Humanities & Social Sciences Communications

COMMENT

https://doi.org/10.1057/s41599-021-00750-9 OPEN

Mind the gap! On the future of AI research

Emma Dahlin [1,2]✉

Research on AI tends to analytically separate technical and social issues, viewing AI first as a technical object that only later, after it has been implemented, may have social consequences. This commentary paper discusses how some of the challenges of AI research relate to the gap between technological and social analyses, and it proposes steps ahead for how to practically achieve prosperous collaborations for future AI research. The discussion draws upon three examples to illustrate the analytical gap in different phases of the development of AI systems. Attending to the planning phase, the first example highlights the risk of oversimplifying the task for an AI system by not incorporating a social analysis at the outset of the development. The second example illuminates the issue of system acceptance, where the paper elaborates on why acceptance is multi-faceted and need not be approached as merely a technical problem. With the third example, the paper notes that AI systems may change a practice, suggesting that a continuous analysis of such changes is necessary for projects to maintain relevance as well as to consider the broader impact of the developed technology. The paper argues that systematic and substantial social analyses should be integral to AI development. Exploring the connections between an AI's technical design and its social implications is key to ensuring feasible and sustainable AI systems that benefit society. The paper calls for further multi-disciplinary research initiatives that explore new ways to close the analytical gap between technical and social approaches to AI.

[1] Department of Thematic Studies (Technology and Social Change), Linköping University, Linköping, Sweden. [2] Department of Anthropology, University of California, Santa Cruz, CA, USA. ✉email: emma.dahlin@liu.se

Introduction

With the emergence of AI, researchers have argued for more collaborations across disciplines to better understand AI in a social context (Theodorou and Dignum, 2020; Tomašev et al., 2020; Jobin et al., 2019; Perc et al., 2019; Sloane and Moss, 2019; Courtland, 2018). To bring about such collaborations, it will be of great importance to address the current gap between technological and social analyses of AI.

In the scientific community, research on AI is commonly divided into technological concerns (connected to natural sciences and engineering) and social concerns (connected to social sciences and humanities). These two strands have been largely disconnected from each other in research. Even when the social impact of AI is recognised, there is typically a sequential separation in that AI is viewed first as a technical object that only later, after it has been implemented, may have social consequences.

This disconnection is contradictory and creates practical and analytical problems for the simple reason that technology is always already social (Latour and Woolgar, 1979). For example, if attempting to dissect an AI system, it would be difficult to distinguish human material from nonhuman material. For the same reason, it is too simplistic to say that humans cooperate with material objects when they encounter AI. Technology can therefore not be approached as a neutral object, separated from things referred to as social. To better understand AI technology in the context in which it operates, the inseparability of these two concerns needs to be reflected in AI research.

This commentary paper discusses how some of the challenges of AI research relate to the gap between technological and social analyses, and it proposes steps ahead for future AI research to practically achieve prosperous collaborations.

Oversimplifying the task to be automated

A critical step in any AI development project is the identification of a task to be automated. This entails a clear understanding of the technical as well as the social capabilities a system requires. These capabilities are often not developed in sync, which can lead to malpractice. For example, in Poland, an AI system was designed to advance efficiency in Public Employment Services (PES) through algorithmic decision-making (Sztandar-Sztanderska and Zielenska, 2018). The purpose of the system was to profile unemployed individuals to determine which programmes they were eligible for, but the case counsellors did not inform the unemployed about the data they collected and how it was used (Sztandar-Sztanderska and Zielenska, 2020). Information about individual characteristics of the unemployed was gathered in the profiling system, which then applied an algorithm to classify the unemployed into one of three categories, without the unemployed knowing into which category they were placed or why (Niklas et al., 2015). The system essentially categorised the unemployed as good or bad investments, leading the Human Rights Commissioner to establish the algorithm's decisions as unjust, and in the end the system was banned (Kuziemski and Misuraca, 2020). A central issue here was that there was no clear plan for how human judgement and algorithmic decision-making could be joined and enacted in a public service context (Zejnilovic et al., 2020).

To understand why the system failed, scholars in social sciences and humanities would not only turn to the system itself, but also analyse the social context in which the system was set to operate (Akrich, 1992). The exclusion of such analysis in the development of the system reveals that the assignment (specification of needs and purpose of the system) was oversimplified. In this case, the social implications of the system were underestimated. Here it is reasonable to suggest that some of the problems that led to the system's failure could have been detected prior to implementation if the developers had incorporated a social analysis conducted by researchers with expertise in ethics, social science, and law. Applying social analysis of the planned technology during the idea and innovation stage would offer developers a better chance at forecasting and managing potential social challenges. This would mean widening the assignment of AI development to incorporate a thorough analysis of the broader context in which the system is expected to operate.

The example from the PES in Poland highlights that the social impact of an AI system cannot be reduced to a separate issue to be dealt with after the technical development. In this regard, engineers need to recognise that in designing AI models they are involved in social practices that shape society. To improve the chances of acceptable and trusted AI systems, a lesson for future projects may therefore be to incorporate social analyses of the technical design at the outset.

Reducing system acceptance to a technical question

Recent studies in radiology have shown how AI can decrease bias (Miller, 2018) and even outperform human radiologists in medical-image analysis (McKinney et al., 2020; Ardila et al., 2019; Topol, 2019). Although this technical progress is promising, developers struggle to incorporate such innovations into practice. For example, it is not uncommon for AI systems to be so complicated that even their developers cannot explain precisely how their creation reached a specific result (Riley, 2019). It is therefore not surprising that many systems are still black-boxed to their intended users (Castelvecchi, 2016). A common user request is the ability to assess an AI system's analysis (Ostherr, 2020; He et al., 2019; Rahwan et al., 2019; Vinuesa et al., 2020). In response to such demands, data scientists and developers of AI technologies have been asked to prioritise the explainability and transparency of AI systems (Watson et al., 2019). Here it is important not to limit the challenge of producing acceptable systems to improving explainability or transparency. To create truly acceptable systems, it is equally important that human–AI interaction be given social attention. For example, no AI systems operate in complete isolation from humans, and the functionality of an AI system therefore, to some extent, depends on being tolerated by humans. To build systems that users can accept is a task that involves multifaceted issues of coordinating human–AI interaction.

An illuminating example is the Probot, capable of autonomously carrying out surgical tasks during prostate resection (Harris et al., 1997a; Cosio and Davies, 1999). First, surgeons place the system in the correct starting position, and the system then autonomously carries out tasks such as removing conical segments of tissue, with the surgeon's role reduced to controlling an emergency stop button (Mei et al., 1996; Harris et al., 1997b). The Probot had been tried on patients with satisfactory results (Mei et al., 1999) and surgeons thought that such automation features were wanted (Rodriguez y Baena and Davies, 2009). However, problems arose when it came to the implementation of the Probot. The surgeons felt passivised as they were largely reduced to observers, and they therefore expressed unease with the system (Yip and Das, 2017). The system was ultimately rejected since surgeons perceived a greater than anticipated need for continuous interaction with their patients during procedures (Rodriguez y Baena and Davies, 2009).

This example illustrates how the acceptance of a system entails not only an understanding of its technical capabilities but also of the multifaceted social concerns that may arise from it. While it is indeed important to improve explainability and transparency in AI systems in order for their intended users to fully understand the system (especially in clinical practice), the Probot example shows that such understanding in no way guarantees system

acceptance. Other factors, such as the people working with the system feeling comfortable, here played a significant role when the Probot failed. On the other hand, for other AI systems, some users may accept and use a system without fully understanding its technical design. It is therefore important that the scope of analysis, regarding what makes an AI acceptable, is widened beyond transparency and explainability. As demonstrated in this example, gaining users' acceptance of an AI system in their practice is an achievement that reaches well beyond the technical capabilities of the system. To address the multifaceted issue of system acceptance, future development projects could engage in more thorough social analyses of trials carried out in the environment where the AI system is going to operate. Incorporating a social analysis of users' interaction with the technology, in real-life settings, could generate important insights into what is required for system acceptance.

Assuming that practices are stable

An AI system often changes the premises upon which a practice is based. It is therefore somewhat misleading to think of AI systems as merely tools. Take the stock-trading practice for example. The introduction of AI systems has entirely reshaped the conditions of the stock market, and the current stock-trading practices would not exist without AI systems (Callon and Muniesa, 2005). The fact that human traders now work alongside automated trading systems has significantly impacted how trading is carried out (Rundle, 2019; Brynolfsson and McAfee, 2018). The role of financial analysts has also changed; in addition to analysing how different human actors may respond to different events, they now need to anticipate how the automated trading systems will act (Lenglet, 2011). Since such systems have become central and fully fledged actors on the stock market, they cannot be viewed simply as tools. The trading algorithms are not passive tools but are actively involved in shaping the market (Callon and Muniesa, 2005). The stock market example illustrates how people may adjust their behaviour to AI systems, and it illuminates how AI technology may shift the socio-technical relations in a practice.

Such changes to socio-technical relations are also a concern in the training of AI systems. Training models using data on how people behave in a practice *without* AI do not reflect how people may behave in a practice with a new AI system implemented. That is, the underlying reality for an AI system is subject to change as soon as the system is introduced in the practice. An important lesson here is that since AI systems intervene with the underlying conditions of a practice, strategies are needed to cope with the fact that the validity of training data is obstructed in the moment a system is implemented in a real-life setting. That is, real-life humans may change their behaviour and reasoning in response to the implemented AI system, which challenges the relevance of the training data.

Acknowledging that practices are not stable but subject to change, AI development projects should incorporate a continuous analysis of changes to the practice. This means that trials need to be executed in real-life settings, since it is in the meeting between the *imagined* user and the *actual* user that one can explore *how* the AI comes into play. A social analysis during long-term implementation of an AI system can attend to changes in socio-technical relations. By recognising that AI systems may change the practices for which they are built, future projects can have a better chance at ensuring that AI systems bring changes that are desirable.

Directions forward

The three examples draw attention to some of the problems with a disconnection between substantial analyses of technological and social concerns, such as: (1) oversimplifying the task being automated, (2) reducing system acceptance to a technical question, and (3) assuming that practices are stable. For future research and practice to overcome such issues, this commentary paper suggests that systematic and substantial social analyses should be integral to future projects that develop AI systems—from early innovation, to technical design, to long-term implementation. Here, the paper sheds light on the need for future projects working to develop AI systems to continuously synchronise their attention to social and technological concerns, investigating how their technology is built, and how it could be built differently with different social consequences. Exploring the connections between an AI's technical design and its social implications will be key in ensuring feasible and sustainable AI systems that benefit society and that people want to use.

Regarding directions forward—trying to join human and AI work in practice—disciplines and sciences need to be better prepared and collaborations need to take place beyond disciplines. Funders and universities alike play an important role in supporting and facilitating such efforts. What is needed includes university initiatives that are truly multi-disciplinary and span boundaries between natural and social sciences. Such initiatives are important in order to create environments through which researchers from different fields can connect, initiate collaborations, and work together. To meet such future needs, engineering, natural sciences, and social sciences will be required to work together in new ways. The great potential here lies in coordinating different types of expertise, such as in-depth theoretical knowledge in social sciences with a wide range of natural sciences and engineering knowledge. It will simply not be enough for engineers, for example, to learn basic skills in social sciences, or for social scientists to learn basic skills of algorithms. Researchers' front-edge expertise from a wide range of knowledge areas should be woven together to combine and cross-fertilise ideas and insights from different disciplines across the sciences (such as law, engineering, social sciences, political science, philosophy, psychology, anthropology, and other disciplines). It is in such environments that new theories and methods can be developed. Projects that find new ways to connect technological and social analyses will be better equipped to understand and influence how AI changes society.

This commentary paper has focused on drawing attention to an analytical gap in AI development and a related scarcity of multi-disciplinary research. The identified issues call for more research and further discussion on how multi-disciplinary research collaborations could be coordinated to approach AI development more comprehensively.

Received: 2 October 2020; Accepted: 26 January 2021;
Published online: 15 March 2021

References

Akrich M (1992) The de-scription of technical objects. In: Law J, Bijker W (eds) Shaping technology/building society. MIT Press, Cambridge, MA: 205–224

Ardila D, Atilla PK, Bharadwaj S, Choi B, Reicher JJ, Peng L, Tse D, Etemadi M, Ye W, Corrado GC, Naidich DP, Shetty S (2019) End-to-end lung cancer screening with three-dimensional deep learning on low-dose shest computed tomography. Nat Med 25:954–961

Brynolfsson E, McAfee A (2018) The business of artificial intelligence: what it can -and cannot do - for your organization. Harvard Bus Rev 7: 3-11

Callon M, Muniesa F (2005) Peripheral vision: Economic markets as calculative collective devices. Organ Stud 26:1229–1250

Castelvecchi D (2016) Can we open the black box of AI? Nature 538:20–23

Cosio AF, Davies BL (1999) Automated prostate recognition: a key process for clinically effective robotic prostatectomy. Med Biol Eng Comput 37:236–243

Courtland R (2018) Bias detectives: the researchers striving to make algorithms fair. Nature 558:357–360

Harris SJ, Arambula-Cosio Q, Mei Q, Hibberd RD, Davies BL, Wickham JEA, Kundu B (1997a) The Probot—an active robot for prostate resection. Proc Inst Mech Eng 211:317–325
Harris SJ, Mei Q, Hibberd BL, Davies BL (1997b) Experiences using a special purpose robot for prostate resection. In: Proceedings of the 8th International Conference on Advanced Robotics, 1997, ICAR'97. IEEE, pp. 161–166
He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K (2019) The practical implementation of artificial intelligence technologies in medicine. Nat Med 25:30–36
Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat. Mach Intell 1:389–399
Kuziemski M, Misuraca G (2020) AI governance in the public sector: three tales from the frontiers of automated decision-making in democratic settings. Telecommun Policy 44:1–13
Latour B, Woolgar S (1979) Laboratory life: the construction of scientific facts. Princeton University Press, Princeton
Lenglet M (2011) Conflicting codes and codings: how algorithmic trading is reshaping financial regulation. Theory Cult Soc 28:44–66
McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GC, Darzi A, Etemadi M, Garcia-Vicente F, Gilbert FJ, Halling-Brown M, Hassabis D, Jansen S, Karthikesalingam A, Kelly CJ, King D, Ledsam JR, Melnick D, Mostofi H, Peng L, Reicher JJ, Romera-Parades B, Sidebottom R, Suleyman M, Tse D, Young KC, De Fauw J, Shetty S (2020) International evaluation of an AI system for breat cancer screening. Nature 577:89–94
Mei Q, Harris SJ, Hibberd RD, Wickham JEA, Davies BL (1996) PROBOT – a computer integrated prostatectomy system. Vis Biomed Comput 1131: 581-590
Mei Q, Harris SJ, Hibberd RD, Wickham JEA, Davies BL (1999) Optimising operation process for computer integrated prostatectomy. Springer, Berlin, Heidelberg
Miller AC (2018) Want less-biased decisions? Use algorithms. Harvard Bus Rev. https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms.
Niklas J, Sztandar-Sztanderska K, Szymielewicz K (2015) Profiling the unemployed in Poland: social and political implications of algorithmic decision making. Fundacja Panoptykon, Warsaw
Ostherr K (2020) Artificial Intelligence and Medical Humanities. J Med Humanit https://doi.org/10.1007/s10912-020-09636-4
Perc M, Ozer M, JH (2019) Social and juristic challenges of artificial intelligence. Palgrave Commun 5(1):1–7
Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon J-F, Breazeal C, Crandall JW, Christakis NA, Couzin ID, Jackson MO (2019) Machine behaviour. Nature 568:477–486
Riley P (2019) Three pitfalls to avoid in machine learning. Nature 572:27–29
Rodriguez y Baena F, Davies B (2009) Robotic surgery: from autonomous systems to intelligent tools. Robotica 28:163–170
Rundle J (2019) Wall Street Braces for Imapct of AI. Wall Street J. https://www.wsj.com/articles/wall-street-braces-for-impact-of-ai-11575887402
Sloane M, Moss E (2019) AI's social sciences deficit. Nat Mach Intell 1:330–331
Sztandar-Sztanderska K, Zielenska M (2018) Changing social citizenship through information technology. Soc Work Soc 16:1–13
Sztandar-Sztanderska K, Zielenska M (2020) What makes an ideal unemployed person? Values and norms encapsulated in a computerized profiling tool. Soc Work Soc 18:1–16
Theodorou A, Dignum V (2020) Towards ethical and socio-legal governance in AI. Nat Mach Intell 2:10–12
Tomašev N, Cornebise J, Hutter F, Mohamed S, Picciariello A, Connelly B, Belgrave DCM, Ezer D, Cachat van der Haert F, Mugisha F, Abila G, Arai H, Almiraat H, Proskurnia J, Snyder K, O'take-Matsuure M, Othman M, Glasmachers T, de Wever W, Whye Teh Y, Emitiyaz Khan M, De Winne R, Schaul T, Clopath C (2020) AI for social good: unlocking the opportunity for positive impact. Nat Commun 11:1–6
Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25:44–56
Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, Felländer A, Langhans SD, Tegmark M, Nerini FF (2020) The role of artificial intelligence in achieving the sustainable development goals. Nat Commun 11:1–10
Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, Floridi L (2019) Clinical applications of machine learning algorithms: beyond the black box. BMJ 364:2–9
Yip M, Das N (2019) Robot autonomy for surgery. Preprint at *arXiv*. https://arxiv.org/abs/1707.03080
Zejnilovic L, Lavado S, Martinez de Rituerto de Troya I, Sim S, Bell A (2020) Algorithmic Long-Term Unemployment Risk Assessment in Use: Counselors' Perception and Use Practices. Global Perspectives. https://doi.org/10.1525/gp.2020.12908

Acknowledgements

This research was funded by the Swedish Research Council under the International Postdoc Grant no. 2019-00697.

Funding

Open access funding provided by Linköping University.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.D.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.