Collective Action on Artificial Intelligence: A Primer and Review

Robert de Neufville and Seth D. Baum
Global Catastrophic Risk Institute
https://gcrinstitute.org

Published in *Technology in Society*, 66:101649 (2021),
DOI 10.1016/j.techsoc.2021.101649.
This version 26 November 2021.

Abstract
Progress on artificial intelligence (AI) requires collective action: the actions of two or more individuals or agents that in some way combine to achieve a result. Collective action is needed to increase the capabilities of AI systems and to make their impacts safer and more beneficial for the world. In recent years, a sizable but disparate literature has taken interest in AI collective action, though this literature is generally poorly grounded in the broader social science study of collective action. This paper presents a primer on fundamental concepts of collective action as they pertain to AI and a review of the AI collective action literature. The paper emphasizes (a) different types of collective action situations, such as when acting in the collective interest is or is not in individuals' self-interest, (b) AI race scenarios, including near-term corporate and military competition and long-term races to develop advanced AI, and (c) solutions to collective action problems, including government regulations, private markets, and community self-organizing. The paper serves to bring an interdisciplinary readership up to speed on the important topic of AI collective action.

Keywords: artificial intelligence, catastrophic risk, collective action, governance

1. Introduction

Collective action "arises when the efforts of two or more individuals or agents… are required to accomplish an outcome" (Sandler, 2015, p. 196). The overall development of artificial intelligence (AI) requires collective action, as do efforts to ensure that AI development results in good outcomes for society, both because it requires individuals to coordinate their actions and because it is simply too large and complex a task for any single individual acting alone to accomplish.

This paper is a primer and review of AI collective action. By *AI collective action*, we mean the collective action humans take to improve the outcomes of AI development, not the collective action of groups of AIs. The paper is a primer in that it serves as an introduction to the topic for a diverse audience of social scientists, computer scientists, policy analysts, government officials, organizers, activists, concerned citizens, and anyone else with an interest in the topic. The paper is a review in that it reviews the literature that has thus far accumulated on the topic.

Many aspects and applications of AI will require collective action. In particular, as we show below, collective action will be needed to reach agreement on AI rules and standards, to develop AI that is broadly socially beneficial rather than merely being profitable for particular developers, and to avoid competition or conflict that could lead to

AI being developed or used in a less safe and beneficial way. AI is a potentially transformative technology that could shape the way people live, work, and communicate. This raises the question of how AI can contribute to or hinder good outcomes for society —or, phrased differently, how AI can contribute to or hinder the building of a good society.[1] As Coeckelbergh (2018) notes, the question of the appropriate role of technology in society is a political question of concern to the general public rather than a purely technical question of concern only to private individuals. Addressing that question will require collective action.

AI collective action is a relatively new field. A large portion of the work on it has been written in just the last few years. The topic has also attracted interest from a wide range of thinkers from both inside and outside academia. This paper takes an inclusive approach to the literature, drawing on everything from full-length peer-reviewed research papers to short blog posts and commentaries. The intent is to provide a broad review of existing ideas about AI collective action, regardless of where they are published.

The core method of this paper is a nonsystematic review of the AI collective action literature. The selection of literature reviewed is not necessarily comprehensive or representative of the full body of work on the topic, though we believe we have identified the overwhelming majority of relevant work. The literature reviewed was identified via keyword searches, our own prior knowledge of the literature, and citation tracking. The Google Scholar and Google Web Search tools were queried with phrases such as "artificial intelligence collective action", "artificial intelligence governance", and "artificial intelligence race". "Artificial intelligence race" was included because a large number of the articles that concern AI collective action focus in particular on AI races. Most of the literature identified was published after 2010 because this is a new literature, but keyword searches were not limited to articles published after 2010. All academic publications were included. Non-academic publications were included if in the authors' judgment they contained a well-supported and compelling discussion of AI collective action.

Much of the AI collective action literature identified in this paper focuses on long-term scenarios in which competitive pressures increase the risk of catastrophic outcomes, especially via the development of advanced forms of AI such as artificial general intelligence (AGI) and superintelligence.[2] Another important focus of the literature is on military applications and the potential dangers of near-term AI arms races. This paper has some emphasis on the literature on near-term and long-term AI races (Section 3), though much of the discussion is more general. Indeed, despite the wide range of forms and applications that AI technology has now and could have in the future, the collective action issues they raise are similar both in the nature of the issues and in their potential solutions. In this regard, the paper contributes to a broader analysis of synergies between near-term and long-term AI issues (Baum, 2018a; Baum, 2020; Cave & Ó hÉigeartaigh, 2019; Prunkl & Whittlestone, 2020).

[1] See discussion of the idea of the good society in Griffy-Brown et al. (2018).

[2] AGI is AI that is capable of thinking across a wide range of cognitive domains. In contrast, current AI is narrow in the sense that it is only capable of thinking across a relatively narrow range of domains. Superintelligence is AI that is significantly smarter than humans in many important respects. AGI is often considered an important precursor to superintelligence. A concern is that AGI and/or superintelligence could outsmart humanity, take over the world, and potentially kill everyone in the process (Yudkowsky, 2008; Miller; 2012; Bostrom, 2014). This scenario is referred to in this paper as an "AI catastrophe".

Following a general overview of AI collective action, the paper presents discussions of AI race scenarios and collective action that could be taken to address them. Much of the literature to date focuses on the collective action issues raised by races. It should be stressed that races are not the only scenarios that raise AI collective action issues, though they are important ones.

2. A Primer on Collective Action

This section presents a general background on collective action, with emphasis on aspects of relevance to AI. It is intended mainly for an interdisciplinary non-specialist audience; much (though not all) will be familiar to readers who are versed in the study of collective action.

2.1 Divergent vs. Convergent Interests

The concept of collective action is perhaps most closely associated with "collective action problems", in which individual interests and group interests diverge. Where individual interests are at odds with group interests, individuals' pursuit of their own self-interest may lead to outcomes that are worse for not only for the group as a whole but also for each individual considered separately. The prisoner's dilemma is a simple and well-known form of collective action problem.[3]

Collective action situations also include scenarios in which individual interests and group interests converge. In these cases, individuals' pursuit of their own self-interest can lead to outcomes that are collectively optimal. In markets, for example, individuals can make transactions that benefit themselves and improve group outcomes by allocating resources more efficiently.[4] Individual interests also converge in coordination problems when individuals would benefit by agreeing on a common set of rules or standards. Where individual interests align, the challenge is to coordinate individual actions rather than to resolve conflicts among individual interests.

AI collective action includes cases of both divergent and convergent interests. AI races are, in the literature identified in this paper, typically associated with a mix of convergent and divergent interests (Section 3). Divergent interests also arise for AI work done in the public interest, including basic research published in the open literature and the development of standards and techniques for AI safety and ethics.[5] In both of these cases, individuals have limited incentives to do the work, so it is often funded by governments or private philanthropies. In general, public funding of basic research has a range of benefits in scientific and technical fields besides AI as well (Salter & Martin, 2001). Convergent interests can arise when AI developers would benefit from using the same standardized platforms. It can be valuable for everyone to use the same programming languages, code repositories, operating systems, etc. so that programs can interoperate and developers do not have to start from scratch each time they join a new

[3] For a detailed description of the prisoner's dilemma and related scenarios as they pertain to AI collective action, see Askell et al. (2019).

[4] An important exception is when there are externalities, i.e. when benefits or harms of a market transaction are not built into the price of the good or service being sold. For example, the harms from global warming are in most instances an externality that is not built into the price of fossil fuel. In these situations, individual and collective interests can diverge.

[5] Hughes (2007) and AI Impacts (2016) both discuss these points in the context of AI safety and ethics.

project. In some cases, it may not particularly matter which platform is used, as long as everyone uses the same one.

Collective action is more difficult to achieve when individual and group interests diverge, because individuals' pursuit of their own interests will not lead on its own to collectively optimal outcomes.[6] The study of collective action therefore tends to focus on the more difficult challenges posed by divergent interests. This paper has a similar emphasis. Nonetheless, it is important to recognize that AI collective action does not always entail competition among individuals with divergent interests. Furthermore, there are often opportunities to align individual and group interests. This can be a powerful way to improve AI outcomes.

One important way individual and group interests can converge is when individuals value group outcomes. Collective action problems like the prisoner's dilemma assume individuals are narrowly self-interested, but individuals may also care about others, in which case they may choose to act in the group interest even if doing so goes against their self-interest. In the extreme case, individuals may value only group outcomes, so that they always act in the group interest regardless of the implications for themselves.

Even if all individuals value only group outcomes, they could still be in competition if they value different aspects of group outcomes. For example, a survey by Baum (2017a) finds that AGI developers typically value either humanity as a whole or intellectual progress for its own sake, but not both. Since these goals might be at odds, one can imagine competition between pro-humanity developers and pro-intellectual progress developers. Similarly, the Armstrong, Bostrom and Shulman (2016) model of dangerous AI races uses an "enmity" parameter to model AI projects' preference for their own project winning the race. Enmity could derive from either self-interest or differing preferences about which group values to favor. In the model, higher enmity results in less cooperation and more dangerous races.

Efforts to promote and reach consensus on group values could increase the convergence of interests and make collective action more likely. In recent years, different versions of AI ethics principles have proliferated (Zeng et al., 2018), which may be a constructive step in this direction (though for a critical analysis of this, see Whittlestone et al., 2019). Establishing ethical norms among AI developers could also increase the convergence of interests. Examples of attempts to establish ethical norms for AI development include senior AI researcher Stuart Russell's call for people in AI to care more about the societal consequences of the technology (Bohannon, 2015) and employee activism among workers in the field of AI (Gasser & Schmitt, 2019). A relevant historical precedent is the moratorium on recombinant DNA in the 1970s, which succeeded in part because the relevant scientific community had a shared culture of political activism that left them predisposed to accepting a moratorium on their work (Grace, 2015).

2.2 The Excludability and Rivalry of Goods

In public choice economics, collective action is often evaluated in terms of the type of goods that are being produced.[7] "Goods" in this context are anything that one or more individuals benefit from. Goods can include tangible material objects like computers, and intangible things like information.

[6] For discussion of this issue see e.g. Olson (1965) and Ostrom (1990).

[7] See e.g. Olson (1965) and Ostrom 1990.

A common scheme for classifying forms of collective action depends on whether the goods at issue are excludable and rivalrous (Samuelson, 1954; Ostrom & Ostrom, 2017). Goods are excludable if it is possible to prevent people from benefiting from them, and they are rivalrous if enjoying them reduces how much they benefit others. Table 1 presents a 2x2 matrix of (non)excludable and (non)rivalrous goods.

	EXCLUDABLE	**NON-EXCLUDABLE**
RIVALROUS	Private goods	Common goods
NON-RIVALROUS	Club goods	Public goods

Table 1. Goods classified by rivalry and excludability.

The four types of goods can be summarized as follows:

- Private goods (excludable, rivalrous): goods that can be used exclusively by a single agent or subset of agents. An example is labor: a particular hour of a person's time can only be used once, and the worker and her employers can readily exclude others from benefiting from it.
- Club goods (excludable, non-rivalrous): goods that are not depleted when they are used, but nevertheless are exclusive. Examples include proprietary software and research: the owners can exclude others from using them, for example by making them available on a password-protected website, but using them does not diminish other people's ability to access and benefit from them.
- Common goods (non-excludable, rivalrous): goods that are diminished or depleted when they are used but that everyone has access to. An example is open electric grids: anyone can access the electricity by plugging into an outlet, but using the electricity depletes global energy resources.[8]
- Public goods (non-excludable, non-rivalrous): goods that can be enjoyed without being depleted and that anyone can benefit from. Examples include open-source software and openly published research, which anyone can access and use without diminishing other people's ability to benefit from it as well.

Public choice theory holds that collective action problems, in which individual and group interests diverge, arise mainly for non-excludable (common and public) goods because individuals can benefit from them without contributing to their production or maintenance (Sandler, 2015). When goods are excludable, access can be restricted (e.g., via setting higher prices for access) so as to align individual and group interests. This is controversial because higher prices can exclude poorer people and result in outcomes that are worse for the group once considerations of equity are taken into account. In any case, non-excludable goods face the basic problem of free-riding: since individuals can benefit from these kinds of goods whether or not they have contributed to providing or maintaining them, they have an incentive to free-ride on the contributions of others. (This assumes that individuals are self-interested.) As a result, individual community members

[8] It is possible to build electric grids with access restrictions, though this is often not done, especially for the outlets found in most buildings. An example of a common good in which access restrictions cannot be built in is ocean fisheries: oceans cannot be redesigned to prevent certain people from accessing them.

have an incentive to produce less of these types of goods than they would if they were required to contribute, and less than would be optimal for the community as a whole.

Common goods face the additional challenge of maintaining an adequate supply given the incentive to use up the goods. Unlike with public goods, free-riding diminishes the supply of common goods available for other individuals. Classic discussions of common good resources (such as open pastures for grazing livestock animals) warned of the "tragedy of the commons" due to the incentive to deplete the resources (Hardin, 1968). Modern scholarship speaks of the "drama of the commons" because this type of situation does not necessarily end in tragedy (Dietz, Ostrom, & Stern, 2003).[9] How the tragedy can be avoided is a primary focus of collective action research, as discussed in Section 4. Note that the provision of common goods has the same basic structure as the prisoner's dilemma.

Important aspects of AI development have the structure of public and common goods. The avoidance of an AI catastrophe has the structure of a public good in that the enjoyment of the absence of AI catastrophe is neither excludable nor rivalrous. Likewise, there may be an undersupply of efforts to avoid AI catastrophe, with individuals hoping to free-ride on the efforts of others (AI Impacts, 2016).

Cooperation between different AI projects can have the structure of the provision of a common good, especially in scenarios like dangerous races in which competition increases risks. Just as competing individuals can diminish the supply of a common good by overusing it, competing individual AI projects can increase the risk of an AI catastrophe by taking avoidable risks. In both cases, individuals pursuing their own self-interest cause harms to the overall population. This incentive structure has been recognized in some prior AI literature (Bostrom, 2014; AI Impacts, 2016; Armstrong et al., 2016; Askell et al., 2019). AI races are discussed further in Section 3.

2.3 Distribution of Contributions

In certain situations, successful collective action can depend on how the contributions of different actors are distributed. A helpful typology of these situations comes from Hirshleifer (1983) and Barrett (2007). Work by AI Impacts (2016) applies the Barrett (2007) typology to AI. This prior literature frames the typology in terms of public goods instead of the distribution of contributions. However, the situations described by the typology are not restricted to public goods, and in our view are more productively expressed in terms of distributions of contributions.

The core of the typology features four types of collective action situations:

- Aggregate effort: when the result depends on the summed contributions of all actors rather than on how much any given actor contributes, such as when a project needs a certain amount of financing but it does not matter where it comes from.
- Single best effort: when the result depends on the effectiveness of the best effort to address the problem, such as when a solution to a technical problem is needed.
- Mutual restraint: when the result depends on the extent to which actors all refrain from taking certain actions, such as when developing or using a technology is so dangerous that a single failure could be catastrophic.

[9] For recent discussion, see Boyd et al. (2018).

- Weakest link: when the result depends on the effectiveness of the worst effort to address the problem, such as when the security of a system will fail if it fails at any single point.

Aggregate effort, single best effort, and weakest link situations are all variants of situations requiring different distributions of contribution across all the actors. In aggregate effort, what matters is the total contribution of all actors rather than how that effort is distributed among them. An example of an aggregate effort situation is one in which what matters is how much money is raised for a project, not where the money comes from. In single best effort situations, what matters is the work done on a single top-performing project, which could include contributions from any number of actors. An example of a single best effort situation was the attempt to send humans to the moon, which succeeded when one project succeeded, but would have failed if the same effort had been divided among lots of projects that did not reach the moon. In weakest link situations, all actors must make some minimum contribution. Weakest link situations are ones in which every actor is responsible for maintaining some essential part of a joint activity. An example of a weakest link situation is when the failure of a single actor to follow cybersecurity protocols would compromise an entire network. Mutual restraint situations are similar to weakest link situations in that every actor must meet at least some minimum standard, but in mutual restraint situations what matters is not what actors do, but what they refrain from doing. An example of a mutual restraint situation is one in which every actor needs to refrain from a risky behavior that could affect them all, like performing an experiment that could have catastrophic consequences. One can readily imagine other variants of these situations, in which for example success requires two projects to meet some performance threshold (a "double best effort", such as for shipping custom products to two clients), or in which success requires all actors except one to meet some performance threshold (a "second-weakest link", such as for the cybersecurity of a computer network in which one computer can be air gapped to store backup files).

The Barrett (2007) typology also includes coordination situations, in which the result depends on the extent to which actors act in the same way. Coordination systems can also involve a variety of distributions of contribution. For example, weakest link and mutual restraint situations are situations in which every single actor must either act or refrain from acting in the same way (e.g., if all drivers have to act the same way at stoplights); aggregate effort situations are situations in which actors must collectively coordinate their activities closely enough but not perfectly (e.g., if drivers have to stop at stoplights only reliably enough for other drivers to have confidence they can safely go when the light is green).

Cooperation may be easier in single best effort situations than in aggregate effort situations because efforts can be focused on a single top-performing project. Cooperation likewise may be harder in weakest link and mutual restraint situations than in aggregate effort situations because each actor has to meet some minimum standard of action. Cooperation is potentially easier to achieve in coordination situations because actors have an interest in coordinating and may be indifferent as to which coordination scheme is used.

Different aspects of AI development conform to different collective action situations. AI Impacts (2016) observes that refraining from developing an AI that would favor some

groups over others requires mutual restraint, while the development of a safe ethical AI design may depend on a single best effort. The extent to which safe, ethical AI design is funded may depend on the aggregate effort of actors. The elimination of bugs or security problems in a jointly developed system may depend on the contribution of a weakest link actor. The development of standards that allow AI systems to work together may require coordination.

3. Dangerous AI Race Scenarios

Broadly speaking, an AI race is when AI development proceeds more quickly than it otherwise would, especially when multiple development teams compete to develop AI more quickly than each other. AI races are not necessarily dangerous and might even hasten the arrival of socially beneficial forms of AI (Athimoolam, 2018). Similarly, efforts to slow AI development can be harmful by limiting progress (Gurkaynak et al., 2016; Castro, 2019). Nonetheless, a substantial body of literature concerns the potential dangers of AI races.

The main reason AI races could be dangerous is that developers might cut corners on safety in order to develop more quickly. This theme has been articulated, for example, by Danzig (2018) in the context of military AI races and by many authors[10] in the context of long-term AGI development. There can be tension between the value of racing ahead for relative advantage (or perhaps for other reasons) and the value of exercising due caution with respect to harmful unintended consequences. This tension can be found for a variety of technologies, certainly including AI.

In collective action terms, dangerous AI races may be especially worrisome if racing is in the individual interests of developers but not in the collective interest. In such cases these divergent interests create a collective action problem (Section 2.1). Each developer decides whether to participate in a race to develop an AI system. Each would benefit from winning the race, as long as their winning AI system is safe. If the first AI system developed is safe, the benefits of winning accrue primarily to the winning party. However, if the first system is not safe, for example if the winning system is one that causes global catastrophe, everyone might collectively bear the costs. In this situation, it may be in the individual interest of any given AI group to participate in a race even though it is in the collective interest of the group as a whole to avoid a race and develop AI more cautiously. This situation has the same basic incentive structure as prisoner's dilemma or common goods situations (Section 2.2). In the extreme case, these situations could require mutual restraint (Section 2.3) if an AI catastrophe can be avoided only if no developers engage in a race.

Dangerous AI races could occur when it is not in the interest of developers to engage in an AI development race, if developers wrongly believe it is in their interest to do so. Choices about AI development may be made under conditions of "bounded rationality",[11] in which developers' ability to determine the best course of action may be limited in practice. The complexity of AI systems makes their risks difficult to evaluate with any certainty. Cultural norms or a cognitive biases may also make developers inclined to

[10] Shulman (2009); Armstrong et al. (2016); Tomasik (2016); Aldana (2017); Han et al. (2020); Naudé & Dimitri (2018).

[11] "Bounded rationality" refers to the practical limits on actors' ability to make optimal decisions in real world conditions.

downplay risks, even where evidence suggests the risks are significant. As a result, dangerous AI races can arise when developers' perceived interests diverge from the collective interests of the group (e.g., when they imagine the reward for being the winner of an AI development race would be larger than it actually is). In such cases, the collective action problem that results could potentially be resolved by providing developers with accurate information about their individual interests (Askell et al., 2019).[12]

In contrast, actors in non-dangerous AI races may tend to have convergent interests, or at least interests that do not diverge. (An AI race is non-dangerous if the benefits of engaging in the race would outweigh or at least balance the harms.) Many common AI development races may be non-dangerous rather than dangerous. In some cases, such as when competition speeds the development of beneficial technology, the benefits of engaging in a development race may substantially outweigh the harms. Some races may be coordination situations, in which developers collectively benefit from working on the same problems at the same time. For example, the ImageNet database and ImageNet Large Scale Visual Recognition Challenge allow groups working on image recognition to pool their efforts and learn from one another (Russakovsky et al., 2015).

It is also possible that developers could have an incentive *not* to engage in an AI race that would *benefit* the public by spurring valuable innovation. This is also a collective action problem because the interests of the individual developers diverge from the collective interest. This could occur if winning the race would depend on the production of a public good (Section 2.2), like basic research and development (R&D) that is likely to end up in the public domain. In such a case developers would have individual incentives to collectively underinvest in AI development and free-ride on the work of others.

The distinction between dangerous and non-dangerous races is conceptually important for the reasons outlined above. but it may also be a rhetorically important distinction. When AI races are not identified as dangerous, they may be seen as harmless contests that should be played to win, rather than the risky competitions they can be (Baum, 2017b; Cave & Ó hÉigeartaigh, 2018). As a result, using the unqualified term "AI races" could increase the likelihood of irrational and dangerous AI races. Arguably, simply framing AI development as a "race" may make it sound less risky than it is, exaggerate the extent to which it is necessarily a competition, and minimize the potential for beneficial collaboration.[13] We therefore advise identifying dangerous AI races as "dangerous". (Identifying non-dangerous AI races may be less important, but potentially worthwhile nonetheless.)

3.1 Dangerous Near-Term AI Races

AI race scenarios can broadly be split into private and public sector races. It is not a sharp distinction, since there are important interconnections between the private and public sectors, including public funding for private R&D and government use of AI developed

[12] In theory, it might be possible to resolve collective action problems in which individual and collective interests do diverge by misleading actors into believing their individual interests actually align with collective interests, although spreading false information could create other problems.

[13] Similarly, Roff (2019) argues against framing AI competition as an arms race on grounds that it could "could escalate rivalry between states and increase the likelihood of actual conflict".

by the private sector. Nonetheless, private corporations and national governments have different competitive dynamics.

The private sector is currently the driving force behind AI development. The computer technology industry has a reputation for product development that can be rushed and risky, as epitomized by the former Facebook slogan "move fast and break things". Some of it is driven by competition, with rival groups seeking to gain market share, profit, and other advantages. There is intense competition to hire the most talented AI researchers and to build computer systems with superior performance in various tasks; both of these competitions have sometimes been referred to as "arms races" even though they do not involve military armaments (The Economist, 2015; Rabesandratana, 2018; Byrne, 2015). A different sort of private sector "arms race" occurs between AI teams at social media companies tasked with removing inappropriate content and people who post the content (Metz & Isaac, 2019). Despite the ubiquity of near-term private sector AI races, the matter has not yet received substantial scholarly attention.

AI races in the public sector—especially military AI races—have attracted more attention. Geist (2016) traces military AI competition as far back as the 1960s Cold War competition between the Soviet Union and the US. The 1960s initiatives focused on advancing basic research. Today, AI is increasingly being used in operational military systems. A significant concern is the near-future prospect of arms races for autonomous weapons (Altmann & Sauer, 2017; Rickli, 2017; Russell, 2018), though Scharre (2018) documents that countries have not been as quick to embrace autonomous weapons as one might think. More generally, the extent of a military AI arms race may be overstated (Roff, 2019). Nonetheless, there are clear reasons for rival militaries to race one another to improve their AI capabilities. For example, fighter aircraft can gain a relative advantage over adversary planes by using AI to make faster and better combat decisions (Byrnes, 2014). AI arms races may be more likely to occur with (a) weapon systems that attack other systems of the same kind (like fighters that are designed to engage other fighters in dogfights) than with (b) weapon systems that attack something else (like drones that are designed to engage human targets). For (a) but not (b), improving the weapon systems' AI gives the other side a reason to improve its own similar systems' AI.

It is beyond the scope of this paper to assess the danger of near-term AI race scenarios. It is clear that near-term AI could pose risks, and it is plausible that some of these risks may be sufficient to render certain near-term AI races dangerous. This matter has not yet been clearly established in the literature and is a worthy focus of future research.

3.2 Dangerous Long-Term AI Races

The dangers of long-term AI races have received more attention. This literature generally focuses on scenarios involving extremely capable AI and extremely high stakes. It is often argued that the first AI to reach some capability threshold could become enormously powerful, perhaps even powerful enough to effectively take over the world.[14] If that argument is right, then global outcomes could largely be determined by which AI project wins a development race. If the AI is built safely, then the result is an extreme case of "winner takes all". Alternatively, if the AI is not safe, then "winning" the AI

[14] See e.g. Good (1966), Vinge (1993), Kurzweil (2005), Omohundro (2008), Yudkowsky (2008), Chalmers (2010), Barrat (2013), and Bostrom (2014).

development race would be the ultimate Pyrrhic victory, with catastrophic aggregate harms up to and potentially including human extinction.

Because of the extreme stakes it is unambiguously in the collective interest to avoid such catastrophes. It is presumably very much in the self-interest of an AI developer to be the first to develop a safe AI. It may or may not be in the collective interest for that developer to be the first to develop safe AI, depending on whether the AI would make the AI broadly beneficial, since an immensely capable AI might be able to work wonders, both in the service of its developers and the world in general. The exact calculus depends on the probabilities of beneficial and catastrophic outcomes for each potential developer, as well as how much the developer and the collective value these outcomes. Resolving this calculus involves a suite of difficult philosophical and empirical issues that have to date not received attention in the AI collective action literature.

The enormous stakes also blur the distinction between public and private actors. A private actor that built and controlled such an AI could have power rivaling the largest states. The substantial stakes also give states a reason to intervene in AI development and potentially even to nationalize parts of the AI industry (Aldana, 2017, p.10). While some analyses of long-term AI have focused on development in one sector or another,[15] the impact of powerful AI is likely to transcend specific sectors and affect the collective interests of a broad group of actors (though the sector in which development occurs may matter for other reasons).

Several studies develop mathematical models of the dynamics of dangerous long-term AI races. Armstrong et al. (2016) model races in which the reward for winning is large and teams can improve their chances of winning by skimping on safety precautions, which also increases the probability of catastrophe. They find increased risk when there are more competing teams, when the teams have a stronger preference for winning, when taking risks increases the odds of winning, and when teams know each other's capabilities. Aldana (2017) uses a variety of two-player games to explore opportunities to alter incentives toward cooperation, for example by highlighting the potentially catastrophic consequences of failing to cooperate. Han et al. (2020) model competition over successive rounds of AI development, finding that cooperation is less likely when advanced AI can be built in the relatively near future. Finally, Naudé and Dimitri (2018) model the cost of building AGI, finding that if it were expensive, relatively few groups would compete, but that public funding could incentivize cooperation. While some of these findings may seem self-evident, these models offer a means of exploring some of the subtler nuances of races and opportunities to increase cooperation. On the other hand, these models make sweeping mathematical assumptions about complex socio-technological processes and need to be supplemented with empirical studies.

3.3 Long-Term Effects of Near-Term Races

Finally, it is worth briefly discussing the idea that near-term AI races can affect the long-term development of AI. In particular, it has been argued that near-term AI races could slow the long-term development of AI.

One mechanism for this is by generating public backlash. If near-term AI is not developed with sufficient caution, it could cause problems that lead to regulations and other initiatives that impede further AI development. This view was recently expressed

[15] For example, AI Impacts (2016) and Tan & Ding (2019) focus on races between countries.

by Mounir Mahjoubi, the French minister for digital affairs and an architect of France's AI policy. In Mahjoubi's words, "If you don't invest in responsibility around AI, you will create resistance and resentment in the population" (Simonite, 2018). An example of an incident that could create an enduring backlash is a 2018 incident in which a self-driving Uber killed a pedestrian. Following this incident, the US National Highway Traffic Safety Administration and National Transportation Safety Board launched probes into autonomous vehicle safety (Knight, 2018). While it is not yet clear whether the incident will ultimately slow the roll out of autonomous vehicles, it is nonetheless indicative of the possibility.

Another mechanism is that near-term races could focus resources on maximizing near-term performance at the expense of long-term progress. As is common with many areas of technology, long-term advances in AI may require new techniques based on fundamental breakthroughs that derive from basic research. In contrast, optimizing near-term performance may primarily involve the application of existing techniques. Hence, Marcus (2017) argues that focusing on established machine learning techniques has left the field of AI at a long-term disadvantage. In other words, near-term races may not incentivize people to come up with new and possibly more effective AI techniques.

4. Solutions to AI Collective Action Problems

The social science literature on collective action has identified three broad types of approaches to solving collective action problems. Each approach uses different strategies to encourage individuals to act in the collective interest even when their immediate incentives are to act in against the collective interest. The three approaches involve top-down government policy, bottom-up community governance, and private ownership (Ostrom, 1990).

4.1 Top-Down Government Solutions

Governments have some capacity to compel collective action. Governments can set policies that require individuals to act in accordance with the collective interest. Governments also have the authority to enforce compliance with policies and to punish noncompliance. For these reasons, governments are often seen as the appropriate institutions for encouraging collective action, both in general (Ostrom, 1990) and with respect to AI (AI Impacts, 2016).

The AI collective action literature has produced a wide range of proposals for top-down government solutions. The range of these proposals shows the tradeoff that exists between the ambitiousness of a proposal and its feasibility. The more ambitious proposals would probably do more to advance AI collective action *if* they were implemented, but may be difficult to implement. The more modest proposals would do less to advance AI collective action but are probably more feasible.

The most ambitious proposals call for no less than a world government that would monitor AI development and force rogue AI projects to comply with ethics and safety standards. This bold idea has been repeatedly proposed in the literature (Tomasik, 2013; AI Impacts, 2016; Bostrom, 2019). A related proposal is for an "AI nanny" that uses advanced AI to govern humanity and guide the development of even more advanced AI (Goertzel, 2012). Somewhat less ambitious proposals call for international institutions that would house or otherwise govern the global development of AI. These proposals

leave national sovereignty intact except with respect to AI development. Specific proposals include a "global watchdog agency" (Ramamoorthy & Yampolskiy, 2018), an international AI project with broad authority to regulate the development and use of advanced AI (Bostrom, 2014, pp. 104-106; Dewey, 2016; Dafoe, 2019; Nindler, 2019), and publicly funding a limited number of groups that have the exclusive right to develop advanced AI on the condition that it is in the public interest (Naudé & Dimitri, 2018).

The most feasible proposals require relatively modest tweaks to existing governance schemes. For example, AI could be included in existing international arms control agreements (Hughes, 2007). New arms control agreements could be made for AI-based weapons. The feasibility of the proposed ban on autonomous weapons has been questioned, but may nonetheless be possible (Wallach, 2017), and other arms control measures that stop short of a ban would be more feasible. International institutions can also assist in setting the agenda on AI and facilitating dialog. Indeed, this is already occurring on a limited scale via the UN High-Level Panel on Digital Cooperation. Another international body that could guide global AI development is the Global Partnership on AI, a recently formed coalition of states with the mission of supporting "the responsible and human-centric development and use of AI in a manner consistent with human rights, fundamental freedoms, and [their] shared democratic values" (US Department of State, 2020). Similar to this is a proposal for an intergovernmental organization that brings together stakeholders from the public sector, industry, and academia to develop non-binding recommendations for how to increase international cooperation on AI (Erdelyi & Goldsmith, 2018). Since binding "hard law" rules can be difficult to enact, Marchant (2019) proposes a range of non-binding "soft law" measures that create expectations but are not formally enforced by government, including "private standards, voluntary programs, professional guidelines, codes of conduct, best practices, principles, public-private partnerships and certification programs". Finally, an international organization could sponsor, host, or serve as a clearinghouse for research into AI, and play a role similar the European Organization for Nuclear Research (CERN) (Castel & Castel, 2016; Marcus, 2017; Slussalek, 2018) in physics or the Intergovernmental Panel on Climate Change (IPCC) in climate science (Miailhe, 2018). Such an organization could potentially address the collective action problem of underinvestment in basic research, as well as the problem of underinvestment in AI ethics and safety research.

Other literature has explored national government-led solutions. A major focus is on liability schemes for harms caused by privately developed AI and robotics systems.[16] Liability schemes can encourage collective action by changing individuals' incentives so that they align with the collective interest. Other proposals call for national governments to sponsor research on AI safety (McGinnis, 2010), to establish national panels to develop guidelines for AI R&D and use (Daley, 2011), or to create a National Algorithm Safety Board similar to the US National Transportation Safety Board to provide independent oversight of algorithms used to make decisions that impact the public (Shneiderman, 2016).

[16] See e.g. Karnow (1996), Asaro (2007), Marchant & Lindor (2012), Funkhouser (2013), Gurney (2013), LeValley (2013), Scherer (2015), Wu (2015), and Zohn (2015). Note that liability schemes apply mainly for near-term AI; the catastrophic harm from long-term AI may be so severe that it destroys the liability system (White & Baum, 2017).

Outside of the extensive literature on liability, most of the proposed government solutions have involved action at the international level. The reason appears to be that since AI can be developed anywhere in the world, comprehensive collective action requires a global scope. Some even worry that piecemeal national regulations could push AI development underground to "rogue nations" (Goertzel & Pitt, 2012). Nevertheless, national governments have substantial authority even within the largely decentralized international system. Action at the national level is often more feasible than action at the international level, and successful action at the national level can serve as a model for international action.

Top-down government action, whether at the national or international level, is not a perfect solution for AI collective action problems. Governments may struggle to regulate AI due to its complexity and rapid change (Athimoolam, 2018; Askell et al., 2019; Marchant, 2019). Governments themselves may promote corporate or other special interests over the collective interest (Goertzel, 2017). International proposals, especially ambitious ones, require a high degree of international cooperation, which may be hard to achieve given the difficulty of monitoring compliance, the incentives each state would have to defect (Dafoe, 2019, p. 46), the number of political jurisdictions and industries that would be involved, and the speed at which AI technology changes (Marchant, 2019). Even the most ambitious global agency might still fail to prevent dangerous projects from advancing (McGinnis, 2010; Dewey, 2016; Tomasik, 2016). In addition, any government scheme that lacks support from AI communities could create resentment and lead to pushback, making collective action more difficult (Baum, 2017a). Therefore, while top-down government solutions may be able to play a role in advancing AI collective action, they probably cannot resolve all AI collective action issues.

4.2 Private Ownership Solutions

Privatization is a common approach to solving collective action problems. One prominent context in which privatization is often used is in the management of natural resources. A private actor who owns a resource has an incentive to use it optimally. For example, private ownership of pasture may make overgrazing less likely, because the owners have an interest in preserving pasture for their own future benefit.

Private ownership schemes are difficult to apply to AI development, since AI technology has no single owner. Much of the software and AI development techniques are publicly available. Code can be proprietary, but it is relatively difficult to keep code private since it is often easy to copy software and other digital information. Even if AI development were entirely in private hands, it would still have enormous public impacts, creating externalities private owners would not have a direct incentive to address. Private firms might develop AI only in their own interest rather than in the public interest (Goertzel, 2017) underinvest in safety and ethics research that would primarily benefit the public (Hughes, 2007; Miller, 2012; AI Impacts, 2016), and misinform the public about AI risks in order avoid regulation or scrutiny (Baum, 2018b).[17] Tan and Ding (2019) call for a global AI market to mitigate safety risks, but concede that government regulation may be necessary to ensure that AI markets are globally integrated, standardized, and egalitarian.

[17] For a discussion of how to counter such misinformation, see Baum (2018c).

While the ease of copying software may make it difficult to enforce private ownership schemes, hardware may be more susceptible to private ownership schemes. Hwang (2018) describes several attributes of hardware manufacturing that make it easier to govern, including the relatively small number of large, fixed facilities involved in producing the high-end hardware used in cutting-edge AI systems. Hardware manufacturers could conceivably play a role in encouraging AI collective action. However, hardware manufacturing has the same externality as software development: the benefits of safe, ethical practices are spread widely across the public, creating an incentive to underinvest in safety and ethics.

4.3 Bottom-Up Community Solutions

The third type of solution to collective action problems is bottom-up community self-organizing. In bottom-up community solutions, private actors work with one another in the collective interest in the absence of any overarching authority with the capacity to enforce cooperation. Bottom-up community solutions are appealing because they may be more feasible than top-down solutions. Cooperation in the absence of an enforcement authority might seem theoretically inelegant, but empirical studies of real-world collective action find that community self-organizing is often effective (National Resource Council, 2002; Ostrom, 1990).[18]

Soft law instruments like private standards, voluntary programs, and professional guidelines should arguably be considered examples of community self-organizing. Other soft law measures blur the distinction between top-down government solutions discussed in Section 4.1 and community self-organizing. Institutions like the Institute of Electrical and Electronics Engineers (IEEE) and the Partnership on AI are already bringing AI groups together to craft common ethical principles and promote cooperation. These processes are new, and it remains to be seen how successful they will be. Nonetheless, there is at least some chance that they will be successful, just as previous initiatives have succeeded at promoting collective action in other contexts.

Community self-organizing does not require individuals to be altruistic as long as they recognize that cooperating to achieve common goals is in their own private interests. Community self-organizing may be most likely to succeed when individuals are willing to make personal sacrifices for the greater good, the way employee activists who oppose controversial but profitable applications of AI technology are. However, communities may be able to self-police adherence to reasonable norms even if individuals are not willing to sacrifice their private interests. Some have argued that simply fostering norms among people in the field of AI could improve outcomes (Baum, 2017a; Baum, 2018a; Cave & Ó hÉigeartaigh, 2019). Strengthening norms about near-term AI development could also lay the groundwork for collaborating on long-term AI development (Baum, 2018a; Cave & Ó hÉigeartaigh, 2019).

Psychological factors can play an important role in determining the effectiveness of community solutions. For example, AI Impacts (2016) and Baum (2017b) propose that it may be possible to cultivate a taboo against the development of dangerous AI. A taboo is

[18] Baum (2017a) distinguishes between extrinsic constraints imposed on AI communities from the outside and intrinsic measures that are developed by the AI communities themselves. Baum (2017a) argues that while efforts to improve AI outcomes commonly focus on extrinsic measures, intrinsic measures can often be effective. Government and market collective action solutions are generally extrinsic, whereas community solutions are generally intrinsic.

an informal social norm against some action. Taboos can be effective. For example, the taboo against nuclear weapon use may be a major reason no nuclear weapon has been used in violence since 1945 (Tannenwald, 2005; Schelling, 2006). What kind of taboos would be appropriate is a matter of debate—it might be going too far to treat the development of some forms of AI as unacceptable as the use of nuclear weapons—but some kind of taboo against dangerous AI development could facilitate AI collective action.

Community self-organizing may be especially important for AI developed using open-source software. Closed-source/proprietary software could be confined within a single institution, which may be able to make sure the software complies with safety and ethics standards. However, open-source software can be developed by anyone anywhere in the world, which may make top-down enforcement of standards extremely difficult. In the absence of oversight by or accountability to some outside authority, ethical codes can create the appearance of responsibility without having much impact on behavior (Whittaker et al., 2018, p. 29-32). This concern may motivate some of the more draconian proposals for global surveillance regimes to prevent the development of dangerous AI (e.g., Goertzel, 2012; Bostrom, 2019). On the other hand, Goertzel (2017) proposes that open-source AI development might be more attuned to the public interest than corporate or government AI development, since the government and corporations could act in a corrupt and self-interested way. Whether open-source AI development would in fact be more attuned to the public interest is an open question. Regardless, the difficulty of governing open-source software development via top-down regulations makes bottom-up community solutions more compelling.

Community self-governance is not a silver-bullet solution. The empirical social science literature on collective action identifies a range of circumstances in which community self-organizing is more likely to be successful, such as when communities are geographically bounded, when there are at most a few thousand individuals or groups involved, when it is clear to actors how their choices directly affect their collective interest, when the benefits of collective action mostly accrue to the population whose actions determine outcomes, when the actors share a common culture and institutions, and when there are opportunities for actors to learn from experience.[19] Unfortunately, AI collective action situations do not all meet all these conditions. Indeed, no AI collective action situation may meet some of these conditions (e.g. being geographically bounded). This does not mean bottom-up community solutions will necessarily fail for AI, but it does suggest creative approaches may be needed to overcome these challenges. Additionally, some have argued that the high stakes of long-term AI development merit government (and especially international government) solutions (Hughes, 2007; Bostrom, 2019). However, arguably what matters here is not the size of the stakes but the efficacy of the solution. Government intervention is not a silver-bullet solution either (Section 4.1).

In general, there are reasons to think that no single solution or approach can completely solve the collective action problems of AI development or ensure that AI will be safe and beneficial. Global collective action may require a polycentric system of

[19] This list is from Stern (2011, p.215), discussing Ostrom (1990). We recommend Stern (2011) as perhaps the only discussion of the empirical collective action literature in the context of governing risky technologies.

governments, market, and community organizations that address AI issues in different ways and at different scales (Dietz et al., 2003). There may be no way to guarantee AI developers will produce beneficial designs, but, as Baum (2017b) writes, "given the stakes involved in AI, all effective measures for promoting beneficial AI should be pursued" (p. 551).

4.4 Transparency

Transparency is not a solution to collective action problems *per se*, but rather a governance mechanism that can affect the form and extent of AI collective action.

Some general arguments in favor of transparency about AI development have been advanced. It has been proposed that transparency could encourage goodwill and collaboration among AI developers (Bostrom, 2017), foster trust between AI developers and potential AI users (Askell et al., 2019), improve cooperation between AI developers and government regulators (*ibid.*) and even help avoid unreasonable attempts to regulate AI research (Afanasjeva et al., 2017). In practice, these arguments might not necessarily hold. For example, transparency could reduce goodwill, trust, and cooperation if AI developers are seen to be behaving poorly or acting in bad faith. Transparency could also give unscrupulous or incompetent regulators more opportunity to impose counterproductive or unnecessary regulations. However, transparency could still be beneficial on balance if it creates opportunities to address genuinely bad behavior and incentivizes AI developers and other stakeholders to behave well in the first place.

A more contentious matter is whether AI developers should be open about the capabilities of their AI, including by reporting the latest results and openly publishing code. One concern is that sharing new algorithms and capabilities could increase the tools available to malevolent actors (Brundage et al., 2018), although it could also increase the tools available to counter malevolent actors. Another concern is that transparency about AI capabilities could make developers aware of one another's progress, which could prompt them to take shortcuts on safety in order to try to win a perceived AI race (Armstrong et al., 2016; Bostrom, 2017). Again, the converse could be true: if transparency reveals that AI developers are not making substantial progress, then they could focus more on safety and feel less pressure to engage in a race.[20] Finally, transparency could level the playing field for AI developers by enabling them to build on one another's work. This could increase risks by making it harder for a careful and benevolent developer group to dominate the process (Bostrom, 2017), but it could also decrease risks by making it harder for a reckless and malevolent group to dominate. Overall, we concur with Bostrom (2017) that the case for openness about AI capabilities is complicated and mixed.

A clearer case can be made in favor of transparency on AI safety issues. Transparency would create opportunities for outside experts to contribute to any AI project's safety measures, thereby reducing the risks created by the project (Bostrom, 2017). Additionally, transparency could create opportunities for outside observers to check for bugs and other problems with an AI group's work (Askell et al., 2019), although it would also give malevolent actors an opportunity to look for vulnerabilities

[20] For comparison, during the initial race to build nuclear weapons, the US overestimated German progress and may have consequently paid less attention to safety issues in order to be the first to develop nuclear weapons (Groves, 1975).

they could exploit. An important challenge is how to ensure that the beneficial aspects of transparency outweigh the potential harms.

The AI transparency issue was at the center of a recent debate about OpenAI's decision in 2019 to release its Generative Pre-Trained Transformer 2 (GPT-2) language model in stages out of concern that it could be used for malicious purposes (Radford et al., 2019). While some applauded this decision (Mak, 2019), others criticized it for undermining open source norms and denying outside groups the opportunity to mitigate problematic aspects of the code (Zhang, 2019). This incident demonstrates the controversies that can accompany actions with respect to AI transparency.

5. Conclusion

Ensuring that AI generally contributes to good outcomes for society will require collective action. The development and use of AI involve a variety of particular situations in which collective action is required to achieve good outcomes. These situations include AI races, determination of AI development and use standards, and decisions about investment in public goods like basic AI research, AI safety, and AI ethics. A background in collective action can be valuable for understanding these situations and improving AI outcomes. Although AI collective action is a relatively new field of study, it has already produced a range of insights. The primer and review presented in this paper introduces collective action concepts, relates them to issues in AI, and summarizes the existing literature so that readers from a variety of backgrounds can get up to speed on this important topic.

Because this paper is a nonsystematic review, it cannot draw definitive conclusions about the existing literature on AI collective action. However, the research presented in the paper did involve a variety of searches to identify relevant literature. Furthermore, the searches did not identify a large body of literature. Unless the searches failed to identify a significant additional body of literature on AI collective action, which we believe is unlikely, then the trends in the literature identified this paper are indeed reflective of the trends in the actual body of literature on AI collective action. Whether this is in fact the case could be assessed in future research that conducts a systematic review. Any such review would need to account for the significant body of literature that is published outside of traditional academic outlets."

One clear limitation of the AI collective action literature reviewed in this paper is that it makes relatively little use of the insights of the rich social science literature on collective action in other contexts. Human society has quite a lot of experience with collective action in other contexts, and scholars of it have learned a great deal that is relevant to AI collective action. We did not find any studies drawing on the empirical study of AI collective action situations, though we are aware of one study drawing on the empirical literature for a more general discussion of risky emerging technologies (Stern, 2011). Empirical studies of the effectiveness of collective action with respect to different aspects of AI development and use are a promising path future research could take. In particular, it would be valuable to study how institutional design can shape the outcomes of collective action situations.

As Section 3 documents, there has also been relatively little research on competition between private AI groups. Government competition (and especially military competition) has gotten much more attention. While government AI competition is

clearly important, private AI competition is too. Indeed, for now at least, the private sector is the main driver of AI R&D. More detailed studies of private sector AI competition are another promising path for future research.

Although the study of AI collective action is in its infancy, the subject is increasingly pressing. The trajectory of AI development is uncertain, but R&D conducted today may have broad impacts on society. Governments and communities are beginning to formulate policies and institutions that if enacted could be long-lasting. Without further research into how to work together to ensure that AI development leads to collectively more optimal outcomes, society may stumble blindly into outcomes that are collectively worse.

Acknowledgements
Marisa Jurczyk, Jia Yuan Loke, Matthijs Maas, Steven Umbrello, Jun Hong Yap, Nell Watson, Rosie Campbell, and two anonymous reviewers provided helpful feedback on earlier versions of this paper. McKenna Fitzgerald assisted with manuscript preparation. Any remaining errors are the authors' alone.

References

Afanasjeva, O., Feyereisl, J., Havrda, M., Holec, M., Ó hÉigeartaigh, S., & Poliak, M. (2017, September 20). Avoiding the precipice: Race avoidance in the development of artificial general intelligence. *Medium*. https://medium.com/ai-roadmap-institute/avoiding-the-precipice-db720a805190

AI Impacts. (2016, August 8). Friendly AI as a global public good. *AI Impacts.* https://aiimpacts.org/friendly-ai-as-a-global-public-good/

Aldana, E. L. (2017). *A Theory of International AI Coordination: Strategic Implications of Perceived Benefits, Harms, Capacities, and Distribution in AI Development*. Yale University.

Altmann, J., & Sauer, F. (2017). Autonomous weapon systems and strategic stability. *Survival*, *59*(5), 117–142. https://doi.org/10.1080/00396338.2017.1375263

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI & Society*, *31*(2), 201–206. https://doi.org/10.1007/s00146-015-0590-y

Asaro, P. M. (2007). Robots and responsibility from a legal perspective. *Proceedings of the IEEE Conference on Robotics and Information, Workshop on Roboethics*, *4*, 20–24.

Askell, A., Brundage, M., & Hadfield, G. (2019). The role of cooperation in responsible AI development. *ArXiv:1907.04534 [Cs]*. http://arxiv.org/abs/1907.04534

Athimoolam, K. (2018). Solving the artificial intelligence race: Mitigating the problems associated with the AI race. http://mirror.goodai.com/judges/Solving_the_AI_race_KA.pdf

Barrat, J. (2013). *Our Final Invention: Artificial Intelligence and the End of the Human Era* (First Edition). Thomas Dunne Books.

Barrett, S. (2007). *Why Cooperate? The Incentive to Supply Global Public Goods*. Oxford University Press.

Baum, S. D. (2017a). A survey of artificial general intelligence projects for ethics, risk, and policy. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3070741

Baum, S. D. (2017b). On the promotion of safe and socially beneficial artificial intelligence. *AI & Society*, *32*(4), 543–551. https://doi.org/10.1007/s00146-016-0677-0

Baum, S. D. (2018a). Reconciliation between factions focused on near-term and long-term artificial intelligence. *AI & Society*, *33*(4), 565–572. https://doi.org/10.1007/s00146-017-0734-3

Baum, S. D. (2018b). Superintelligence skepticism as a political tool. *Information*, *9*(9), 209. https://doi.org/10.3390/info9090209

Baum, S. D. (2018c). Countering superintelligence misinformation. *Information*, *9*(10), 244. https://doi.org/10.3390/info9100244

Baum, S. D. (2020). Medium-term artificial intelligence and society. *Information*, *11*(6), 290. https://doi.org/10.3390/info11060290

Bohannon, J. (2015). Fears of an AI pioneer. *Science*, *349*(6245), 252–252. https://doi.org/10.1126/science.349.6245.252

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

Bostrom, N. (2017). Strategic implications of openness in AI development. *Global Policy*, *8*(2), 135–148. https://doi.org/10.1111/1758-5899.12403

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, *10*(4), 455–476. https://doi.org/10.1111/1758-5899.12718

Boyd, R., Richerson, P. J., Meinzen-Dick, R., De Moor, T., Jackson, M. O., Gjerde, K. M., Harden-Davies, H., Frischmann, B. M., Madison, M. J., Strandburg, K. J., McLean, A. R., & Dye, C. (2018). Tragedy revisited. *Science*, *362*(6420), 1236–1241. https://doi.org/10.1126/science.aaw0911

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., … Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *ArXiv:1802.07228 [Cs]*. http://arxiv.org/abs/1802.07228

Byrne, M. (2015, January 3). The slow race to solve chess. *Vice*. https://www.vice.com/en_us/article/gvy8wq/the-race-to-solve-chess

Byrnes, M. W. (2014). Nightfall: Machine autonomy in air-to-air combat. *Air & Space Power Journal*, *28*(3), 48–75.

Castel, J.-G., & Castel, Matthew E. (2016). The road to artificial super-intelligence: Has international law a role to play? *Canadian Journal of Law and Technology*, *14*(1).

Castro, D. (2019, March 5). The U.S. may lose the AI race because of an unchecked techno-panic. Center for Data Innovation. https://www.datainnovation.org/2019/03/the-u-s-may-lose-the-ai-race-because-of-an-unchecked-techno-panic/

Cave, S., & Ó hÉigeartaigh, S. S. (2018). An AI race for strategic advantage: Rhetoric and risks. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18*, 36–40. https://doi.org/10.1145/3278721.3278780

Cave, S., & Ó hÉigeartaigh, S. S. (2019). Bridging near- and long-term concerns about AI. *Nature Machine Intelligence*, *1*(1), 5–6. https://doi.org/10.1038/s42256-018-0003-2

Chalmers, D. J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, *17*(9–10), 7–65.
Coeckelbergh, M. (2018). Technology and the good society: A polemical essay on social ontology, political principles, and responsibility for technology. *Technology in Society 52*, 4-9. https://doi.org/10.1016/j.techsoc.2016.12.002
Dafoe, A. (2019). *AI Governance: A Research Agenda* (pp. 1–53). Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf
Daley, W. (2011). Mitigating potential hazards to humans from the development of intelligent machines. *Synesis: A Journal of Science, Technology, Ethics and Policy*, *2*(1), 44–50.
Danzig, R. (2018). *Managing Loss of Control as Many Militaries Pursue Technological Superiority* (Technology & National Security, p. 40). Center for New American Security. https://www.cnas.org/publications/reports/technology-roulette
Dewey, D. (2016). Long-term strategies for ending existential risk from fast takeoff. In *Risks of Artificial Intelligence* (1st ed., pp. 242–266). CRC PRESS.
Dietz, T. (2003). The struggle to govern the commons. *Science*, *302*(5652), 1907–1912. https://doi.org/10.1126/science.1091015
Erdelyi, O. J., & Goldsmith, J. (2018). Regulating artificial intelligence proposal for a global solution. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 95–101.
Funkhouser, K. (2013). Paving the road ahead: Autonomous vehicles, products liability, and the need for a new approach. *Utah Law Review*, *2013*(1), 437–462.
Gasser, U., & Schmitt, C. (2019). The role of professional norms in the governance of artificial intelligence. In *The Oxford Handbook of Ethics of AI*. Oxford University Press. https://doi.org/10.2139/ssrn.3378267
Geist, E. M. (2016, September 2). It's already too late to stop the AI arms race—We must manage it instead. *Bulletin of the Atomic Scientists*. https://thebulletin.org/2016/09/its-already-too-late-to-stop-the-ai-arms-race-we-must-manage-it-instead/
Goertzel, B. (2012). Should humanity build a global AI nanny to delay the singularity until it's better. *Journal of Consciousness Studies*, *19*(1–2), 96–11.
Goertzel, B., & Lawrence, C. (2017, July 21). The corporatization of AI is a major threat to humanity. *H+ Media*. https://hplusmagazine.com/2017/07/21/corporatization-ai-major-threat-humanity/
Goertzel, B., & Pitt, J. (2012). Nine ways to bias open-source AGI toward friendliness. *Journal of Evolution and Technology*, *22*(1), 116–131.
Good, I. J. (1966). Speculations concerning the first ultraintelligent machine. *Advances in Computers*, *6*, 31–88. https://doi.org/10.1016/S0065-2458(08)60418-0
Grace, K. (2015). *The Asilomar Conference: A case study in risk mitigation* (No. 2015–9; p. 68). Machine Intelligence Research Institute. https://intelligence.org/files/TheAsilomarConference.pdf
Griffy-Brown, C., Earp, B. D., & Rosas, O. (2018). Technology and the good society. *Technology in Society 52*, 1-3.
Groves, L. R. (1975). *Now It Can Be Told: The Story of the Manhattan Project*. Da Capo Press.

Gurkaynak, G., Yilmaz, I., & Haksever, G. (2016). Stifling artificial intelligence: Human perils. *Computer Law & Security Review*, *32*(5), 749–758. https://doi.org/10.1016/j.clsr.2016.05.003

Gurney, J. K. (2013). Sue my car not me: Products liability and accidents involving autonomous vehicles. *Journal of Law, Technology & Policy*, *2013*(2), 247–277.

Han, T. A., Pereira, L. M., Santos, F. C., & Lenaerts, T. (2020). To regulate or not: A social dynamics analysis of the race for AI supremacy. *ArXiv:1907.12393 [Physics]*. http://arxiv.org/abs/1907.12393

Hardin, Garrett. (1968). The tragedy of the commons. *Science*, *162*(3859), 1243–1248. https://doi.org/10.1126/science.162.3859.1243

Hirshleifer, J. (1983). From weakest-link to best-shot: The voluntary provision of public goods. *Public Choice*, *41*, 371–386. https://doi.org/10.1007/BF00141070

Hughes, J. (2007). Global technology regulation and potentially apocalyptic technological threats. In *Nanoethics: The Ethical and Social Implications of Nanotechnology* (pp. 201–214). John Wiley & Sons.

Hwang, T. (2018). Computational power and the social impact of artificial intelligence. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3147971

Karnow, C. E. A. (1996). Liability for distributed artificial intelligences. *Berkeley Tech Law Journal*, *11*(1), 147–204.

Knight, W. (2018, March 19). What Uber's fatal accident could mean for the autonomous-car industry. *MIT Technology Review*. https://www.technologyreview.com/2018/03/19/241022/what-ubers-fatal-accident-could-mean-for-the-autonomous-car-industry/

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. Viking.

LeValley, D. (2013). Autonomous vehicle liability—Application of common carrier liability. *Seattle University Law Review Supra*, *36*(5). https://digitalcommons.law.seattleu.edu/sulr_supra/5

Mak, A. (2019, February 22). When Is Technology Too Dangerous to Release to the Public? *Slate Magazine*. https://slate.com/technology/2019/02/openai-gpt2-text-generating-algorithm-ai-dangerous.html

Marchant, G. (2019, January 25). "Soft Law" Governance of Artificial Intelligence. *AI Pulse*. https://aipulse.org/soft-law-governance-of-artificial-intelligence/

Marchant, G., & Lindor, R. (2012). The Coming Collision Between Autonomous Vehicles and the Liability System. *Santa Clara Law Review*, *52*(4), 1321–1340.

Marcus, G. (2017, July 29). Artificial intelligence is stuck. Here's how to move it forward. *The New York Times*. https://www.nytimes.com/2017/07/29/opinion/sunday/artificial-intelligence-is-stuck-heres-how-to-move-it-forward.html

McGinnis, J. O. (2010). Accelerating AI. *Northwestern University Law Review Colloquy*, *104*(366), 366–381.

Metz, C., & Isaac, M. (2019, May 17). Facebook's A.I. whiz now faces the task of cleaning it up. Sometimes that brings him to tears. *The New York Times*. https://www.nytimes.com/2019/05/17/technology/facebook-ai-schroepfer.html

Miailhe, N. (2018, December 20). AI & global governance: Why we need an intergovernmental panel for artificial intelligence. *United Nations University Centre*

for Policy Research. https://cpr.unu.edu/ai-global-governance-why-we-need-an-intergovernmental-panel-for-artificial-intelligence.html
Miller, J. D. (2012). *Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World*. Benbella Books.
National Research Council. (2002). *The Drama of the Commons*. National Academies Press. https://doi.org/10.17226/10287
Naudé, W., & Dimitri, N. (2018). The race for an artificial general intelligence: Implications for public policy. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3235276
Nindler, R. (2019). The United Nation's capability to manage existential risks with a focus on artificial intelligence. *International Community Law Review*, *21*(1), 5–34. https://doi.org/10.1163/18719732-12341388
Olson, M. (1965). *The Logic of Collective Action*. Harvard University Press.
Omohundro, S. M. (2008). The basic AI drives. *AGI*, *171*, 483–492.
Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.
Ostrom, E., & Ostrom, V. (2017). Public economy organization and service delivery. In *Elinor Ostrom and the Bloomington School of Political Economy* (Vol. 3, A Framework for Policy Analysis). Lexington Books.
Prunkl, C., & Whittlestone, J. (2020). Beyond near- and long-term: Towards a clearer account of research priorities in AI ethics and society. *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20),* 6. http://arxiv.org/abs/2001.04335
Rabesandratana, T. (2018, March 30). Emmanuel Macron wants France to become a leader in AI and avoid 'dystopia.' *Science*. https://www.sciencemag.org/news/2018/03/emmanuel-macron-wants-france-become-leader-ai-and-avoid-dystopia
Radford, A., Wu, J., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019, February 14). Better language models and their implications. OpenAI. https://openai.com/blog/better-language-models/
Ramamoorthy, A., & Yampolskiy, R. (2018). Beyond MAD?: The race for artificial general intelligence. *ICT Discoveries*, *Special Issue 1*, 1–8.
Rickli, J.-M. (2017). The impact of autonomous weapons systems on international security and strategic stability. In *Defence Future Technologies: What We See on the Horizon* (pp. 61–64). Armasuisse Science & Technologies.
Roff, H. M. (2019). The frame problem: The AI "arms race" isn't one. *Bulletin of the Atomic Scientists*, *75*(3), 95–98. https://doi.org/10.1080/00963402.2019.1604836
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y
Russell, S. J. (2018). The new weapons of mass destruction? *The Security Times*, 40–41.
Salter, A. J. & Martin, B. R. (2001). The economic benefits of publicly funded basic research: A critical review. *Research Policy 30*(1), 509-532.
Samuelson, P. A. (1954). The pure theory of public expenditure. *The Review of Economics and Statistics*, *36*(4), 387. https://doi.org/10.2307/1925895

Sandler, T. (2015). Collective action: Fifty years later. *Public Choice*, *164*(3–4), 195–216. https://doi.org/10.1007/s11127-015-0252-0

Scharre, P. (2018). *Army of None: Autonomous Weapons and the Future of War*. W. W. Norton & Company.

Schelling, T. C. (2006). An astonishing 60 years: The legacy of Hiroshima. *Proceedings of the National Academy of Sciences*, *103*(16), 6089–6093. https://doi.org/10.1073/pnas.0600437103

Scherer, M. U. (2015). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2609777

Shneiderman, B. (2016). Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proceedings of the National Academy of Sciences*, *113*(48), 13538–13540. https://doi.org/10.1073/pnas.1618211113

Shulman, C. (2009). Arms control and intelligence explosions. Paper presented at the 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July 2-4.

Simonite, T. (2018, December 6). Canada, France plan global panel to study the effects of AI. *Wired*. https://www.wired.com/story/canada-france-plan-global-panel-study-ai/

Slusallek, P. (2018, January 8). Artificial intelligence and digital reality: Do we need a CERN for AI? The OECD Forum Network. https://www.oecd-forum.org/posts/28452-artificial-intelligence-and-digital-reality-do-we-need-a-cern-for-ai

Stern, P. C. (2011). Design principles for global commons: Natural resources and emerging technologies. *International Journal of the Commons*, *5*(2), 213. https://doi.org/10.18352/ijc.305

Tan, J. Z., & Ding, J. (2019). AI governance through AI markets. http://www.joshuatan.com/wp-content/uploads/2019/08/AI_governance_through_AI_markets.pdf

Tannenwald, N. (2005). Stigmatizing the bomb: Origins of the nuclear taboo. *International Security*, *29*(4), 5–49. https://doi.org/10.1162/isec.2005.29.4.5

The Economist. (2015, May 9). Rise of the machines. *The Economist*. https://www.economist.com/briefing/2015/05/09/rise-of-the-machines

Tomasik, B. (2013). International cooperation vs. AI arms race. Foundational Research Institute, Center on Long-Term Risk. https://longtermrisk.org/files/international-cooperation-ai-arms-race.pdf

U.S. Department of State. (2020, June 15). Joint statement from founding members of the Global Partnership on Artificial Intelligence. *United States Department of State*. https://www.state.gov/joint-statement-from-founding-members-of-the-global-partnership-on-artificial-intelligence/

Vinge, V. (1993, March 30). The coming technological singularity: How to survive in the post-human Era. VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute. https://ntrs.nasa.gov/citations/19940022856

Wallach, W. (2017). Toward a ban on lethal autonomous weapons: Surmounting the obstacles. *Communications of the ACM*, *60*(5), 28–34. https://doi.org/10.1145/2998579

White, T. N., & Baum, S. D. (2017). Liability for present and future robotics technology. In *Robot Ethics 2.0* (Vol. 1, pp. 66–79). Oxford University Press. https://doi.org/10.1093/oso/9780190652951.003.0005

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. (December 2018). *AI Now Report 2018* (p. 62). AI Now Institute. https://ainowinstitute.org/AI_Now_2018_Report.pdf

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 195–200. https://doi.org/10.1145/3306618.3314289

Wu, S. S. (2015). Product liability issues in the U.S. and associated risk management. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomes Fahren* (pp. 575–592). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-45854-9_26

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In *Global Catastrophic Risk* (pp. 308–345). Oxford University Press. https://intelligence.org/files/AIPosNegFactor.pdf

Zeng, Y., Lu, E., & Huangfu, C. (2018). Linking artificial intelligence principles. *ArXiv:1812.04814 [Cs]*. http://arxiv.org/abs/1812.04814

Zhang, H. (2019, February 29). Dear OpenAI: Please open source your language model. *The Gradient*. https://thegradient.pub/openai-please-open-source-your-language-model/

Zohn, J. R. (2015). When robots attack: How should the law handle self-driving cars that cause damages. *University of Illinois Journal of Law, Technology & Policy*, *2015*, 461–485.