Reconciliation Between Factions Focused on Near-Term and Long-Term Artificial Intelligence

Seth D. Baum
Global Catastrophic Risk Institute

AI & Society 33(4), 2018, 565-572. This version 26 November 2018.

Abstract
Artificial intelligence (AI) experts are currently divided into "presentist" and "futurist" factions that call for attention to near-term and long-term AI, respectively. This paper argues that the presentist-futurist dispute is not the best focus of attention. Instead, the paper proposes a reconciliation between the two factions based on a mutual interest in AI. The paper further proposes a realignment to two new factions: an "intellectualist" faction that seeks to develop AI for intellectual reasons (as found in the traditional norms of computer science) and a "societalist faction" that seeks to develop AI for the benefit of society. The paper argues in favor of societalism and offers three means of concurrently addressing societal impacts from near-term and long-term AI: (1) advancing societalist social norms, thereby increasing the portion of AI researchers who seek to benefit society; (2) technical research on how to make any AI more beneficial to society; and (3) policy to improve the societal benefits of all AI. In practice, it will often be advantageous to emphasize near-term AI due to the greater interest in near-term AI among AI and policy communities alike. However, presentist and futurist societalists alike can benefit from each others' advocacy for attention to the societal impacts of AI. A reconciliation between the presentist and futurist factions can improve both near-term and long-term societal impacts of AI.

Keywords: artificial intelligence, near-term artificial intelligence, long-term artificial intelligence, societal impacts of artificial intelligence, artificial general intelligence, artificial superintelligence

1. Introduction

Artificial intelligence (AI) experts are—to generalize—of two minds about long-term AI, especially regarding the potential for radically transformative types of future AI such as artificial superintelligence (ASI) and brain emulation. One faction views long-term AI as a profound issue, warranting extensive attention and immediate action. The other faction views long-term AI as an unhelpful distraction from the more pressing near-term AI issues.

The purpose of this paper is to explore the potential for reconciliation between these two seemingly divergent factions. A reconciliation would enable them to expend less energy debating each other and more energy on activities that both agree are positive. In particular, if there are certain activities that simultaneously address issues associated with both near-term and long-term AI, then there is no need to engage in a near-term vs. long-term AI debate, because either position yields the same prescriptions.

I will refer to the two factions as the *futurists* and the *presentists*. Each can be said to have a core claim:

The futurist AI claim: Attention should go to the potential for radically transformative long-term AI.

The presentist AI claim: Attention should go to existing and near-term AI.

This division is a generalization in that some AI experts support both claims; there may even be some who support neither claim, though the rise of near-term AI applications makes it increasingly difficult for even non-experts to argue that AI is unworthy of attention.

The futurist faction can be traced to the early days of AI. Early AI researchers such as I.J. Good and Marvin Minsky predicted that AI would soon be capable of human-level thinking. Foreshadowing contemporary futurist concerns, Good (1965) warned of self-improving ultraintelligent machines that would undergo "intelligence explosion" in which "the intelligence of man would be left far behind", forming "the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control" (p.33).

When the early predictions failed to pan out, the field entered into a more muted phase known as AI winter. Here began the presentist faction. Most AI researchers shifted from trying to build human-level general AI to building narrow, domain-specific AI. Thus "mainstream AI" came to be associated with developing narrow, near-term AI and repudiating AI futurism. While a minority of researchers continued work on human-level general AI (e.g., Goertzel and Pennachin 2007) and a few high-profile futurists brought long-term AI some public attention (e.g., Kurzweil 2006), most AI researchers worked on narrow, near-term AI (see e.g. Bostrom 2014, p.18; Goertzel 2014, p.1).

Within the last few years, two parallel developments have raised the prominence of the futurist-presentist AI divide. One is a breakthrough in the performance of certain AI systems, in particular those involving "deep learning". Major applications of AI are rapidly appearing, from medical diagnosis to self-driving cars. This breakthrough has generated considerable interest in near-term AI while making dramatic long-term AI projections appear more plausible.

The other development is a spike in attention going to long-term AI, sparked mainly by the publication of the book *Superintelligence* (Bostrom 2014), public comments by several major celebrities (e.g., Hackett 2016; Hern 2016), and accompanying efforts by futurist researchers and activists (e.g., Future of Life Institute no date). The sober, philosophical tone of *Superintelligence* and the perceived credibility of some new outspoken futurists has brought the futurist perspective a burst of serious consideration. This has in turn sparked backlash from the presentist faction. The following, from an op-ed in *The New York Times* by Microsoft principal researcher Kate Crawford, is representative:

> According to some prominent voices in the tech world, artificial intelligence presents a looming existential threat to humanity: Warnings by luminaries like Elon Musk and Nick Bostrom about "the singularity" — when machines become smarter than humans — have attracted millions of dollars and spawned a multitude of conferences. But this hand-wringing is a distraction from the very real problems with artificial intelligence today, which may already be exacerbating inequality in the workplace, at home and in our legal and judicial systems. Sexism, racism and other forms of discrimination are being built into the machine-learning algorithms that underlie the technology behind many "intelligent" systems that shape how we are categorized and advertised to (Crawford 2016).

Crawford emphasizes inequality/discrimination as the near-term AI issue to focus on. Other presentists have emphasized such issues as military applications (e.g., Arkin 2009), safety issues with medical applications or self-driving cars, disruptions to labor markets ("technological unemployment"), or simply the technical capabilities of the AI itself. The unifying theme for this group is its insistence on a focus on current and near-term AI.

With growing interest in both near-term and long-term AI, the futurist-presentist divide has intensified. In addition to Crawford (2016), the divide can also be found in: a series of events on AI in 2016 sponsored by the White House, in which "many of the speakers… emphasized the need to focus on short-term concerns over long-term concerns of artificial general intelligence" (Conn 2016a); remarks by AI researcher Andrew Ng, arguing that worrying about future AI is as unimportant as worrying about "overpopulation on Mars" (Garling 2015); and an article by AI researcher Oren Etzioni, suggesting that AI that would not be built within the next 25 years is not worth paying attention to (Etzioni 2016), which prompted a reply by AI researcher Stuart Russell and political scientist Allan Dafoe arguing AI built more than 25 years from now can still be worth attention (Dafoe and Russell 2016); see also Bostrom (2014, p.18).

This sort of futurist-presentist divide is not unique to AI. For example, Selin (2007) documents a similar divide within the field of nanotechnology, in which a "mainstream" majority of researchers advocate for attention to (and funding for) existing and near-term nanotechnology at the expense of potential long-term transformative nanotechnology. Similarly, one explanation for the low priority commonly given to global warming is the fact that the harmful impacts of climate change will accrue mainly in the future (Weber 2006). In recognition of the challenge of motivating action on long-term issues, Baum (2015) proposes to address a wide range of long-term catastrophic risks—including long-term AI, nanotechnology, and climate change—by leveraging interest in synergistic near-term co-benefits.

The present paper follows in a similar vein as Baum (2015), though it is not intended as a critique or an advocacy of either the futurist or presentist AI factions. Instead, it seeks to pragmatically pursue reconciliation between the factions so that they may be able to pursue mutually agreeable activities. Three such activities are described in Section 4. Leading up to that, Section 2 explores the ideas at the root of the futurist and presentist perspectives, while Section 3 proposes a reconciliation based on a general concern for the societal impacts of AI.

2. Root Perspectives

This section attempts to articulate the perspectives at the root of the futurist and presentist factions. The descriptions derive from readings from, observations of, and interactions with the two factions. Root perspectives can help explain why the two factions disagree with each other and where reconciliation may be possible.

2.1 AI Futurists

Two distinct root perspectives can be found among AI futurists. One is centered on the intellectual goals of AI research, the other on the implications of a certain ethical perspective.

The first perspective is rooted in an intellectual aspiration of building advanced AI that have general intelligence at a human or superhuman level. "General" in this context means that they can think across a wide range of domains, or "have a broad capability to self-adapt to changes in their goals or circumstances" (Goertzel 2014, p.1). Building such advanced AI has long been "the grand dream of artificial intelligence" (Legg 2008, p.125), but it is only pursued by a small portion of AI researchers. These are the researchers who are willing to accept that their work

may be of little immediate practicality in order to work towards a more grand future achievement. They seek "to promote the idea that intelligent machines, even super intelligent machines, is a topic that is both important and one that can be scientifically studied, even if just theoretically for now" (Legg 2008, p.126).

Goertzel (2014, p.3) presents what it calls the *core hypothesis of artificial general intelligence*:

> The creation and study of synthetic intelligences with sufficiently broad (e.g. human-level) scope and strong generalization capability, is at bottom qualitatively different from the creation and study of synthetic intelligences with significantly narrower scope and weaker generalization capability.

Thus, advanced future AI requires fundamentally different research than is needed for the narrower contemporary AI. Goertzel (2014) maintains that this hypothesis is widely held among artificial general intelligence researchers. Thus, there is a community of researchers dedicated specifically to developing advanced future AI. (Goertzel 2014 also surveys this research.)

The second AI futurist perspective starts with an ethical concern for long-term effects and pairs this with a belief that long-term AI can have major long-term effects. The ethical concern for long-term effects has rigorous philosophical justification, though as with any ethical position, it is not universally held. It follows directly from a principle of equality: if all human lives (or whatever else we might care about) should be valued equally, then this (arguably) includes future lives. Furthermore, there could be many more future lives than present lives. Thus, philosophers and economists have long recognized the importance of actions that can have ongoing benefits to future generations, such as economic saving (Ramsey 1928) and survival of the population (Koopmans 1974). This same ethical perspective has been articulated more recently by some in the AI futurist faction (e.g., Bostrom 2003). The perspective prompts people to seek out issues and actions that could have significant long-term impact; the "merely near-term" is seen as trivial in comparison. This big-picture, long-term perspective explains why the futurist AI faction disproportionately attracts philosophers (e.g., Bostrom 2014; Price 2013) and cosmologists (e.g., Hawking et al. 2014), and why they give their organizations names like Future of Humanity Institute and Future of Life Institute.

The long-term ethical perspective is not fundamentally about AI. Instead, it is a guide for all human action; any role for AI is circumstantial. It relates to AI via the belief that certain AI could have major long-term effects, sending human civilization on fundamentally different trajectories. Only AI that could be so transformative is worth paying attention to; the rest is trivial detail. Thus, the prospect of future transformative AI is seen as important, and indeed is important even if there is only a small probability of it being built and even if it would be built decades or centuries from now. As long as there is some nonzero probability that future AI could, say, take over the world and kill everyone (Bostrom 2014) or dominate the global economy while remaking the future of civilization (Hanson 2016), then it is worth paying attention to. Indeed, such transformative long-term AI competes for attention not with near-term AI but with other potentially transformative issues such as global warming, nuclear war, and long-term nanotechnology.

The long-term ethical perspective can seem harsh in its relentless prioritization of long-term issues. For example, Crawford (2016) laments racial bias in AI evaluation of criminals in the United States, resulting in harsher prison sentences for African-Americans than for Americans of

European descent.[1] Racial bias in law enforcement would seem to be an important issue, and indeed, at the time of this writing, it is at the very top of the public agenda in the United States. However, the attention is on racial bias as it affects people today, not on how it affects long-term outcomes for humanity. Indeed, it is possible that present cases of racial bias will have limited long-term effect, especially relative to a possible future transformative AI. The same holds for most, if not all, other near-term AI issues. Ironically, it is the same ethical principle of equality that underlies opposition to racial bias that also motivates the ignoring of racial bias in favor of long-term issues.

2.2 AI Presentists

In contrast, AI presentists are motivated by a desire to work on contemporary and near-term activities and issues. This holds for presentist AI researchers as well as those focused on societal issues raised by AI.

Most AI research is presentist. The focus on contemporary and near-term AI systems is rooted in a practical, down-to-Earth perspective that seeks to work on actual systems and avoid speculation about future possibilities, especially when those possibilities seem strange and fantastic. Thus, for example, AI researcher Andrew Ng dismisses concerns about advanced future AI because "the future is so uncertain" that it cannot be meaningfully considered, and instead attention should be on problems that are more immediate and less ambiguous (Garling 2015). Similarly, Nilsson describes how AI researchers shy away from work on advanced future AI because they worry it would "risk losing our respectability" (Nilsson 2010, p.399).

This near-term focus of AI research is not unusual. Looking across all fields of science and engineering, one finds few people motivated by long-term effects for humanity or other considerations deriving from theoretical philosophy. Indeed, they often believe ethical considerations to be fundamentally outside their professional domain, so much that they "accept radical restrictions on having consequential opinions about *what ought to be done*" (Shapin 2010, p.388; emphasis original). This can be seen, for example, in the United States National Science Foundation's two criteria for evaluating funding proposals: intellectual merit and broader impacts. Intellectual merit is what scientists and engineers consider intrinsically important; it is what wins grants and publications in top journals. Broader impacts is the societal significance of the research, which is often considered less important (e.g., Schienke et al. 2009).

AI researchers generally focus on near-term AI because that is what they consider to have more intellectual merit. Long-term AI is speculative and untestable, which places it outside the bounds of normal science and engineering. The potential for long-term AI to have larger societal impacts is not a factor. From this perspective, even talking about the long-term impacts can be counterproductive, leading to concerns about respectability, as noted above, and to boom-bust cycles of hype and disillusionment, as has been seen over the years of AI "summer" and "winter". Hence presentists can resent the mere mention of long-term AI.

The norm of intellectual merit further explains why some AI researchers prefer to focus on the capabilities of the AI itself and not on societal issues.[2] Coming from computer science or related backgrounds, they view their role as one of advancing the science and technology of AI; attending to any related societal issues is, if anything someone else's job. However, many AI

[1] The racial bias Crawford describes comes from the investigative journalism of Angwin et al. (2016).

[2] One can argue that the AI itself is situated within society, and thus that AI researchers inevitably work on societal issues, even when they believe that they are working only on the AI itself. However, for present purposes, what matters is that the AI researchers believe that they are focusing on the AI itself and not on societal issues, even if it can be argued that they are inadvertently working on societal issues.

researchers are attentive to societal issues associated with AI. Crawford (2016) is not a rare exception.[3]

When presentists do turn to societal issues, they turn to those that relate to existing and near-term AI. This focus is reinforced by the fact that the general public and political leadership also tend to emphasize near-term issues. Hence even seemingly major long-term issues like global warming commonly get drowned out by near-term issues like economic growth (Scruggs and Benegal 2012). And so, when presentist AI researchers seek to discuss societal issues related to their AI, they find a welcome audience in social policy circles. This is seen, for example, in the heavy focus on near-term issues in the recent White House Symposia on AI Research (Conn 2016a).

3. Reconciliation

Despite the divergent perspectives between the presentists and futures, as well as the sometimes bitter debate between them, the two factions share a common interest in AI and belief of its importance. This commonality may seem banal, but it nonetheless distinguishes them from the large population that gives AI little attention or none at all. Thus, for starters, the two factions can at least unite in promoting AI as a locus of attention.

But there can be much more than that. There are significant parts of both factions that are concerned about the societal impacts of AI. Their concerns may come from different places, the futurists starting with concern about a certain type of societal impact and then applying this to AI, and the presentists starting with an interest in a certain type of AI and then applying this to societal issues. The fact that they work in opposite directions explains their divergent viewpoints and AI emphases. But they can at least agree that the societal impacts of AI are important, worth paying attention to, and worth trying to improve, even while they disagree on which impacts to focus on.

Meanwhile, there are those in both factions who are not so concerned about societal impacts, who instead prefer to focus on the AI itself. Thus, there is potential for a realignment of AI factions from futurist vs. presentist to those who are concerned about societal impacts vs. those who are not. This would seem to be the crucial distinction: between those who wish to pursue AI for its own sake or for narrow conceptions of intellectual merit, on the one hand, and on the other hand those who wish to pursue AI for the benefit of society. I will call these realigned factions the *intellectualist faction* and the *societalist faction*, rooted in the following core claims:

> *The intellectualist AI claim*: AI should be developed for its own sake, i.e. for its intellectual merit.
>
> *The societalist AI claim*: AI should be developed for its impacts for society.

It should be clarified that societalists may often emphasize intellectual aspects of AI en route to pursuing societal benefits. This is to say that belonging to the societalist faction does not preclude one from pursuing intellectually advanced AI. There can be many opportunities to be a "good" AI researcher according to traditional science and engineering notions of intellectual merit while still working towards societal benefits.

Another important clarification is that, among societalists, there will always be disagreements about which societal impacts to focus on; this is normal ethical and political

[3] As just one of many other examples, see Arkin (2009) on societal issues associated with military robotics.

debate. And just as conventional politics—the art of the possible—seeks areas of aligned interests, *quid pro quo*, and compromise, so too can those who disagree about the societal impacts of AI.

This realignment has a big advantage in that it enables focus on promoting attention to societal impacts in general. It is straightforward to argue that presentists and futurists alike *should be* concerned about societal impacts. AI researchers, like other scientists and engineers or any other people, are not above the law or above morality. It is everyone's responsibility to help society.[4] The realignment has further advantages in the synergistic opportunities to be found between presentist and futurist societalists.

4. Opportunities

Within a realigned societalist faction, the best opportunities will generally be those in which the interests of presentists and futurists are aligned. Such opportunities require no compromise or other sacrifice. Several such opportunities are apparent.

4.1 Social Norms

A first opportunity is to encourage more AI researchers to care about the societal impacts of their work. This assumes—quite reasonably, one would think—that the societal impacts of AI tend to be better when more AI researchers aim to achieve better societal impacts. In other words, the pro-societal efforts of AI researchers tend to be productive, not counterproductive.

While it is possible to persuade AI researchers one at a time to care about societal impacts, it will often be more effective to persuade many of them all together. This can be achieved by establishing societalist social norms among AI communities, i.e. by making it normal for AI researchers to seek to benefit society through their work. Societalist social norms are seen, for example, in AI researcher Stuart Russell's call for the AI field to switch from a norm of "building pure intelligence for its own sake, regardless of the associated objectives and their consequences" to a norm of not just caring about societal impacts, but also making that be "how practitioners define what they do" (Bohannon 2015:252).

The advantage of advancing societalist social norms is that it creates more AI researchers available to work on pro-societal designs, policies, etc., and likewise fewer people pushing back against such activities. An expanded societalist faction should benefit both near-term and long-term societalist agendas: a rising tide lifts all boats.

Given that AI research communities tend to focus on near-term AI, it may be advantageous to emphasize norms associated with the societal impacts of near-term AI. In practice, this means engaging AI researchers on such issues as inequality/discrimination (as in Crawford 2016), military applications (as in Arkin 2009), the safety of medical AI or self-driving cars, etc. It is relatively easy to sell the importance of these issues because either they already are important issues or they clearly will be in the near future, and because they do not depend on the arrival of any speculative future forms of AI but instead are rooted in existing and near-term AI.

Social norms surrounding the societal impacts of near-term AI can be of indirect benefit to the societal impacts of long-term AI. One mechanism for this is the durability of norms. What is now long-term AI may one day become near-term AI. If societalist norms remain intact, then the near-term societalists would switch over to what is now long-term AI. There is reason to believe that societalist norms may be so durable. For comparison, norms against slavery or racism or

[4] Again, as with any ethical position, the claim that people should help society is not universally held. A broader defense of this claim is beyond the scope of this paper.

sexism have proved durable;[5] within academia, the norm in favor of interdisciplinary research has gradually advanced for many years. If AI societalist norms could be similarly durable, they would help with what is now near-term and long-term AI.

A second mechanism is for societalist norms to prompt AI researchers to become more interested in long-term impacts. This could happen via attention to the ethics underlying which societal impacts one might care about. Ethical reflection is a natural fit for societalist AI, i.e. societalists are likely to engage in ethical reflection. That reflection could include the ethics of future generations, which is precisely how many existing futurists came to be futurists.

Therefore, a case can be made for promoting near-term societalist norms among AI communities. Because the issues are near-term, they are more likely to attract interest from AI communities. This increases the size of the near-term societalist faction. Then, some portion of these societalists may come to appreciate the ethical argument for caring about long-term impacts. This increases the size of the long-term societalist faction. Furthermore, the expanded societalist faction may remain in place when the long-term gets nearer. So, this is a win-win for both near-term and long-term societalists.

4.2 Technical Research

Societalist social norms expand the population of AI experts who can pursue the lines of technical AI research that benefits society. Some technical research may pertain only to specific AI applications and is thus only of interest to narrow presentist or futurist concerns. However, there is also some technical research that is of wide relevance to both near-term and long-term AI. Progress on this research is another win-win for near-term and long-term societalists.

An example of this sort of "timeless" technical research can be found in Amodei et al. (2016). This paper describes a range of technical problems oriented towards avoiding unintended harms from AI systems. The paper is expressed in terms of unintended harms from near-term AI. However, in an interview, Amodei describes how he sees the technical problems as being relevant for near-term and long-term AI alike, such that the near-term/long-term distinction is unnecessary (Conn 2016b). Thus, all societalists could support this technical research agenda, regardless of whether they are ultimately concerned about near-term or long-term impacts. Indeed, in the same interview, Amodei further notes that his paper has received broad praise, including from people who fit into both the presentist and futurist factions (Conn 2016b).

4.3 Policy

Given the general orientation of policy communities towards near-term issues, it will typically be easier for policy to address issues associated with near-term AI. However, there are at least two ways that AI policy can concurrently support both near-term and long-term AI, making for a third win-win for near-term and long-term societalists.

First, the process of developing near-term AI policy can create processes and competencies of relevance for long-term AI. Already, near-term AI policy issues have prompted legal scholars to familiarize themselves with some technical details of AI (e.g., Calo 2011; Funkhouser 2013; Hammond 2015). Such scholarship, and accompanying policy discussions, prepares policy communities for addressing long-term AI.[6]

[5] The world still has slavery and racism and sexism, but not as much as it once did. For example, while the United States continues to grapple with a variety of racial biases, it has become unthinkable to support the "separate but equal" racial segregation of the former "Jim Crow" laws.

[6] There has also been some legal scholarship focused on long-term AI (e.g., McGinnis 2010; Wilson 2013), but this is a relatively small minority of AI legal scholarship.

Near-term AI policy will also create demand for people with AI backgrounds working in policy positions. Such in-house AI expertise can help ensure that AI policy matches actual issues associated with AI technology and does not inadvertently restrict beneficial AI or enable harmful AI. Good models can be found, for example, in the Science & Technology Policy Fellowships offered by the American Association for the Advancement of Science and the IEEE-USA Government Fellowships offered by the Institute of Electrical and Electronics Engineers. These are both temporary fellowships; there should also be AI experts in permanent government positions. As long as the private sector offers abundant and lucrative employment for AI experts, it may be difficult to attract them to government positions. This challenge could be overcome via sufficiently strong societalist norms, which could motivate some societalists to forgo salary and pursue government positions simply because it is the right thing to do for society.

Second, some specific policy measures may improve outcomes for both near-term and long-term AI. This may seem counterintuitive, given the rapidly changing nature of AI. Indeed, a common concern about policy for emerging technologies like AI is that policy will fail to keep up with changes in the technology. In response to this concern, Moses (2007) proposes to phrase policy in more general terms, such as "choose the AI design that is more beneficial to society"; such policy could remain relevant even if AI technology undergoes major changes. Other policies of enduring relevance can be policies that establish institutions for AI governance, including policies to establish support for pro-societal AI efforts.

5. Conclusion

The current division between AI futurists and presentists—i.e. those concerned mainly about long-term AI vs. those concerned mainly about near-term AI—is not the best focus of attention. Instead of bickering over which AI impacts are most important, both groups would be wise to focus on addressing the AI impacts in general. There are ample opportunities to concurrently address both near-term and long-term AI impacts, including by establishing societalist social norms within AI communities, by advancing technical research that improves societal outcomes for all types of AI, and by advancing public policy that improves the governance of all types of AI. Making progress on these three fronts will take effort, which is precisely why those who worry about AI impacts should not drain their energy on internal disputes.

The Introduction to this paper quoted a passage from Crawford (2016) that argues for attention to near-term inequality/discrimination issues instead of to long-term threats. The passage is a representative example of the presentist vs. futurist dispute. However, in light of this paper's proposed reconciliation, the passage could be rewritten to instead take on the societalist vs. intellectualist dispute:

> According to some prominent voices in the tech world, artificial intelligence presents a major intellectual achievement, a triumph of science and technology. They say this achievement is to be celebrated, full stop. Indeed, research into the science of artificial intelligence has attracted millions of dollars and spawned a multitude of conferences. But this intellectual triumphalism is a distraction from the problems that artificial intelligence creates for society. It may already be exacerbating inequality in the workplace, at home and in our legal and judicial systems. It is also creating safety challenges for medical products and self-driving cars, as well as profound ethical dilemmas in its military applications. Looking further into the future, there may be some chance that artificial intelligence presents a looming existential

threat to humanity. These are serious issues that demand serious attention from society as a whole and especially from the technologists whose work is at the heart of these issues.

Expositions along these lines promise to reconcile the differences between presentists and futurists and instead shift attention to where it is desperately needed: on the societal issues posed by AI and on the positive steps can be taken to address these issues.

Acknowledgments
Tony Barrett and two anonymous reviewers provided helpful comments on earlier versions of this paper. Any remaining errors are the author's alone. The views in this paper are the author's and are not necessarily the views of the Global Catastrophic Risk Institute.

References

Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016). Concrete problems in AI safety. arXiv:1606.06565.

Angwin J, Larson J, Mattu S, Kirchner L (2016). Machine Bias. ProPublica, 23 May.

Arkin RC (2009). Ethical robots in warfare. IEEE Technology and Society Magazine 28(1):30-33.

Baum SD (2015). The far future argument for confronting catastrophic threats to humanity: Practical significance and alternatives. Futures 72:86-96.

Bohannon J (2015). Fears of an AI pioneer. Science 349(6245):252.

Bostrom N (2003). Astronomical waste: The opportunity cost of delayed technological development. Utilitas 15(3):308-314.

Bostrom N 2014. Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford.

Calo R (2011). Open robotics. Maryland Law Review 70(3):571–613.

Conn A (2016a). The White House considers the future of AI. Future of Life Institute, 15 June.

Conn A (2016b). Transcript: Concrete problems in AI safety with Dario Amodei and Seth Baum. Future of Life Institute, 31 August.

Crawford K (2016). Artificial intelligence's white guy problem. The New York Times.

Dafoe A, Russell S (2016). Yes, we are worried about the existential risk of artificial intelligence. MIT Technology Review, 2 November.

Etzioni O (2016). No, the experts don't think superintelligent AI is a threat to humanity. MIT Technology Review, 20 September.

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.

Future of Life Institute (no date). AI Activities. https://futureoflife.org/ai-activities (accessed 2 May 2017)

Garling C (2015). Andrew Ng: Why 'deep learning' is a mandate for humans, not just machines. Wired, May.

Goertzel B (2014). Artificial general intelligence: Concept, state of the art, and future prospects. Journal of Artificial General Intelligence 5(1): 1-48.

Goertzel B, Pennachin C (eds) (2007). Artificial General Intelligence. Springer Verlag, New York.

Good IJ (1965). Speculations concerning the first ultraintelligent machine. Advances in Computers 6:31-88.

Hackett R (2016). Watch Elon Musk divulge his biggest fear about artificial intelligence. Fortune, 17 August.
Hammond DN (2015). Autonomous weapons and the problem of state accountability. Chicago Journal of International Law 15:652-687.
Hanson R (2016). The Age of Em: Work, Love, and Life when Robots Rule the Earth. Oxford University Press, Oxford.
Hawking S, Tegmark M, Russell S, Wilczek F (2014). Transcending complacency on superintelligent machines. The Huffington Post, 19 April.
Hern A (2016). Stephen Hawking: AI will be 'either best or worst thing' for humanity. The Guardian, 19 October.
Koopmans TC (1974). Proof for a case where discounting advances the doomsday. Review of Economic Studies 41:117-120.
Kurzweil R (2006). The Singularity Is Near: When Humans Transcend Biology. Viking, New York.
Legg, S. (2008). Machine Super Intelligence. Doctoral dissertation, University of Lugano.
McGinnis JO (2010). Accelerating AI. Northwestern University Law Review 104:366-381
Moses LB (2007). Recurring dilemmas: The law's race to keep up with technological change. University of Illinois Journal of Law, Technology & Policy 2007(2):239-285.
Nilsson NJ (2010). The Quest for Artificial Intelligence: A History of Ideas and Achievements. Cambridge, UK, Cambridge University Press.
Price H (2013). Cambridge, cabs and Copenhagen: My route to existential risk. The New York Times, 27 January.
Ramsey FP (1928). A mathematical theory of saving. Economic Journal 38(152):543-559.
Schienke EW, Tuana N, Brown DA, Davis KJ, Keller K, Shortle JS, Stickler M, Baum SD (2009). The role of the NSF Broader Impacts Criterion in enhancing research ethics pedagogy. Social Epistemology 23(3-4):317-336
Scruggs L, Benegal S (2012) Declining public concern about climate change: can we blame the great recession? Global Environmental Change 22(2):505-515
Selin C (2007). Expectations and the emergence of nanotechnology. Science, Technology & Human Values 32(2):196-220.
Shapin S (2010). Never Pure: Historical Studies of Science as if It Was Produced by People with Bodies, Situated in Time, Space, Culture, and Society, and Struggling for Credibility and Authority. Johns Hopkins University Press, Baltimore.
Weber EU (2006). Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). Climatic Change 77(1-2):103-120.
Wilson G (2013). Minimizing global catastrophic and existential risks from emerging technologies through international law. Virginia Environmental Law Journal 31:307-364.