AI's social sciences deficit

To create less harmful technologies and ignite positive social change, AI engineers need to enlist ideas and expertise from a broad range of social science disciplines, including those embracing qualitative methods, say Mona Sloane and Emanuel Moss.

Mona Sloane and Emanuel Moss

Computer scientists are building a vast array of machine learning systems (often called AI) that can perform human tasks reliably, making us believe that AI can be a judge, a shopkeeper, a chauffeur, a financial analyst, a receptionist, a security guard, a doctor or a paralegal. Many of us enjoy the fruits of this labour, from our e-mail folder that is kept clear of spam to the improved chances we have of detecting cancer early. But AI doesn't make everybody's life easier or safer. There is mounting evidence that AI can exacerbate inequality, perpetuate discrimination and inflict harm: Virginia Eubanks[1] has demonstrated how automated systems in government services can increase stigma, exacerbate poverty and inflict harm based on social class; Safiya Noble[2] has shown how search engines discriminate against women of colour; Joy Buolamwini and Timnit Gebru[3] have provided evidence for discrimination in image data bases and automated gender classification systems; Wilson, Hoffman and Morgenstern[4] have proven that there are higher error rates for pedestrians with darker skin tones in object detection systems; Bolukbasi et al.[5] highlighted gender stereotypes in word embeddings; and Os Keyes[6] has shown how automated gender recognition systems perpetuate violence against trans identities. These new streams of work show that as AI is increasingly used in the organization of society and its basic institutions, the stakes are high. It is impossible to build an equitable and prosperous future with decision-making machines that amplify historical patterns of oppression.

Quantitative approaches are insufficient

Technologists are increasingly looking to social scientists to help fix the problem of harmful or biased AI through a focus on ethics or safety[7,8], or to develop AI that is aligned with human values[9,10]. But they would do better by making full use of the social sciences, not just embracing quantitative methods.

Take, for example, the case of autonomous vehicles and the 'moral machine' dilemma[11]. The narrative is that if a truly autonomous vehicle is to be exposed to real-world traffic, it must at the very least be equipped with some ethical guiding principles that make its decision-making process socially acceptable, particularly in the context of potential crash situations. To address this and other problems, technologists are exploring the concept of value alignment — the idea that machines should act in accordance with human values.

But computationally, it is exceptionally difficult to define and encode something as fluid and contextual as 'human values' into a machine. Where researchers have tried to do so, they have set out to represent human relations, in all their complexity, through highly formalized games: for example, in OpenAI's recent work on AI safety that attempts to codify human values for machine learning by judging the winner of debates[12]. Anyone who has had a conversation about an important decision knows that there is never a clear winner, and many times there are not even clear sides to take. Game theory has given us interesting thought experiments, but human values are not the same as the probabilistic outcomes of contrived situations.

As technologists tend to prefer ideas that can easily be translated back into the language of mathematics, there is a risk that qualitative social science disciplines such as sociology and anthropology, which are committed to observing the complexity of social life and making sense of it, are filtered out. Instead, quantitative approaches such as analytic philosophy, behavioural economics and evolutionary psychology will dominate[13,14]. But we argue that it is misguided to only frame social issues in quantitative ways, where they can be studied and measured in abstract lab situations or thought experiments: AI does not fail people in a lab; it fails them in real life, with real consequences.

Four cues for technologists

There is a wide gulf separating the corpus of knowledge and methods that the qualitative social sciences have honed over the past century and the computational policies that guide AI design. This gulf can be bridged by taking the following four cues.

Quantitative data can be problematic. Qualitative data are often seen as too subjective to be useful in replicable experiments. But this is a strength. Qualitative data account for themselves, whereas quantitative data often erase the conditions of their collection. They appear as if from nowhere and cannot account for any gaps or biases that might undermine how completely they can be said to represent the phenomena being studied. This lack of accountability can be used to dodge ethical research practices, too, because 'historical' data do not often prompt informed consent from research subjects. Often, the proper degree of scrutiny is impossible. Only later, when and if 'bias' in datasets is revealed through cases of AI harm — and by those affected — do these kinds of considerations surface.

Individuals are not the locus of truth, values or culture. The notion that truths, values and culture can be codified, quantified and extracted from how individuals behave under artificial conditions assumes that these amorphous concepts are external to lived experiences or social context. This is a dangerously simplified understanding of how our social world works. It also leads to a conclusion that individuals' behaviours can unproblematically be manipulated by technological means. Social norms and cultural values do not live inside people's heads. At the same time, culture does not exist apart from people either. Rather, culture is continuously being constituted through the ways people interact with each other and the physical world.

Context matters. If individuals are seen as the locus of culture, the assumption is embedded that individuals remain the same when they move from one context to the next. But clearly, they do not. Social life is deeply contextual, and behavioural

responses vary across social contexts. Few patterns can be universalized, and a lesson learned in one context cannot be neatly applied to a different situation even if it seems similar. Individuals' behaviours only make sense in context, and any experiments should attempt to account as fully as possible for that context, and should be informed by rich, embedded qualitative studies of those social contexts.

The future need not look like the past. Quantified social data are produced in a world that is deeply unjust, with a host of power dynamics and historical legacies of inequity. Quantitative social sciences utilize fossilized data that enshrine these dynamics. Using statistical methods to extrapolate from these data, even when aligned with human values, cannot be expected to lead to a less unjust world. Understanding these histories and dynamics is a necessary, but perhaps not entirely sufficient, step toward creating a more equitable world for all.

Asking the right questions

Pursuing these ways forward can empower AI designers to create less harmful technology and ignite positive social change. To aide that process, we suggest including three concrete questions drawn from qualitative social research[15,16] into iterative processes of AI design.

What do we know about society, and why?. Qualitative social research can help us take stock of and understand the categories through which we make sense of social life, and which are being used in AI. For example, technologists are not trained to understand how racial categories in machine learning[17] are reproduced as a social construct that has real-life effects on the organization and stratification of society. These questions are discussed in depth in the social sciences, which can help create the socio-historical backdrop against which the (oppressive) history of ascribing categories like 'race' can be made explicit[18]. They can also provide technologists with the conceptual tools for understanding how race in technology itself is a tool "designed to stratify and sanctify social injustice"[19] and to consider when a machine learning solution is, in fact, not appropriate[20]. A qualitative viewpoint is, therefore, key for what Daniels, Nkonde and Mir[21] call "racial literacy in tech": "An intellectual understanding of how structural racism operates in algorithms … and technologies not yet developed … and a commitment to take action to reduce harms to communities of color."

How do we know what we know (about society)?. Data always reflect the biases and interests of those doing the collecting. Qualitative research is explicit about the data collection, whereas quantitative research practices in AI are not. For example, technologists tend to source AI training data from readily available datasets, such as Wikipedia. But there is evidence that Wikipedia entries display subtle forms of gender bias[22]. Similarly, Richardson, Schultz and Crawford[23] have shown that predictive policing systems designed to forecast criminal activity in the United States often build on data that come from racially biased and unlawful practices. Assuming that these datasets contain some forms of 'universal truth' that can be extracted shows that AI design practices routinely deprioritize questions around biased data collection and the harm associated with it. A qualitative approach can tackle this by introducing protocols for more ethical data research practices in data science, as, for example, in the Pervasive Data Ethics for Computational Research (PERVADE) project[24].

Who is designing a technological intervention for a social setting, who participates, and who is affected by it?. A quantitative approach does not require the researcher or AI designer to locate themselves in the social world and consider their own position of power and privilege. It, therefore, does not require an assessment of who is included in vital AI design decisions, and who is not. Neither does it call for including their research subjects into the production of knowledge. Qualitative research, on the other hand, is more prone to requiring the researcher to reflect on how their interventions affect the world in which they make their observations[25]. Qualitative research can also build on the idea of co-producing knowledge[26] and the observation that research subjects deeply care about the work they participate in[27].

Onwards

To achieve socially just technology, we need to include the broadest possible notion of social science, one that includes disciplines that have developed methods for grappling with the vastness of the social world, and that helps us understand how and why AI harms arise as part of a large, complex and emergent techno-social system. This is not about achieving parity between disciplines. It is about recognizing the strengths of disciplines for what they are, and about levelling the playing field. There is something deeply unjust about ignoring qualitative ways of knowing the sociotechnical world: privileging quantitative knowledge is complicit with perpetuating a framing of technology as superior to human judgement. As we continue to weave together social, cultural and technological elements of our lives, we must integrate different types of knowledge into technology development. A more socially just and democratic future for AI in society cannot merely be calculated or designed; it must be lived in, narrated and drawn from deep understandings about society. ❐

Mona Sloane[1]* and Emanuel Moss[2]

[1]*New York University, New York, NY, USA.* [2]*The Graduate Center, CUNY, New York, NY, USA.*

**e-mail: mona.sloane@nyu.edu*

Published online: 9 August 2019

https://doi.org/10.1038/s42256-019-0084-6

References

1. Eubanks, V. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's, 2018).
2. Noble, S. U. *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York Univ. Press, 2018).
3. Buolamwini, J. & Gebru, T. *Proc. Mach. Learn. Res.* **81**, 77–91 (2018).
4. Wilson, B., Hoffman, J. & Morgenstern, J. Preprint at https://arxiv.org/abs/1902.11097 (2019).
5. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. & Kalai, A. Preprint https://arxiv.org/abs/1606.06121 (2016).
6. Keyes, O. in *Proc. ACM on Human-Computer Interaction* **2**, 88 (ACM, 2018).
7. Amodei, D. et al. Preprint at https://arxiv.org/abs/1606.06565 (2016).
8. Greene, D., Hoffmann, A. L. & Stark, L. in *Proc. 52nd Hawaii International Conference on System Sciences* 2122–2131 (HICSS, 2019).
9. Sloane, M. in *Proc. Weizenbaum Conference 2019 'Challenges of Digital Inequality - Digital Education, Digital Work, Digital Life'* https://doi.org/10.34669/wi.cp/2.9 (2019).
10. Metcalf, J., Moss, E. & boyd, d. *Soc. Res.* **86**, 449–476 (2019).
11. Awad, E. et al. *Nature* **563**, 59–64 (2018).
12. Irving, G. & Askell, A. *Distill* https://doi.org/10.23915/distill.00014 (2019).
13. Katz, Y. Preprint at https://doi.org/10.2139/ssrn.3078224 (2017).
14. Stark, L. *Soc. Stud. Sci.* **48**, 204–231 (2018).
15. boyd, d. & Crawford, K. *Inform. Commun. Soc.* **15**, 662–679 (2012).
16. Elish, M. C. & boyd, d *Commun. Monogr.* **85**, 57–80 (2017).
17. Benthall, S. & Haynes, B. D. in *Proc. ACM Fairness, Accountability, and Transparency Conference (FAT*)* 289–298 (ACM, 2019).
18. Bowker, G. C. & Star, S. L. *Sorting Things Out: Classification and Its Consequences* (MIT Press, 2000).
19. Benjamin, R. *Race After Technology: Abolitionist Tools for the New Jim Code* (Polity Books, 2019).
20. Stark, L. *XRDS Crossroads* **25**, 50–55 (Spring, 2019).
21. Daniels, J., Nkonde, M. & Mir, D. *Advancing Racial Literacy in Tech: Why Ethics, Diversity in Hiring and Implicit Bias Trainings Aren't Enough* (Data & Society's Fellowship Program, 2019).
22. Wagner, C., Garcia, D., Jadidi, M. & Strohmaier, M. in *The International AAAI Conference on Web and Social Media* 454–463 (AAAI, 2015).
23. Richardson, R., Schultz, J. & Crawford, K. *NYU Law Rev.* **94**, 192–233 (2019).
24. Metcalf, J. et al. *Medium* https://medium.com/pervade-team/the-study-has-been-approved-by-the-irb-gayface-ai-research-hype-and-the-pervasive-data-ethics-ed76171b882c (2017).
25. Back, L. *The Art of Listening* (Berg, 2007).
26. *Nature* **562**, 7 (2018).
27. Howard, D. & Irani, L. *in Proc. 2019 CHI Conference on Human Factors in Computing Systems* 97 (ACM, 2019).