NeurIPS 2022 Workshop on AI Cultures

Cultural Work in 'AI Ethics and Society' Pedagogy: Some Early Reflections. By Maya Indira Ganesh and Rune Nyrup

Many articles about AI, algorithmic bias and fairness, and AI Ethics, will end with some version of the following: "What if our response to bias in AI wasn't just to fix the computers, but the society that trains them?" This sentence is from an essay by digital media theorist Ethan Zuckerman about the field-shaping work of his then-doctoral student, Dr. Joy Buolamwini, that demonstrated how facial recognition systems amplify pre-existing racial and social discrimination.[1] Scholarship by Cathy O'Neil, Safiya Noble, Virgina Eubanks, and Ruha Benjamin, among others, have brought Zuckerman's-and Buolamwini's- point to broader public attention in the US, UK, and Europe. Scandals such as the Cambridge Analytica revelations, and the circumstances around the very public termination of Dr. Timnit Gebru and Dr. Margaret Mitchell from Google's Ethical AI Lab have served as flashpoints. Thus there has been a recognition in the discourse of 'AI ethics' as the publication of ethics principles, that correcting, arresting, improving, shaping, or otherwise transforming the course of AI technologies requires a critical engagement with the political-economic and social relations underpinning it. In other words, correcting algorithmic bias and making AI ethical is contingent *inter alia* on the acknowledgment that computers and computation are: always already social; cultural artefacts with distinctly situated histories and politics; neither neutral, nor deterministically 'acting' on the world; socio-technical. Yet, these ideas are not widely recognised nor legible across multiple academic fields and disciplines that are studying and shaping this technology as being about 'ethics'. However, there are people to whom this *is* legible as matters of ethics; some of them have enrolled for a Master's level (MSt)[2] program on AI Ethics and Society that we teach. This paper shares early reflections from our teaching practice, about how advancing a 'critical engagement' with AI and ethics requires work in negotiating the cultural differences within and across academic disciplines, and outside it. We begin by situating our pedagogy in 'AI ethics', and then briefly sketch out dimensions of cultures of translation and transformation.

'AI ethics' education

There are 'upstream' initiatives to amplify 'AI ethics' concerns in Big Tech, entrepreneur, and government ecosystems, such as tech journalism, 100 Women in AI Ethics List[3], or ethics

[1] Zuckerman, E (2022). How to make better algorithms: start with the people who train the machines. *Prospect Magazine* 27 January 2022, https://www.prospectmagazine.co.uk/science-and-technology/how-to-make-better-algorithms-start-with-the-people-who-train-the-machines

[2] A MSt is a Master's level program usually offered in a continuing education context and distinct to the Universities of Oxford and Cambridge. The first author is the co-lead of this MSt, along with Dr. Jonnie Penn and Dr. Henry Shevlin who share in the work of shaping this course, and these reflections. The second author, Dr. Rune Nyrup, teaches on the MSt, marks assignments, and is involved in developing other AI ethics educational programmes. Both authors work at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge, which is the intellectual hub of the MSt. The University's Institute for Continuing Education manages the administrative side of the MSt.

[3] https://womeninaiethics.org/the-list/

toolkits for tech founders[4]. There are 'downstream' initiatives focused on educating undergraduate students, such in Computer Science or Engineering who, it is imagined, will someday swim upstream and join industry or government.[5] Germane to our provocation on the translation required between different disciplinary and pedagogic cultures, is the finding that undergraduate CS/Engineering students struggle with styles of thinking and learning in the Humanities, which offer skills in tolerating - even generating - contradictions associated with the human condition.'Authority-based' knowledge domains, education for solutionist problem solving, and universalisms in a culturally plural world governed by 'Universal' Declarations of Human Rights, are sites of pedagogical challenge in AI ethics education.[6]

There might also be a 'mid-stream' approach, like the MSt we teach: a two-year, remotely-taught, part-time continuing-education program launched in September 2021 with an inaugural cohort chiefly from the UK, with a small number from other countries. Our students work in Big Tech, government, law, military start-ups, public policy, retail and regulatory banking, finance, among other sectors. This MSt centres the political-economic, moral, historic, and social relations mediated by AI technologies, imaginaries, and industries. Concepts like transhumanism, long term AGI risk, and AI consciousness are sites of serious intellectual engagement, as are Crip Technoscience and gig workers' rights. Rather than 'train AI ethicists', or advance techniques to 'de-bias' algorithms, we encourage agnosticism and ambivalence about these concepts. Not everyone we teach shares our ambivalence, however. And, fully cognizant of the limits of 'AI ethics', and its "uselessness"[7], what unites us as teachers is a commitment to supporting our students to learn how to navigate their doubts and despair about unethical AI.

There is also a 'drive thru' approach to AI Ethics education: some universities offer shorter executive-level courses from a few days to a few weeks long, such as a masterclass in the ethics of AI[8], or in a distinct application, such as AI in healthcare.[9] These courses are opportunities for working professionals to get exposure to the topic, understand the technology and liabilities, and frameworks for the safe and ethical deployment of AI.

Cultural work in 'AI Ethics' pedagogy

Our experience of teaching and designing this MSt has generated the following reflections on the kinds of cultural tensions and work required to bring a critical perspective to the study and shaping of AI as an ethical technology.

Cultures of translation across disciplines. The Humanities and Social Sciences (HSS) are not a uniform nor singular 'perspective' that might serve as a corrective to CS, a discipline

---

[4] https://omidyar.com/news/omidyar-network-partners-with-institute-for-the-future-to-launch-the-ethical-operating-system-a-guide-to-anticipating-the-future-impact-of-todays-technology/

[5] Casey Fiesler's live document demonstrates the extensive nature of university ethics education for undergraduate level Computer Science students ethics. The Mozilla Foundation has awarded grants to universities to develop 'Responsible Computer Science' education.

[6] Burton, E., Goldsmith, J., Mattei, N. (2018) How to teach computer ethics through science fiction. *Commun. ACM* 61, 8 (August 2018), 54–64. https://doi.org/10.1145/3154485

[7] Munn, L. The uselessness of AI ethics. *AI Ethics* (2022). https://doi.org/10.1007/s43681-022-00209-w

[8] https://www.lse.ac.uk/study-at-lse/online-learning/courses/ethics-of-ai-masterclass

[9] https://www.hsph.harvard.edu/ecpe/programs/ai-for-health-care-concepts-and-applications/

assumed to resist complexity or abstraction. Our work is in fact shaped by frictions *within and across* the organisation of knowledge practices in HSS. The MSt is oriented towards a slow transformation in the imaginaries, social relations, and political values associated with AI. It seeks to influence communities of practitioners to bring a critical, practical, measured approach to how AI might be designed and adopted in society. How we go about this is a challenging and complex ongoing struggle for us as educators and researchers from different disciplinary cultures and traditions. We actively foreground conflicts in our (the three course co-leads') values, histories of disciplinary enculturation, political affiliations, and differing levels of uneasiness about the epistemes and commercial products emerging through AI. So substantial work must go into translating the disciplinary cultures we bring to 'the Humanities'- terms, methods, approaches to knowledge and scholarship - in order to be legible to each other, and to our students. The curriculum reflects our own research interests and training across History, Analytical Philosophy, Cognitive Neuroscience, Behavioural Economics, Media and Cultural Studies, and Science and Technology Studies. This strains how we select materials for study, set and mark essay questions, how we develop syllabi that are resonant with each other's, and bring a coherent, critical identity to our approach to 'AI ethics'. Hence, the work of AI Ethics involves unpacking and reconciling with the structure and divisions that comprise knowledge-making and its political economies within the contemporary University.

'*Tech translator' culture.* The 'tech translator' in software development contexts refers to people who perform product or client management roles negotiating approaches and values between software and product development, and clients. We find that our students attempt to do this in their workplaces as a response to exposure to new ideas and perspectives on the MSt. They share their written assignments, organise reading groups in their organisations, or give talks on topics discussed in the classroom. They complain about paywalled academic work that they believe their colleagues will want to read. We understand this to be a challenge of translation related to the different institutional cultures that our students are immersed in both inside and outside the academy. There is the translation work we do-and our students do-in making critical academic texts accessible and legible to non-academics; then, our students must translate those ideas back into academic written assignments on which they are marked, by academics, to attain their Master's qualification. The significant work by academics in advancing a critical engagement with AI is inaccessible at the level of language and ideas, as well quite literally in terms of paywalled scholarship.

Cultures of transformation. Some proponents of 'AI Ethics' assume that 'ethics' will somehow magically 'occur' to improve the rollout of AI, usually as a direct, top-down 'application' of something a philosopher determines. But, the transformation of AI into ethical technology requires work by numbers of everyday people in industry, society, and government, within supportive, or hostile, organisational cultures. 'AI Ethics' neither imagines these individuals nor thinks about their futures as people experiencing personal and collective transformation through a course of study; or how they will transform their organisations. For instance, one of our students working in a leading UK retail bank has engaged his colleagues in the Data Protection team about their use of algorithmic nudges based on an essay he wrote for the course; he is now trying to negotiate ways around the limited advertising budget that necessitates such nudges. The resources required to sustain and build community to support this individual and others like him in our inaugural cohort are limited.

Conclusion

"Critical technical awakenings", by Malik and Malik (2021)[10] is inspired by the legendary life and work of Phil Agre, and most notably, his essay, *Towards a Critical Technical Practice: Lessons Learned in Trying to Reform AI*.[11] Central to the Maliks' paper, and to Agre's, is that a critical consciousness or 'awakening' is a recognition that the world and human social relations, as flawed as they are, are not problems to be framed in terms legible to, nor solved or improved on by computational formalisation and abstractions. This is antithetical to 'AI Ethics', a project that is largely business ethics, and a means to advance AI with minimal oversight. To make these points as critical arguments in multi-disciplinary teaching is neither easy nor straightforward, and requires the kinds of cultural work outlined here. We hope to advance and deepen these ideas through future situated pedagogic research and practice.

[10] Malik, M. and Malik, M. (2021) Critical technical awakenings, *Journal of Social Computing* ISSN 2688-5255 06/06 pp365−384 Volume 2, Number 4, December 2021 DOI: 10.23919/JSC.2021.0035

[11] Agre, P.E. (1997) Toward a critical technical practice: Lessons learned in trying to reform AI, in *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work*, G. Bowker, L. Star, B. Turner, and L. Gasser, (Eds) Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc pp. 131–157.