Philosophy & Technology (2021) 34 (Suppl 1):S1-S5
https://doi.org/10.1007/s13347-021-00466-3

RESEARCH ARTICLE

'AI for Social Good': Whose Good and Who's Good? Introduction to the Special Issue on Artificial Intelligence for Social Good

Josh Cowls[1,2]

Received: 14 July 2021 / Accepted: 31 July 2021 / Published online: 13 August 2021

Abstract
This introduction sets out the aims and scope of the Special Issue and provides an overview of each of the research articles and commentaries that follow.

Keywords AI · Social good · AI for social good

Over the past decade, research into artificial intelligence (AI) has emerged from the shadow of a long winter of disregard into a balmy summer of hope and hype. Whilst scholars and advocates have studiously documented the risks and potential harms of deploying AI-based tools and techniques in an array of societal domains, the idea nonetheless persists that the promised power of AI functionally could and ethically should be harnessed for, or at least (re-)oriented towards, 'socially good' purposes.

The twin aims of this Special Issue, simply stated, are to interrogate the plausibility of this notion and to consider its implications. The case that AI may — if developed carefully and deployed sensitively (Floridi et al., 2020) — contribute to net-positive outcomes in (some) socially relevant spheres is not without foundation. An array of efforts and initiatives are already underway, for example, to develop AI-based responses to help meet the UN's Sustainable Development Goals (SDGs) by 2030 (Cowls et al., 2021; Vinuesa et al., 2020). And as the suddenness of the Covid pandemic and the rapidity of global climate change both serve to remind us, global society faces challenges so stark and all-encompassing that an 'all-of-the-above' attitude towards potential solutions — especially those with a silicon sheen — may prove irresistible. Yet the susceptibility to the allure of technological solutions that increasing societal vulnerability engenders ought to give all of us pause, for reasons of both practice and of principle.

✉ Josh Cowls
josh.cowls@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, Oxford, UK

[2] Alan Turing Institute, London, UK

In practical terms, it is only when the haze of hype that surrounds AI is cut through that we can begin to consider the actual benefits and costs of specific AI-based tools, deployed in specific domains, developed for specific purposes. AI as a term operates as useful rhetorical flypaper for politicians looking to establish their twenty-first century policymaking credentials, as much as for start-ups on the search for seed-funding — and overstating the potential of AI serves the interests of both constituencies, and many others, well. It therefore seems prudent to adopt a cautious, evidence-based attitude towards claims made about the potential positive impact of AI, whether in the private, public, or non-profit spheres.

The prospect of AI for social good also invites more principled questions. The recent wave of AI hype has coincided not only with historically stark social challenges, but also at a time in which a small handful of private technology companies occupy a dominant position in many walks of life. Many of these companies have developed AI for use, first and foremost, in the day-to-day operation of their products and services, from predicting consumer preferences and recommending videos to driving autonomous vehicles and detecting atrial fibrillation. Several of them have also created units explicitly dedicated to using AI for socially good purposes. The active involvement of these for-profit companies in avowedly 'good' initiatives, as well as, for example, the embrace of similar efforts by governments both democratic and autocratic, points to the broader questions of whose 'good' is being served by projects branded as AI for social good, and who may be called 'good' as a result of such efforts.

Such questions are not novel. Moore (2019), for example, has explored the meaning and implications of 'AI for social good' and argues instead for 'AI for not bad', citing the vagueness and critical inadequacy of the term, bringing to mind Taylor's (2016) question regarding whose good is meant to be served by the use of big data as a 'public good'. But addressing these questions more fully — about whose good is served by, and who ought to be thought of as good as a result of, AI for social good — benefits from the series of empirically and ethically grounded contributions that are assembled in this Special Issue.

In 'Artificial Moral Agents Within in Ethos of AI4SG', Bongani Andy Mabaso asks which ethical framework artificial agents should be obliged to follow. Mabaso makes the case for exemplarism, a theory based on virtue ethics, as just such a framework, identifying several key features of exemplarism that 'fit the ethos of AI4SG'. These include exemplarism's conceptual grounding within existing exemplars of moral goodness, and the framework's responsiveness to societal expectations. Mabaso provides an example of an artificial agent deployed in an educational context, notes ongoing challenges to the teaching and learning of moral behaviour from exemplars and concludes by arguing that an exemplarist-based artificial agent may already be technologically possible if developed for carefully selected social contexts where relevant data is available.

The article 'In the Frame: the Language of AI', by Bones and co-authors, focuses attention on the discourses and practices associated with AI4SG. They employ a feminist epistemology to engage critically with the language most frequently used to characterise AI. Through a hybrid of historical, textual and corpus linguistic approaches, they show how the affordances and constraints of

several key terms adjacent to AI, like 'data', 'memory' and 'intelligence', help to shape non-expert understanding of what AI is and what it could be. Reframing AI for social good, they argue, is therefore in part a matter of making more careful, sensitive choices with respect to the language used to convey AI's potential and limitations, eschewing euphemism, hype and dogma. True AI4SG rests, therefore, on the facilitation of more inclusive and representative conversations which draw on an improved vocabulary and a more informed sense of what direction the deployment of AI could and should take.

In 'Artificial Interdisciplinarity: Artificial Intelligence for Research on Complex Societal Problems', Seth D. Baum explores a potential intersection between AI, interdisciplinary research and complex social problems. As Baum notes, much present AI work is oriented towards goals of expanding AI's technological capacity and increasing the profits of technology companies, whilst interdisciplinary research is far from a panacea for solving social challenges and can sometimes be outright harmful to societal interests. Whilst, as Baum acknowledges, there are already several ways in which AI systems facilitate interdisciplinary research, such as search engines and recommendation systems, there are several more tasks they could be designed to do, especially in the medium term, in support of interdisciplinary research. Yet Baum also anticipates the difficulties and risks that may arise from such a programme of development, some of which emerge from the risks of AI more generally, and others which arise as result of the high societal stakes of interdisciplinary research in areas such as global climate change and nuclear power.

In their research article 'Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias', Ying-Tung Lin, Tzu-Wei Hung and Linus Ta-Lun Huang explore AI in the context of implicit bias. Whilst acknowledging existing evidence suggesting that AI can perpetuate bias, the authors provide a framework within which to consider the use of AI to actively reduce the harms of implicit bias. Using recruitment processes as a case study, they highlight several areas in which using AI could potentially be of benefit, both with respect to the different information that AI provides to human users, and regarding the interventions that AI systems can be designed to make to reduce harms. Whilst the interventions proposed here apply primarily at the level of individual cognitive biases, the authors argue that they could form part of a package of responses that — along with structural changes — can serve the common interest of reducing the harms of implicit bias.

The Issue concludes with a series of commentaries considering other elements of AI in the context of social good. In 'Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea', Nathalie A. Smuha considers the ethical and political implications of taking a human rights-based approach to a governance framework for good AI. Writing against the backdrop of the European Commission's High-Level Expert Group and its charge to develop guidelines and recommendations for EU policymaking, Smuha discusses the search for a moral compass to guide the future directions of AI governance. Though acknowledging that human rights are neither flawless nor free of contestation, Smuha nonetheless advocates, persuasively, for the adoption of a human rights framework as the compass to steer AI governance in societally 'good' directions.

Meanwhile, Trooper Sanders' commentary, 'Testing the Black Box: Institutional Investors, Risk Disclosure, and Ethical AI', considers the role of institutional investors in advancing the development of responsible AI. Sanders notes that pushing for greater transparency in how AI is designed and deployed on the part of investors can involve both emboldening regulators and assessing the ethical fitness of companies in their portfolio. Drawing on lessons from the environmental, social and governance investing movement, Sanders outlines several ways in which 'institutional investors could give the ethical AI field some essential oomph'. Sanders thus identifies a potentially impactful front in the push to ensure that AI benefits society.

Finally, in their commentary, 'How to Handle Armed Conflict Data in a Real-World Scenario?', Trivedi and co-authors assess the utility of deep natural language processing to transform data from armed conflicts for the benefit of conflict resolution practitioners. Introducing a model initially trained on conflict data from Syria, the authors report a high degree of accuracy in the automated classifications of specific events that occurred between 2018 and 2019. Noting that there is no 'one-size-fits-all' model for classification tasks of this sort, given the complexity and context-specificity of armed conflict, the authors argue that sensitively constructed models like the one they introduce could give conflict resolution practitioners more time to conduct other essential analyses and collect more data.

Taken together, the research articles and commentaries that constitute this Special Issue provide an array of distinct perspectives that address both the plausibility and implications of AI4SG in particular settings and contexts. The contributions make clear that neither 'AI' nor 'social good' should be thought of as uncontested or incontestable terms, and we should remain wary of the twin dangers of unjustified hype and unseen harm arising from the continued growth of interest in, and application of, AI. Yet at what is a time of great vulnerability for a great many, undoubtedly there remain numerous domains in which sensitively designed systems utilising some form of artificial intelligence may help meet the pressing needs of societies and communities of practice around the world.

References

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). A definition, benchmark and database of AI for social good initiatives. *Nature Machine Intelligence, 3*(2), 111–115. https://doi.org/10.1038/s42256-021-00296-0

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Science Engineering and Ethics*. https://doi.org/10.1007/s11948-020-00213-5

Moore, Jared. (2019). AI for not bad. *Front Big Data, 2*, 32. https://doi.org/10.3389/fdata.2019.00032

Taylor, L. (2016). The ethics of big data as a public good: Which public? Whose good? *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences, 374*(2083), 20160126. https://doi.org/10.1098/rsta.2016.0126

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Nerini, F. F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications, 11*(1), 1–10. https://doi.org/10.1038/s41467-019-14108-y