Lessons for Artificial Intelligence from Other Global Risks

Seth D. Baum,[1] Robert de Neufville,[1] Anthony M. Barrett,[1] and Gary Ackerman[1,2]

1. Global Catastrophic Risk Institute, http://gcrinstitute.org
2. University at Albany, State University of New York , College of Emergency Preparedness, Homeland Security and Cybersecurity, https://www.albany.edu/cehc

Forthcoming in Maurizio Tinnirello (Editor), *The Global Politics of Artificial Intelligence*, Boca Raton: CRC Press. This version 19 November 2019.

Abstract

The prominence of artificial intelligence (AI) as a global risk is a relatively recent phenomenon. Other global risks have longer histories and larger bodies of scholarship. The study of these other risks can offer considerable insight to the study of AI risk. This paper examines four risks: biotechnology, nuclear weapons, global warming, and asteroid collision. Several overarching lessons are found. First, the extreme severity of global risks is often insufficient to motivate action to reduce the risks. Second, perceptions of global risks can be influenced by people's incentives and by their cultural and intellectual orientations. Third, the success of efforts to address global risks can depend on the extent of buy-in from parties who may be negatively affected by the efforts. Fourth, global risks and risk reduction initiatives can be shaped by broader socio-political conditions, such as the degree of policy influence of private industry within a political jurisdiction. The paper shows how these and other lessons can inform efforts to reduce risks from AI.

1. Introduction

Those who do not learn from history are doomed to repeat it—or so the saying goes. The progression of artificial intelligence (AI) technology is pushing human society in new directions, but not all of the dynamics are entirely new. Many features of the AI issue have arisen in other contexts. That holds both for AI as it exists today and as it may exist sometime in the future. Likewise, efforts to manage the progression of AI and improve outcomes for society have much to learn from past experience with other issues that have similar features. A little history can go a long way.

The process of learning from history is broadly similar to the concept of transfer learning in the computer science of AI. Transfer learning refers to the process of saving knowledge gained from solving one problem and then applying it to solving another, different problem.[1] Transfer learning is a significant challenge in AI and a major focus of ongoing computer science research. Humans are in many respects substantially more capable at transfer learning than current AI systems—AI transfer learning is currently grappling with tasks such as recognizing different types of features within an image.[2] Nonetheless, even among humans, transfer learning can take dedicated effort, especially for complex tasks such as addressing major global issues.[3]

The focus of this paper is to present some insights from the study of global risks, especially (but not exclusively) global catastrophic risks. These insights are often of particular relevance for AI catastrophe scenarios, especially (but again not exclusively) long-term runaway AI scenarios in which humanity is unable to control the AI and catastrophe ensues. These AI catastrophe scenarios have several similarities to other

global risks, including but not limited to their potential extreme catastrophic severity. In many cases, these other risks have been addressed and studied much more extensively than AI has. The paper aims to accelerate the study of AI risk by leveraging the substantial body of experience and scholarship on other global risks.

The field of AI is of course not new. It has a rich history over many decades, as is documented in several excellent histories.[4] The idea of runaway AI catastrophe is also not new—it can be traced to early work in the 1960s[5] and even the 1860s.[6] What is relatively (but again not completely) new is the treatment of AI as a social, risk, and policy issue. Much of this is driven by the considerable recent successes of AI technology and its many applications across society. Some of it is also driven by a specific interest in the more dramatic long-term AI scenarios.[7] Now is a good time for the AI issue to learn from other issues.

There is a vast universe of insight available from other global risks, and one paper can only survey a small portion of it. The portion presented here is a mix of what the present authors are most familiar with and what we believe is most important for improving AI outcomes. That includes some emphasis on cases from US history, though much of it is of international relevance. Many prior studies have also applied insights from other global risks to specific aspects of the challenge of managing AI.[8] In one similar study, Allen and Chan[9] survey four emerging technologies to derive insights for AI as a US national security issue. The present paper also surveys multiple sources of lessons, covering four global risks of relevance for improving overall AI outcomes, especially with respect to catastrophic risks.

It should be noted that transferring lessons from one global risk to another is not the only reason to study multiple risks. Another reason is to address important questions that span multiple risks, such as how to prioritize scarce resources across multiple risks and how to address tradeoffs in potential actions that could increase one risk but decrease another. Cross-risk tradeoffs may be of particular relevance to AI due to its potential to affect many sectors of society, including sectors implicated in other global risks. While AI poses certain risks, if it is developed safely and responsibly, it could bring a range of benefits, including reductions in other global risks.[10] Cross-risk allocation and tradeoff decisions provide a compelling reason to study multiple global risks; the potential to transfer lessons across risks provides another. We therefore believe that cross-risk research should be emphasized in programs to understand and address AI and other global risks.

After an overview of definitions of key terms, the paper proceeds with discussions of four global risks, each embodying a specific theme of relevance to AI. First, biotechnology is a field of emerging technology with numerous important social benefits and also major risks. Second, nuclear weapons are technologies of paramount strategic importance across the international community. Third, global warming is a major risk that derives from profit-seeking activity by some of the largest corporations in the world and widespread consumer use. Fourth, asteroid collision is an extreme global risk that has garnered substantial international scientific and policy attention despite its very low probability. The paper transfers insights from the histories of these four cases to the study of how to effectively manage AI.

2. Definitions

Before proceeding, it is worth briefly pausing to define some key concepts used in this paper, especially for the benefit of an interdisciplinary audience. The concepts of AI, risk, and global catastrophic risk all have multifaceted and contested definitions. To a large extent, the substance of this paper is not sensitive to the particular definition used, but it is nonetheless worth elaborating on the definitions.

A common definition of AI is any artificial agent that can "achieve goals in a wide range of environments".[11] However, this emphasis on goal achievement is contested. For example, Goertzel[12] explores definitions of intelligence rooted in self-organization and involving more than just the achievement of goals. Also contentious is exactly how wide the range of environments must be for an artificial system to qualify as AI. Indeed, a common observation is that once computers can achieve some task (e.g., defeating humans at chess), then this task is no longer seen as requiring AI, which is a matter of moving goalposts.[13] An inclusive definition of AI could potentially include, for example, those "intelligent" doors at the supermarket that open when people walk up to them. An exclusive definition could potentially exclude any system that cannot perform the same set of cognitive tasks as the human mind in the same way the human mind would perform them.

For the present paper, the exact scope of AI is less important than the potential societal consequences of some AI systems (assuming, of course, that these systems qualify as AI). The introduction refers to dramatic long-term AI scenarios including catastrophic runaway AI. That is shorthand for a hypothesized form of future AI whose intelligence reaches the point where it can outsmart humans and wrest control from humanity. Such AI may undergo a process called recursive self-improvement, in which the AI makes a smarter AI, which makes a smarter AI, *ad infinitum*, potentially resulting in an "intelligence explosion" or "Singularity".[14] The forms of AI involved in this process are often known as strong AI, artificial general intelligence (AGI), or superintelligence. In contrast, current AI is "weak" and narrow, with capabilities only within a relatively narrow portion of intellectual tasks, in contrast with human minds, which can succeed across a relatively wide range of tasks.

It should be stressed that the primary arguments of this paper do not depend on the exact form of the AI, but only on its potential to pose a significant global risk. While a runaway AI presumably could pose a global risk, other forms of AI may as well. Likewise, it should also be noted that a number of analysts have expressed skepticism about runaway AI scenarios.[15] Even if these scenarios can be dismissed, then the lessons in this paper may be applicable to other AI scenarios.

This raises the question of what qualifies as a risk. To start from the basics, a risk is a "possibility of loss or injury".[16] Risk is commonly quantified as the probability of some loss multiplied by the severity of the loss if it occurs. Attention to global risks such as runaway AI are commonly predicated on the view that a risk of extreme loss can be important even if its probability is low. The exact probability of runaway AI is a controversial matter and not essential for this paper.

This paper makes some use of the concept of global catastrophic risk. Global catastrophe has been defined in a variety of ways, including as the death of at least 25% of the human population[17] or as significant undesirable change in the state of the global human system.[18] Some treatments emphasize the risk of human extinction, on grounds

that extinction would entail loss of all future generations,[19] whereas others argue that sub-extinction catastrophes involving significant permanent harm can be of comparable importance.[20]

The present paper takes a wider view on global risk. Its focus is on the lessons that can be learned from global risks, not the severity of the global events that the risks entail. The severity is important for the lessons in some cases but not all of them. The paper includes cases in which there are fruitful lessons even if the lesson is coming from a global risk whose severity does not meet the standards for global catastrophe outlined above. The point here is the insightfulness of the lesson, not the size of the risk.

3. Biotechnology

Biotechnology has been defined as "the application of science and technology to living organisms as well as parts, products, and models thereof to alter living or nonliving materials for the production of knowledge, goods, and services".[21] Biotechnology is in several respects a close analog to AI. Both are classes of emerging technology with many applications that can either decrease or increase global risks. Biotechnology can help counter pandemics, such as by enabling the rapid synthesis of vaccines[22] or using "gene drives" to propagate disease resistance among the population of insect vectors like mosquitoes.[23] Alternatively, biotechnology can exacerbate or even cause pandemics, such as by laboratory accidents that inadvertently release deadly pathogens[24] and by making it easier for dangerous pathogens to be weaponized.[25]

Infectious diseases and AI are also similar in that a small source can readily self-replicate and spread worldwide. Biological pathogens self-replicate within their host organisms (e.g., within human bodies) and can jump from host to host. Similarly, some computer software can self-replicate via "copying" and "pasting" within a computer and via transmission from computer to computer—hence the term "computer virus". Computer viruses may not all involve AI; whether they do may depend on one's definition of AI. Computer security firm Malwarebytes contends that "there are currently no examples of AI-enabled malware in the wild" but the possibility is "realistic" with existing AI techniques or fairly straightforward extensions of them.[26] Such self-replication and propagation would presumably be even more feasible for more advanced forms of AI. (An important exception would be for forms of AI that require specialized hardware to run on.)

These two similarities—both are emerging technologies with diverse applications and potential for propagation via self-replication—make biotechnology a valuable case study for AI. Biotechnology is of further value because it has a relatively extensive history as a societal issue.

One early episode was the 1975 Asilomar Conference on Recombinant DNA Molecules, aimed at assessing and managing risks from recombinant DNA research. The applicability of the 1975 Asilomar Conference to AI governance is explored in detail by Grace.[27] Grace finds that while recombinant DNA turned out to not be as dangerous as some scientists initially feared, the conference nonetheless had some success in getting this portion of the scientific community to take precautions in their research. Specifically, the US National Institutes of Health issued safety guidelines that it required for the researchers it funded, including a moratorium on recombinant DNA. Industry labs also voluntarily opted to comply with the NIH guidelines.

As Grace[28] documents, a major contributor to the success of the 1975 Asilomar Conference was substantial buy-in from the relevant scientific community. A majority of the relevant scientists, including many leaders in the field, indicated concern about the risk. The moratorium was widely—perhaps universally—followed, despite the lack of an enforcement mechanism, especially in industry. It may have helped that in the 1970s, academia already had a culture of political activism regarding issues such as environmental degradation and the Vietnam War. The success was thus due less to institutional requirements and more to communities of people. As Grace[29] puts it, "informal social mechanisms played an overwhelming part in producing a pause in research and triggering further action".

Likewise, social buy-in may be essential for successfully addressing issues in AI. Baum[30] distinguishes between extrinsic measures for addressing AI, which are imposed on AI communities from the outside, and intrinsic measures, which originate from within. AI governance conversations sometimes emphasize extrinsic measures, such as in calls to ban dangerous AI technologies,[31] restrict development of certain forms of AI to a single lab with United Nations oversight,[32] or install research review boards to assess which AI research can proceed.[33] However, measures such as these may fail if they do not have substantial buy-in from the AI community, especially if the relevant forms of AI are widely pursued. Therefore, it is important to seek buy-in from AI communities, so that any rules or guidelines would be self-enforcing. The 1975 Asilomar Conference is an important case in point.

The prospect of widespread AI community buy-in is of particular relevance to the idea of relinquishing dangerous AI technology proposed by Joy.[34] This proposal has been criticized for requiring universal buy-in among computer scientists, especially in the face of commercial or national security pressures.[35] However, if sufficient buy-in could be achieved, relinquishment might succeed even without substantial institutional pressure; the relinquishment of recombinant DNA is a case in point. Furthermore, several studies have expressed concern that even a global AI institution with enforcement powers could fail to prevent dangerous secret AI projects.[36] Thus, achieving informal community buy-in may be more important than establishing formal governance institutions, though there may nonetheless be a constructive role for both. (The challenge of monitoring AI projects is discussed further in the section on nuclear weapons.)

A more recent and controversial case is the debate over gain of function (GOF) research on potential pandemic pathogens (PPP). This research manipulates existing pathogens to make them more pathogenic, in order to learn more about the pathogen and advance the medical response to outbreaks. The research is controversial due to concerns that the enhanced pathogen could accidentally or intentionally be released from the lab, enter the human population, and spread. The enhancements could result in a more severe pandemic than what would occur from the naturally occurring pathogens. An active debate has emerged on the merits of GOF-PPP experiments. Some argue that the benefits exceed the risks, including people involved in the experiments.[37] Others argue the opposite, including people who commonly emphasize security perspectives on biotechnology.[38]

In 2012, GOF-PPP researchers agreed to a voluntary pause on these experiments.[39] Then, in 2014, the US government announced a moratorium on GOF-PPP experiments in order to assess whether the potential public health benefits of the experiments were worth

the risks. The US National Science Advisory Board for Biosecurity commissioned a risk-benefit analysis to inform the debate. The ensuing report[40] spans 1,006 pages but does not reach a definitive conclusion on whether the benefits exceed the risks. One proposed explanation for the report's inconclusiveness is that "the areas that separate pro- and anti-GOF advocates fall into areas of judgment and belief, and these differences cannot be adjudicated by risk-benefit analysis".[41] Despite the lingering disagreements and ambiguities, the US lifted the moratorium in 2017.[42]

The controversy surrounding the GOF-PPP case makes it arguably a more relevant case study for AI than the case of recombinant DNA. As with GOF-PPP experiments, experts are divided on the risks and benefits of AI, especially for the prospect and risk of long-term runaway AI. Achieving consensus on an AI moratorium may thus be more difficult, and there may be pressure to end the moratorium even before the risks and potential benefits are conclusively evaluated and consensus is reached on how best to proceed. Therefore, the GOF-PPP case may be a fertile ground for further study on advancing AI debates and reaching clear conclusions and consensus on how to manage potentially dangerous AI research.

An additional line of research on biotechnology worth examining is an assessment of biotechnology stakeholder reactions to efforts to promote responsible research and innovation, or RRI.[43] RRI aims to promote an inclusive and reflective technology research and development process, with one aim being for "societal alignment".[44] However, Kuzma and Roberts[45] find significant reluctance to adopt RRI among biotechnologists. Notably, academic biotechnologists expressed more reluctance than their industry counterparts, due to concerns about intellectual freedom and skepticism about non-expert outsiders imposing unwise restrictions. This finding resembles that of a recent survey of AGI research and development projects, in which academic projects were more likely to articulate intellectual values and industry projects were more likely to articulate values based on benefiting humanity.[46]

While the RRI biotechnology study of Kuzma and Roberts[47] is ongoing, they meanwhile call for a "more practical RRI" in which RRI researchers and advocates are actively engaged with technology projects instead of advocating RRI from the sidelines in research journals and other venues. This would appear to be wise advice for all fields of technology, including AI, and for all paradigms for improving technology development processes and outcomes, including but not limited to RRI.

Grotto[48] reviews the history of the governance of genetically modified organisms (GMOs) in agriculture and derives implications for the potential regulation of AI. Grotto contrasts the treatment of GMOs in Europe, where strict regulations drastically curtailed the use of GMOs, to the US, where a business-friendly regulatory regime led to widespread cultivation of GMOs. Grotto notes that these divergent regulatory environments were not inevitabilities, but instead were linked to historical coincidences such as European concern about food safety derived from the outbreak of mad cow disease in the United Kingdom. Additionally, the initial regulatory decisions have had lasting effects over several decades in both jurisdictions. The same could potentially apply for regulations of AI.

Finally, it is worth noting the difficulties inherent in controlling even the most pernicious forms of biotechnology. The use of biological weapons was outlawed by the Geneva Protocol of 1925, and over 180 states have signed the 1972 Biological and Toxin

Weapons Convention that prohibits the acquisition and stockpiling of these weapons. Yet, the fundamentally dual-use nature of biotechnology, where the same technique or equipment could be used for beneficial or harmful ends, enabled several states—most notably the Soviet Union in the 1970s and 1980s—to flout the international bioweapon ban and embark on massive bioweapons programs. Despite multiple attempts over several decades, the international community has been unable to craft a verification regime for the Convention that is both practically enforceable and politically acceptable. This might serve as a negative lesson in the difficulties of exercising international control over technologies—like some forms of AI—that have inherently dual-use applications. For example, image recognition techniques that can enhance public web search engines can also enhance target recognition in weapon systems. This is not the case with respect to nuclear weapons (discussed in the next section), where the underlying technologies are far less dual-use in nature.

4. Nuclear Weapons

Biotechnology is similar to AI as risky, potentially self-propagating, dual-use technologies. However, the vast majority of biotechnology applications (with the few exceptions mentioned above) are in the civilian sector. In contrast, the destructive power of nuclear weapons is emblematic of a military technology with unequivocal and paramount strategic importance on the international stage. Potentially, AI could have a similar importance, especially for more advanced forms of AI. Therefore, the extensive history and study of nuclear weapons may be a fruitful source of insights for AI. (Nuclear technology also has civilian applications, though these are not explored given that the attendant dual-use issues are substantially similar to those of biotechnology, which has already been considered.)

A potential distinction between AI and nuclear weapons is that whereas concerns about nuclear weapons often focus on intentional harm to geopolitical adversaries, concerns about AI (especially runaway AI) often focus on accidental harm to everyone.[49] However, this distinction is at most a matter of degree. It is true that countries generally do not aim to attack themselves with their own nuclear weapons, but AI developers also generally do not aim to harm themselves with their own AI systems. Furthermore, nuclear weapons must also be handled with great care to avoid accidental detonation on home soil.[50] Therefore, while nuclear weapons are not a perfect analog for AI, the similarities may be sufficient to apply lessons from the former to the latter.

One AI topic for which nuclear weapons lessons may be especially salient is on the prospect of a race to be the first to build advanced AI. It is sometimes proposed that a sufficiently advanced AI, such as a strong AI or AGI, could confer extreme "winner takes all" advantages to whomever builds it first.[51] This could occur in particular if the AI undergoes a rapid intelligence explosion but remains under the control of its builders, who then may obtain a high degree of power over all global affairs. If control is lost, catastrophe could ensue. A recent survey found no significant evidence of a race to build AGI, and instead found significant cooperation between projects.[52] Meanwhile, there is some competition on other, more modest forms of AI, such as between the US and China[53] and between companies within sectors such as autonomous vehicles.[54] The prospect of an advanced AI race is plausible due to the strategic implications of the

technology, and it is also a concern because it could preclude sufficient caution with respect to the safety of the technology.[55]

Several studies propose that the Cold War nuclear arms race may be a good analog to an initial race to build AI,[56] though a better analog would be the initial race to build nuclear weapons. An essential feature of the initial development of nuclear weapons is the extreme geopolitical tensions of that era. It appears to be a historical coincidence that the relevant scientific breakthroughs in nuclear physics occurred during the run-up to and fighting of WWII, since there is no clear link between the initial development of nuclear physics and increasing tensions in Europe. This geopolitical context may, however, explain the very rapid progression from the 1939 discovery of nuclear fission[57] to the establishment of nuclear weapons development projects in each of Germany, Japan, the Soviet Union, the UK, and the US (1939 to 1942; exact project start dates are ambiguous), and finally to the first detonations of nuclear weapons (1945). Indeed, reading this history today, it is remarkable how quickly the events unfolded, and how extensive were the contacts between scientists and high-level government officials.[58] The fast pace was motivated by a desire to build nuclear weapons first and therefore achieve major—perhaps decisive—advantage in WWII and its aftermath.[59]

An implication of this is that an AI race could be avoided or at least managed more carefully if major geopolitical tensions can also be avoided. This point applies not just for advanced AI, but also for near-term AI and other technologies of military significance. The matter is well-documented by Scharre[60] in interviews with military officials and experts and accompanying analysis. Scharre documents that militaries have thus far largely abstained from deploying autonomous weapons, or have at most proceeded rather cautiously, due to a variety of concerns including cost, safety, and ethics. (Autonomous weapons can be defined as weapons capable of selecting and firing on targets without human input.[61]) However, Scharre finds that militaries are much more likely to use autonomous weapons if a major war breaks out and countries find themselves compelled to do whatever it takes to win. One expert compares the situation to the US abstention from unrestricted submarine warfare prior to the bombing of Pearl Harbor on 7 December 1941. In regards to whether the US would deploy autonomous weapons, the expert asks, "Is it December 6 or December 8?" It follows that if a race for advanced AI is to be avoided, it may be important, perhaps even crucial, to avoid major wars between the countries that could build advanced AI.

Another important lesson from the nuclear weapons race for a potential AI race concerns secrecy. The nuclear weapons projects were highly secretive, and espionage was sometimes but not always successful. Notably, the US and its allies did not know how little progress the German nuclear weapons program was making until August 1944.[62] Because beating the German program was a primary goal of the US program, had the US known earlier, it is possible that it would have ended its own program, or at least pursued its program more carefully. Similarly, if future AI programs learn of rival programs' struggles and cessations, then they may also stop or proceed more carefully. This possibility runs counter to the proposed idea that information about rival AI projects increases risks.[63]

If an AI race is won, such that there is only one group in possession of advanced AI, then the situation may resemble the period spanning from 1945 to 1949 in which only the US possessed nuclear weapons—the so-called nuclear monopoly period.[64] It is sometimes

proposed that advanced AI may involve a strong first-mover advantage, sometimes referred to as "winner takes all".[65] While the extent to which the nuclear weapon monopoly involved the same dynamic is unclear and may have been more limited, there were nonetheless serious proposals for using the power of nuclear weapons to maintain a monopoly.

One proposal sought to keep the nuclear monopoly under US control. The US would have threatened nuclear attack against any country that attempted to build nuclear weapons, or, if need be, executed such an attack.[66] One view held that this would be preferable to permitting catastrophic nuclear proliferation.[67] It also would have given the US a strong and potentially dominant position in global politics. US President Truman ultimately declined to follow this proposal, which may suggest that the first party to build advanced AI may likewise decline to use it to maintain monopoly and a dominant global position.[68] However, there is no guarantee that other leaders would have made the same choice as Truman.

Another proposal—the Baruch Plan—called for an International Atomic Development Authority that would consolidate nuclear expertise and oversee global use of nuclear power for both peaceful and military purposes. Such an arrangement could have maintained much of the geopolitical status quo; in particular, it would theoretically not have required US dominance. The Soviet Union nevertheless rejected the Baruch Plan, apparently out of concern the US and its allies would in practice dominate the new international atomic authority.[69] Potentially, had the US threatened nuclear war if the Soviet Union refused to terminate its nuclear weapons program, it might have been more inclined to accept the Baruch Plan, though it is unclear how the Soviets would have reacted in this circumstance.

The Baruch Plan is perhaps the best historical precedent for several proposals for global AI governance backed by the power of AI.[70] The essence of these proposals is to first build an AI capable of monitoring for rogue AI development projects, and then to use this AI as the basis for enforcing global compliance with safety and ethics standards. In some variants, the AI itself could conduct the enforcement. Such a scheme could leave humans in charge, and could perhaps buy humans the time needed to carefully reflect on how best to build a more powerful AI, including an AI that humans could not or would not control. However, just as the Baruch Plan struggled to gain international consensus, so too could a comparable plan for AI. Indeed, consensus on AI may be more elusive due to important and potentially divisive questions about which types of AI to build.[71] (A milder variant of this scheme is for an international AI research center that consolidates resources for AI development, modeled after CERN.[72] However, the CERN model aims for scientific breakthrough, not safe development of technology in the public interest, and it may likewise be more applicable to the initial development of AI than the subsequent monopoly.)

The US nuclear monopoly ended in 1949 following the Soviet development of nuclear weapons. Since then, the world has persisted with multiple nuclear powers. Similarly, some AI scenarios involve multiple advanced AIs. Such scenarios have been considered especially in the context of AIs based on the digitization or emulation of human brains,[73] though they could also occur for other forms of AI.

A central feature of the ongoing era of multiple nuclear powers is the doctrine of nuclear deterrence, in which the threat of nuclear attack dissuades rival countries from

waging major wars. The absence of a global war since WWII arguably affirms the effectiveness of nuclear deterrence,[74] though this is controversial: other factors may explain international stability after WWII, including a desire to avoid any major war, nuclear or non-nuclear, and the general satisfaction of the Soviet Union and the United States with their positions in global affairs.[75] Similarly, deterrence could potentially facilitate the nonviolent and reasonably peaceable coexistence of rival AI powers. Scholarship on and experience with nuclear deterrence suggests that an AI deterrence regime may be most successful if (1) no side has the ability to destroy rivals or disable their AI systems without suffering devastating retaliation, an ability sometimes referred to in the nuclear weapons literature as "primacy"[76] and in the AI literature as "decisive strategic advantage",[77] (2) rival parties have incentives to avoid crises, or to deescalate crises if they occur, a condition sometimes referred to in the nuclear weapons literature as "crisis stability",[78] and (3) miscalculations on the intentions and activities of rivals can be avoided. These and other aspects of nuclear deterrence could prove valuable for managing a world of multiple rival advanced AI powers.

5. Global Warming

Out of all the global risks, global warming has probably been the subject of the most extensive interdisciplinary scholarly inquiry—indeed, it is probably the most extensive by a large margin. There are robust literatures on the psychology of global warming,[79] the economics,[80] the epistemic and policy implications of catastrophic risk,[81] military dimensions,[82] and much more. The voluminous scope of global warming research makes it a rich source of insight for many other global risks, including AI.

Existing AI studies have just begun to scratch the surface of insight from global warming literature. One study draws on the psychology of global warming to inform the design of both formal regulations and informal community-based measures to improve AI outcomes.[83] Another draws on the politics and psychology of skepticism and misinformation about global warming to explore how similar dynamics could play out with AI.[84] These are important topics, but there is a lot more lurking in the extensive global warming literature.

The global warming literature may be of particular relevance for scenarios in which AI is developed in the private sector. National AI development projects are plausible (and more closely related to the pursuit of nuclear weapons), but AI is currently developed primarily in the private sector. Indeed, AI is an important technology for some of the largest corporations in the world. It is therefore worth studying cases in which corporate activity poses a global risk. The case of global warming and the fossil fuel industry serves this purpose well. What follows is a very brief history to illustrate some major dynamics.

For many years, the fossil fuel industry has sought to downplay the importance of global warming and dispute the underlying science.[85] However, this was not always the case. Initially, some fossil fuel companies were active in the mainstream science of global warming. This early history is of particular relevance for the current state of affairs in AI and is worth exploring in some detail.

In 1979, Exxon installed on its Esso Atlantic supertanker custom scientific equipment for measuring air and ocean carbon dioxide concentrations. The project assessed the ocean's uptake of atmospheric carbon dioxide, which at the time was an important uncertainty in the science of global warming. The supertanker project was part of a

broader engagement by Exxon in the mainstream scientific study of global warming during the decade 1977-1987, as was recently documented in an investigative journalism project by InsideClimate News.[86]

1988 marked the beginning of serious policy interest in addressing global warming, at least in the US. Prompted in part by a severe drought and heat wave, the US Senate Energy and Natural Resources Committee held a hearing in which NASA's James Hansen delivered a now-famous testimony expressing 99% certainty that global warming had begun. As reported in a *New York Times* article, which ran at the top of the front page, several Senators on the Committee concurred that global warming was a threat and that action should be taken to counteract it.[87]

At around the same time, Exxon began supporting efforts to amplify uncertainty about the science of global warming, apparently as a strategy to stymie policy restrictions on its fossil fuel business. This change in practice is seen, for example, in the Global Climate Coalition, an industry lobbyist group that Exxon co-founded in 1989. Exxon continued its scientific research on global warming, much of which continued to support the mainstream scientific consensus, but its public-facing communications tended to question the science and oppose policy action.[88]

The divergent content of its scientific research and public communications served different purposes. As reported by InsideClimate News, Exxon wanted its own sound science to guide its internal planning, confer it legitimacy to help it influence policy, and adhere to scientific standards.[89] In contrast, the public communications were a business strategy aimed at avoiding costly regulations. This strategy has a long history, dating to 1950s tobacco industry efforts to question the science linking tobacco to cancer, and it remains in use across multiple industries, including fossil fuels.[90] Exxon's science/public divergence permits it to claim it accepts the reality of global warming while actively thwarting efforts to seriously address it.

The AI issue may now be where global warming was in the late 1970s to early 1980s: public recognition has begun, but policy regulations are not yet in serious consideration. This may explain why AI corporations are active in efforts on AI ethics: acknowledging that AI poses serious ethical issues is not yet a threat to their core business model. Indeed, the corporations may wish to demonstrate that they are responsible actors on AI and therefore do not need to be regulated. (They may even want to show that they are more responsible than their competitors, such that their competitors need to be regulated and they do not.)

The history of global warming shows that if corporations view the issue as a significant threat to their profits, then addressing the issue becomes quite a lot more difficult. The corporate ethics statements may continue insofar as it improves the corporations' public image without committing them to any costly restrictions on their business activities. Meanwhile, the companies may seek to publicly downplay the risks associated with their technology, and to lobby governments to prevent regulations. This is what the fossil fuel industry did, despite global warming posing a significant risk of global catastrophe that has long been backed by extensive mainstream science. The risk of global catastrophe from AI has a much more tenuous scientific basis and thus may be considerably easier for industry to sow doubts about.[91] (Conversely, improving expert consensus on AI risk could help counteract industry obfuscation).[92]

An important difference between global warming and AI is that whereas all fossil fuel can increase global warming, not all AI technology poses a global risk. For example, contemporary AI systems designed to play games like chess and Go may be a significant cultural phenomenon but they are not significant threats to human welfare. In order to avoid AI catastrophe, only certain forms of AI may need to be restricted, specifically those that could cause catastrophe. It thus follows that a key question for AI governance is whether the restrictions need to avoid catastrophe would cover forms of AI that are also profitable. In this context, Baum[93] coins the term "AGI profit-R&D synergy", defining it as "any circumstance in which long-term AGI R&D delivers short-term profits". If there is AGI profit-R&D synergy, then corporations may resist restrictions on the development of AGI, even though the technology could pose a global risk. The extent of AGI profit-R&D synergy could be an important—perhaps even crucial—factor in the safe governance of AI.

Some arguments against regulating fossil fuels may also apply to AI. First, it is sometimes argued that regulation stifles innovation and economic growth and restricts consumer lifestyles. This has been a common refrain in global warming debates[94] and is starting to be heard for AI.[95] Second, it is sometimes argued that regulations should be delayed until the risks are adequately understood. This has also been a common refrain in global warming debates, although the argument is sometimes, though not always, made disingenuously.[96] The same argument might also be made for AI, potentially, but not necessarily, disingenuously. Whether any particular regulation would bring net benefits (by reducing the risks from a technology more than it restricts the potential benefits) and when regulations should be introduced are important matters for policy analysis but are beyond the scope of this paper.

It is important to note that the pathologies of global warming governance do not apply equally across the globe. The case of Exxon as discussed above applies in particular to the US. Overall, the US has been relatively susceptible to corporate influence on global warming due to a variety of political, economic, and cultural factors. For example, Sheldon Whitehouse, a US Senator and strong advocate for environmental protection, attributes much of the problem to the 2010 US Supreme Court decision in the case of *Citizens United v. Federal Election Commission*, which permitted unlimited corporate spending on election-related communications.[97] Whitehouse reports that after this court case, many politicians abstained from supporting action on global warming out of concern that the fossil fuel industry would support their political opponents. In countries with more restrictive campaign finance rules, the fossil fuel industry may tend to have less influence on global warming policy. The same could hold for AI policy as well.

Finally, the history of global warming also provides a more general lesson regarding the role of scholarly expertise in public debates about science and technology issues. In public debates about global warming, corporate messaging has diminished the influence of the scientific consensus. Similarly, recent public debates about AI have given extensive attention to science and technology celebrities with limited AI expertise, such as Bill Gates, Elon Musk, and the late Stephen Hawking.[98] The history of global warming shows that public debates can diverge from expert opinion for an extended period of time. Public debates have different dynamics and epistemic standards. Efforts to improve the quality of public debates about AI should proceed accordingly.

6. Asteroid Collision

In several respects, asteroid collision and AI are very different types of issues. They differ in their origin (outer space vs. technology), their empirical basis (which is much stronger for asteroid collision), and their degree of social consensus (AI is much more controversial). Indeed, asteroid collision is notable for being perhaps the most well characterized global catastrophic risk in terms of the probabilities and severities of the risk.[99] Nonetheless, both asteroid collision and AI involve the prospect of extreme global catastrophe. Concern about the risk of global catastrophe has motivated high-level efforts to address asteroid collision by both the international scientific community and major national governments. These successes have only been partial—more work to address asteroid collision remains to be done—but they nonetheless suggest a pathway for high-level attention to AI risk even if AI catastrophe is perceived as unlikely.

Asteroid collision should be a quintessential case of what Jonathan Wiener[100] calls "the tragedy of the uncommons": a risk so rare that it is overlooked by the lay public and policymakers. Yet the history of the risk shows that this has not been the case. (The history below draws heavily on Chapman.[101])

Scientific awareness of the asteroid collision threat began in the 1940s, but was largely dormant until the early 1980s, following the landmark Alvarez et al.[102] study of the Cretaceous-Paleogene extinction and an important workshop in 1981. Public interest grew in the late 1980s via a trade press book[103] and the "near miss" of asteroid 1989FC (it was "near" in astronomical terms but not in terms of its danger to Earth). Policy interest was sparked by a position paper published by the American Institute of Aeronautics and Astronautics.[104] This outreach culminated in the 1990 US House NASA Authorization Report Language calling for NASA attention to the asteroid threat. The text of the Report Language is illuminating:

> The chances of the Earth being struck by a large asteroid are extremely small, but since the consequences of such a collision are extremely large, the Committee believes it is only prudent to assess the nature of the threat and prepare to deal with it. We have the technology to detect such asteroids and to prevent their collision with the Earth.[105]

This text shows the US House of Representatives reaching the conclusion that an extreme catastrophic risk should be taken seriously and addressed even if its probability is extremely low. The logic here mirrors the logic found throughout academic studies advocating attention to global catastrophic risks,[106] including the risk of runaway AI.[107] The asteroid threat therefore offers an important precedent, one that may be worth revisiting in policy debates about AI.

The 1990 US House NASA Authorization Report Language is no anomaly. The US government has remained engaged on the asteroid threat. Most recently, the US National Science and Technology Council, an Executive Branch advisory group, published the *National Near-Earth Object Preparedness Strategy and Action Plan*.[108] The US has also sponsored astronomy studies to detect asteroids, as have other countries. Scientists report the detection of over 90% of large asteroids, none of which are found to be on Earthbound trajectories.[109] Ongoing detection programs scan for smaller (and thus harder to detect) asteroids. The US and other countries are also developing techniques for

deflecting away Earthbound asteroids. The US has taken at least some formal steps toward operationalizing those techniques: the US National Nuclear Security Administration is holding onto an important component of nuclear explosives for "potential use in planetary defense against earthbound asteroids".[110] While more could be done, it is nonetheless clear that the asteroid threat has significant high-level policy recognition and support for efforts to address it.

The nature of the asteroid threat may have made it easier for governments to recognize it than it would be for the AI threat. The Cretaceous-Paleogene extinction provides what appears to be a clear proof of principle, and the overall science of asteroids is relatively well understood. Scientists routinely publish figures graphing the frequency of collision as a function of asteroid size based on well-established empirical data.[111] This makes it easier for government officials to believe in the validity of the threat. Furthermore, the asteroid threat involves no human enemies whose livelihood may be put at risk by efforts to address the threat—the only thing put at risk is the asteroid itself. Likewise, there is likely to be less in the way of institutions lobbying against asteroid risk reduction.

However, the history of the asteroid threat shows that it did in fact struggle to gain serious recognition, and it did also have to overcome institutional opposition. Early media coverage included a significant "giggle factor" and portrayed concerned astronomers as "Chicken Littles" playing up concern to generate funding for their research.[112] Additionally, many scientists, including those in leadership at NASA, pushed back against efforts to address the asteroid threat. The scientists did not want the "giggle factor" tarnishing their reputations, and they did not want the applied mission of the asteroid threat to pull scarce funds away from pure (non-applied) scientific research.[113] Recognition of the AI threat faces very similar challenges. For the asteroid threat, these challenges have been overcome with at least some modest success. This fact should provide some encouragement to efforts to gain serious attention for the AI threat.

Furthermore, the human dimensions of asteroid risk are not as well understood as the physical and environmental dimensions. This holds in particular for potentially globally catastrophic human harm.[114] While asteroid risk is probably the most well-characterized global catastrophic risk, the exact risk estimates are nonetheless uncertain. This uncertainty has not precluded policy action; either the uncertainty has gone unnoticed by policymakers, or the policymakers opted to act anyway. The willingness of policymakers to act despite uncertainties in the risk is an encouraging precedent for AI, which is a considerably more uncertain risk.

7. Lessons Learned

Several overarching lessons for the study of AI can be drawn from the four global risks surveyed in the preceding sections. First and foremost, the extreme severity of global risks does not on its own ensure they will be addressed successfully. The severity of global risks does sometimes move key actors to take action, such as US Congressional action on asteroid risk. Other actors have not been persuaded by the severity, such as academic biotechnologists reluctant to adopt RRI and the fossil fuel industry opposing global warming policy. There are compelling theoretical reasons to prioritize reducing global risks, but these reasons are not always persuasive in practice.

Second, perceptions of global risks can be strongly influenced by people's incentives and by their cultural and intellectual orientations, especially where the size of the risk is uncertain. Global risks are highly uncertain due to the complexity of global events and the rarity of (and thus lack of data on) global catastrophes. Even the risk of asteroid collision, which derives from relatively simple and well-understood astronomical processes, has significant uncertainties with respect to human consequences. Communicating asteroid risk has also been challenging due to the risk's "giggle factor". Other risks are more contentious. Global warming risk is disputed, perhaps disingenuously, by a fossil fuel industry that has an incentive to avoid regulation. GOF-PPP risk is disputed by different populations of experts, with those conducting GOF-PPP experiments sometimes finding the risk to be lower than those who emphasize the security dimensions of biotechnology. The size of AI risk is also currently disputed within expert communities, could also come to be disputed by industry, and might be difficult to communicate due to its own distinct "giggle factor". Efforts to characterize and raise awareness about AI risk should be mindful of these dynamics to mitigate biases in analysis and public discourse.

Third, whether the response to global risks is successful may depend on buy-in especially from those who stand to lose as a result of risk reduction measures. Out of all the cases studied in this paper, two stand out as relatively successful stories of risk reduction: recombinant DNA and asteroid collision. The former involved a moratorium that had broad buy-in from the relevant scientific community. The latter involves response measures that do not implicate or restrict anyone to any significant extent. Contrasting examples abound. Biotechnology RRI initiatives face resistance from academics concerned about intellectual freedom. Biological and nuclear weapons arms control initiatives face resistance from states concerned about losing strategic advantage. Initiatives to reduce greenhouse gas emissions face pushback from the fossil fuel industry. Obtaining buy-in for AI risk reduction may be especially challenging because many key actors, including academics, states, and industry, could stand to lose as a result of risk reduction initiatives. AI risk reduction initiatives may need an unusually large and multifaceted effort to achieve buy-in in order to succeed.

Finally, risks and risk reduction initiatives can be heavily shaped by broader socio-political conditions. GMO regulation has been stricter in Europe than in the US due to Europe's less business-friendly political culture and its recent experience with mad cow disease. Fossil fuel regulation has also been relatively lax in the US, perhaps due to its relatively permissive campaign finance laws (which are closely tied to the business-friendly US political culture). Nuclear weapons technology was developed extremely quickly because certain breakthroughs in nuclear physics happened to coincide with the extreme international competition of the 1930s and 1940s. Likewise, AI risk reduction initiatives will not take place in a vacuum. To succeed, the initiatives should account for the particular socio-political conditions and the (possibly unforeseen) circumstances in which they will take place.

8. Conclusion

History is not doomed to repeat itself. Past failures to manage global risks do not necessarily portend future failures—especially if important lessons are learned. At the same time, past successes do not necessarily portend future successes. While AI is

relatively new as a social, risk, and policy issue, it has much to learn from other global risks.

This paper proposes that it is possible to accelerate the study of AI as a social, risk, and policy issue by leveraging the existing scholarship on and experience with other global risks. To demonstrate this possibility, the paper presents examples from four other classes of global risk: biotechnology, nuclear weapons, global warming, and asteroid collision. Although it would be valuable to expand the study of these four cases to broader international contexts, they shine considerable light on how to understand the prospects for AI catastrophe, how such an outcome could be avoided, and how AI outcomes can be improved more generally.

In addition, these sorts of historical case studies may hold some rhetorical value for efforts to improve AI outcomes. The history may help some people take certain AI scenarios more seriously, especially scenarios involving long-term, high-stakes AI. Many people in academia, government, and other sectors may be dismissive of such scenarios,[115] instead preferring to focus their attention on more near-term and empirically robust issues. The history of other global risks can provide at least an indirect empirical basis for some important aspects of long-term AI, and can likewise demonstrate that similar issues have often gotten substantial high-level attention. The history may be of particular value for relating long-term AI to people with background in other global risks, because it can help to make long-term AI seem more familiar.

At the heart of this paper is a claim that transferring lessons from other global risks can be an efficient and productive means of advancing progress on AI. In putting forward this claim, we do not mean to imply that lessons from other global risks are sufficient for studying issues in AI. To the contrary, AI will inevitably pose some novel challenges that require dedicated original analysis. Furthermore, we do not mean to claim that transferring lessons from other risks will be the most efficient and productive research strategy for all groups working on issues in AI. This approach will tend to work best for research groups, such as our own, that already have background in other risks. The merits of this approach for research groups that are more narrowly specialized on AI is an important question and is beyond the scope of this paper. Instead, this paper serves to demonstrate the intellectual and practical benefits that can be gained from transferring lessons from other global risks to the study of AI.

Acknowledgements
Maurizio Tinnirello and two anonymous reviewers provided helpful comments on an earlier version of this paper. Jake Stone assisted in formatting the manuscript. Any remaining errors are the authors' alone.

[1] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering* 22 no. 10 (2010): 1345-59.
[2] A. R. Zamir et al., "Taskonomy: Disentangling Task Transfer Learning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018): 3712-22.
[3] D. N. Perkins and G. Salomon, "Transfer of learning," In *International Encyclopedia of Education* 6452-57. (Oxford: Pergamon Press, 1992); Perkins and Salomon, "Knowledge to Go: A Motivational and Dispositional View of Transfer," *Educational Psychologist* 47 no. 3 (2012): 248-58.
[4] D. Crevier. *AI: The Tumultuous History of the Search for Artificial Intelligence*. (New York: Basic Books, 1993); J. Markoff. *Machines of Loving Grace: The Quest for Common Ground Between Humans and Robots*. (New York: HarperCollins, 2016); P. McCorduck, *Machines Who Think: 25th Anniversary Edition*. (Natick, MA: AK Peters, 2004).
[5] N. Wiener, "Some Moral and Technical Consequences of Automation," *Science*, 131 no. 3410 (1960): 1355–8; I. J. Good, "Speculations Concerning the First Ultraintelligent Machine," *Advances in Computers* 6 (1965): 31-88.
[6] S. Butler, "Darwin Among the Machines," *The Press*, June 13, 1863.
[7] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. (Oxford: Oxford University Press, 2014); Callaghan et al. *Technological Singularity*. (Berlin: Springer, 2017); Eden et al. *Singularity Hypotheses*. (Berlin: Springer, 2015); K. Sotala and R.V. Yampolskiy, "Responses to Catastrophic AGI Risk: A Survey," *Physica Scripta* 90 no. 1 (2015): 018001. https://doi.org/10.1088/0031-8949/90/1/018001.
[8] Some notable examples include K. Grace. *The Asilomar Conference: A Case Study in Risk Mitigation*. (MIRI Technical Report 2015-9, 2015); K. Grace. *Leó Szilárd and the Danger of Nuclear Weapons: A Case Study in Risk Mitigation*. (MIRI Technical Report 2015-10, 2015); J. Altmann, and F. Sauer, "Autonomous Weapon Systems and Strategic Stability," *Survival* 59 no. 5 (2017): 117-42; S. D. Baum, "On the Promotion of Safe and Socially Beneficial Artificial Intelligence," *AI & Society* 32 no. 4 (2017): 543-51; S. D. Baum, "Countering Superintelligence Misinformation," *Information* 9 no. 244 (2018); A. Grotto, "Genetically Modified Organisms: A Precautionary Tale for AI," *AI Pulse*, January 24, 2019. https://aipulse.org/genetically-modified-organisms-a-precautionary-tale-for-ai-governance-2; M. Maas, "How Viable is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons," *Contemporary Security Policy* 40 no. 3 (2019): 285-311.
[9] G. Allen and T. Chan. *Artificial Intelligence and National Security*. (Belfer Center for International Affairs, Harvard Kennedy School, July 2017).
[10] As is argued, for example, by Bostrom, *Superintelligence*. For a counterargument, see S. Pueyo, "Growth, Degrowth, and the Challenge of Artificial Superintelligence," *Journal of Cleaner Production* 197 no. 2 (2018): 1731-6.
[11] S. Legg and M. Hutter, "Universal Intelligence: A Definition of Machine Intelligence," *Minds & Machines* 17 no. 4 (2007): 391-444.
[12] B. Goertzel, "Superintelligence: Fears, Promises and Potentials," *Journal of Evolution and Technology* 25 no. 2 (2015): 55-87.
[13] McCorduck, *Machines Who Think*.
[14] Bostrom, *Superintelligence*; A. H. Eden, J. H. Moor, J. H. Søraker, and E Steinhart (eds), *Singularity Hypotheses*. (Berlin: Springer, 2012); Sotala and Yampolskiy, "Catastrophic AGI Risk,"; Callaghan, *Technological Singularity*.
[15] R. Brooks, "I, Rodney Brooks, am a Robot." *IEEE Spectrum,* June 1, 2008, https://spectrum.ieee.org/computing/hardware/i-rodney-brooks-am-a-robot; J. J. Bryson and P. P. Kime "Just an Artifact: Why Machines are Perceived as Moral Agents." In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 1641-46, ed by Toby Walsh (Vol. 2. Barcelona, July 2011); for an overview, see Baum, "Countering Superintelligence Misinformation".
[16] S. Kaplan and J. Garrick, "On the Quantitative Definition of Risk. *Risk Analysis*." 1 no.1 (1981): 11-27.
[17] A. Atkinson. *Impact Earth: Asteroids, Comets and Meteors—The Growing Threat*. (London: Virgin, 1999).
[18] S. D. Baum and I. C. Handoh, "Integrating the Planetary Boundaries and Global Catastrophic Risk Paradigms." *Ecological Economics* 107 (2014): 13-21.
[19] e.g., D. Parfit. *Reasons and Persons*. (Oxford: Oxford University Press, 1984); J. G. Matheny, "Reducing the Risk of Human Extinction," *Risk Analysis* 27 no. 5 (2007): 1335-44.
[20] N. Bostrom, "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9 no. 1 (2002); S. D. Baum et al., "Long-term Trajectories of Human Civilization." *Foresight* 21 no. 1 (2019): 53-83.
[21] A. Wyckoff. *OECD Directorate for Science, Technology and Industry Committee for Scientific and Technological Policy Report*, DSTI/EAS/STP/NEST1/RD(2001)30, 2001, as discussed in A. S. Dahms, "Biotechnology: What it is, What it is Not, and the Challenges in Reaching a National or Global Consensus," *Biochemistry and Molecular Biology Education* 32 no. 4 (2006): 271-8.
[22] P. R. Dormitzer, "Rapid Production of Synthetic Influenza Vaccines," *Current Topics in Microbiology and Immunology* 386 (2015): 237-73.

[23] National Academies of Sciences, Engineering, and Medicine. *Gene Drives on the Horizon: Advancing Science, Navigating Uncertainty, and Aligning Research with Public Values*. (Washington, DC: The National Academies Press, 2016).
[24] Gryphon Scientific. *Risk and Benefit Analysis of Gain of Function Research*. (Washington, D.C: Gryphon Scientific LLC, 2015).
[25] Biological and Toxin Weapons Convention 2011. *Scientific and Technological Developments that May Be Relevant to the Convention*.
[26] P. Arntz et al., "When Artificial Intelligence Goes Awry: Separating Science Fiction from Fact," (Malwarebytes 2019) https://resources.malwarebytes.com/resource/artificial-intelligence-goes-awry-separating-science-fiction-fact.
[27] Grace, *The Asilomar Conference*.
[28] Grace, *The Asilomar Conference*.
[29] Grace, *The Asilomar Conference*, 20.
[30] Baum, "Socially Beneficial Artificial Intelligence."
[31] R. A. Posner. *Catastrophe: Risk and Response*. (Oxford: Oxford University Press, 2004); G. Wilson "Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law." *Virginia Environmental Law Journal* 31 (2013): 307–64.
[32] Bostrom, *Superintelligence*.
[33] R. Yampolskiy and J. Fox, "Safety Engineering for Artificial General Intelligence," *Topoi* 32 no. 2 (2013): 217–26.
[34] B. Joy, "Why the Future Doesn't Need Us," *Wired*, April 1, 2000.
[35] J. J. Hughes, "Global Technology Regulation and Potentially Apocalyptic Technological Threats." In *Nanoethics: The Ethical and Social Implications of Nanotechnology*, ed. F. Allhoff et al., 201-14, (Hoboken, NJ: John Wiley, 2007); J. O. McGinnis, "Accelerating AI," *Northwestern University Law Review* 104 no. 366 (2010): 366–81.
[36] D. Dewey, "Long-term Strategies for Ending Existential Risk from Fast Takeoff," In *Risks of Artificial Intelligence*, ed. V. C. Müller, 243-66, (Boca Raton: CRC, 2015); McGinnis "Accelerating AI,"; Tomasik *International Cooperation vs. AI Arms Race*. (Foundational Research Institute, 2016) https://foundational-research.org/files/international-cooperation-ai-arms-race.pdf.
[37] For example, R. A. M. Fouchier, "Studies on Influenza Virus Transmission Between Ferrets: The Public Health Risks Revisited." *mBio* 6 no. 1 (2015): e02560-14, https://doi.org/10.1128/mBio.02560-14.
[38] For example, M. Lipsitch and T. V. Inglesby, "Moratorium on Research Intended to Create Novel Potential Pandemic Pathogens," *mBio* 5 no. 6 (2014): e02366-14, https://doi.org/10.1128/mBio.02366-14; M. Lipsitch and T. V. Inglesby, "Reply to 'Studies on Influenza Virus Transmission Between Ferrets: The Public Health Risks Revisited,'" *mBio* 6 no.1 (2015): e00041-15, https://doi.org/10.1128/mBio.00041-15.
[39] Fouchier et al., "Pause on Avian Flu Transmission Research," *Science* 335 no. 6067 (2012): 400-1.
[40] Gryphon Scientific, *Risk and Benefit Analysis*.
[41] M. J. Imperiale and A. Casadevall, "Zika Virus Focuses the Gain-of-function Debate." *mSphere* 1 no. 2 (2016): e00069-16, https://doi.org/10.1128/mSphere.00069-16.
[42] D. Reardon, "Ban on Pathogen Studies Lifted." *Nature* 553 (2018): 11.
[43] J, Kuzma and P. Roberts, "Cataloguing the Barriers Facing RRI in Innovation Pathways: A Response to the Dilemma of Societal Alignment," *Journal of Responsible Innovation* 5 no 3 (2018): 338–46.
[44] B. Ribeiro et al., "Introducing the Dilemma of Societal Alignment for Inclusive and Responsible Research and Innovation," *Journal of Responsible Innovation*, 5 no. 3 (2018): 316–31.
[45] Kuzma and Roberts, "Cataloguing the Barriers."
[46] S. D. Baum. *A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy*. (Working Paper Global Catastrophic Risk Institute 17-1, 2017).
[47] Kuzma, "Cataloguing the Barriers."
[48] Grotto, "Genetically Modified Organisms."
[49] For example, S. Armstrong et al., "Racing to the Precipice: A Model of Artificial Intelligence Development." *AI & Society* 31 no. 2 (2016): 201–6; C. Shulman, "Arms Control and Intelligence Explosions," (paper presented at the 7th European Conference on Computing and Philosophy, Bellaterra, Spain, July 2–4, 2009).
[50] E. Schlosser. *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety*. (New York: Penguin, 2013).
[51] For example, Shulman, "Intelligence Explosions"; Armstrong, "Racing to the Precipice."
[52] Baum, "A Survey of Artificial Intelligence."
[53] N. Thompson and I. Bremmer, *"The AI Cold War That Threatens Us All." Wired,* October 23, 2018, https://www.wired.com/story/ai-cold-war-china-could-doom-us-all.
[54] D. Welch and E. Behrmann, "Who's Winning the Self-driving Car Race?" *Bloomberg*, May 7, 2018, https://www.bloomberg.com/news/features/2018-05-07/who-s-winning-the-self-driving-car-race.

[55] Shulman, "Intelligence Explosions"; S. Cave and S. S. Ó hÉigeartaigh, "An AI Race for Strategic Advantage: Rhetoric and Risks." In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18*, 36–40 (New Orleans: ACM Press, 2018); W. Naudé and N. Dimitri, "The Race for an Artificial General Intelligence: Implications for Public Policy," *AI & Society* (2019) https://doi.org/10.1007/s00146-019-00887-x.
[56] A. Ramamoorthy and R. Yampolskiy, "Beyond MAD?: The Race for Artificial General Intelligence." *ICT Discoveries* no. 1 (February 2, 2018); Shulman, "Intelligence Explosions"; Tomasik, "*International Cooperation*."
[57] O. R. Frisch, "Physical Evidence for the Division of Heavy Nuclei Under Neutron Bombardment," *Nature* 143 no. 3616 (1939): 276; L. Meitner and O. R. Frisch, "Disintegration of Uranium by Neutrons: A New Type of Nuclear Reaction," *Nature* 143 no. 3615 (1939): 239.
[58] R. W. Clark. *The Birth of the Bomb*. (New York: Horizon Press, 1961); M. Gowing, *Britain and Atomic Energy 1939-1945*. (London: Macmillan, 1964); K. Macrakis, *Surviving the Swastika: Scientific Research in Nazi Germany*. (New York: Oxford University Press, 1993).
[59] H. L. Stimson, "The Decision to Use the Atomic Bomb," *Harper's Magazine*, February 1947, 98-101.
[60] P. Scharre. *Army of None: Autonomous Weapons and the Future of War*. (New York: Norton, 2018).
[61] Scharre and Horowitz, "An Introduction to Autonomy in Weapon Systems" (Working paper, Center for a New American Security, February 13, 2015) https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-weapon-systems.
[62] S. A. Goudsmit, *Alsos*. (New York: Henry Schulman, 1947).
[63] Armstrong et al., "Racing to the Precipice."
[64] G. H. Quester, *Nuclear Monopoly*. (New Brunswick: Transaction Publishers, 2000).
[65] Shulman, "Intelligence Explosions"; Cave, "An AI Race"; Naudé, "Artifical General Intelligence."
[66] Quester, *Nuclear Monopoly*.
[67] B. Russell, "The Bomb and Civilization," *Forward* 39 no. 33, (August 18, 1945).
[68] Dewey, "Ending Existential Risk."
[69] Dewey, "Ending Existential Risk."
[70] Dewey, "Ending Existential Risk"; Shulman "Intelligence Explosion"; Goertzel "Superintelligence."
[71] Dewey, "Ending Existential Risk."
[72] G. Marcus, "Artificial Intelligence is Stuck. Here's How to Move it Forward," *New York Times*, July 29, 2017, https://nytimes.com/2017/07/29/opinion/sunday/artificial-intelligence-is-stuck-heres-how-to-move-it-forward.html.
[73] R. Hanson. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. (Oxford: Oxford University Press, 2016); R. A. Koene, "Embracing Competitive Balance: The Case for Substrate-Independent Minds and Whole Brain Emulation." In *Singularity Hypotheses:* A *Scientific and Philosophical Assessment*, ed. A. H. Eden et al., 241-67, (Berlin: Springer, 2012).
[74] K. Waltz, "The Spread of Nuclear Weapons: More May Better," *Adelphi Papers* 21 no. 171 (1981); R. Rauchhaus, "Evaluating the Nuclear Peace Hypothesis: A Quantitative Approach," *Journal of Conflict Resolution* 53 no. 2 (2009): 258-77.
[75] J. Mueller, "The Essential Irrelevance of Nuclear Weapons: Stability in the Postwar World," *International Security* 13 no. 2 (1988): 55-79; W. Wilson, *Five Myths About Nuclear Weapons*. (Boston: Houghton Mifflin Harcourt, 2013).
[76] For example, K. A. Lieber and D. G. Press, "The End of MAD? The Nuclear Dimension of US Primacy," *International Security* 30 no 4 (2006): 7-44.
[77] For example, Bostrom, *Superintelligence*.
[78] For example, Altmann, "Autonomous Weapon Systems."
[79] For example, Centre for Research on Environmental Decisions, *The Psychology of Climate Change Communication: A Guide for Scientists, Journalists, Educators, Political Aides, and the Interested Public*. (New York: Columbia University Center for Research on Environmental Decisions, 2009).
[80] For example, N. Stern, *The Economics of Climate Change: The Stern Review*. (Cambridge, UK: Cambridge University Press, 2007).
[81] For example, M. L. Weitzman, "On Modeling and Interpreting the Economics of Catastrophic Climate Change," *Review of Economics and Statistics* 91 no. 1 (2009): 1–19.
[82] For example, CNA Military Advisory Board 2014.
[83] Baum, "Socially Beneficial Artificial Intelligence."
[84] S. D. Baum, "Superintelligence Skepticism as a Political Tool," *Information* 9 no. 209 (2018); Baum "Countering Superintelligence Misinformation."
[85] N. Oreskes and E. M. Conway. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. (New York: Bloomsbury, 2010).
[86] N. Banerjee, L. Song, and D. Hasemyer, "Exxon: The Road Not Taken." *InsideClimate News*, September 16, 2015, https://insideclimatenews.org/content/Exxon-The-Road-Not-Taken.
[87] P. Shabecoff, "Global Warming has Begun, Expert Tells Senate," *New York Times*, June 24, 1988.

[88] G. Supran and N. Oreskes, "Assessing ExxonMobil's Climate Change Communications (1977–2014)," *Environmental Research Letters* 12 no. 8 (2017): 084019, https://doi.org/10.1088/1748-9326/aa815f.
[89] InsideClimate News, "The Road Not Taken."
[90] Oreskes, *Merchants of Doubt*.
[91] Baum, "Superintelligence Skepticism."
[92] Baum, "Countering Superintelligence Misinformation."
[93] Baum, "A Survey of Artificial Intelligence," 19.
[94] J. D. Collomb, "The Ideology of Climate Change Denial in the United States," *European Journal of American Studies* 9 no. 1 (2014) https://doi.org/10.4000/ejas.10305.
[95] e.g., Gurkaynak et al., "Stifling Artificial Intelligence: Human Perils," *Computer Law and Security Review: The International Journal of Technology Law and Practice* 32 no. 5 (2016): 749-58, https://doi.org/10.1016/j.clsr.2016.05.003 2016; D. Castro, "The U.S. May Lose the AI Race Because of An Unchecked Techno-Panic," *Center for Data Innovation*, March 5, 2019, https://www.datainnovation.org/2019/03/the-u-s-may-lose-the-ai-race-because-of-an-unchecked-techno-panic.
[96] Oreskes, *Merchants of Doubt*.
[97] S. Whitehouse. *Captured: The Corporate Infiltration of American Democracy*. (New York: New Press, 2017).
[98] V. Galanos, "Exploring Expanding Expertise: Artificial Intelligence as an Existential Threat and the Role of Prestigious Commentators, 2014-2018," *Technology Analysis & Strategic Management* 31 no. 4 (2019): 421-32.
[99] S. D. Baum, "Uncertain Human Consequences in Asteroid Risk Analysis and the Global Catastrophe Threshold," *Natural Hazards* 94 no. 2 (2018): 759-75.
[100] J. Wiener, "The Tragedy of the Uncommons: on the Politics of Apocalypse," *Global Policy* 7 no. 1 (2016): 67–80.
[101] C. R. Chapman, "History of the Asteroid/Comet Impact Hazard," *Southwest Research Institute*, https://www.boulder.swri.edu/clark/ncarhist.html.
[102] L. W. Alvarez et al., "Extraterrestrial Cause for the Cretaceous-Tertiary Extinction," *Science* 208 no. 4448 (1980): 1095–108.
[103] C. R. Chapman and D. Morrison. *Cosmic Catastrophes*. (New York: Plenum, 1989).
[104] American Institute of Aeronautics and Astronautics, *Dealing with the Threat of an Asteroid Striking the Earth*. (Reston, VA, 1990).
[105] US House (1990). Report Language to H.R.5649, National Aeronautics and Space Administration Multiyear Authorization Act of 1990. (Washington, DC: United States House of Representatives). Quote at p. 30.
[106] e.g., Posner, *Catastrophe*; Wiener, "Tragedy of the Uncommons."
[107] Sotala, "Catastrophic AGI Risk."
[108] National Science and Technology Council, *National Near-Earth Object Preparedness Strategy and Action Plan*. (Washington, DC: US National Science and Technology Council, 2018).
[109] A. Mainzer et al., "The Population of Tiny Near-Earth Objects Observed by NEOWISE," *Astrophysical Journal* 784 no. 2:110 (2014), https://doi.org/10.1088/0004-637X/784/2/110.
[110] Government Accountability Office, *Nuclear Weapons: Actions Needed by NNSA to Clarify Dismantlement Performance Goal*. (Washington, DC: Government Accountability Office, 2014).
[111] For example, A. W. Harris et al., "Asteroid Impacts and Modern Civilization: Can we Prevent a Catastrophe?" In *Asteroids IV*, ed. P. Michel et al., (Tucson: University of Arizona Press, 2015).
[112] Chapman, "History of the Asteroid."
[113] Chapman, "History of the Asteroid."
[114] Baum, "Uncertain Human Consequences."
[115] For example, Brooks, "I, Rodney Brooks,"; Bryson, "Just an Artifact."