Current and Near-Term AI as a Potential Existential Risk Factor

Benjamin S. Bucknall*
Department of Information Technology
Uppsala University, Sweden
ben.s.bucknall@gmail.com

Shiri Dori-Hacohen†
Reducing Information Ecosystem Threats (RIET) Lab
Computer Science & Engineering Department
University of Connecticut, USA
shiridh@uconn.edu

ABSTRACT

There is a substantial and ever-growing corpus of evidence and literature exploring the impacts of Artificial intelligence (AI) technologies on society, politics, and humanity as a whole. A separate, parallel body of work has explored existential risks to humanity, including but not limited to that stemming from unaligned Artificial General Intelligence (AGI). In this paper, we problematise the notion that current and near-term artificial intelligence technologies have the potential to contribute to existential risk by acting as intermediate risk factors, and that this potential is not limited to the unaligned AGI scenario. We propose the hypothesis that certain already-documented effects of AI can act as existential risk factors, magnifying the likelihood of previously identified sources of existential risk. Moreover, future developments in the coming decade hold the potential to significantly exacerbate these risk factors, even in the absence of artificial general intelligence. Our main contribution is a (non-exhaustive) exposition of potential AI risk factors and the causal relationships between them, focusing on how AI can affect power dynamics and information security. This exposition demonstrates that there exist causal pathways from AI systems to existential risks that do not presuppose hypothetical future AI capabilities.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → *Social networking sites*; *Social recommendation*; • **Social and professional topics** → *Surveillance*; *Governmental regulations*; • **Applied computing** → *Cyberwarfare*.

KEYWORDS

AI safety; Existential risk; Societal impacts of AI

ACM Reference Format:
Benjamin S. Bucknall and Shiri Dori-Hacohen. 2022. Current and Near-Term AI as a Potential Existential Risk Factor. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3514094.3534146

*Corresponding Author. This work was conducted while BSB was a research intern at the Existential Risk Observatory, Amsterdam.
†Senior Corresponding Author

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08.
https://doi.org/10.1145/3514094.3534146

1 INTRODUCTION

As early as the 1950's, leading academics of the time, including Albert Einstein and Bertrand Russell, were warning of risks of human extinction due to the use of nuclear weapons [9]. More recently, the study of risks that may threaten the very existence of our species has grown as an academic discipline following Nick Bostrom's introduction of the topic of existential risks [10]. Various processes have since been proposed through which our species could go extinct, along with potential approaches to reduce this risk. Crucially, one does not need to estimate any of the existential risk scenarios as highly probable in order to determine that working to prevent them is an extremely valuable prospect; an extremely large negative value with a very low probability, still leads to an incredibly high expected value for even the smallest degree of reduction in risk [10]. Within the existential risk research community, one of the most discussed risks is that of misaligned artificial intelligence (AI), of which many proposed scenarios rely on the assumption of at least 'human-level' artificial general intelligence (AGI), if not outright superintelligence. While we do not deny that some such risks are valid and deserve attention, we feel that less powerful AI systems, including those that are present at the time of writing, ought to also be included in the discussion of existential risks.

In parallel, much concern and attention is being paid to the shorter-term harms of contemporary AI systems, including in digital humanities [see, e.g., 22, 26] and in computer science by the Fairness, Accountability, Transparency and Ethics (FATE) community [see, e.g., 16]. We find these contributions to be long overdue and highly merited. At the same time, these conversations and studies rarely, if ever, discuss how contemporary AI systems could pose existential threats. This position paper aims to highlight a possible extension to current digital humanities research into the domain of existential risk research, and bridge the gap between the two fields by suggesting plausible pathways through which current and near-term AI systems could impact the existential risk stemming from previously identified sources.

It is important to note that, despite much of the following discussion primarily addressing how AI could increase existential risk, there is also potential for AI to act as a significant positive factor in how we are able to address existential risks. If utilised responsibly, AI is an incredibly powerful tool which would doubtless have many potential applications to devising, coordinating, and implementing responses to the dangers that we face. However, identifying both positive and negative impacts of AI with regards to existential risk, and assessing the net affect of these impacts are not the aims of the present paper, and so discussion of the benefits of AI will be minimal.

This paper will take the following structure. We will begin by briefly summarising relevant prior work on existential risk as well

as introducing a number of key pre-existing concepts in the AI and existential risk literatures to which we will refer throughout the paper. The following two sections will each discuss a broad trend of AI, those of AI's impacts on power dynamics and information security respectively, going into detail as to how these trends are currently, and could potentially manifest. We then discuss how each of the commonly discussed sources of existential risk is impacted by the trends discussed in the preceding sections. Following this, a graphical representation of the interacting AI impacts, sources of risks, and risk factors is given and discussed before the final section concludes.

2 PRIOR WORK ON EXISTENTIAL RISK

We begin by briefly reviewing the relevant background of existential risk research with a focus on existential risk from AI. As mentioned in the introduction, the study of existential risks to humanity as an academic discipline is said to have begun with Bostrom's influential article *'Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards'* in 2002 [10], though some discussion was taking place before this [9, 53]. In his article, Bostrom offers a definition of existential risk to be *'[o]ne where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential.'* He also uses a two-dimensional classification of risks based on both *who* is affected by a risk as well as *how severe* the impact is on each individual that is affected. In terms of these two dimensions, Bostrom takes existential risks to be those that are both *pan-generational*, in that they affect all or almost all living people as well as future generations, and *crushing*, which is the greatest severity that Bostrom considers. Following Bostrom, attention paid to existential risk increased noticeably with several books being published on the subject including Martin Rees' *'Our Final Century?'* [69], Richard Posner's *'Catastrophe'* [67], and Nick Bostrom and Milan Ćirković's *'Global Catastrophic Risks'* [12].

More recently, in his 2020 book *'The Precipice: Existential Risk and the Future of Humanity'*, Toby Ord summarises much of the work on existential risks carried out since Bostrom's seminal paper [64]. In it he specifies a classification of identified risk sources consisting of: natural risks, such as those from asteroids or supervolcanoes; risks from use of nuclear weapons; biological risks, including pandemics and biotechnology; environmental risks, including those from runaway climate change and biodiversity loss; risks from unaligned artificial intelligence; and unknown future risks. It is Ord's definition of an existential risk as *'a risk that threatens the destruction of humanity's longterm potential'* [64] that we will use in this paper. Note that this definition, in a similar fashion to Bostrom's, includes not only extinction of the human species, but also scenarios of technological or social stagnation, unrecoverable collapses of society or dystopias, all of which would result in the non-fulfilment of our species' potential [6]. Furthermore, we will make use of another definition that Ord highlights, that of an *existential risk factor* (or simply *risk factor*), being a situation, state of affairs, or event that, while may not directly constitute an existential risk, can increase the probability of an existential catastrophe occurring, or reduce our abilities to respond to one [64]. A commonly cited illustrative example is that of a great power war, that is, a war between technologically advanced and powerful nations. While the war itself might not be a source of existential risk, it is certainly plausible that it would contribute to overall risk during, and perhaps after, the war if, for example, it pushes involved nations towards use or development of nuclear, chemical, or biological weapons of mass destruction, or incentivises greater funding towards developing novel, more dangerous weapons technologies.

One of the sources of risk that is estimated to contribute the most to the total amount of risk currently faced by humanity is that of unaligned artificial intelligence, which Ord estimates to pose a one-in-ten chance of existential catastrophe in the coming century. Several books including Bostroms *'Superintelligence'* [11], Stuart Russell's *'Human Compatible'* [72], and Brian Chrisitian's *'The Alignment Problem'* [20], as well as numerous articles [see, e.g., 2, 32, 41] have addressed the *alignment problem* of how to ensure that the values of any advanced AI systems developed in the coming years, decades, or centuries are aligned with those of our species.

Much of this literature has started from the assumption of currently hypothetical levels of AI capability, introducing terms such as *superintelligence*, *artificial general intelligence*, or *human-level machine intelligence*, and proceeding to explore the inherent difficulties in controlling such systems and presenting potential strategies for doing so. In recent years however, more focus has been payed to the impacts of AI that are not assumed to have such strong capabilities, and the ways in which such systems could still pose an existential risk. For example, K. Eric Drexler [30] considers risks from pluralities of interacting AI services, shedding light on a new trajectory to existential catastrophe from AI. Furthermore, researchers such as Paul Christiano [21] and Joseph Carlsmith [18] are now considering risks from *'power-seeking AI'*, that is, systems that despite not being assumed to be at some specific level of generality or intelligence, nonetheless could seek to increase their level of influence for their own strategic advantage. Finally, scholars within the field of AI governance are considering ways in which even weaker conceptions of future AI can have dramatic impacts [see, e.g., 39, 88]. For example, Zwetsloot and Dafoe introduce the concept of *structural* risks, complementing the notions of *accident* and *misuse* risks from AI [28, 91]. Whereas misuse risks are those caused by the deliberate use of AI in a harmful manner, and accident risks occur due to some glitch, fault, or oversight that causes an AI system to exhibit unexpected harmful behaviour, structural risks of AI are those caused by how a system *'shapes the broader environment in ways that could be disruptive or harmful'*. Dafoe also defines a set of perspectives through which to view the challenges of governing AI. These are the *ecology* and *general purpose technology (GPT)* perspectives, which consider a potential global ecosystem of interacting AI systems, and view AI as a broad GPT comparable to the combustion engine, or electricity, respectively, as well as the more familiar *superintelligence* perspective, which focuses on the challenges of governing AI with cognitive abilities far greater than our own. Compared to the superintelligence perspective, the ecology and GPT perspectives lend themselves to a broader outlook of AI and thus lend themselves to studying potential structural risks from AI.

This shift towards less dramatic and speculative failure scenarios due to advanced AI is accompanied by a similarly growing body that moves away from characterising other existential risks solely as singular events and towards descriptions of risks as the result of

the complex interactions between multiple, more mundane vulnerabilities in our social and political systems [see, e.g., 3, 24, 55]. For example, in their in-depth exploration of hazards, vulnerabilities and exposures to existential risk, Liu et al. emphasise the need for expanding our understanding of the variety of paths that might lead to existential risk [55]. Among their arguments, they provide examples of how narrow AI could be sufficient to pose existential risk. This work aims to contribute further to this perspective by focusing on the potential indirect pathways from narrow AI to existential risk, without assuming AGI.

2.1 A note on terminology

Throughout this paper we make use of Baum's definitions of near-term, mid-term, and long-term AI [5]. As defined by Baum, near-term AI is *'AI that already exists or is actively under development with a clear path to being built and deployed'*, while long-term AI is *'AI that has at least human-level general intelligence.'* Mid-term is then loosely defined as AI that falls in between these two categories, with potential for overlap. While some researchers have criticised such distinctions [68], we believe that these terms can be of instrumental use if defined precisely. Finally, this paper will take a holistic view of AI which considers systems as being situated in, and inseparable from, their wider physical and digital environment. This may include (but is not limited to) political, corporate, and social structures, and the technological systems (software and hardware) through which users interact with AI.

3 AI CAN SHIFT OR STRENGTHEN EXISTING POWER DYNAMICS

In this section we consider the ways that AI, as a general purpose technology, can affect the power dynamics between different pairs of actors from nation states, multinational corporations, and the public. These actors were selected as particularly prominent in relation to AI based on current trends: large, multinational technology companies develop AI for use in their services to be used by the public; the public clearly has an important relationship with nation states; and states interact with multinational technology companies through, for example, regulation. States are also involved in AI research and development through their military and intelligence agencies, and governmental funding of research.

3.1 State-State Power Relationships

AI has the potential to disturb the relationship between a pair of nation states. Some examples of such disturbance already exist, with voices in the USA and the West more generally expressing increasing concern towards China and Russia's respective goals regarding AI. For example, the National Security Commission on Artificial Intelligence claimed earlier this year that *'America's technological predominance - the backbone of its economic and military power - is under threat'* and that *'AI is deepening the threat posed by cyber attacks and disinformation campaigns that Russia, China, and others are using to infiltrate our society, steal our data, and interfere in our democracy'* [76].

In particular, China has undergone staggering development and economic growth since 1979, with its economy doubling in size on every eight years [60]. Furthermore, the publication in 2017 of the *'New Generation Artificial Intelligence Development Plan'* (AIDP), China's central document outlining it's targets for future AI development, marks a considerable step towards its desired position as a global technology superpower [86]. This dramatic rise in technological capabilities of China and other Asian nations has been interpreted as a broader movement of 'Easternisation', that is, a shift of global power towards the East. For example, it has been claimed that *'[i]n the 19th century, the world was Europeanized. In the 20th century, it was Americanized. Now, in the 21st century, the world is being irreversibly Asianized'* [51].

If one takes the view that such a shift in global power is a primary cause for concern based on moral, political, or social reasoning, then it is clear that this could constitute a risk factor. For example, the AI Now Institute has observed that *'[t]he urgency of "beating" China [in terms of the development of AI] is commonly justified based on the nationalist assumption that the US would imbue its AI technologies... with better values than China would. China's authoritarian government is presumed to promote a more dystopian technological future than Western liberal democracies'* [27]. If one ascribes to this (admittedly Euro- and US-centric) view, it could be argued that there are possible realisations of such a 'dystopian technological future' that would constitute an existential catastrophe, regardless of how likely they are.

Nevertheless, it is possible for this global shift to constitute a risk factor even if 'Easternisation' is not viewed as a primary risk. The mere existence of such views could itself be a risk if it contributes to a growing sense of an AI arms race between technological superpowers. The emergence of such a competitive dynamic may go on to inhibit international coordination within and without the field of AI, incentivise against AI safety precautions, or apply pressure for investment in intentionally harmful AI technologies such as lethal autonomous weapons (LAWs); some of these trends are already emerging. It is conceivable that the coming decades might witness an 'AI cold war dynamic', increasing the chances of physical conflict between involved states, and diverting resources away from other pressing issues including existential risks.

3.2 State-Corporation Power Relationships

We now consider the evolving nature of relationships between nation states and multinational corporations with significant interests in AI technology.

The past two-decades have seen a monumental rise of private corporations in the technology sector, including Facebook/Meta, Google/Alphabet, Amazon, Microsoft, Netflix, and Twitter, to the extent that, as an example, in the fiscal year ending April 2021, Alphabet had a revenue of $183bn, which would make it the 54th wealthiest country in terms of GDP.[1] In addition to their wealth, these corporations interact with the general public to extents unprecedented for private bodies, with a large proportion of our social, professional, and commercial activities being facilitated by services provided by these companies. Technology giants are increasingly coming under scrutiny for the negative externalities they are imposing on their users, such as disinformation, extreme polarization and mental health risks, to name a few [see, e.g., 56]. However, the

[1] This is a conservative comparison. If market capitalisation is used rather than revenue, the argument is even stronger.

massive scale and disproportionate influence of these corporations can put nation-states in a difficult position when it comes to interacting with them. For example, a proposal made by the Australian government in February 2021 for a new law that would require platforms such as Facebook and Google to pay for the media content that they distribute led to a high-profile dispute. Eventually an agreement was reached, but not before Facebook restricted the actions of Australian news outlets on the site and Google threatened to withdraw its web search service from the country [7].

Another difficulty faced by states results from the global reach of these corporations. This is exemplified in the practice of legal tax-avoidance amongst such multinationals, and the resulting attempts of the OECD to adjust international corporation tax prices to better distribute the taxes of multinationals amongst the nations in which they operate. This will be achieved through a plan that will *'ensure a fairer distribution of profits and taxing rights among countries with respect to the largest [multinational enterprises], including digital companies'* as well as *'put a floor on competition over corporate income tax, through the introduction of a global minimum corporate tax rate'* [62].

Inevitably, many of the goals and decisions of such corporations are profit-driven, and as such can often be misaligned with those of wider society. It is thus important for states to devise and enact sufficient regulatory processes in order to ensure that the well-being of the public is prioritised, the functionality of political processes is maintained, and the power of wealthy multinational corporations is used for the benefit of society as a whole; however, this is easier said than done in a rapidly changing technological environment.

Finally, it is worth noting that, much like the international 'AI arms race' discussed above, it is conceivable that a similar situation could occur between a state and a multinational with significant spending in AI research.[2] As in the case of state-state arms races, this could lead to corner-cutting with regards to AI safety, though in contrast to international conflict, at this date it seems unlikely that an arms race involving a private corporation could directly lead to armed conflict.

3.3 State-Citizen Relationships

Next, we consider the changing dynamic between states and their citizens as a result of adopting AI surveillance systems.[3] This can take the form of a number of technologies and purposes, including both automatic facial and voice recognition, smart/predictive policing, or the nascent practice of affect recognition, which aims to automatically 'read' an individuals emotions from facial micro-expressions. The rapid increase in the ubiquity of such AI systems has been well documented and represents a major increase in the ability that nation-states have to gain knowledge of, and power over, individuals within their borders. The 2018 annual report by the AI Now Institute at New York University claimed that *'[t]he role of AI in widespread surveillance has expanded immensely in the U.S., China, and many other countries worldwide'* [87]. They further backed this up a year later in their 2019 report, claiming that *'... despite growing public concern and regulatory action, the rollout of facial recognition and other risky AI technologies has barely slowed down'* [27]. These findings are also supported by Steven Feldstein at the Carnegie Endowment for International Peace who found that *'AI surveillance technology is spreading at a faster rate to a wider range of countries than experts have commonly understood'* [34]. Feldstein further emphasises that uptake of such systems is not limited to a particular class of state having found example countries in every major world region, and with *'political systems [that] range from closed autocracies to advanced democracies.'* Many of these systems are already highly criticised in the academic community [see, e.g., 22, 26] and are in extensive use also by corporations, who have in turn served to normalise this formerly unpalatable[4] level of surveillance [26, 90].

Worryingly, the rise in uptake of AI surveillance technologies is not accompanied by a similarly-paced development in the ethical and legal frameworks that such technologies are embedded in. The AI Now Institute argue there is a growing divide between the theory and practice of ethics in this area [27]. While there has recently been an increase in the awareness of these issues among governments, corporations, and many societal groups, little concrete progress is being made in addressing them. Furthermore, while Feldstein reiterates that some applications of AI surveillance technology are legal, he also states: *'Even democracies with strong rule of law traditions and robust oversight institutions frequently fail to protect individual rights in their surveillance programs'* [34]. This sets a troubling precedent. If even the most stable and traditionally well-functioning democracies are failing to meet their own regulatory standards for current systems, how will they deal with the more-powerful technologies on the horizon - and furthermore, what hope do we have that less democratic countries will abide by international treaties aiming to enforce ethical use of AI in surveillance?

This rapid development of AI surveillance technology, and the lagging ethical and legal frameworks, almost certainly raises the possibility of a 1984-style, Orwellian repressive autocracy (regardless of the actual probability of such a scenario playing out in practice). Under Ord's definition of existential risk the emergence and stabilisation of such a regime would constitute an existential catastrophe.

Finally, it is worth mentioning an interesting connection between the above discussion and the previous subsection addressing state-corporation relations. Many of the hardware and software components of AI surveillance systems are developed by private companies that sell the technology to governments and police departments; privatization also legitimises surveillance that would be considered extra-legal if run by the state [26]. The AI Now Institute draws a connection between smart city projects and the consolidation of further power in the hands of corporations [27]. A concerning example is the ongoing partnerships between Amazon's *'Ring'* doorbell and more than 2,000 U.S. police departments [26, 57]. Emails released by VICE News show *Ring* employees encouraging police departments to share social media posts advertising *Ring* and its partner app *Neighbor*, as well as advising them on how to best persuade hesitant residents to share footage from their *Ring* doorbells [40]. The implications to, and extent of, citizen privacy

[2] A comparable arms race could also occur between two or more multinational corporations, and may arguably be occurring already.

[3] Other AI systems may change the state-citizen dynamic in other ways; we focus here on surveillance which is especially salient for many at the time of writing.

[4] At least in the West, if not globally.

infringements due to public-private surveillance partnerships are only beginning to be understood.

4 AI AFFECTS INFORMATION TRANSFER AND ACCESS

In this section we look at the ways that current AI systems, as situated in their wider environments, can affect individuals' and states' access to information and the effect that this could have on social and political systems.

4.1 Information Ecosystem Threats

The advertising-based revenue model that underlies most of the internet today is driven and propelled by incredibly advanced AI systems, leading to well-documented and troubling phenomena such as disinformation,[5] extreme polarization, hate speech, and self-radicalization [42, 90]. Much has been written about how web search and social media, two of the predominant modes of interaction with online information, both rely on AI for their technological and financial success. Underlying the explosive growth of search engine and social media are state-of-the-art AI approaches, including chiefly search and recommender systems, trained on unprecedented amounts of user data and optimised to maximise revenue.

The connections between the ad-tech, AI-powered giants such as Alphabet (Google) and Meta (Facebook) and the meteoric rise of information campaigns, bots, and weaponised controversy have been well documented [e.g., 42, 61]. As the first entry in their *'Ledger of Harms'*, the Center for Humane Technology has listed *'Making Sense of the World: Misinformation, conspiracy theories, and fake news'* [35]. 'Filter bubbles' refer to the feedback loops formed whereby a user is shown search or social media results that align with their existing preferences, leading to ever-increasing levels of confirmation bias [66, 81], which significantly exacerbate the pre-internet phenomenon of echo chambers [23]. During their 2016 information campaigns, for example, the Internet Research Agency (IRA) utilised certain features of the Facebook feed algorithm in order to create large-scale groups of like-minded users which were then gradually primed to be more vulnerable to disinformation [61]. Optimization of metrics such as 'time spent' and 'engagement' leads to the prioritization and artificial amplification of emotionally-charged, evocative content, often focused on negative emotions such as anger [29] and habits such as 'doomscrolling' [79], serving as a form of 'reward hacking' for the human mind, whereby the users are dehumanised and converted into mere inputs to the AI's profit machine [26]. Such optimizations may additionally lead the AI to shift the preferences of its users [49, 52], whether to make them more easily predictable [73] or simply to better serve the company's profit motive [29]. The extreme polarization that results contributes significantly to the spread of misinformation [85], and is actively exploited by states and other players sowing disinformation [61], leading to a vicious cycle [4, 29]. This process has led to a breakdown of intersubjectivity in the US and other countries, arguably driven by social media and its underlying AI systems. The plot thickens when these challenges coincide and intersect with other, related issues of hate speech, self-radicalization online, incentives for brevity [38, 44], and trolling behaviors [45], making nuanced conversations all but impossible [29]. Indeed, a recent and prominent paper published in the Proceedings of the National Academy of Sciences has argued that *'seemingly minor algorithmic decisions'* may be reshaping our long-evolved information-foraging and decision-making processes in as of yet undetermined but potentially harmful ways [4].

On top of all these, AI-powered synthetic media – including, but not limited to, deep fakes – pose an entirely new level of challenge that society has yet to fully reckon with [19], simultaneously allowing the creation of incredibly convincing false narratives, while providing cover frequently referred to as the 'liar's dividend' - whereby true and damaging evidence can be waved off as deep fakes [75]. Given the extraordinary success of contemporary disinformation campaigns relying chiefly on so-called "cheap fakes," an overactive imagination is not required in order to see the feasibility of wide-scale AI-powered manipulation in the near future.

Taken together, these are significant and growing threats to the information ecosystem, and by extension, to the collective decision making capacity of humanity. These threats have the potential to act as a mediating risk factor, particularly regarding those sources of existential risk for which collective action of the public can directly affect the severity of outcomes, such as pandemics, nuclear war and climate change, as we will discuss in further detail below.

Finally, some scholars have argued that, over and above whatever multiplier effect they may have to other existential risks, information ecosystem threats constitute an existential risk of their own right, painting a dire image of a dystopian, 1984-esque world where truth ceases to carry meaning [54].

4.2 Cybersecurity and International Cyber Warfare

One of the more natural applications of current and near-term AI is in cybersecurity, as reflected by it's dramatic current and projected growth as an industry worth $1 billion in 2016 to $34.8 billion in 2025 [59]. However, there are many open questions regarding the large-scale effects of this trend, not least regarding how the current balance between offence and defence in cybersecurity will be affected by the increased use of AI. Though this is still a disputed topic, though there are persuasive reasons to believe that cybersecurity (including its AI developments) leans towards offense [15, 36, 46, 77].

There are many ways in which the use of AI in cybersecurity may make it easier to successfully carry out attacks. Matteo *et al.* note that the use of machine learning methods in the creation of defense systems creates a further vulnerability in the system itself, if attackers are able to influence the training of the system in use [82]. They point to studies showing that carefully constructed changes to training data, imperceptible to human overseers, can result in unexpected behaviour of the trained system[6]; a recent survey details many techniques to train machine learning to be robust to such challenges [31]. Furthermore, Johnson points out

[5] Disinformation campaigns refer to coordinated attempts to spread false information; misinformation is the unintentional sharing of false information [see, e.g., 29, 63].

[6] Matteo *et al.* give the example of how adding 8% of faulty data during training of an AI system intended to recommend drug dosages can result in over a 75% change in the doses for 50% of patients.

that AI can be used by adversaries to design and implement complex and customised cyber-attacks with unprecedented accuracy and efficiency [46].

On the other hand, AI also provides advanced methods for detecting and responding to cyber-attacks. Wirkuttis and Klein observe that the abilities of AI systems to handle large data sets make them prime candidates for the automation of cybersecurity related tasks, network monitoring, and malicious intrusion identification [89].

Cybersecurity between nation-states is now being referred to as 'cyberwarfare' [see, e.g., 31] and actively discussed as 'the new cold war' [1, 70]. To summarise, AI is now a major part of cybersecurity, and an arms race has the potential to lead to potential catastrophes and heightened tensions between nations.

5 HOW THESE TRENDS CONSTITUTE RISK FACTORS

Having discussed some of the potential impacts that current and near-term AI can have, and is already having, we now move on to see how these impacts can act as risk factors. We will begin by discussing general factors that can affect total existential risk from any source, before briefly considering each of the commonly studied sources of risk separately.[7]

5.1 General Risk Factors

Many of the AI trends and impacts discussed above have direct implications on the functioning of social and political structures. For example, the use of recommender systems and incentives for brevity on social media may lead to an increase in political polarisation and a less informed public respectively. In democratic societies this may have a knock-on effect in terms of the political process. An increase in political polarisation in the general public, for example, may be reflected within political institutions, whereby elected representatives are unwilling to compromise due to a rise in political partisanship. This would lead to an increase in difficulty of passing key legislature aimed at addressing pressing problems, including existential risks, thereby acting as a risk factor. Furthermore, a less informed public may be more susceptible to coordinated mis- and dis-information campaigns, such as the Russian state interference with the 2016 U.S. presidential election [61]. Such effects act as risk factors by eroding trust in traditional democratic political structures and processes, thus inhibiting their ability to respond to existential risks. Political processes may also be hindered by changes to state-corporation relations that result in corporations having greater political lobbying power, or political division over government regulation of multinational corporations.

Secondly, increased state-state competition, in the form of an arms race, AI or otherwise, can also seriously impede responses to existential and catastrophic risks by diverting attention and resources towards the addressing the competition dynamic. While not relating to state-state competition, we can see an example in how more immediately pressing issues can divert attention away from longer-term concerns in countries' reactions to the COVID-19 pandemic. Some writers have observed and pointed out the dangers of tunnel vision towards such scenarios at the expense of other, less immediate but no less serious issues such as climate change and nuclear proliferation, among others [37]. It is reasonable to expect that the emergence of an AI arms race, or other cold war scenario, could have similar impacts on our responses to other dangers. Finally, as mentioned above, the development and deployment of AI-powered surveillance systems increases the probability of Orwellian dystopias of global, sustainable, repressive autocracy.

[7]These sources are: nuclear weapons, biotechnology and pandemics, climate change, natural risks, and unaligned artificial general intelligence.

5.2 Nuclear Weaponry

We now consider near-term AI's impacts on specific sources of risk, starting with that from nuclear weaponry. This is most likely the source of risk that would be most affected by heightened state-state tensions. If competitive dynamics resulting from an AI arms race between two nuclear states grow uncontrolled and become military in nature, this would increase the probability of a first-strike nuclear attack. The evolving international cybersecurity landscape could also play a role in making such an event more likely, if a state is better able to protect its intelligence, or gain greater access to its adversary's. Use of nuclear weapons has been widely discussed as a possible source of existential risk due to the possibility of 'nuclear winter' [25], which has the potential to lead to widespread famine, and potential human extinction [64]. Further in-depth discussion of AI's impact on nuclear security can be found in the Stockholm International Peace Research Institute's three-volume report *'The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk'* [13, 74, 83], as well as in [58] and [33].

5.3 Pandemics and Biotechnology

The emergence of a deadly infectious virus, be it naturally occurring or engineered, has been predicted to be a major sources of existential risk [43, 64]. There are a number of ways in which near-term AI trends can act as a risk factor with regards to pandemics. First, as we have seen from the ongoing COVID-19 pandemic, an adequate and coordinated response to such a scenario requires not only political bodies capable of dealing with the threat, but also a public that will play their part to protect themselves and other individuals from the disease. In this regard, it is likely that our ability to collectively respond to a global pandemic depends on public trust in the relevant political systems, in order to maximise the effectiveness of measures such as quarantining, mask-wearing, and mass vaccination. We have seen waves of COVID misinformation and anti-vaccination conspiracy theories throughout the course of the pandemic, with much of the discourse taking place on social media. If we are to have any hope of responding well to another, potentially more dangerous pandemic in the future, it is crucial that mis- and disinformation and conspiracy theories do not disseminate successfully at anywhere near the scale they have for COVID-19.

Additionally, and specifically to engineered pandemics, the availability of powerful AI systems may make it easier for malevolent actors to get hold of the technologies and techniques required to design and produce dangerous pathogens to cause a pandemic. While this is admittedly speculative, recent developments demonstrate AI's powerful role in solving problems in molecular biology, namely protein folding [17]. Similarly powerful neural architectures, in malicious hands, may cause untold damage [84].

Finally, heightened international political tensions and state-state rivalry could spur some states to develop and potentially deploy advanced bioweaponry. Indeed, despite the foreseeable widespread condemnation that such an act would provoke, frameworks for regulating such practices is severely under-resourced and underfunded with Ord giving the example that the UN's Biological Weapons Convention operates on an annual budget of only $1.4 million – less than the average McDonald's restaurant [64].

5.4 Climate Change

Similarly to pandemics, the cumulative impact of individuals' actions are able to impact existential climate change risk either positively [71] or negatively. Also like pandemics, climate change has a history of being clouded in mis- and disinformation, often perpetuated by those considered to be doing the most damage [50, 65]. If we are to be able to address the ongoing climate crisis, the public and political spheres must be able to agree on the dangers that we are facing and the strategies that we can employ to alleviate them. In a society where much of the public's information gathering takes place on social media and where AI algorithms actively promote misleading and hazardous information, AI can increase the likelihood of climate catastrophe.

Additionally, the practice of developing and training AI systems has a drastic first-hand climate impact, due to the large amounts of energy needed to run the hardware system on which the AI is running [80], as well as possible secondary effects through, for example, applications to the exploration and extraction of oil and natural gas [48]. It has been estimated that the technology industry as a whole contributed between 3% and 3.6% of global greenhouse gas emissions in 2020 [8]. Furthermore, it has been reported that training a single natural language processing (NLP) AI model produces 300,000 kilograms of carbon dioxide emissions [80]. According to the AI Now Institute's 2019 report, this amounts to 125 New York to Beijing round-trip flights [27]. This is an often-overlooked aspect of the growth of AI that must also be addressed if we are to sustain technological development whilst avoiding a climate catastrophe [47].

5.5 Natural Risks

Natural extinction risks are those which humans are not responsible for causing or exacerbating, such as asteroid impacts or supervolcano eruptions. Thus we do not foresee AI to have any significant magnifying impact on such risks aside from the general impacts discussed above.[8]

5.6 Unaligned AGI

Finally, we note that developments in near-term AI will have impacts on how the field will proceed far into the future. Thus, both successes and failures of current AI systems may affect the architecture of a potential artificial general intelligence, the society in which it is developed, or our ability to align it with humanity. For example, if an AI arms race dynamic emerges between states or technology companies, there may be incentives for corner-cutting when it comes to implementing safety practices in order to maximise the capabilities of the system under construction. If such practices continue until the stage at which an AGI is being developed, this could have catastrophic outcomes at the time of deployment of the AGI system.

[8]However, as noted in the introduction, AI may be useful in mitigating such natural, as well as man-made, risks.

6 DISCUSSION

Figure 1 shows a diagrammatic representation of the effects of near-term AI, established sources of existential risk, and the potential risk factors identified in this paper that constitute the causal relations between them.

Table 1: Key for edge colours used in Figure 1

Edge Colour	Description	References
Cyan	AI arms-race scenario	Section 3.1
Orange	Great power war	Sections 5.2 & 5.3
Red	Deliberate malicious use of current AI systems	Section 5.3 [14]
Green	Carbon emissions of training large ML models	Section 5.4 [8, 27, 47, 48, 80]
Yellow	AI-enabled surveillance	Section 3.3 [22, 26, 27, 34, 87, 90]
Magenta	Corporate lobbying and government regulation	Section 5.1
Blue	Modified collective behaviour due to effects on the information ecosystem	Section 4.1 [4, 29, 78]

In this graph we can see explicit causal pathways from current and near-term AI on the left; through its effects on various power relationships, cybersecurity, and the information ecosystem – represented as blue nodes; to identified sources of existential risk – given as red nodes. The central yellow box represents general risk factors as given in Section 5.1. Edge colours correspond to specific causal relations between nodes, as detailed in Table 1.

We summarise each of these causal relations in turn, starting with cyan edges, representing an AI arms-race scenario (Section 3.1). Such a scenario could result from heightened state-state tensions and could increase the risk posed by unaligned AGI if the resulting development lacks adequate safety considerations. Furthermore, an arms race could act as a general risk factor through the diversion of resources away from existential risk as states increasingly focus on AI research and development. Accordingly, cyan edges connect state-state relations to the risk from unaligned AGI as well as the general risk factor of diverted resources.

Secondly, orange edges represent a great power war scenario. Such a scenario could also result from heightened geopolitical tensions, and could provide incentive for the development or use of nuclear or biological weapons. Hence, orange edges connect state-state relations to both nuclear and biological risk.

Red edges show the potential malicious use of AI. As noted in Section 5.3, AI systems could be applied to the engineering of dangerous novel pathogens, and thus a red edge directly connects current and near-term AI to biological risks. Furthermore, AI could be maliciously applied to aid in targeted disinformation campaigns

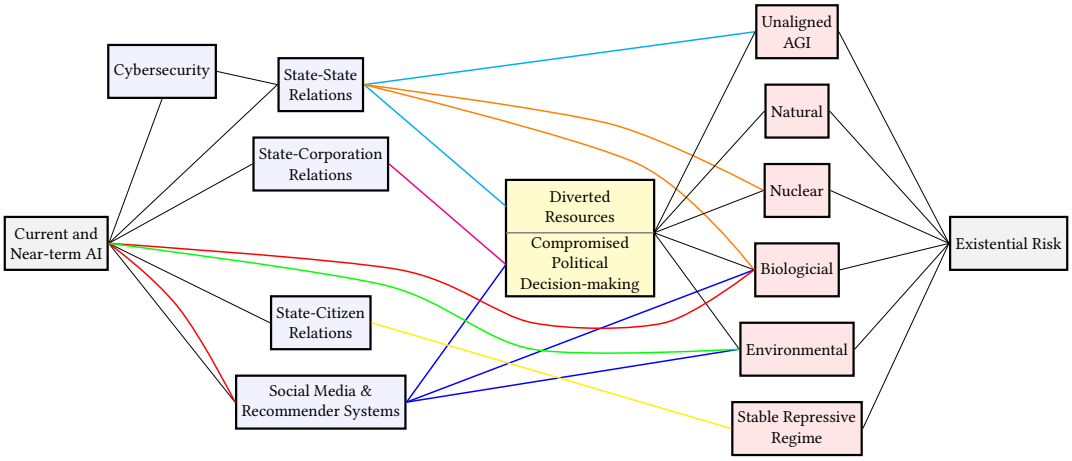

Figure 1: A graphical representation of the causal pathways from current and near-term AI to existential risk identified in this paper. Blue nodes represent effects of current and near-term AI, whereas red nodes represent identified existential risks [64]. The yellow box represents the general risk factors discussed in Section 5.1. Coloured edges represent causal connections as given in Table 1.

on social media, with the aim of sowing division or confusion in public discourse. Thus, a red edge also connects AI directly to social media. This is alongside a black edge representing effects of AI within social media that are not the result of malicious intent, such as the occurrence of filter bubbles or the spread of misinformation.

There are three causal pathways that only appear as a single edge in Figure 1. Firstly, the green edge represents the direct carbon emissions that result from the training of large ML systems, as discussed in Section 5.4. This edge thus connects current and near-term AI directly to environmental risks. Secondly, the yellow edge denotes the use of AI in surveillance technologies, which, as discussed in Section 3.3, raises the possibility of stable repressive regimes. Thus it connects the node representing state-citizen relations to the node representing such regimes. Lastly, the magenta edge shows the practices of corporate lobbying and government regulation of corporations, as mentioned in Section 5.1. This edge connects state-corporation relations to the general risk factor of compromised political decision-making.

Finally, the blue edges represent effects on collective behaviour as a result of AI's impact on the information ecosystem. We have discussed how these changes, largely occurring through AI applications in social media, could act as a general risk factor through compromising political decision-making, and can also have effects on the spread of pandemics and the severity of the ongoing climate crisis. Accordingly, blue edges connect social media to general risk factors, as well as biological and environmental risks.

With the complex structure of the interactions between AI effects, risk sources, and risk factors somewhat elucidated by the figure, a few features worthy of discussion are brought to the fore. Firstly, the position of cybersecurity considerations is fairly unique when compared to the other AI impacts addressed, in that we do not deem it to be a primary risk factor for any of the risks considered. Instead, we consider it to act solely through its impacts on state-state relations through actions noted above such as shifting the offense-defence balance. In this respect, AI's impact on cybersecurity could be considered purely as a constituent component of its impact on state-state relations. While effects on cybersecurity may also impact other parts of the diagram, such as state-citizen relations,[9] we do not deem these to be of the scale of constituting an existential risk factor and so are not explicitly included.

Furthermore, Figure 1 allows us to crudely compare the relative potential magnitudes that each AI impact contributes to total existential risk. If an AI impact acts as multiple risk factors (corresponding to multiple colours of edges), or acts upon multiple risk sources, this could indicate that this impact affects the magnitude of total existential risk that we face more than those impacts only acting as a single factor, or that only affect a single source. This clearly depends on the magnitudes of each risk factor, as well as the estimated contribution of each risk source to total existential risk, and so any such deductions from this graph are speculative at best. Future work aiming to make this diagram more quantitative by including numerical estimates of risk could be of value, though making such an extension would clearly be both challenging and speculative. That being said, preliminary judgement would deem the effects of AI on state-state relationships to be particularly important for future research as a result of state-state tensions acting on five of the six identified sources of risk through the potential scenarios of an AI arms-race and use of either nuclear or biological weaponry. On the other hand, AI's impact on state-citizen relations seems to be of less concern due to its affecting only the risk of a stable repressive regime through the use of AI in surveillance technologies. We reiterate that such conclusions are highly uncertain and could benefit from quantitative extensions to this framework.

7 CONCLUSIONS AND FUTURE WORK

In this paper we have argued that societal and political issues surrounding contemporary AI systems can have far-reaching impacts

[9] We thank an anonymous review for this observation.

on humanity through their potential acting as existential risk factors, rather than solely through the development of unaligned AGI. We argued that short-term harms from extant AI systems may magnify, complicate, or exacerbate other existential risks, over and above the harms they are inflicting on present society. In this manner, we have offered a bridge connecting two seemingly distinct areas of study: AI's present harms to society and AI-driven existential risk. By proposing concrete mechanisms for current and near-term AI to act as an intermediate factor to existential risk, we have taken the first step in demonstrating the connection between its primary impacts and existential risk. We believe that more research with the purpose of shedding light on these connections would be incredibly valuable. In particular, we see opportunities to extend the framework presented in Figure 1 either through quantitative estimates of relative likelihoods, or qualitative extensions to other AI impacts or sources of existential risk. The current position paper serves only as a first step in identifying and addressing risks of this nature.

ACKNOWLEDGMENTS

The authors would like to thank José Hernández-Orallo and Gabriel Pedroza for valuable comments and discussion, and Otto Barten for his continued support via the Existential Risk Observatory. We also thank our anonymous reviewers at SafeAI workshop and AIES for detailed feedback, comments, and further reading suggestions. Finally, we thank AI Safety Support for facilitating our initial connection.

REFERENCES

[1] James M Acton. 2020. Cyber Warfare & Inadvertent Escalation. *Dædalus* 149, 2 (2020), 133–149.
[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. arXiv:1606.06565 [cs.AI]
[3] Shahar Avin, Bonnie C. Wintle, Julius Weitzdörfer, Seán S. Ó hÉigeartaigh, William J. Sutherland, and Martin J. Rees. 2018. Classifying global catastrophic risks. *Futures* 102 (2018), 20–26. https://doi.org/10.1016/j.futures.2018.02.001
[4] Joseph B. Bak-Coleman, Mark Alfano, Wolfram Barfuss, Carl T. Bergstrom, Miguel A. Centeno, Iain D. Couzin, Jonathan F. Donges, Mirta Galesic, Andrew S. Gersick, Jennifer Jacquet, Albert B. Kao, Rachel E. Moran, Pawel Romanczuk, Daniel I. Rubenstein, Kaia J. Tombak, Jay J. Van Bavel, and Elke U. Weber. 2021. Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences* 118, 27 (2021). https://doi.org/10.1073/pnas.2025764118
[5] Seth D. Baum. 2020. Medium-Term Artificial Intelligence and Society. *Information* 11, 6 (2020). https://doi.org/10.3390/info11060290
[6] Seth D. Baum, Stuart Armstrong, Timoteus Ekenstedt, Olle Häggström, Robin Hanson, Karin Kuhlemann, Matthijs M. Maas, James D. Miller, Markus Salmela, Anders Sandberg, Kaj Sotala, Phil Torres, Alexey Turchin, and Roman V. Yampolskiy. 2019. Long-term trajectories of human civilization. *Foresight* 21, 1 (2019), 53–83. https://doi.org/10.1108/FS-04-2018-0037
[7] BBC News. 2021. Australia News Code: What's this row with Facebook and Google all about? https://www.bbc.com/news/world-australia-56107028. Accessed: 2021-08-26.
[8] Lotfi Belkhir and Ahmed Elmeligi. 2018. Assessing ICT global emissions footprint: Trends to 2040 & recommendations. *Journal of Cleaner Production* 177 (2018), 448–463. https://doi.org/10.1016/j.jclepro.2017.12.239
[9] Max Born, Percy W. Bridgman, Albert Einstein, Leopold Infeld, Frederic Joliot-Curie, Herman J. Muller, Linus Pauling, Cecil F. Powell, Joseph Rotblat, Bertrand Russell, and Hideki Yukawa. 1955. Russell-Einstein Manifesto. Available at https://www.atomicheritage.org/key-documents/russell-einstein-manifesto. Accessed: 2022-03-01.
[10] Nick Bostrom. 2002. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology* 9 (2002). https://www.nickbostrom.com/existential/risks.pdf
[11] Nick Bostrom. 2014. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.
[12] Nick Bostrom and Milan M. Ćirković. 2008. *Global Catastrophic Risks.* Oxford University Press.
[13] Vincent Boulanin, Shahar Avin, Frank Sauer, John Borrie, Dimitri Scheftelowitsch, Justin Bronk, Page O. Stoutland, Martin Hagström, Petr Topychkanov, Michael C. Horowitz, Anja Kaspersen, Chris King, S.M. Amadae, and Jean-Marc Rickli. 2019. *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk, Volume I, Euro-Atlantic Perspectives.* Technical Report. SIPRI.
[14] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán ÓhÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. 2018. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.* Technical Report. https://maliciousaireport.com/
[15] Ben Buchanan and Andrew Imbrie. 2022. *The New Fire: War, Peace, and Democracy in the Age of AI.* MIT Press.
[16] Toon Calders, Eirini Ntoutsi, Mykola Pechenizkiy, Bodo Rosenhahn, and Salvatore Ruggieri. 2021. Introduction to The Special Section on Bias and Fairness in AI. *ACM SIGKDD Explorations Newsletter* 23, 1 (2021), 1–3.
[17] Ewen Callaway. 2021. DeepMind's AI predicts structures for a vast trove of proteins. *Nature* 595 (Jul 2021), 635. https://doi.org/10.1038/d41586-021-02025-4
[18] Joseph Carlsmith. 2021. *Is power-seeking AI an existential risk?* Technical Report. Open Philanthropy. Draft report available at https://www.lesswrong.com/posts/HduCjmXTBD4xYTegv/draft-report-on-existential-risk-from-power-seeking-ai.
[19] Bobby Chesney and Danielle Citron. 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.* 107 (2019), 1753.
[20] Brian Christian. 2021. *The Alignment Problem: How Can Machines Learn Human Values?* Atlantic Books Ltd.
[21] Paul Christiano. 2019. What Failure Looks Like. https://www.lesswrong.com/posts/HBxe6wdjxK239zajf/what-failure-looks-like. Accessed: 2022-03-02.
[22] Wendy Hui Kyong Chun. 2021. *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition.* MIT Press.
[23] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021). https://doi.org/10.1073/pnas.2023301118
[24] Owen Cotton-Barratt, Max Daniel, and Anders Sandberg. 2020. Defence in Depth Against Human Extinction: Prevention, Response, Resilience, and Why They All Matter. *Global Policy* 11, 3 (2020), 271–282. https://doi.org/10.1111/1758-5899.12786
[25] Joshua Coupe, Charles G. Bardeen, Alan Robock, and Owen B. Toon. 2019. Nuclear Winter Responses to Nuclear War Between the United States and Russia in the Whole Atmosphere Community Climate Model Version 4 and the Goddard Institute for Space Studies ModelE. *Journal of Geophysical Research: Atmospheres* 124, 15 (2019), 8522–8543. https://doi.org/10.1029/2019JD030509
[26] Kate Crawford. 2021. *The Atlas of AI.* Yale University Press.
[27] Kate Crawford, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker. 2019. *AI Now 2019 Report.* Technical Report. New York: AI Now Institute. https://ainowinstitute.org/AI_Now_2019_Report.html
[28] Allan Dafoe. 2020. AI Governance: Oppotunity and Theory of Impact. https://www.allandafoe.com/opportunity. Accessed: 2021-08-26.
[29] Shiri Dori-Hacohen, Keen Sung, Jengyu Chou, and Julian Lustig-Gonzalez. 2021. *Restoring Healthy Online Discourse by Detecting and Reducing Controversy, Misinformation, and Toxicity Online.* Association for Computing Machinery, New York, NY, USA, 2627–2628.
[30] K. Eric Drexler. 2019. *Reframing Superintelligence: Comprehensive AI Services as General Intelligence.* Technical Report. Future of Humanity Institute.
[31] Vasisht Duddu. 2018. A survey of adversarial machine learning in cyber warfare. *Defence Science Journal* 68, 4 (2018), 356.
[32] Tom Everitt, Gary Lea, and Marcus Hutter. 2018. AGI Safety Literature Review. arXiv:1805.01109 [cs.AI]
[33] Marina Favaro. 2021. *Weapons of Mass Distortion: A new approach to emerging technologies, risk reduction, and the global nuclear order.* Technical Report. Centre for Science and Security Studies. https://www.kcl.ac.uk/csss/assets/weapons-of-mass-distortion.pdf
[34] Steven Feldstein. 2019. The Global Expansion of AI Surveillance. https://carnegieendowment.org/files/WP-Feldstein-AISurveillance_final1.html. Accessed: 2021-08-26.
[35] Center for Humane Technology. 2021. Ledger of Harms. https://ledger.humanetech.com/.
[36] Ben Garfinkel and Allan Dafoe. 2019. How does the offense-defense balance scale? *Journal of Strategic Studies* 42, 6 (2019), 736–763. https://doi.org/10.1080/01402390.2019.1631810

[37] Peter Giger. 2020. COVID-19 could distract the world from even greater threats. https://www.weforum.org/agenda/2020/10/covid-19-distract-world-greater-threats/. Accessed: 2021-08-26.

[38] Kristina Gligorić, Ashton Anderson, and Robert West. 2019. Causal Effects of Brevity on Style and Success in Social Media. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 45 (2019), 23 pages. https://doi.org/10.1145/3359147

[39] Ross Gruetzemacher and Jess Whittlestone. 2022. The transformative potential of artificial intelligence. *Futures* 135 (2022). https://www.sciencedirect.com/science/article/pii/S0016328721001932

[40] Caroline Haskins. 2019. Amazon Is Coaching Cops on How to Obtain Surveillance Footage Without a Warrant. https://www.vice.com/en/article/43kga3/amazon-is-coaching-cops-on-how-to-obtain-surveillance-footage-without-a-warrant. Accessed: 2021-08-26.

[41] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. Unsolved Problems in ML Safety. arXiv:2109.13916 [cs.LG]

[42] Tim Hwang. 2020. *Subprime Attention Crisis: Advertising and the Time Bomb at the Heart of the Internet.* FSG originals.

[43] Thomas V. Ingelsby and Amesh A. Adalja. 2019. *Global Catastrophic Biological Risks.* Springer.

[44] Kokil Jaidka, Alvin Zhou, and Yphtach Lelkes. 2019. Brevity is the Soul of Twitter: The Constraint Affordance and Political Discussion. *Journal of Communication* 69, 4 (2019), 345 – 372. https://doi.org/10.1093/joc/jqz023

[45] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2 (2018), 1–33. https://doi.org/10.1145/3185593

[46] James Johnson. 2019. Artificial intelligence & future warfare: implications for international security. *Defense & Security Analysis* 35, 2 (2019), 147–169. https://doi.org/10.1080/14751798.2019.1600800

[47] Lynn Kaack, Priya Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick. 2021. Aligning artificial intelligence with climate change mitigation. (2021). hal-03368037.

[48] Lynn Kaack, Priya Donti, Emma Strubell, and David Rolnick. 2020. Artificial Intelligence and Climate Change: Opportunities, considerations, and policy levers to align AI with climate change goals. https://eu.boell.org/en/2020/12/03/artificial-intelligence-and-climate-change Heinrich-Böll-Stiftung, Ecology.

[49] Dimitris Kalimeris, Smriti Bhagat, Shankar Kalyanaraman, and Udi Weinsberg. 2021. Preference Amplification in Recommender Systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 805–815. https://doi.org/10.1145/3447548.3467298

[50] Mark Kaufman. 2020. The carbon footprint sham. https://mashable.com/feature/carbon-footprint-pr-campaign-sham. Accessed: 2021-08-26.

[51] Parag Khanna. 2019. *The Future is Asian: Commerce, Conflict, and Culture in the 21st Century.* Simon & Schuster.

[52] David Krueger, Tegan Maharaj, Shane Legg, and Jan Leike. 2019. Misleading Meta-Objectives and Hidden Incentives for Distributional Shift. *Workshop on Safe Machine Learning at the 7th International Conference on Learning Representations (ICLR 2019)* (2019), 1–7.

[53] John Leslie. 1996. *The End of the World: The Science and Ethics of Human Extinction.* Routledge.

[54] Herbert Lin. 2019. The existential threat from cyber-enabled information warfare. *Bulletin of the Atomic Scientists* 75, 4 (2019), 187–196. https://doi.org/10.1080/00963402.2019.1629574

[55] Hin-Yan Liu, Kristian Cedervall Lauta, and Matthijs Michiel Maas. 2018. Governing Boring Apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research. *Futures* 102 (2018), 6–19.

[56] Amanda Lotz. 2019. Amazon, Google and Facebook warrant antitrust scrutiny for many reasons – not just because they're large. https://theconversation.com/amazon-google-and-facebook-warrant-antitrust-scrutiny-for-many-reasons-not-just-because-theyre-large-118370

[57] Kim Lyons. 2021. Amazon's Ring now reportedly partners with more than 2,000 US police and fire departments. https://www.theverge.com/2021/1/31/22258856/amazon-ring-partners-police-fire-security-privacy-cameras. Accessed: 2021-08-26.

[58] Matthijs M. Maas, Kayla Matteuci, and Di Cooke. 2022. Military Artificial Intelligence as Contributor to Global Catastrophic Risk. In *Cambridge Conference on Catastrophic Risks 2020.* Draft available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4115010.

[59] Markets & Markets. 2019. AI in Cybersecurity Market. https://www.marketsandmarkets.com/market-reports/ai-in-cybersecurity-market-224437074.html. Accessed: 2021-08-26.

[60] Wayne M. Morrison. 2019. *China's Economic Rise: History, Trends, Challenges, and Implications for the United States.* Technical Report. Congressional Research Service.

[61] Robert S Mueller. 2019. *The Mueller report: Report on the investigation into Russian interference in the 2016 presidential election.* WSBLD.

[62] OECD. 2021. 130 countries and jurisdictions join bold new framework for international tax reform. https://www.oecd.org/newsroom/130-countries-and-jurisdictions-join-bold-new-framework-for-international-tax-reform.htm. Accessed: 2021-08-26.

[63] Jonathan Corpus Ong and Jason Vincent A. Cabañes. 2018. *Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines.* https://doi.org/10.7275/2cq4-5396

[64] Toby Ord. 2020. *The Precipice: Existential Risk and the Future of Humanity.* Hachette Books.

[65] Naomi Oreskes and Erik M. Conway. 2010. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming.* Bloomsbury Publishing.

[66] Eli Pariser. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think.* Penguin.

[67] Richard Posner. 2004. *Catastrophe: Risk and Response.* Oxford University Press.

[68] Carina Prunkl and Jess Whittlestone. 2020. Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* Association for Computing Machinery, New York, NY, USA, 138–143. https://doi.org/10.1145/3375627.3375803

[69] Martin Rees. 2003. *Our Final Century: Will the Human Race Survive the Twenty-First Century?* William Heinemann.

[70] Cheerala Rohith and Ranbir Singh Batth. 2019. Cyber Warfare: Nations Cyber Conflicts, Cyber Cold War Between Nations and its Repercussion. In *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE).* IEEE, 640–645.

[71] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. 2022. Tackling Climate Change with Machine Learning. *ACM Comput. Surv.* 55, 2 (2022), 96 pages. https://doi.org/10.1145/3485128

[72] Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking Books.

[73] Stuart J. Russell. 2019. Stuart J. Russell on Filter Bubbles and the Future of Artificial Intelligence. https://www.youtube.com/watch?v=ZkV7anCPfaY&ab_channel=LongNowFoundation. Accessed: 2021-08-26.

[74] Lora Saalman, Hwang Ji-Hwan, Su Fei, Jiang Tianjiao, Vasily Kashin, Kim Ji-Sun, Vadim Kozyulin, Arie Koichi, Li Xiang, Cai Cuihong, Liu Yangyue, Hwang Il-Soon, and Nishida Michiru. 2019. *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk, Volume II, East Asian Perspectives.* Technical Report. SIPRI.

[75] Kaylyn Jackson Schiff, Daniel S. Schiff, and Natália S Bueno. 2021. The Liar's Dividend: The Impact of Deepfakes and Fake News on Trust in Political Discourse. (2021). https://doi.org/10.17605/OSF.IO/QPXR8.

[76] Eric Schmidt, Robert Work, Safra Catz, Eric Horvitz, Steve Chien, Andrew Jassy, Clyburn Mignon, Gilman Louie, Chris Darby, Willian Mark, Kenneth Ford, Jason Matheny, José-Marie Griffiths, Katharina McFarland, and Andrew Moore. 2021. *Final Report.* Technical Report. National Security Commission on Artificial Intelligence. https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf

[77] Bruce Schneier. 2018. *Click Here to Kill Everybody: Security and Survival in a Hyper-connected World.* W. W. Norton & Company.

[78] Elizabeth Seger, Shahar Avin, Gavin Pearson, Mark Briers, Seán Ó hÉigeartaigh, and Helena Bacon. 2020. *Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologicall-advanced world.* Technical Report. The Alan Turing Institute, Defence and Security Programme. https://www.cser.ac.uk/resources/epistemic-security/

[79] Bhakti Sharma, Susanna S. Lee, and Benjamin K. Johnson. 2022. The Dark at the End of the Tunnel: Doomscrolling on Social Media Newsfeeds. *Technology, Mind, and Behavior* 3, 1 (2022). https://tmb.apaopen.org/pub/nn9uaqsz.

[80] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. arXiv:1906.02243 [cs.CL]

[81] Cass R. Sunstein. 2018. *# Republic: Divided Democracy in the Age of Social Media.* Princeton University Press.

[82] Mariarosaria Taddeo, Tom McCutcheon, and Luciano Floridi. 2019. Trusting Artificial Intelligence in Cybersecurity is a Double-Edged Sword. *Nat Mach Intell* 1 (2019), 557–560. https://doi.org/10.1038/s42256-019-0109-1

[83] Petr Topychkanov, Kritika Roy, Saima Aman Sial, Dmitry Stefanovich, Maaike Verbruggen, Sanatan Kulshrestha, Yanitra Kumaraguru, and Malinda Meegoda. 2030. *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk, Volume III, South Asian Perspectives.* Technical Report. SIPRI.

[84] Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence* 4 (2022), 189–191. https://doi.org/10.1038/s42256-022-00465-9

[85] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of*

Sciences 113, 3 (2016), 554–559. https://doi.org/10.1073/pnas.1517441113

[86] Graham Webster, Rogier Creemers, Paul Triolo, and Elsa Kania. 2017. Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017). https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/. Accessed: 2021-08-26.

[87] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI Now 2018 Report.* Technical Report. New York: AI Now Institute. https://ainowinstitute.org/AI_Now_2018_Report.html

[88] Jess Whittlestone, Kai Arulkumaran, and Matthew Crosby. 2021. The Societal Implications of Deep Reinforcement Learning. *J. Artif. Int. Res.* 70 (2021), 1003–1030. https://doi.org/10.1613/jair.1.12360

[89] Nadine Wirkuttis and Hadas Klein. 2017. Artificial intelligence in cybersecurity. *Cyber, Intelligence, and Security* 1, 1 (2017), 103–119.

[90] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* Hachette UK.

[91] Remco Zwetsloot and Allan Dafoe. 2019. Thinking About Risks From AI: Accidents, Misuse and Structure. *Lawfare* (2019). https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.