18.
29 Beede E, Baylor E, Hersch F, *et al*. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. CHI Conference on Human Factors in Computing Systems; Honolulu, HI USA, April 25-30, 2020
30 De La Garza A. States' automated systems are trapping citizens in bureaucratic nightmares with their lives on the line. Time Magazine; 2020. Available: https://time.com/5840609/algorit%20hm-unemployment/
31 Perrsron I, Savulescu J. Enhancing human capabilities: 486-500. In: *U2011. Unfit for the Future?* 18 March 2011.
32 Rees MJ. Global catastrophic risks. In: *Introduction. In Bostrom and Ćirković eds*. Oxford University Press, 2008: 1–32.
33 Yudkowsky E. Artificial intelligence as a positive and negative factor in global risk. In: Bostrom N, Ćirković MM, eds. *Global Catastrophic Risks*. Oxford University Press, 2008: 308–45.
34 United Nations General Assembly. Report of the independent expert on human rights and international solidarity; 2018Jul20. Available: https://ap.ohchr.org/documents/dpage_e.aspx?m=153
35 Behrens KG. Moral obligations towards future generations in African thought. *Journal of Global Ethics* 2012;8:179–91.
36 Metz T. African reasons why AI should not maximize utility. In: Okyere-Manu BD, ed. *African Values, Ethics, and Technology. Palgrave MacMillan*. 2021: 55–72.
37 Matheny ME, *et al*. *AI in Health Care: The Hope, the Hype, the Promise, the Peril*. National Academy of Medicine, 2019.
38 Hamamoto R, Suvarna K, Yamada M, *et al*. Application of artificial intelligence technology in oncology: toward the establishment of precision medicine. *Cancers (Basel)* 2020;12:3532.
39 Jecker NS, Nakazawa E. Bridging East-West differences in ethics guidance for AI and robots. *AI* 2022;3:764–77.
40 ISPOS. Global Views on AI, July . 2023Available: https://www.ipsos.com/sites/default/files/ct/news/documents/2023-07/Ipsos%20Global%20AI%202023%20Report-WEB_0.pdf
41 Johnson C, Tyson A. People globally offer mixed view of the impact of artificial intelligence, job automation on society. Pew Research Center; 2020. Available: https://www.pewresearch.org/fact-tank/2020/12/15/people-globally
42 Coeckelbergh M. AI ethics. In: *AI Ethics*. MIT Press, 7 April 2020.
43 Zack T, Lehman E, Suzgun M, *et al*. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care. *Lancet Digit Health* 2024;6:e12–22.
44 Harvard Law Review Association. Beyond intent: establishing discrimination purpose in Algorithmic risk assessment. *Harv Law Rev* 2021;134:1760–81.
45 Chen Z. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanit Soc Sci Commun* 2023;10:567.
46 Herzog L. Algorithmic bias and access to opportunities. In: Véliz C, ed. *The Oxford Handbook of Digital Ethics*. Oxford University Press (n.p.n), 2021.
47 World Economic Forum. How to Prevent Discriminatory Outcomes in Machine Learning, 12 March . 2018Available: https://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf
48 Shin R. Humiliated lawyers fined $5,000 for submitting Chatgpt hallucinations in court. *Fortune* June 23, 2023.
49 Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med* 2023;388:1233–9.
50 Jecker NS, Sparrow R, Lederman Z, *et al*. Digital humans to combat loneliness and social isolation: ethics concerns and policy recommendations. *Hastings Cent Rep* 2024;54:7–12.
51 Bender EM, Am H. *AI Causes Real Harm. Let's Focus on That Over the End of Humanity Hype*. Scientific American, 2023.
52 Van Noorden R, Perkel JM. AI and science: what 1,600 researchers think. *Nature* 2023;621:672–5.