Synthese
https://doi.org/10.1007/s11229-021-03316-z

S.I. : DECTHEORY&FUTOFAI

Editorial to "Decision theory and the future of AI"

Yang Liu[1] · Stephan Hartmann[2] · Huw Price[3]

In the long run, the development of artificial intelligence (AI) is likely to be one of the biggest technological revolutions in human history. Even in the short run it will present tremendous challenges as well as tremendous opportunities. The more we do now to think through these complex challenges and opportunities, the better the prospects for the kind of outcomes we all hope for, for ourselves, our children, and our planet.

Thinking through these challenges needs a new kind of interdisciplinary research community. Many sources of expertise and insight are likely to be relevant, and this community needs to be very well-connected in several dimensions—'horizontally' between academic disciplines, 'vertically' to the policy and technology worlds, and of course geographically. AI is a global technology, and many of the challenges and opportunities of AI will be global in nature. Accordingly, getting AI right is not just an engineering challenge, but also a challenge for many other societal and academic sectors, including the humanities. Put another way, there is an engineering challenge of a 'sociological' kind, about how best to foster the necessary research community.

The field of decision theory is ideally placed to make contributions here, at several levels. AI innovations, including techniques from machine learning, are increasingly used to make *decisions* with significant social and ethical consequences, ranging from determining the news feeds on social media to making sentencing and parole recommendations in the criminal justice system. Decision theory provides and studies the standards by which such decisions are evaluated and improved. What is a rational decision? How can we train machines to make rational decisions? What is the relationship between human decision-making and machine decision-making? How can one make machine decision-making transparent (i.e. understandable to a human agent)? Which role does cognitive science play in these developments?

✉ Yang Liu
yl587@cam.ac.uk

[1] Faculty of Philosophy and Leverhulme Centre for the Future of Intelligence (CFI), University of Cambridge, Level 1, 16 Mill Lane, Cambridge CB2 1SB, UK

[2] Munich Center for Mathematical Philosophy of Science, LMU Munich, Geschwister-Scholl-Platz 1, 80539 Munich, Germany

[3] Leverhulme Centre for the Future of Intelligence (CFI), University of Cambridge, Level 1, 16 Mill Lane, Cambridge CB2 1SB, UK

Published online: 12 September 2021

Perhaps even more importantly, the field of decision theory itself is highly interdisciplinary, with a strong presence in disciplines such as philosophy, mathematical logic, economics, psychology, and cognitive science, amongst others. In addition, of course, it has foundational links to computer science and machine learning. So it is ideally placed to contribute to the sociological project, offering fertile ground in which to foster the interdisciplinary community needed for the challenges of AI, short term and long term.

This special issue stems from a conference series established with these goals in mind. *Decision Theory and the Future of AI* began in 2017 as a collaboration between the Leverhulme Centre for the Future of Intelligence (CFI) and the Centre for the Study of Existential Risk (CSER) at Cambridge, and Munich Center for Mathematical Philosophy (MCMP) at LMU Munich. The first two conferences were held at Trinity College, Cambridge in 2017 and LMU Munich in 2018. The first meeting outside Europe was held at ANU, Canberra, in 2019, in conjunction with ANU's Humanising Machine Intelligence project. A fourth conference was planned at PKU, Beijing, in 2020, before Covid intervened. We will be back!

Several of the papers in this special issue were presented at one of these conferences, while others were submitted in response to an open call for papers. The range of topics, and even more so the range of authors and their home disciplines and affiliations, is a tribute to the richness of the territory, both in intellectual and in community-building terms.

Acknowledgements We would like to thank the Cambridge-LMU Strategic partnership for support for the conference series, and our institutions (CFI, CSER, and the Faculty of Philosophy at Cambridge, and MCMP at LMU Munich) for administrative support. We also thank the authors and referees of the papers for all their work. Yang Liu and Huw Price gratefully acknowledge the support of the Leverhulme Trust, and of a grant from Templeton World Charity Foundation (TWCF0128); the opinions expressed in this publication are those of the authors and do not necessarily reflect the views of TWCF.