An AI Race for Strategic Advantage: Rhetoric and Risks

Dr Stephen Cave
Leverhulme Centre for the Future of Intelligence
University of Cambridge
sjc53@cam.ac.uk

Dr Seán S ÓhÉigeartaigh
Centre for the Study of Existential Risk
University of Cambridge
so348@cam.ac.uk

Abstract

The rhetoric of the race for strategic advantage is increasingly being used with regard to the development of artificial intelligence (AI), sometimes in a military context, but also more broadly. This rhetoric also reflects real shifts in strategy, as industry research groups compete for a limited pool of talented researchers, and nation states such as China announce ambitious goals for global leadership in AI. This paper assesses the potential risks of the AI race narrative and of an actual competitive race to develop AI, such as incentivising corner-cutting on safety and governance, or increasing the risk of conflict. It explores the role of the research community in responding to these risks. And it briefly explores alternative ways in which the rush to develop powerful AI could be framed so as instead to foster collaboration and responsible progress.

Introduction

Global leadership in different areas of fundamental and applied research in artificial intelligence has emerged as a strategic priority, both for major companies and nation states. The last two years have seen heightened competition between industry research groups for talented researchers and startups (Metz 2017, CB Insights 2017), and the release of strategic plans aimed at establishing research leadership in AI by the United States (NSTC, 2016), China (Fa, 2017), and other research-leading nations (Hall and Prescenti 2017, NEDO 2017, CIFAR 2017).

As the applications of AI research become more lucrative, and the impacts of AI on economic, scientific and military applications more transformative, this competition can be expected to intensify. A narrative has begun to emerge, in both policy and public contexts (Simonite 2017, Allen and Husain 2017, Stewart 2017, Allen and Kania 2017, Allen and Chan 2017), framing future trajectories relating to the development of AI in terms typically associated with a race for technological superiority. The China State Council's 2017 "A Next Generation Artificial Intelligence Development Plan" includes in its aims

> To seize the major strategic opportunity for the development of AI, to build China's first-mover advantage in the development of AI

The United States National Science and Technology Council's Research and Development Strategic Plan highlights that the US no longer leads the world in numbers of deep learning-related publications. Strong public statements have been made by political and technology leaders: notable examples include Russia's President Vladimir Putin stating that "whoever becomes the leader in this sphere will become the ruler of the world" (RT 2017) and OpenAI cofounder Elon Musk tweeting that "competition for AI superiority at national level [is the] most likely cause of WW3" (Musk 2017).

These developments reflect a perception that global leadership in AI could confer scientific, infrastructural and economic advantages to frontrunners. Such comparative advantage could also be self-perpetuating, if it helps to recruit the best research talent, or leads to greater wealth and spending power, or to the accumulation of better computing and data resources, or if the application of AI to the process of scientific research itself leads to further breakthroughs. Similarly, although many breakthroughs are likely to be local in application, the perception that some fundamental advances can be applied to a broad range of research challenges further fuels the idea that a winner could take it all.

Such notions of 'winner takes all' or the importance of technological leadership, foster the framing of AI development as a race. We will refer to this framing as that of the 'race for technological superiority' in AI. This kind of race is the primary concern of this paper. But at the same time, it is important to note that there are also concerns about a specifically *military* AI *arms* race. Although this is not the primary concern of this paper, there are some noteworthy interrelations with the broader race for technological superiority: most significantly, if AI fulfils its promise as a widely-applicable and transformational technology, then general superiority in AI will to a significant degree *imply* also military superiority. Many aspects of AI progress are likely to have dual-use relevance to both civilian and military applications (for example, advances enabling greater autonomy and a wider range of capabilities in autonomous vehicles) not least given the deep connection between AI research and the defence community (Geist 2016). And if systems approach general intelligence or superintelligence, they are likely to confer a very significant strategic advantage that could encompass, but go well beyond, conventional weaponry, with radical implications for security and geopolitics (Allen and Chan 2017).

The central concern of this paper is what might happen if the rhetoric noted above around first mover advantage or the pursuit of AI superiority becomes a widespread or even dominant framing of AI development. Our concerns arise both from use of the language of a race for technological superiority (independently of whether such a race is happening) and from such a race becoming a reality.

The Dangers of an AI race for Technological Advantage: in Rhetoric and in Reality

What is so bad about framing the development of AI in terms of a race for technological advantage? After all, it is widely agreed that AI brings enormous potential benefits across many sectors. One recent report estimated that it could add £232 billion by 2030 to the UK economy alone, with healthcare one of the sectors most enhanced, potentially bringing faster, better service to consumers (PwC 2017). There is a widespread belief that competition and other market forces are central to such innovation. So in as much as a race to develop AI technology means these kinds of benefit come sooner, we have reason to view it positively.

But at the same time, the development of such a potentially powerful new technology will need to be steered if it is to be as beneficial as possible while minimising the risks it might pose (Crawford and Calo 2016). In the words of the Future of Life Institute's open letter on 'Research Priorities for Robust and Beneficial Artificial Intelligence', signed by over 8,000 people including many leading figures in AI, work should focus "not only on making AI more capable, but also on maximizing the societal benefit of AI." (Future of Life Institute, 2017a). The danger of an AI race is that it makes exactly this thoughtful steering towards broadly beneficial outcomes more difficult.

Three Sets Of Risks

We want to distinguish between three sets of risks:

i) The dangers of an AI 'race for technological advantage' framing, regardless of whether the race is seriously pursued;

ii) The dangers of an AI 'race for technological advantage' framing and an actual AI race for technological advantage, regardless of whether the race is won;

iii) The dangers of an AI race for technological advantage being won.

(i) Risks Posed by a Race Rhetoric Alone

It is possible that the trend towards 'race for technological advantage' terminology in AI, including suggestions such as Vladimir Putin's that the winner of such a race "will become the ruler of the world," could pose risks even if the race is not pursued in earnest, let alone won. We perceive two main risks here:

(i.a) The kind of thoughtful consideration of how to achieve broadly beneficial AI, as mentioned above, will require subtle, inclusive, multi-stakeholder deliberation over a prolonged period. The rhetoric of the race for technological advantage, with its implicit or explicit threat that dire consequences will follow if some other group wins that race, is not likely to be conducive to such deliberation. Indeed, rhetoric around technological superiority (such as the 'arms race' rhetoric used in the Cold War in the US), played into what has been called a politics of fear, or a politics of insecurity -- that is, a political climate that discourages debate in favour of unquestioning support for a prescribed agenda (Griffith 1987). In *The Politics of Insecurity*, Jef Huysmans argues that use of the language of security (by which he means militarised language, which would include 'arms race' and related rhetoric) "is a particular technique of framing policy questions in logics of survival with a capacity to mobilize politics of fear in which social relations are structured on the basis of distrust" (Huysmans 2006).

(i.b) Second, if the rhetoric of a competitive, 'winner takes all' AI race is used in the absence of an actual race, it could contribute to sparking such a race.

(ii) Risks Posed by a Race Emerging

If the rhetoric of a race for technological advantage became an actual race to develop sophisticated AI, the risks increase further:

(ii.a) First, there is the risk that racing to achieve powerful AI would not be conducive to taking the proper safety precautions that such technology will require (Armstrong, Bostrom, and Shulman 2016). We mentioned above the need for broad consultation about the role AI should play in the life of a community. This might help address important considerations such as avoiding biased systems, or maximising fairness. But in addition to these goals, serious attention must also be given to ensuring humans do not lose control of systems. Such considerations become particularly important if AI approaches general intelligence or superintelligence (Bostrom 2014), but also long before, particularly when AI systems are performing critical functions. The risk is that as the perceived benefit to winning the race increases, so correspondingly does the incentive to cut corners on these safety considerations.

(ii.b) Second, a 'race for technological advantage' could increase the risk of competition in AI causing real conflict (overt or covert). Huysmans argues that militarised language such as this has "a specific capacity for fabricating and sustaining antagonistic relations between groups. In the case of the race for technological advantage, it encourages us to see competitors as threats or even enemies. The belief that a country intends in earnest to win an AI race, and that this would result in technological dominance, could, for example, prompt other countries to use aggression to prevent this (akin to the cyberattacks made against Iranian nuclear facilities attributed to the US and Israel) (Nakashima 2012), or motivate the targeting of key personnel (precedents -- though specific to their historical context -- might include Operation Paperclip, during which over 1,600 German scientists and engineers who had worked on military technology were taken to the US (Jacobsen 2014), or the apparently ongoing operations to encourage the defection of nuclear scientists between nations) (Golden 2017). Such scenarios would also increase the risk that a general race for technological superiority became increasingly a military AI arms race.

(iii) Risks Posed by Race Victory

The third category of risks of an AI race for technological superiority are those that would arise if a race were won. We will not explore these in detail here -- and the forms they take will anyway depend on the precise nature of the technology in question. But as an example, these risks include the concentration of power in the hands of whatever group possesses this transformative technology. If we survey the current international landscape, and consider the number of countries demonstrably willing to use force against others, as well as the speed with which political direction within a country can change, and the persistence of non-state actors such as terrorist groups, we might conclude that the number of groups we would *not* trust to responsibly manage an overwhelming technological advantage exceeds the number we would.

Choices for the Research Community

Given the dangers noted above of talking about AI development as a competitive race, one could argue that it would be better for the community of researchers considering AI and its impacts to avoid this terminology altogether. In this way, the language of the race for dominance would be seen as (in Nick Bostrom's terms) an *information hazard* -- perhaps either an *idea hazard* (a general idea whose dissemination could increase risk) or an *attention hazard* (if we consider that the dangerous idea already exists, but is as yet largely unnoticed) (Bostrom 2011).

But if we believe that the idea is already being disseminated by influential actors, such as states, including major powers, then the information hazard argument is weakened. It might still apply to particularly influential researchers -- those whose thoughts on the future of AI can become headline news around the world. But even in their case, and particularly in the case of lesser-known figures, there is a strong countervailing argument: that if the potentially dangerous idea of an AI race is already gaining currency, then researchers could make a positive impact by publicly drawing attention to these dangers, as well as by pursuing research dedicated to mitigating the risks of such a framing.

Of course, many leading researchers are already speaking out against an AI arms race in the sense of a race to develop autonomous weapons -- see for example the Future of Life Institute's open letter on this, signed by over three thousand AI researchers and over fifteen thousand others (Future of Life Institute 2015b). We believe this community could also usefully direct its attention to speaking out against an AI race in this other sense of a competitive rush to develop powerful general-purpose AI as fast as possible. Both media and governments are currently giving considerable attention to AI, yet are still exploring ways of framing it and its impacts. We believe that there is therefore an opportunity now for researchers to influence this framing[1]. One of the principles on AI agreed at the 2017 Asilomar conference offers a precedent on which to build -- it states:

> Race Avoidance: Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards
> (Future of Life Institute 2017).

Alternatives to a Race Approach

If we are not to pursue (and talk about pursuing) AI development as a race, then how should we pursue (and talk about pursuing) such development? This is too big a topic to consider thoroughly here. But we note these promising directions for alternative narratives around progress in, and benefits of, AI:

AI Development as a Shared Priority for Global Good

As advances in AI find application to an ever-wider range of scientific and societal challenges, there is a burgeoning discussion around harnessing the benefits of AI for global benefit. This reflects a widespread view among AI scientists

[1] For example, few countries have published formal strategies or legislation on AI, but a number have commissioned reviews that have sought expert opinion, eg, the UK Government-commissioned report on AI (Hall and Prescenti, 2017) and the White House report (NSTC, 2016).

and a growing number of policymakers that AI presents tremendous opportunities for making progress on global societal ills, and aiding in tackling some of the biggest challenges we face in the coming century -- among them climate change and clean energy production, biodiversity loss, healthcare, global poverty and education.

Emphasising these benefits could counteract a race approach in a number of ways: First, if there is global scientific consensus that some of the key aims of AI should be to benefit humanity in these ways, then it becomes less important in which companies or countries key breakthroughs occur. Second, it makes clear that *cooperation* on the development of AI stands to result in faster progress on these pressing challenges. Lastly, if the aims of the field are to benefit humanity worldwide, then the global community represent stakeholders in the process of AI development; this narrative therefore promotes inclusive and collaborative development and deployment of AI.

Cooperation on AI as it is Applied to Increasingly Safety-Critical Settings Globally

The next decade will see AI applied in an increasingly integral way to safety-critical systems; healthcare, transport, infrastructure to name a few. In order to realise these benefits as quickly and safely as possible, sharing of research, datasets, and best practices will be critical. For example, to ensure the safety of autonomous cars, pooling expertise and datasets on vehicle performances across as wide as possible a range of environments and conditions (including accidents and near-accidents) would provide substantial benefits for all involved. This is particularly so given that the research, data, and testing needed to refine and ensure the safety of such systems before deployment may be considerably more costly and time-consuming than the research needed to develop the initial technological capability.

Promoting recognition that deep cooperation of this nature is needed to deliver the benefits of AI robustly may be a powerful tool in dispelling a 'technological race' narrative; and a 'cooperation for safe AI' framing is likely to become increasingly important as more powerful and broadly capable AI systems are developed and deployed.

Responsible Development of AI and Public Perception

AI is the focus for a growing range of public concerns as well as optimism (Ipsos MORI 2017, Fast and Horvitz 2017). Many stakeholders, including in industry, recognise the importance of public trust in the safety and benefits offered by AI if it is to be deployed successfully (Banavar 2017). It is the view of the authors that a narrative focused on global cooperation and safe, responsible development of AI is likely to inspire greater public confidence than a narrative focused more on technological dominance or leadership. Other powerful new technologies, such as genetically modified organisms and nuclear power, have in the past proven controversial, with significant communities arguing for a cautious, safety-first approach, to which the rhetoric of the race is antithetical.

Recent Narratives

There have been encouraging developments promoting the above narratives in recent years. 'AI for global benefit' is perhaps best exemplified by the 2017's ITU summit on *AI for Global Good* (Butler 2017), although it also features prominently in narratives being put forward by the IEEE's *Ethically Aligned Design* process (IEEE 2016), the Partnership on AI, and programmes and materials put forward by Microsoft, DeepMind and other leading companies. Collaboration on AI in safety-critical settings is also a thematic pillar for the Partnership on AI[2]. Even more ambitious cooperative projects have been proposed by others, for example the call for a 'CERN for AI' from Professor Gary Marcus, through which participants "share their results with the world, rather than restricting them to a single country or corporation" (Marcus 2017). Finally, the overall narrative of cooperation was clearly expressed in a statement issued by a Beneficial AI conference in Tokyo[3]:

> The challenges of ensuring that [AI] is beneficial are challenges for us all. We urge that these challenges be addressed in a spirit of cooperation, not competition. (Beneficial AI Tokyo, 2017).

Conclusion

Although artificial intelligence has been discussed for decades, only recently has it received serious and sustained attention from governments, industry and the media. Among the various ways in which this technology is framed, we have highlighted one that we consider to be potentially dangerous: that of the race for technological advantage. Partly, we believe that a general race to develop AI would be dangerous because it would also encompass -- given the dual use of this technology -- a literal, military arms race. But even if this were not the case, we believe there would be risks – e.g. from corner-cutting in safety and consultation. Although some might argue that the research community should altogether avoid using AI 'race' terminology for fear of giving it currency, we believe that the idea is already current enough to justify interventions that draw attention to the dangers. Much work remains to be done in understanding the dynamics of a possible race, and in developing alternative framings for AI development -- but there are encouraging examples on which to build.

[2] See https://www.partnershiponai.org/thematic-pillars/

[3] In which both authors were involved.

Acknowledgments

The authors would like to thank Shahar Avin, Seth Baum, Haydn Belfield, Miles Brundage, Ariel Conn, and Kanta Dihal for helpful comments and discussions.

References

Allen, G. and Kania, E. 2017. China is Using America’s Own Plan to Dominate the Future of Artificial Intelligence. *Foreign Policy*, 8 September 2017.

Allen, G., & Chan, T. 2017. Artificial Intelligence and National Security, Technical Report, Harvard Kennedy School, Harvard University, Boston, MA.

Allen, J. R., Husain, A. 2017. The Next Space Race is Artificial Intelligence. *Foreign Policy*, 3 November 2017.

Armstrong, S., Bostrom, N., and Shulman, C. 2016. Racing to the precipice: a model of artificial intelligence development. *AI & Society* 31:201–206.

Banavar, G. 2016. Learning to trust artificial intelligence systems, Report, IBM, Armonk, NY.

Beneficial AI Tokyo, 2017. Cooperation for Beneficial AI. http://conscious-m.sakura.ne.jp/aiandsociety/signatories/tokyo-statement.html

Bostrom, N. 2011. Information Hazards: A Typology of Potential Harms from Knowledge. *Review of Contemporary Philosophy* 10: 44-79.

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

Bostrom, N. 2017. Strategic Implications of Openness in AI Development. *Global Policy* 8: 135-148.

Butler, D. 2017. AI summit aims to help world’s poorest. *Nature* 546(7657):196-197.

Canadian Institute for Advanced Research. 2017. Pan-Canadian Artificial Intelligence Strategy Overview, Report, Canadian Institute for Advanced Research (CIFAR), Toronto, Canada.

CB Insights. 2017. The Race For AI: Google, Baidu, Intel, Apple In A Rush To Grab Artificial Intelligence Startups, Technical Report, CB Insights, New York, NY.

Crawford, K. and Calo, R. 2016. There is a blind spot in AI research. *Nature* 538: 311–313.

Fa, G. 2017 State Council Notice on the Issuance of the Next Generation Artificial Intelligence Development Plan. State Department, China.

Fast, E., & Horvitz, E. 2017. Long-Term Trends in the Public Perception of Artificial Intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* 963-969. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence, Inc.

Future of Life Institute. 2015. Autonomous Weapons: An Open Letter from AI & Robotics Researchers, Report, Future of Life Institute, Cambridge, MA.

Future of Life Institute. 2015. Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter, Report, Future of Life Institute, Cambridge, MA.

Geist, E. M. 2016. It’s already too late to stop the AI arms race—We must manage it instead. *Bulletin of the Atomic Scientists* 72: 318-321.

Golden, D. 2017. The science of spying: how the CIA secretly recruits academics. *The Guardian,* 10 October 2017

Griffith, R. 1987. *The Politics of Fear: Joseph R. McCarthy and the Senate*. Amherst, Mass.: Univ of Massachusetts Press.

Hall, W., and Pesenti, J. 2017. Growing the Artificial Intelligence Industry in the UK, Report, HM Government, London, United Kingdom.

Huysmans, J. 2006. *The Politics of Insecurity: Fear, Migration and Asylum in the EU*. Oxford, UK: Routledge.

Ipsos MORI. 2017. Public views of Machine Learning: Findings from public research and engagement conducted on behalf of the Royal Society, Report, Royal Society, London, United Kingdom.

Jacobsen, A. 2014. *Operation Paperclip: The secret intelligence program that brought Nazi scientists to America*. London, UK: Hachette UK.

Marcus, G. 2017. Artificial Intelligence Is Stuck. Here’s How to Move It Forward. *New York Times*, 29 July 2017.

Metz, C. 2017. Tech Giants Are Paying Huge Salaries for Scarce A.I. Talent. *New York Times*, 22 October 2017.

Musk, E. 2017. *Twitter*, 4 September 2017.

Nakashima, E. 2012. Stuxnet was work of U.S. and Israeli experts, officials say. *The Washington Post*, 2 June 2012.

Networking and Information Technology Research and Development Subcommittee, National Science and Technology Council. 2016. The National Artificial Intelligence Research and Development Strategic Plan, Report, National Science and Technology Council, Washington, D.C.

PwC. 2017. Sizing the prize: What’s the real value of AI for your business and how can you capitalise? Technical Report, PwC, London, United Kingdom.

PwC. 2017. The economic impact of artificial intelligence on the UK economy, Technical Report, PwC, London, United Kingdom.

RT. 2017. 'Whoever leads in AI will rule the world’: Putin to Russian children on Knowledge Day. *RT*, 1 September 2017.

Simonite, T. 2017. AI could revolutionise war as much as nukes. *Wired*, 19 July 2017.

Stewart, P. 2017. U.S. weighs restricting Chinese investment in artificial intelligence. *Reuters*, 13 June 2017.

Strategic Council for AI Technology. 2017. Artificial Intelligence Technology Strategy, Report, New Energy and Industrial Technology Development Organisation (NEDO), Kawasaki, Japan.

The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. 2016. Ethically Aligned Design, Technical Report, IEEE, New York City, NY.

The State Council of the People's Republic of China. 2017. State Council Notice on the Issuance of the Next Generation Artificial Intelligence Development Plan, Report, State Council, Beijing, China. Translated by Creemers, R., Webster, G., Triolo, P., and Kania, E. 2017. New America, Washington, D.C.