Article

Medium-Term Artificial Intelligence and Society

Seth D. Baum (ID)

Global Catastrophic Risk Institute, P.O. Box 40364, Washington, DC 20016, USA; seth@gcrinstitute.org

check for
updates

Abstract: There has been extensive attention to near-term and long-term AI technology and its accompanying societal issues, but the medium-term has gone largely overlooked. This paper develops the concept of medium-term AI, evaluates its importance, and analyzes some medium-term societal issues. Medium-term AI can be important in its own right and as a topic that can bridge the sometimes acrimonious divide between those who favor attention to near-term AI and those who prefer the long-term. The paper proposes the medium-term AI hypothesis: the medium-term is important from the perspectives of those who favor attention to near-term AI as well as those who favor attention to long-term AI. The paper analyzes medium-term AI in terms of governance institutions, collective action, corporate AI development, and military/national security communities. Across portions of these four areas, some support for the medium-term AI hypothesis is found, though in some cases the matter is unclear.

Keywords: near-term AI; long-term AI; medium-term AI; intermediate-term AI; mid-term AI; societal implications of AI

1. Introduction

Attention to AI technologies and accompanying societal issues commonly clusters into groups focusing on either near-term or long-term AI, with some acrimonious debate between them over which is more important. Following Baum [1], the near-term camp may be called "presentists" and the long-term camp "futurists".

The current state of affairs suggests two reasons for considering the intermediate period between the near and long terms. First, the medium term (or, interchangeably, intermediate term or mid term) has gone neglected relative to its inherent importance. If there are important topics involving near-term and long-term AI, then perhaps the medium term has important topics as well. Second, the medium term may provide a common ground between presentists and futurists. Insofar as both sides consider the medium term to be important, it could offer a constructive topic to channel energy that may otherwise be spent on hashing out disagreements.

Rare examples of previous studies with dedicated attention to medium-term AI are Parson et al. [2,3]. (There is a lot of work that touches on medium-term AI topics, some of which is cited in this paper. However, aside from Parson et al. [2,3], I am not aware of any publications that explicitly identify medium-term AI as a topic warranting dedicated attention.) Both studies [2,3] recognize medium-term AI as important and neglected. Parson et al. [2] acknowledges that some prior work in AI covers topics that are important across all time periods, and thus are also relevant to the medium term. It provides a definition of medium-term AI, which is discussed further below, and it provides some analysis of medium-term AI topics. Parson et al. [3] posits that the neglect of the medium term may derive in part from the academic disciplines and methodologies of AI researchers, which may point the researchers toward either the near term or the long term but not the medium term. The present paper extends Parson et al.'s [2] work on definitions and presents original analysis of a different mix of medium-term

AI topics. The present paper also explores the medium term as a potential point of common ground between presentists and futurists.

Several previous attempts have been made to bridge the presentist–futurist divide [1,4,5]. An overarching theme in this literature is that the practical steps needed to make progress are often (though not always) the same for both near-term and long-term AI. Instead of expending energy debating the relative importance of near-term and long-term AI, it may often be more productive to focus attention on the practical steps that both sides of the debate agree are valuable. This practical synergy can arise for two distinct reasons, both with implications for medium-term AI.

First, certain actions may improve near-term AI and the near-term conversation about long-term AI. Such actions will often also improve the near-term conversation about mid-term AI. For example, efforts to facilitate dialog between computer scientists and policymakers can improve the quality of policy discussions for near-, mid-, and long-term AI. Additionally, efforts encouraging AI developers to take more responsibility for the social and ethical implications of their work can influence work on near-, mid-, and long-term AI. For example, the ethics principles that many AI groups have recently established [6] are often quite general and can apply to work on near-term and long-term AI, as can analyses of the limitations of these principles [7]. Here it should be explained that there is near-term work aimed at developing systems that may only become operational over the mid or long term, especially work consisting of basic research toward major breakthroughs in AI capabilities.

Second, certain actions may improve near-term AI, and, eventually, long-term AI. These actions may often also eventually improve mid-term AI. For example, some research on how to design near-term AI systems more safely may provide a foundation for also making mid- and long-term AI systems safer. This is seen in the AI safety study of Amodei et al. [8], which is framed in terms of near-term AI; lead author Amodei describes the work as also being relevant for long-term AI [9]. Additionally, AI governance institutions established over the near term may persist into the mid and long term, given the durability of many policy institutions. Of course, AI system designs and governance institutions that persist from the near term to the long term would also be present throughout the mid-term. Furthermore, evaluating their long-term persistence may require understanding of what happens during the mid-term.

Dedicated attention to the medium term can offer another point of common ground between presentists and futurists: both sides may consider the medium term to be important. Presentists may find the medium term to be early enough for their tastes, while futurists find it late enough for theirs. As elaborated below, the reasons that presentists have for favoring near-term AI are different types of reasons than those of the futurists. Presentists tend to emphasize immediate feasibility, certainty, and urgency, whereas futurists tend to emphasize extreme AI capabilities and consequences. Potentially, the medium term features a widely appealing mix of feasibility, certainty, urgency, capabilities, and consequences. Or not: it is also possible that the medium term would sit in a "dead zone", being too opaque to merit presentist interest and too insignificant to merit futurist interest. This matter will be a running theme throughout the paper and is worth expressing formally:

> *The medium-term AI hypothesis*: There is an intermediate time period in which AI technology and accompanying societal issues are important from both presentist and futurist perspectives.

The medium-term AI hypothesis can be considered in either empirical or normative terms. As an empirical hypothesis, it proposes that presentists and futurists actually consider the medium term to be important, or that they would tend to agree that the medium term is important if given the chance to reflect on it. As a normative hypothesis, it proposes that presentists should agree that the medium term is important, given the value commitments of the presentist and futurist perspectives. Given the practical goal of bridging the presentist–futurist divide, the empirical form is ultimately more important: what matters is whether the specific people on opposite sides of the divide would, upon consideration, find common ground in the medium term. (It is unlikely that they currently do find common ground in the medium term, due to lack of attention to it.) Empirical study of presentist and

futurist reactions to the medium term is beyond the scope of the present paper. Instead, the aim here is to clarify the nature of the presentist and futurist perspectives in terms of the attributes of the medium term that they should consider important and then to examine whether the medium term is likely to possess these attributes. The paper therefore proceeds mainly in normative terms, though grounded in empirical observation of the perspectives articulated by actual presentists and futurists.

More precisely, the medium-term AI hypothesis proposes that the perspectives underlying both groups should rate the medium term as important. This presumes that "perspectives" can rate things as important even when detached from the people who hold them. Such detachment is permitted here simply so that the analysis can proceed without going through the more involved (but ultimately important) process of consulting with the people who hold presentist and futurist perspectives.

Evaluating the medium-term AI hypothesis is one aim of this paper. First, though, more needs to be said on how the medium term is defined.

2. Defining the Medium Term

The medium term is, of course, the period of time between the near term and the long term. However, discussions of near-term and long-term AI often do not precisely specify what constitutes near-term and long-term. Some ambiguity is inevitable due to uncertainty about future developments in AI. Additionally, different definitions may be appropriate for different contexts and purposes—for example, what qualifies as near-term may be different for a programmer than for a policymaker. Nonetheless, it is worth briefly exploring how the near, mid, and long terms can be defined for AI. Throughout, it should be understood that the near, mid, and long terms are all defined relative to the vantage point of the time of this writing (2019–2020). As time progresses, what classifies as near-, mid-, and long-term can shift.

The first thing to note is that near- vs. mid- vs. long-term can be defined along several dimensions. The first is chronological: the near term goes from year A to year B, the mid term from year B to year C, and the long term from year C to year D. The second is in terms of the feasibility or ambitiousness of the AI: the near term is what is already feasible, the long term is the AI that would be most difficult to achieve, and the mid term is somewhere in between. Third, and related to the second, is the degree of certainty about the AI: the near term is what clearly can be built, the long term is the most uncertain and speculative, and the mid term is somewhere in between. Fourth is the degree of sophistication or capability of the AI: the near term is the least capable, the long term is the most capable, and the mid term is somewhere in between. Fifth, and related to the fourth, is with respect to impacts: the near term has (arguably; see below) the mildest impacts on human society and the world at large, the long term has the most extreme impacts, and the mid-term is somewhere in between. Sixth is urgency: the near term is (arguably) the most urgent, the long term the least urgent, and the mid term is somewhere in between.

The dimension of impacts is somewhat complex and worth briefly unpacking. Near-term AI may have the mildest impacts, in the sense that if AI continues to grow more capable and be used more widely and in more consequential settings it will tend to have greater impacts on the human society that exists at that time. Put differently, if A = the impacts of near-term AI on near-term society, B = the impacts of mid-term AI on mid-term society, and C = the impacts of long-term AI on long-term society, then (it is supposed) A < B < C. There are, however, alternative ways of conceptualizing impacts. One could take a certain presentist view and argue that only present people matter for purposes of moral evaluation, such as is discussed by Arrhenius [10], or that future impacts should be discounted, as in many economic cost–benefit evaluations. In these cases, near-term AI may be evaluated as having the largest impacts because the impacts of mid- and long-term AI matter less or not at all. Or, one could consider the impacts of a period of AI on all time periods: the impact of near-term AI on the near, mid, and long terms, the impacts of mid-term AI on the mid- and long-terms, and the impact of long-term AI on the long term. This perspective recognizes the potential for durable impacts of AI technology, and would tend to increase the evaluated size of the impacts of near- and mid-term

AI. While recognizing the merits of these alternative conceptions of impacts, this paper uses the first conception, involving A vs. B vs. C.

There may be no one correct choice of dimensions for defining the near/mid/long term. Different circumstances may entail different definitions. For example, Parson et al. [2] are especially interested in societal impacts and implications for governance, and thus use definitions rooted primarily in impacts. They propose that, relative to near-term AI, medium-term AI has "greater scale of application, along with associated changes in scope, complexity, and integration" [2] (pp. 8–9), and, relative to long-term AI, medium-term AI "is not self-directed or independently volitional, but rather is still to a substantial degree developed and deployed under human control" [2] (p. 9). (One can quibble with these definitions. Arguably, near-term AI is already at a large scale of application, and there may be no clear demarcation in scale between near- and mid-term AI. Additionally, while it is proposed that long-term AI could escape human control, that would not necessarily be the case. Indeed, discussions of long-term AI sometimes focus specifically on the question of how to control such an AI [11].) The medium term is a period with substantially greater use of AI in decision-making, potentially to the point in which "the meaning of governance" is challenged [2] (p. 9), but humans remain ultimately in control. This is a reasonable definition of medium-term AI, especially for impacts and governance purposes.

The present paper is more focused on the presentist/futurist debate, and so it is worth considering the definitions used in the debate. Elements of each of the six dimensions can be found, but they are not found uniformly. Presentists often emphasize feasibility and degree of certainty. Computer scientist Andrew Ng memorably likened attention to long-term AI to worrying about "overpopulation on Mars" [12], by which Ng meant that it might eventually be important, but it is too opaque and disconnected from current AI to be worth current attention. Another presentist theme is urgency, especially with respect to the societal implications of near-term AI. Legal scholar Ryan Calo [13] (p. 27) argues that "AI presents numerous pressing challenges to individuals and society in the very short term" and therefore commands attention relative to long-term AI. For their part, futurists often emphasize capability and impacts. Commonly cited is the early remark of I.J. Good [14] (p. 33) that "ultraintelligent" AI (AI with intelligence significantly exceeding that of humans) could be "the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control". Chronological definitions are less common. One exception is Etzioni [15], who downplays long-term AI on grounds that it is unlikely to occur within 25 years. (In reply, futurists Dafoe and Russell [16] argue that potential future events can still be worth caring about even if they will not occur within the next 25 years.)

Taking the above into account, this paper will use a feasibility definition for near-term AI and a capability definition for long-term AI. The paper defines near-term AI as *AI that already exists or is actively under development with a clear path to being built and deployed.* Per this definition, near-term AI does not require any major research breakthroughs, but instead consists of straightforward applications of existing techniques. The terms "clear", "major", and "straightforward" are vague, and it may be reasonable to define them in different ways in different contexts. (This vagueness is relevant for the medium-term AI hypothesis; more on this below.) Nonetheless, this definition points to current AI systems plus the potential future AI systems that are likely to be built soon and do not depend on research breakthroughs that might or might not manifest.
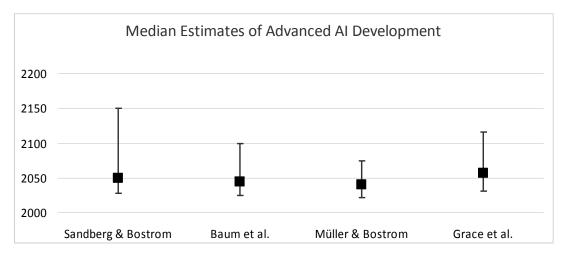
The paper defines long-term AI as *AI that has at least human-level general intelligence.* Interest in long-term AI often focuses on human-level artificial intelligence (HLAI), artificial general intelligence (AGI), strong AI, and artificial superintelligence (ASI). However, there may be narrow AI systems that are appropriate to classify as long-term. For example, Cave and ÓhÉigeartaigh [4] (p. 5) include "wide-scale loss of jobs" as a long-term AI issue separately from the prospect of superintelligence. (Note that the most widespread loss of jobs may require AGI. For example, Ford [17] (p. 3) writes "If, someday, machines can match or even exceed the ability of a human being to think and to conceive new ideas—while at the same time enjoying all the advantages of a computer in areas like computational

speed and data access—then it becomes somewhat difficult to imagine just what jobs might be left for even the most capable human workers".) A plausible alternative definition of long-term AI is *AI that achieves major intellectual milestones and/or has large and transformative effects*. This is more of a catch-all definition that could include sufficiently important narrow AI systems such as those involved in job loss. In this definition, the terms "major", "large", and "transformative" are vague. Indeed, current AI systems arguably meet this definition. Therefore, the paper will define long-term AI in terms of HLAI, while noting the case for the alternative definitions.

The paper's use of a feasibility definition for near-term and a capability definition for long-term may be consistent with common usage in AI discussions. However, the use of a different dimension for near-term (feasibility) than for long-term (capability) can induce some chronological blurring in two important respects.

First, AI projects that are immediately practical may have long time horizons. This may be especially common for projects in which AI is only one component of a more complex and durable system. Military systems are one domain with long lifespans. A 2016 report found that some US nuclear weapon systems were still using 1970s-era 8-inch floppy disks [18]. AI is currently being used and developed for a wide variety of military systems [19]. Some of these could conceivably persist for many decades into the future—perhaps in the B-52H bomber, which was built in the 1960s and is planned to remain in service through the 2050s [20]. (AI is used in bombers, for example, to improve targeting [21]. AI is used more extensively in fighters, which execute complex aerial maneuvers at rapid speeds and can gain substantial tactical advantage from increased computational power and autonomy from human pilots [22].) One can imagine the B-52H being outfitted with current AI algorithms and retaining these algorithms into the 2050s, just as the 8-inch floppy disks have been retained in other US military systems. Per this paper's definitions, this B-52H AI would classify as near-term AI that happens to remain in use over a long time period, well beyond the 25 years that Etzioni [15] treats as the "foreseeable horizon" worthy of attention.

Second, AI systems with large and transformative effects, including AGI, could potentially be built over relatively short time scales. When AGI and related forms of AI will be built is a matter of considerable uncertainty and disagreement. Several studies have asked AI researchers—predominantly computer scientists—when they expect AI with human or superhuman capacity to be built [23–26]. (Note that these studies are generally framed as being surveys of experts, but it is not clear that the survey participants are expert in the question of when AGI will be built. Earlier predictions about AI have often been unreliable [27]. This may be a topic for which there are no experts; on this issue, see Morgan [28].) The researchers present estimates spanning many decades, with some estimates being quite soon. Figure 1 presents median estimates from these studies. Median estimates conceal the range of estimates across survey participants, but the full range could not readily be presented in Figure 1 because, unfortunately, only Baum et al. [23] included the full survey data. If the early estimates shown in Figure 1 are correct, then, by this paper's definitions, long-term AI may be appearing fairly soon, potentially within the next 25 years.

Figure 1. Estimates for when AI will reach superhuman capability (Baum et al.) [23] and human-level capability (Sandberg and Bostrom, Müller and Bostrom, and Grace et al.) [24–26]. Shown are estimates for when the probability that the milestone is reached is 10% (lower mark), 50% (square), and 90% (upper mark). For each study, the median estimates across the survey participants are plotted.

3. The Medium-Term AI Hypothesis

With the above definitions in mind, it is worth revisiting the medium-term AI hypothesis. If presentists are, by definition, only interested in the present, then they would not care at all about the medium term. However, the line between the near term and the medium term is blurry. As defined above, near-term AI must have a clear path to being built and deployed, but "clearness" is a matter of degree. As the path to being built and deployed becomes less and less clear, the AI transitions from near-term to medium-term, and presentists may have less and less interest in it. From this standpoint, presentists may care somewhat about the medium term, especially the earlier portions of it, but not to the same extent as they care about the near term.

Alternatively, presentists might care about the medium term because the underlying things they care about also arise in the medium term. Some presentists are interested in the implications of AI for social justice, or for armed conflict, or for transportation, and so on. Whereas it may be difficult to think coherently about the implications of long-term AI for these matters, it may not be so difficult for medium-term AI. For example, a major factor in debates about autonomous weapons (machines that use AI to select and fire upon targets) is whether these weapons could adequately discriminate between acceptable and unacceptable targets (e.g., enemy combatants vs. civilians) [29,30]. Near-term AI cannot adequately discriminate; medium-term AI might be able to. Therefore, presentists concerned about autonomous weapons have reason to be interested in medium-term AI. Whether this interest extends to other presentist concerns (social justice, transportation, etc.) must be considered on a case-by-case basis.

For futurists, the medium term may be important because it precedes and influences the long term. If the long term begins with the advent of human-level AGI, then this AI will be designed and built during the medium term. Some work on AGI is already in progress [31], but it may be at a relatively early stage. Figure 1 illustrates the uncertainty: the earliest estimates for the onset of AGI (and similar forms of AI) may fall within the near term, whereas the latest estimates fall much, much later. Futurists may tend to be most interested in the period immediately preceding the long term because it has the most influence on AGI. Their interest in earlier periods may depend on the significance of its causal impact on AGI.
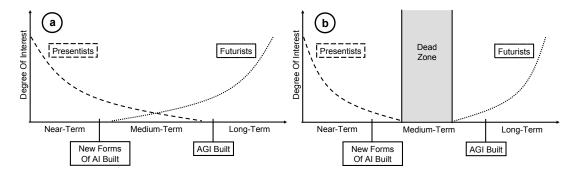
It follows that there are two bases for assessing the medium-term AI hypothesis. First, the hypothesis could hold if AI that resembles near-term AI also influences long-term AI. In that case, the technology itself may be of interest to both presentists and futurists. Alternatively, the hypothesis could hold if the societal implications of medium-term AI raise similar issues as near-term AI, and if the medium-term

societal context also influences long-term AI. For example, medium-term autonomous weapon technology could raise similar target discrimination issues as is found for near-term technology, and it could also feed arms races for long-term AI. (To avoid confusion, it should be understood that discussions of long-term AI sometimes use the term "arms race" to refer to general competition to be the first to build long-term AI, without necessarily any connection to military armaments [32]. Nonetheless, military arms races for long-term AI are sometimes posited [33].)

Both of the above derive from some measure of continuity between the near, mid, and long terms. Continuity can be defined in terms of the extent of change in AI systems and related societal issues. If near-term AI techniques and societal dimensions persist to a significant extent through the end of the medium term (when long-term AI is built), then the medium-term AI hypothesis is likely to hold.

The chronological duration of the medium term may be an important factor. Figure 1 includes a wide range of estimates for the start of the long term. If the later estimates prove correct, then the medium term could be quite long. A long duration would likely tend to mean less continuity across the near, mid, and long terms, and therefore less support for the medium-term AI hypothesis. That is not necessarily the case. One can imagine, for example, that AI just needs one additional technical breakthrough to go from current capabilities to AGI, and that it will take many decades for this breakthrough to be made. One can also imagine that the issues involving AI will remain fairly constant until this breakthrough is made. In that case, near-term techniques and issues would persist deep into the medium term. However, it is more likely that a long-lasting medium term would have less continuity and a larger dead zone period with no interest from either presentists or futurists. If AGI will not be built for, say, another 500 years, presentists are unlikely to take an interest.

Figure 2 presents two sketches of the degree of interest that presentists and futurists may hold in the medium term. Figure 2a shows a period of overlap in which both presentists and futurists have some interest; here, the medium-term AI hypothesis holds. Figure 2b shows a dead zone with no overlap of interest; here, the medium-term AI hypothesis does not hold. Figure 2 is presented strictly for illustrative purposes and does not indicate any rigorously derived estimation of actual presentist or futurist interests. It serves to illustrate how presentists' degree of interest could decline over time and futurists' degree of interest could increase over time, with implications for the medium-term AI hypothesis. Figure 2 shows presentist/futurist interest decreasing/increasing approximately exponentially over time. There is no particular basis for this, and the curves could just as easily have been drawn differently.

Figure 2. Illustrative sketches of presentist and futurist interest in the near, medium, and long term. (**a**) shows overlapping interest: the medium-term AI hypothesis holds; (**b**) shows a dead zone with no overlapping interest: the medium-term AI hypothesis does not hold. The sketches are strictly for illustrative purposes only. The phrase "new forms of AI built" is defined with reference to the definition of near-term AI in the main text.

To sum up, assessing the medium-term AI hypothesis requires examining what medium-term AI techniques and societal dimensions may look like, and the extent of continuity between the near-, mid-, and long-term periods.

4. The Intrinsic Importance of Medium-Term AI

Thus far, the paper has emphasized the potential value of medium-term AI as a point of common interest between presentists and futurists. This "consensus value" will remain a major theme in the sections below. However, it is worth pausing to reiterate that medium-term AI can also be important in its own right, regardless of any implications for presentists and futurists. Assessing the extent to which it is intrinsically important requires having some metric for intrinsic importance. A detailed metric is beyond the scope of this paper. For present purposes, it suffices to consider that medium-term AI and its accompanying societal issues may be important for the world as it exists during the medium term. It is further worth positing that there may be opportunities for people today to significantly influence the medium term, such that the medium term merits attention today due to its intrinsic importance. With that in mind, the paper now turns to the details of medium-term AI and society.

5. Medium-Term AI Techniques

My own expertise is not in the computer science of AI, and so I can say relatively little about what computer science AI techniques may look like over the medium term. Therefore, this section serves as a placeholder to note that the space of potential medium-term AI techniques is a topic worthy of attention for those with the expertise to analyze and comment on it.

6. Medium-Term AI Societal Dimensions

While the medium-term societal dimensions of AI will, to at least some extent, depend on the capabilities of the medium-term AI techniques, it is nonetheless possible to paint at least a partial picture of the societal dimensions, even without clarity on the techniques. What follows is indeed a partial picture, shaped to a significant extent by my own areas of expertise. It aims to illustrate potential medium-term scenarios in several domains and discuss their implications for near-term and long-term AI and their prospects for bridging the presentist/futurist divide.

6.1. Governance Institutions

Governance institutions can be quite durable. For example, the United Nations was founded in 1945, and despite many calls for reform, the UN Security Council retains China, France, Russia, the United Kingdom, and the United States as permanent members. The "P5 countries" are an artifact of World War II that arguably does not match current international affairs, but changing the membership would require a consensus that is quite elusive. For example, a case could be made for adding Brazil and India, but then Argentina and Pakistan may object, so no change is made. Not all governance institutions are this ossified, but many of them are quite enduring. This continuity makes governance institutions a compelling candidate for the medium-term AI hypothesis.

The near-term is an exciting time for AI governance. Institutions are now in the process of being designed and launched. Decisions being made now could have long-lasting implications, potentially all the way through the end of the medium term and the beginning of the long term. (It is harder to predict much of anything if and when AGI/ASI/HLAI is built, including the form of governance institutions. One attempt to make such predictions is Hanson [34].)

One notable example is the International Panel on Artificial Intelligence (IPAI) and Global Partnership on AI (GPAI). The IPAI/GPAI has recently been proposed by the governments of Canada and France, first under the IPAI name and later under the GPAI name [35,36]. Documents on the IPAI/GPAI emphasize issues that are relevant in the near term and may continue to be relevant through the medium term. One set of issues listed for illustrative purposes is: "data collection and access; data control and privacy; trust in AI; acceptance and adoption of AI; future of work; governance, laws and justice; responsible AI and human rights; equity, responsibility and public good" [35].

The documents published on the IPAI/GPAI give no indication of any focus on long-term issues relating to AGI. (The future of work could arguably classify as a long-term issue.) However,

the IPAI/GPAI may nonetheless be relevant for the long term. If the IPAI/GPAI takes hold then it could persist for a long time. For comparison, the Intergovernmental Panel on Climate Change (IPCC) was formed in 1988 and remains an active and important institution. The IPAI/GPAI follows a similar model as the IPCC and may prove similarly durable. Additionally, while long-term issues are not featured in the early-stage documents that have thus far been published on the IPAI/GPAI, that does not preclude the IPAI/GPAI from including long-term issues within its scope once it is up and running. Whether long-term issues are included could come down to whether people interested in the long-term take the initiative to participate in IPAI/GPAI processes. Indeed, one of the most thoughtful discussions of the IPAI/GPAI published to date is by Nicolas Miailhe [37] of The Future Society, an organization explicitly working "to address holistically short, mid and long term governance challenges" in AI [38]. Such activity suggests that the IPAI/GPAI could be an institution that works across the range of time scales and persists significantly into the future.

6.2. Collective Action

An important dynamic for the societal impacts of AI is whether AI development projects can successfully cooperate on collective action problems: situations in which the collective interest across all the projects diverges from the individual interests of the projects. Collective action has been a significant theme in discussions of long-term AI, focused on the prospect of projects cutting corners on safety to be the first to achieve important technological milestones [32,39]. Collective action problems can also arise for near-term AI. One near-term concern is about military AI arms races [40] (though this concern is not universally held [41]).

Social science research on collective action problems identifies three broad classes of solutions for how to get actors to cooperate: government regulation, private ownership, and community self-organizing [42]. Each is worth briefly considering with an eye toward the medium term.

Government regulation is perhaps the most commonly proposed solution for AI collective action problems. While some proposals focus on domestic measures [43], global regimes may be favorable due to AI being developed worldwide. This is reflected in proposals for international treaties [44] or, more ambitiously, global governance regimes with broad surveillance powers and the capacity to preemptively halt potentially dangerous AI projects through the use of force [45]. This more ambitious approach may be theoretically attractive in terms of ensuring AI collective action, though it is also unattractive for its potential for abuse, up to and including catastrophic totalitarianism [46]. Regardless, in practice, an intrusive global government is very likely a nonstarter at this time and for the foreseeable future, probably into the medium term. Nations are too unlikely to be willing to cede their national sovereignty to a global regime, especially on a matter of major economic and military significance. (Perhaps some future circumstances could change this, but the desire to preserve sovereignty, especially from rival and adversarial states, has been a durable feature of the international system.) Even a more modest international treaty may be asking too much. Treaties are difficult to create, especially if universal international consensus is needed (for example, because AI can be developed anywhere), and when access to and capability with the technology is unevenly distributed across the international community (as is very much the case with AI; for general discussion of emerging technology treaty challenges, see [47]). Instead, government regulations are likely to be more modest, and play at most a partial role in facilitating collective action. Whatever it is that governments end up doing, there is strong potential for institutions that are durable across the medium term, as discussed in Section 6.1.

Private ownership is commonly used for natural resource management. An entity that owns a natural resource has an incentive to sustain it and the means to do so by charging users for access at a sufficiently high fee. Private ownership schemes are difficult to apply to AI software due to the difficulty of restricting access. Hardware may offer a more viable option because hardware manufacturing facilities are geographically fixed and highly visible sites of major industrial infrastructure, in contrast with the ephemerality of software (For related discussion, see [48]). Hardware manufacturing is also typically privately owned [49]. AI collective action could conceivably be demanded by the

manufacturers, especially the select manufacturers of the advanced hardware used in the most capable AI projects. However, the benefits of AI collective action are experienced by many entities, and therefore would predominantly classify as externalities from the perspective of hardware manufacturers, in the sense that the benefits would be gained by other people and not by the manufacturers. This reduces the manufacturers' incentives to promote collective action and likewise reduces viability of private ownership schemes for AI collective action. Nonetheless, to the extent that hardware manufacturing can play a role, it could be a durable one. Hardware manufacturing is led by relatively durable corporations including Intel (founded 1968), Samsung Electronics (founded 1969), SK Hynix (formerly Hyundai Electronics, founded 1983), and Taiwan Semiconductor Manufacturing Company (founded 1987). These corporations are likely to remain important over medium-term and potentially also long-term time periods.

Community self-organizing for AI collective action can be seen in several important areas. One is in initiatives to bring AI developers together for promoting ethical principles. The Partnership on AI is a notable example of this. Importantly, the Partnership has recently welcomed its first Chinese member, Baidu [50]. This suggests that its emphasis on human rights (partners include Amnesty International and Human Rights Watch) will not limit its reach to Western organizations. Another area is in the collaborations between AI projects. For example, Baum [31] documents numerous interconnections between AGI projects via common personnel and collaborations, suggesting a cooperative community. Community self-organizing may lack the theoretical elegance of government regulation or private ownership, but it is often successful in practice. Whether it is successful for AI remains to be seen. AI community initiatives are relatively young, making it more uncertain how they will play out over the medium and long term.

6.3. Corporate AI Development

The financial incentives of for-profit corporations could become a major challenge for the safe and ethical development of AI over all time periods. How can companies be persuaded to act in the public interest when their financial self-interest points in a different direction? This is of course a major question for many sectors, not just AI. It is an issue for AI right now, amid a "techlash" of concerns about AI in social media bots, surveillance systems, and weaponry. It could also be an issue for AI over the mid and long term.

With regards to long-term AI, Baum [31] (p. 19) introduces the term "AGI profit–R&D synergy", defined as "any circumstance in which long-term AGI R&D delivers short-term profits". If there is significant AGI profit–R&D synergy, then it could make AGI governance substantially more difficult by creating financial incentives that may not align with the public interest. AGI profit–R&D synergy concerns long-term AI, but it is inherently a medium-term phenomenon because it would occur when AGI is being developed. Assessing the prospect of AGI profit–R&D synergy requires an understanding of the technical computer science details of AI as it transitions from the medium term to the long term, which is beyond the scope of this paper. If the medium-term details have any sort of close relation to near-term AI, that could constitute a significant strengthening of the medium-term AI hypothesis.

If AI companies' financial self-interest diverges from the public interest, how would they behave? Ideally, they would act in the public interest. In some cases, perhaps they will, especially if they are pushed to do so by people both within and outside of the companies. Unfortunately, experience from other sectors shows that companies often opt to act against the public interest, as seen, for example, in pushback by the tobacco industry against regulations aimed at reducing cancer risk; by the fossil fuel industry against regulations aimed at reducing global warming risk [51]; and by the industrial chemicals industry against regulations aimed at reducing neurological disease risk [52]. It is worth considering the prospect that AI companies may (mis) behave similarly.

It has been proposed that AI companies could politicize skepticism about AI and its risks to avoid regulations that would restrict their profitable activities [53]. This sort of politicized skepticism has a long history, starting with tobacco industry skepticism about the link between cigarettes and cancer

and continuing to this day with, for example, fossil fuel industry skepticism about global warming. One mechanism for this work is to fund nominally independent think tanks to produce publications that promote policies and issue stances consistent with the companies' financial self-interest.

Some attributes of this pattern can be seen in recent writing by the think tank the Center for Data Innovation, which warns of an "unchecked techno-panic" that is dampening public enthusiasm for AI and motivating government regulations [54]. The extent to which this constitutes a case of politicized skepticism is unclear. Specifically, the extent of the Center for Data Innovation's industry ties could not be ascertained for this paper. Likewise, it is not the intent of this paper to accuse this organization of conflicts of interest. It is also not the intent to claim the opposite—that there is no conflict of interest in this case. (Indeed, the presence of conflict of interest is often hidden—hence, industry firms fund the work of nominally independent think tanks instead of doing it in-house.) Instead, the intent is merely to provide an example that illustrates some aspects of the politicized skepticism pattern. Importantly, whereas the proposal of politicized AI skepticism focuses on skepticism about long-term AI [53], the skepticism of the Center for Data Innovation is focused on the near term [54]. Likewise, the pattern of politicized AI skepticism has the potential to play out across time periods, especially when there is significant profit–R&D synergy and concurrent prospects of government regulation.

6.4. Militaries and National Security Communities

Advanced militaries have long been involved with the forefront of AI in their capacity as research funders and increasingly as users of the technology. The advanced militaries also often have substantial technical expertise, as do the broader national security policy communities that they interface with. Furthermore, militaries are sometimes tasked with operations and planning across a range of time periods, and national security communities are likewise sometimes oriented toward thinking over such time periods. This is seen in the example cited above of the plan for the B-52H bomber to remain in service through the 2050s. It thus stands to reason that advanced militaries and national security communities could be interested in medium-term AI and its links between the near term and long term.

There is already some military attention to AGI. One clear example is the JASON report *Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD* [55], which was produced in response to a US Department of Defense query about AGI. Another is the excellent book [19], which features a full chapter on AGI and ASI. Both publications provide nuanced accounts of long-term AI. The publications are produced by analysts who are especially technically savvy and are not representative of the entire military and national defense communities. Nonetheless, they are among the publications that people in these communities may consult and do indicate a degree of awareness about long-term AI.

As documented by Baum [31], there are some current AGI R&D projects with military connections. Most of these are US academic groups that receive funding from military research agencies such as DARPA and the Office of Naval Research. One is a small group at the primary national defense research agency of Singapore. None of them have any appearance of the sort of major strategic initiative that is sometimes postulated in literature on long-term AI [33].

Given the current state of affairs, it is highly likely that advanced militaries and national security communities will be engaged in AI throughout the medium term. That raises the question of their likely role. Despite common concerns within AI communities, as manifest for example in Google employee protest over Project Maven, militaries can actually be a constructive voice on ethics and safety. For example, a major theme of the [55] report is that what it calls the "ilities"—"reliability, maintainability, accountability, verifiability, evolvability, attackability, and so forth" [55] (p. 2) are a major concern for military applications and "a potential roadblock to DoD's use of these modern AI systems, especially when considering the liability and accountability of using AI in lethal systems" [55] (p. 27). Militaries are keen to avoid unintended consequences, especially for high-stakes battlefield technologies.

It is also important to account for the geopolitical context in which militaries operate. Militaries can afford to be more restrained in their development and use of risky technologies when their nations

are at peace. In an interview, Larry Schuette of the Office of Naval Research compares autonomous weapons to submarines [19] (pp. 100–101). Schuette recounts that in the 1920s and 1930s, the US was opposed to unrestricted submarine warfare, but that changed immediately following the 7 December 1941 attack on Pearl Harbor. Similarly, the US is currently opposed to autonomous weapons, and on the question of whether it will remain opposed, Schuette replies, "Is it December eighth or December sixth"?

It follows that the role of militaries in medium-term AI may depend heavily on the state of international relations during this period. It stands to reason that the prospects for cautious and ethical AI development are much greater during times of peace than times of war. There is an inherent tension between pushing a technology ahead for strategic advantage and exercising caution with respect to unintended consequences, as is articulated by Danzig [56]. Peaceful international relations tips the calculus toward caution and can empower militaries and national security communities to be important voices on safety and ethics.

7. Conclusions

Parson et al. [2] argued that medium-term AI and its accompanying societal issues are important in their own right. This paper's analysis yields the same conclusion. For each of the issue areas studied here—governance institutions, collective action, corporate development, and military/national security—the medium-term will include important processes. In a sense, this is not much of a conclusion. It is already clear that AI is important in the near term, and there is plenty of reason to believe that AI will become more important as the technology and its applications develop further.

What then of the presentist–futurist debate? This paper proposes the medium-term AI hypothesis, which is that there is an intermediate time period that is important from both presentists and futurist perspectives. With the near term defined in terms of feasibility and the long term in terms of capability, it follows that the medium-term AI hypothesis is more likely to hold if near-term AI techniques and societal dimensions persist to a significant extent through the end of the medium term, when long-term AI is built. To the extent that the hypothesis holds, attention to the medium term could play an important role in bridging the divide that can be found between presentist and futurist communities.

The paper finds mixed support for the medium-term AI hypothesis. Support is strong in the case of AI governance institutions, which are currently in development and may persist through the medium-term, with implications for long-term AI. Support is ambiguous for AI collective action: government initiatives to promote collective action may play relatively little role at any time, private ownership schemes are difficult to arrange for AI, and community self-organizing has potential that might or might not be realized. Each of these three schemes for achieving collective action could potentially play out over near- and medium-term periods, with implications for long-term AI, but whether they are likely to is unclear. Regarding corporate AI development, a key question is whether near-to-medium-term AI technology could serve a profitable precursor to AGI, creating AGI profit–R&D synergy. Whether the synergy would occur is an important question for future research. Finally, advanced militaries and national security communities are already paying attention to AGI and are likely to remain active in a range of AI technologies through the medium term. While it is unclear whether military/national security communities will be important actors in the development of AGI, there is substantial potential, providing support for the medium-term AI hypothesis.

In closing, this paper has shown that at least some important AI processes are likely to play out over the medium term, and that they will be important in their own right and from both presentist and futurist perspectives. The exact nature and importance of medium-term AI is a worthy subject of future research. To the extent that medium-term AI can be understood, this can point to opportunities to positively influence them, resulting in better overall outcomes for society.

References

1. Baum, S.D. Reconciliation between factions focused on near-term and long-term artificial intelligence. *AI Soc.* **2018**, *33*, 565–572. [CrossRef]

2. Parson, E.; Re, R.; Solow-Niederman, A.; Zeide, E. Artificial Intelligence in Strategic Context: An Introduction. *AI Pulse*. 8 February 2019. Available online: https://aipulse.org/artificial-intelligence-in-strategic-context-an-introduction (accessed on 2 February 2020).

3. Parson, E.; Fyshe, A.; Lizotte, D. Artificial Intelligence's Societal Impacts, Governance, and Ethics: Introduction to the 2019 Summer Institute on AI and Society and Its Rapid Outputs. *AI Pulse*. 26 September 2019. Available online: https://aipulse.org/artificial-intelligences-societal-impacts-governance-and-ethics-introduction-to-the-2019-summer-institute-on-ai-and-society-and-its-rapid-outputs (accessed on 2 February 2020).

4. Cave, S.; Ó hÉigeartaigh, S.S. Bridging near and long-term concerns about AI. *Nat. Mach. Learn.* **2019**, *1*, 5–6. [CrossRef]

5. Prunkl, C.; Whittlestone, J. Beyond near and long-term: Towards a clearer account of research priorities in AI ethics and society. In Proceedings of the Third AAAI/ACM Annual Conference on AI, Ethics, and Society, New York, NY, USA, 7 February 2020.

6. Zeng, Y.; Lu, E.; Huangfu, C. Linking artificial intelligence principles. In Proceedings of the AAAI Workshop on Artificial Intelligence Safety, Honolulu, HI, USA, 12 December 2019.

7. Whittlestone, J.; Nyrup, R.; Alexandrova, A.; Cave, S. The role and limits of principles in AI ethics: Towards a focus on tensions. In Proceedings of the Second AAAI / ACM Annual Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27 January 2019.

8. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete Problems in AI Safety. 2016. Available online: https://arxiv.org/abs/1606.06565 (accessed on 2 February 2020).

9. Conn, A. Transcript: Concrete problems in AI safety with Dario Amodei and Seth Baum. *Future of Life Institute*. 2016. Available online: https://futureoflife.org/2016/08/31/transcript-concrete-problems-ai-safety-dario-amodei-seth-baum (accessed on 2 February 2020).

10. Arrhenius, G. The person-affecting restriction, comparativism, and the moral status of potential people. *Ethical Perspect.* **2005**, *10*, 185–195. [CrossRef] [PubMed]

11. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.

12. Garling, C. Andrew Ng: Why 'Deep Learning' Is a Mandate for Humans, Not Just Machines. *Wired*. 2015. Available online: https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans-not-just-machines (accessed on 2 February 2020).

13. Calo, R. Artificial Intelligence Policy: A Primer and Roadmap. Available online: https://www.ssrn.com/abstract=3015350 (accessed on 2 February 2020).

14. Good, I.J. Speculations concerning the first ultraintelligent machine. In *Advances in Computers*; Alt, F.L., Rubinoff, M., Eds.; Academic Press: New York, NY, USA, 1965; pp. 31–88.

15. Etzioni, O. No, the Experts Don't Think Superintelligent AI Is a Threat to Humanity. *MIT Technology Review*. 20 September 2016. Available online: https://www.technologyreview.com/s/602410/no-the-experts-dont-think-superintelligent-ai-is-a-threat-to-humanity (accessed on 20 February 2020).

16. Dafoe, A.; Russell, S. Yes, We Are Worried about the Existential Risk of Artificial Intelligence. *MIT Technology Review*. 2 November 2016. Available online: https://www.technologyreview.com/s/602776/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence (accessed on 2 February 2020).

17. Ford, M. Could artificial intelligence create an unemployment crisis? *Commun. ACM* **2013**, *56*, 1–3. [CrossRef]

18. Federal Agencies Need to Address Aging Legacy Systems. United States Government Accountability Office, GAO-16-468. 2016. Available online: https://www.gao.gov/assets/680/677436.pdf (accessed on 2 February 2020).

19. Scharre, P. *Army of None: Autonomous Weapons and the Future of War*; W. W. Norton: New York, NY, USA, 2018.

20. Mizokami, K. How B-52 Bombers Will Fly Until the 2050s. *Popular Mechanics*. 10 September 2018. Available online: https://www.popularmechanics.com/military/aviation/a23066191/b-52-bombers-fly-until-the-2050s (accessed on 2 February 2020).

21. Roblin, S. Bombs away: Russia's 'New' Tu-22M3M Bomber Might Look Familiar (and Still Deadly). *The National Interest*. 13 October 2018. Available online: https://nationalinterest.org/blog/buzz/bombs-away-russias-new-tu-22m3m-bomber-might-look-familiar-and-still-deadly-33381 (accessed on 2 February 2020).

22. Byrnes, M.W. Nightfall: Machine autonomy in air-to-air combat. *Air Space Power J.* **2014**, *May–June*, 48–75.

23. Baum, S.D.; Goertzel, B.; Goertzel, T.G. How long until human-level AI? Results from an expert assessment. *Technol. Forecast. Soc. Chang.* **2011**, *78*, 185–195. [CrossRef]

24. Sandberg, A.; Bostrom, N. Machine Intelligence Survey. Technical Report #2011-1, Future of Humanity Institute, Oxford University. 2011. Available online: https://www.fhi.ox.ac.uk/wp-content/uploads/2011-1.pdf (accessed on 27 May 2020).

25. Müller, V.C.; Bostrom, N. Future progress in artificial intelligence: A poll among experts. In *Fundamental Issues of Artificial Intelligence*; Müller, V.C., Ed.; Springer: Berlin, Germany, 2016; pp. 555–572.

26. Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; Evans, O. When will AI exceed human performance? Evidence from AI experts. *J. Artif. Intell. Res.* **2018**, *62*, 729–754. [CrossRef]

27. Armstrong, S.; Sotala, K.; Ó hÉigeartaigh, S.S. The errors, insights and lessons of famous AI predictions—And what they mean for the future. *J. Exp. Theor. Artif. Intell.* **2014**, *26*, 317–342. [CrossRef]

28. Morgan, M.G. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 7176–7184. [CrossRef] [PubMed]

29. Arkin, R. Lethal autonomous systems and the plight of the non-combatant. In *The Political Economy of Robots*; Kiggins, R., Ed.; Palgrave Macmillan: Cham, Switzerland, 2018; pp. 317–326.

30. Rosert, E.; Sauer, F. Prohibiting autonomous weapons: Put human dignity first. *Glob. Policy* **2019**, *10*, 370–375. [CrossRef]

31. Baum, S.D. A survey of artificial general intelligence projects for ethics, risk, & policy. *GCRI Work. Pap.* **2017**, *2017*. [CrossRef]

32. Armstrong, S.; Bostrom, N.; Shulman, C. Racing to the precipice: A model of artificial intelligence development. *AI Soc.* **2016**, *31*, 201–206. [CrossRef]

33. Shulman, C. Arms control and intelligence explosions. In Proceedings of the 7th European Conference on Computing and Philosophy, Bellaterra, Spain, 2–4 July 2009.

34. Hanson, R. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*; Oxford University Press: Oxford, UK, 2016.

35. Prime Minister of Canada. Mandate for the International Panel on Artificial Intelligence. *Canada*. 6 December 2018. Available online: https://pm.gc.ca/eng/news/2018/12/06/mandate-international-panel-artificial-intelligence (accessed on 2 February 2020).

36. Kohler, K.; Oberholzer, P.; Zahn, N. *Making Sense of Artificial Intelligence: Why Switzerland Should Support a Scientific UN Panel to Assess the Rise of AI*; Swiss Forum on Foreign Policy: Geneva, Switzerland, 2019; Available online: https://www.foraus.ch/wp-content/uploads/2019/10/20191022_Making-Sense-of-AI_WEB-1.pdf (accessed on 2 February 2020).

37. Miailhe, N. AI & Global Governance: Why We Need an Intergovernmental Panel for Artificial Intelligence. Centre for Policy Research, United Nations University. 20 December 2018. Available online: https://cpr.unu.edu/ai-global-governance-why-we-need-an-intergovernmental-panel-for-artificial-intelligence.html (accessed on 2 February 2020).

38. The AI Initiative. The Future Society. 2018. Available online: http://thefuturesociety.org/the-ai-initiative (accessed on 2 February 2020).

39. Cave, S.; Ó hÉigeartaigh, S.S. An AI race for strategic advantage: Rhetoric and risks. In Proceedings of the AAAI/ACM Annual Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018.

40. Geist, E.M. It's already too late to stop the AI arms race—We must manage it instead. *Bull. At. Sci.* **2016**, *72*, 318–321. [CrossRef]

41. Roff, H.M. The frame problem: The AI "arms race" isn't one. *Bull. At. Sci.* **2019**, *75*, 95–98. [CrossRef]

42. Ostrom, E. *Governing the Commons: The Evolution of Institutions for Collective Action*; Cambridge University Press: Cambridge, UK, 1990.

43. Scherer, M.U. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harv. J. Law Technol.* **2016**, *29*, 353–400. [CrossRef]

44. Wilson, G. Minimizing global catastrophic and existential risks from emerging technologies through international law. *Va. Environ. Law J.* **2013**, *31*, 307–364.

45. Bostrom, N. The vulnerable world hypothesis. *Glob. Policy* **2019**, *10*, 455–476. [CrossRef]

46. Caplan, B. The totalitarian threat. In *Global Catastrophic Risks*; Bostrom, N., Ćirković, M.M., Eds.; Oxford University Press: Oxford, UK, 2008; pp. 504–519.

47. Picker, C.B. A view from 40,000 feet: International law and the invisible hand of technology. *Cardozo Law Rev.* **2001**, *23*, 149–219.

48. Hwang, T. Computational Power and the Social Impact of Artificial Intelligence. 2019. Available online: https://www.ssrn.com/abstract=3147971 (accessed on 2 February 2020).

49. List of Semiconductor Fabrication Plants. *Wikipedia*. Available online: https://en.wikipedia.org/wiki/List_of_semiconductor_fabrication_plants (accessed on 2 February 2020).

50. Introducing Our First Chinese Member to the Partnership on AI. Partnership on AI. 16 October 2018. Available online: https://www.partnershiponai.org/introducing-our-first-chinese-member-to-the-partnership-on-ai (accessed on 2 February 2020).

51. Oreskes, N.; Conway, E.M. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*; Bloomsbury: New York, NY, USA, 2010.

52. Grandjean, P. *Only One Chance: How Environmental Pollution Impairs Brain Development—And How to Protect. the Brains of the Next Generation*; Oxford University Press: Oxford, UK, 2013.

53. Baum, S.D. Superintelligence skepticism as a political tool. *Information* **2018**, *9*, 209. [CrossRef]

54. Castro, D. The U.S. May Lose the AI Race Because of an Unchecked Techno-Panic. Center for Data Innovation. 5 March 2019. Available online: https://www.datainnovation.org/2019/03/the-u-s-may-lose-the-ai-race-because-of-an-unchecked-techno-panic (accessed on 2 February 2020).

55. Potember, R. *Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD*; The MITRE Corporation: McLean, VA, USA, 2017.

56. Danzig, R. Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority. Center for a New American Security. 30 May 2018. Available online: https://www.cnas.org/publications/reports/technology-roulette (accessed on 2 February 2020).