Bridging near- and long-term concerns about AI

Debate about the impacts of AI is often split into two camps, one associated with the near term and the other with the long term. This divide is a mistake — the connections between the two perspectives deserve more attention, say Stephen Cave and Seán S. ÓhÉigeartaigh.

Stephen Cave and Seán S. ÓhÉigeartaigh

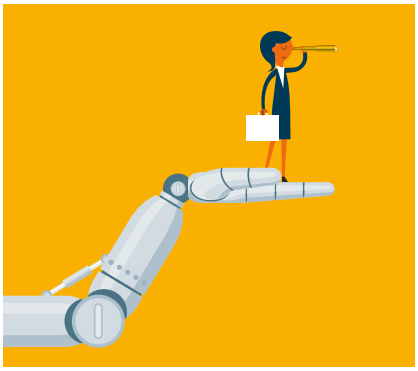
Research on the challenges posed by artificial intelligence (AI) has often been divided into two sets of issues, associated with two seemingly separate communities of researchers and technologists[1]. One set of issues relates to the near term — that is, immediate or imminent challenges involving fairly clear players and parameters, such as privacy, accountability, algorithmic bias[2] and the safety of systems that are close to deployment[3,4]. A second set of issues relates to longer-term concerns and opportunities that are less certain, such as wide-scale loss of jobs[5], risks of AI developing broad superhuman capabilities that could put it beyond our control[6], and fundamental questions about humanity's place in a world with intelligent machines[7].

These two sets of issues are often seen as entirely disconnected[1]. Researchers working on near-term issues see longer-term issues as a distraction from real and pressing challenges[8], or as too distant, uncertain or speculative to allow for productive work now[9]. On the other hand, those focused on longer-term challenges argue that their potential impact dwarfs that of present-day systems[7], and that these issues therefore deserve a proportionate share of research attention.

We believe that this perception of disconnect is a mistake. There are in reality many connections between near- and long-term issues, and researchers focused on one have good reasons to take seriously work done on the other. Those focused on the long term should look to the near term because research directions, policies and collaborations developed on a range of issues now could significantly affect long-term outcomes. At the same time, those focused on the near term could benefit from considering work on long-term forecasting and contingency planning, which takes seriously the disruptive potential of this powerful new technology.

Connected research priorities

First, on research directions: it is perhaps surprising how many of the central issues of AI ethics and safety span different time horizons. Immediate concerns such as robustness and reliability are crucial challenges for existing real-world systems, but will also grow in importance as increasingly powerful systems are deployed. Technology can exhibit strong path dependence: decisions people face in the future might be heavily constrained by decisions made now. For example, Greg Brockman, chief technology officer and co-founder of OpenAI, said recently[10]: "The Internet was built with security as an afterthought, rather than a core principle. We're still paying the cost for that today … With AI, we should consider safety, security and ethics as early as possible, and bake these into the technologies we develop."



Credit: Sorbetto/DigitalVision Vectors/Getty images

Some critics have argued that long-term concerns about artificial general intelligence (AGI), or superintelligence, are too hypothetical and (in theory) too far removed from current technology for meaningful progress to be made researching them now[9]. However, a number of recent papers have illustrated not only that there is much fruitful work to be done on the fundamental behaviours and limits of today's machine learning systems, but also that these insights could have analogues to concerns raised about future AGI systems[11,12]. Although there is no guarantee that current AI techniques will play a role in the development of AGI, it is reasonable to work on the assumption that they could. If so, technical safety research done now could both provide practical benefits for systems emerging in the near term, and fundamental frameworks for future systems.

Policy turning points

Second, policy measures enacted now could also influence the trajectories and impact of AI systems in ways that are highly relevant to longer-term concerns. Take explainability (the extent to which the decisions of autonomous systems can be understood by relevant humans): if regulatory measures make this a requirement, more funding will go to developing transparent systems, while techniques that are powerful but opaque may be deprioritized.

To take another example, much longer-term thinking has focused on whether AI-enabled automation will result in the elimination of enough jobs to significantly disrupt current social and economic structures. While experts remain divided on the likely scale of the problem, it is realistic at least to expect substantial changes to some sectors (such as lorry driving). A range of policies could help manage this, including educational measures to facilitate transferable skills or financial safety nets for those made redundant. Even while much remains uncertain, very significant disruption is a real possibility — and early steps could both ease immediate suffering and help manage long-term volatility.

Norms and institutions

Third, precedents set and collaborations developed now could reap benefits far into the future. For example, one very significant application of AI is to warfare. In the near term, we could see many such applications, including drones that navigate environments, select targets and carry out attacks all autonomously. But these might only be precursors of more powerful technologies, such as microdrone swarms

with sophisticated coordination capabilities, which could be used to carry out large-scale, finely targeted attacks. The decisions we make now, for example, on international regulation of autonomous weapons, could have an outsized impact on how this field develops. A firm precedent that only a human can make a 'kill' decision could significantly shape how AI is used — for example, putting the focus on enhancing instead of replacing human capacities.

The challenges we will face are likely to require deep interdisciplinary and intersectoral collaboration between industries, academia and policymakers, alongside new international agreements. Developing these structures will require actors to be willing to trust and compromise, and perhaps forego some technologies and opportunities in favour of others. All this is much easier while the stakes are still relatively low, at least compared to what they might become as AI advances. It is likely to prove far more difficult to establish rules when that would mean withdrawing widely used technologies, as opposed to establishing rules that shift development trajectories now. We need only to consider the difficulties of addressing climate change: foresightful planning and agreements a century ago would probably have been much less painful than trying now to restructure economies long dependent on fossil fuels.

Current international collaborations such as the Partnership on AI and the International Telecommunication Union (ITU)'s AI for Good Global Summits are a good start. The Partnership on AI brings together leading companies, non-governmental organizations and academic institutes from across the world to work together on challenges such as algorithmic fairness, and the application of AI in safety-critical systems. Its cross-sectoral and international buy-in makes it well-placed to begin developing best practices. The ITU (an agency of the United Nations) has a similarly global scope, and focuses on the application of AI to the sustainable development goals, such as combating climate change, poverty and hunger. These and similar initiatives may provide the building blocks needed to engage in a globally coordinated way with future opportunities and challenges of AI, which will inevitably cross sectoral and national lines.

Learning from the long term

These three points relate to ways in which addressing near-term issues could contribute to solving potential long-term problems. But what can taking the long term seriously offer to those focused on the near term? Perhaps the most important point is that the medium to long term has a way of becoming the present. And it can do so unpredictably: we cannot assume that either the power of the technology or its impact on society will develop linearly.

In particular, the impacts — even of current systems — might depend more on tipping points than even progressions. For example, we are currently seeing thresholds being passed in the accuracy of voice recognition and machine translation, leading to these technologies moving from being novelties to wide-scale, everyday use, with broad economic ramifications. Similarly, passing thresholds in the safe performance of self-driving cars could trigger a major shift towards their wide-scale adoption with rapid societal consequences. What the mainstream perceives to be distant-future speculation could therefore become reality sooner than expected.

Those currently focused on the near term might also learn from the long-term thinkers' techniques of foresight, contingency and scenario planning. Putting in place more systematic processes for measuring and forecasting progress in the many aspects of AI, combined with interdisciplinary forecasting processes to anticipate the consequences of such progress and preparing or responding to these consequences, will make us better placed to predict and manage possible tipping points[13].

We may be at a particularly influential point in the history of AI. It is being deployed at a much greater scale, across a broader range of industries, than ever before, and far more resources and talent are going into both fundamental and applied research. Therefore, the decisions we make now, in terms of research priorities and governance, are likely to have a major influence on the trajectories of AI — now and far into the future. Researchers focused on longer-term challenges should give careful consideration to how progress on technical near-term questions, and on broader societal issues, could shed light on the challenges that could come with more powerful, future forms of AI. In parallel, researchers focused on the near-term impacts of AI must recognize that the problems we see now represent a snapshot in time of a technology whose capacities and impacts are developing rapidly. ❐

Stephen Cave* and Seán S. ÓhÉigeartaigh*
Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK.
**e-mail: sjc53@cam.ac.uk; so348@cam.ac.uk*

Published online: 7 January 2019
https://doi.org/10.1038/s42256-018-0003-2

References

1. Baum, S. D. *AI Soc.* https://doi.org/10.1007/s00146-017-0734-3 (2017).
2. Crawford, K. et al. *The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term* (AI Now, 2016); https://ainowinstitute.org/AI_Now_2016_Report.pdf
3. Brundage, M. et al. Preprint at https://arxiv.org/abs/1802.07228 (2018).
4. Dietterich, T. G. & Horvitz, E. J. *Commun. ACM* **58**, 38–40 (2015).
5. Frey, C. B. & Osborne, M. A. *Technol. Forecast. Soc. Change* **114**, 254–280 (2017).
6. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies* (Oxford Univ. Press, Oxford, 2014).
7. Tegmark, M. *Life 3.0. Being Human in the Age of Artificial Intelligence* (Allen Lane, New York, 2017).
8. Calo, R. *Artificial Intelligence Policy: A Roadmap* (UC Davis, Davis, 2017).
9. Williams, C. *The Register* https://www.theregister.co.uk/2015/03/19/andrew_ng_baidu_ai/ (2015).
10. *The Dawn of Artificial Intelligence* (US Government Publishing Office, 2016); https://www.gpo.gov/fdsys/pkg/CHRG-114shrg24175/html/CHRG-114shrg24175.htm
11. Russell, S., Dewey, D. & Tegmark, M. *AI Magazine* **36**, 105–114 (Winter, 2015).
12. Amodei, D. et al. Preprint at https://arxiv.org/abs/1606.06565 (2016).
13. Owen, R. et al. in *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society* (eds Owen, R., Bessant, J. & Heintz, M.) 27–50 (Wiley, Chichester, 2013).

Competing interests

The authors declare no competing interests.