AI, education and ethics – starting a conversation

Neil Selwyn
Monash University, Melbourne

**

This short piece is a foreword written for:

Holmes, W. and Porayska-Pomsta, K. [eds] (2022). *The ethics of artificial intelligence in education: practices, challenges, and debates*. Routledge

**

Regardless of how 'inter-disciplinary' this field might appear to the casual observer, academic discussions of AI remain beset by intellectual schisms and fundamentally different beliefs. On one hand, those working on AI design and development understandably feel that only they have adequate understanding (let alone direct experience) of the technical complexities and underpinning computational theory involved. From this perspective, it can seem that AI innovation has begun to attract unwarranted scrutiny from 'uncredentialled commentators' and 'non-experts' who peddle "a curious 'counter-hype' of critical stances towards AI" (Galanos 2019, p.421).

In contrast, we see mounting frustration amongst critics working in the humanities, arts and social sciences who consider themselves more attuned to the complex social contexts within which AI technologies are implemented. From this perspective, then, it might be reasoned that "the AI community suffers from not seeing how its work fits into a long history of science being used to legitimize violence against marginalized people, and to stratify and separate people" (Chelsea Barabas, cited in Van Noorden 2020, p.358).

All told, a distinct sense of 'them and us' continues to pervade academic discussions of AI, especially as AI tools move out of R&D phases and into 'real world' contexts. This inevitably provokes distinctly different claims and counter-claims - all rooted in very specific approaches to making sense of what 'AI' is, and contrasting understandings of how AI technologies become part of everyday life.

On one hand, these very different outlooks reflect distinct differences in ontology – in other words, how one sees the existence of the social world. As might be expected, there are plenty of computer science researchers and AI developers who immerse themselves in the immediate challenges of improving the functionality of "technologies in use today" (Krafft *et al*. 2020, p.72). This leads to pragmatic expectations that AI technologies can function in social settings in ways that are

broadly quantifiable, calculable and capable of operating effectively given the correct inputs (Wajcman 2019).

In contrast, are many others who feel that these assumptions do *not* extend to the complex social contexts within which AI technologies are now being implemented. From this point of view, any 'real-world' implementation of AI is severely compromised by the 'problem-solving mindset' that prevails within the computer sciences (Berendt 2019). From this point of view, we are seeing growing calls to imbue AI development with a heightened sense of the social, political and cultural dimensions of this work.

So, where does this leave our own specific interest in AI and education? Is education simply another domain in which AI experts and their critics are destined to continue talking at cross-purposes? While it might be argued that these different mindsets and belief systems are defining elements of what is means to work in 'STEM' as distinct to 'HASS', there is clearly need for increased dialogue and mutual understanding – especially as AI systems and processes begin to pervade mainstream education systems and settings.

Thus, while educational AI will undoubtedly continue to be a hotbed of technically-focused problem-solving, we need to take seriously Bettina Berendt's (2019) point about the dangers of seeing AI in purely computational terms. The recent push for more discussion of 'AI ethics' is one such response to this predicament – perhaps offering common-ground on which different academic factions involved in the educational implementation of AI can meet.

AI ETHICS - A CONVERSATIONAL STARTING POINT

This book is a good opportunity to bring about such cross-disciplinary encounters and conversations. As many of its chapters illustrate, considerable progress has already been made in crafting a sense of what AI ethics might look like with regards to education. For example, there are burgeoning discussions of educational AI in terms of privacy, explicability, respect for human autonomy, and so on. This has translated into emerging debates over how educational issues fit with broader discussions of 'fairness, accountability and transparency', 'trustworthy AI', 'humane AI', and so on.

That said, this book also reminds us that these are not discussions that can be wrapped up quickly, or perhaps *ever* addressed to everyone's complete satisfaction. The coming-together of AI, education and ethics immediately raises tricky questions and confronting ideas. In short, AI ethics is not a topic that can be entered into half-heartedly!

Indeed, unlike many other discussions around AI, talk of 'AI ethics' will quickly push many of us into unfamiliar and unsettling territory. When applied to a real-life setting such as education, questions of ethics are inherently normative – concerned with developing shared principles of "how we should live and what we morally ought to

do" (Driver 2005, p.31). These are complex questions of what should be considered 'good' or 'bad' conduct, what constitutes 'right' or 'wrong' behaviour, and similar value judgements. These are tricky negotiations over what we think should (and should not) be done, what is collectively acceptable … and what is not.

So, this book needs to be read as a starting point for a number of different conversations around education and AI that hopefully will evolve over the next few years. Nothing that is written here will provide a definitive guide to how anyone might be able to best 'do' ethics. Instead, this book should provoke more questions than it will provide answers, and raise more contentions than conclusions. It is perhaps best to approach these chapters in a spirit of problem-raising rather than problem-solving. To get things going, then, here are three preliminary sets of contentions …

i) AI ethics are not clearly defined and easily 'fixed'

Despite everything I have just written, it might still seem tempting to hope that AI ethics is something that can be neatly bundled up and dealt with – ideally in a similar manner to how matters of 'ethics' are actioned in domains such as medicine, journalism and business. This might be described as a matter of applied ethics – i.e. identifying how to practically apply moral questions and normative judgements to the development and application of AI in education.

There are certainly many people in the AI community pursuing this line of thought - framing ethical issues in terms of technical challenges that can be addressed through better design and development of AI. This mindset is evident in current enthusiasms for notions such as 'privacy by design', or addressing complex issues of social bias and discrimination through 'correcting' statistical bias, under-representation and variance in datasets

At first glance, such efforts might well seem to be an effective pragmatic way to address a tricky problem. Yet, the moral conundrums and challenges that underpin AI ethics in education are obviously not *wholly* reducible to sets of discrete procedural challenges that can be codified and then 'solved' through better design and programming. As Brent Mittelstadt (2019, p.505) puts it, "the risk is that complex, difficult ethical debates will be oversimplified to make the concepts at hand computable and implementable in a straightforward but conceptually shallow manner".

As such, when talking about AI ethics and education we need to remain sceptical of 'technical fixes' that convey promises of being able to engineer achievable 'ethical' action. We certainly need to remain vigilant for surface-level responses that slip into corporate obfuscation and 'ethics-washing' (Wagner 2018). This has certainly proven the case with Big Tech efforts to set up ethics frameworks and ethics boards to no great effect (other than as an attempt to avoid regulation). Instead, as Mittelstadt (2019, p.505) concludes:

> "Ethics is not meant to be easy or formulaic. Intractable principled disagreements should be expected and welcomed, as they reflect both serious ethical consideration and diversity of thought. They do not represent failure, and do not need to be 'solved'. Ethics is a process, not a destination".

ii) AI ethics are not an intuitive matter of doing 'good'

Of course, this framing of ethics as a 'process' of ongoing moral reflection runs the risk of pushing some folk to disengage completely from *any* sort of grounded, systematic approach to engaging with AI ethics. Oftentimes, this sees discussions descend into the presumption that AI ethics might be best tackled through personal intuition and/or individual commitments to working out how one's own work with AI might be aligned with 'good' outcomes. This inevitably leads to efforts that are underpinned by flimsy and unarticulated political assumptions about what 'AI for good' might constitute (let alone the question of whether an unproblematic 'good' might be achievable at all).

At best, this approach falls into what Ben Green (2018) describes as a non-politicized "know it when you see it" approach to deciding what constitutes fair/humane/good AI and data science. In terms of discussions around AI and education, this can result in crude equivalencies such as 'Poverty=Bad' or 'Staying enrolled on a university course=Good".

Of course, deciding what constitutes 'good' involves complex normative judgements, which ideally need to be worked out through sustained dialogue amongst *all* those implicated in any particular technology use. Crucially, this dialogue should be driven by a strong guiding political philosophy. The lack of such a systematic approach and grounding principles means that any identified ethical 'goods' can become dangerous over-simplifications of politically complex and long-contested issues. This form of 'AI ethics' therefore runs the risk of what Green (2018, p.2) describes as blithely "wading into hotly contested political territory" and resulting in contestable (perhaps regressive) actions.

iii. Our conversations around AI ethics need to be inclusive and far-reaching

Both of these approaches (the over-codified and the over-vague) relate to a third important point of contention. In short, we need to call out the tendency for discussions of AI ethics to be driven by already privileged and dominant voices. Talk of AI ethics rarely originates from the people and groups who are *most* disadvantaged by AI. Instead, discussion of 'AI ethics' to date has been something of a closed shop – dominated by those who are already invested in (and advantaged by) AI.

This leads to narrow and unimaginative discussions about what AI ought to be. For example, as Kate Crawford (2021) observes, discussions tend to focus on idealised ethical *ends* for AI, rather than more messy questions over the actual *means* through

which AI might be applied in an ethical manner. This also leads to a limited set of ideas about how AI should be overseen and how key protagonists might be held accountable. For example, we continue to see prominent calls for industry 'self-regulation' that hold little weight in the face of Big Tech actors for whom multi-million dollar fines are accepted as minor collateral damage. As such, AI ethics codes and guidelines are rarely reinforced by effective mechanisms to ensure that companies and/or their employees actually adhere to stated principles or else are held accountable for any transgressions.

We also need to call out is the tendency for AI ethics to be discussed almost exclusively in terms of Western European and North American understandings of ethics – thereby overlooking philosophical traditions from non-Western contexts. There is much to learn from viewing AI development through the lenses of Buddhist ethics, moral thinking in the Chinese/Daoist tradition, views on ethics from Persian, African and Indian thought. Similarly, framing AI through Indigenous knowledges opens up numerous different ways of thinking about how AI and education might come together (or not).

Above all, discussions of AI ethics need to steer well clear of any belief that these are novel issues and concerns that are best dealt with by 'AI experts'. Instead, when talking about 'AI ethics in education' we are primarily talking about educational ethics and societal ethics – issues and debates that have engaged diverse groups and communities for centuries. The implementation of AI in education is inevitably entangled with long-standing ethical and political dimensions of educational professions, processes and practice. In short, "AI Ethics is effectively a microcosm of the political and ethical challenges faced in society" (Mittelstadt 2019, p.505).

CONCLUSIONS

All told, 'AI ethics' is a useful starting-point (rather than obvious end-point) from which to advance: (i) the commitment that many AI developers have to making better products, as well as (ii) address the serious concerns that critics are now raising around AI and education. This book's interrogation of 'AI ethics' should be seen as an opening gambit that will hopefully lead on to more complex conversations. As argued earlier, AI ethics is not something that can be wrapped up quickly, or ever satisfactorily decided upon and concluded. Engaging with ethics is a morally reflective process that needs to be ongoing.

Crucially, these conversations around AI, education and ethics should be framed by explicit sets of values, and ready to embrace the politics of negotiating between competing perspectives, goals, and agendas. Indeed, many of the most critical issues surrounding the (mis)use of AI in education are profoundly political in nature, and entangled with broader issues of power, disadvantage and marginalisation (see Verdegem 2021). The forms of AI that are beginning to pervade education, and the infrastructures they are embedded in, all "skew strongly toward the centralisation of power" (Crawford 2021, p.223). As such, it could be argued that AI can never be engineered to be completely 'fair' or 'democratised'.

As such, I hope that this book (and others that follow) act as a catalyst for collective discussions of how the educational AI community is co-engaged in political action that has varying impacts on different groups of people in various educational contexts. This might not sound like an attractive proposition for any technically-minded innovator looking to develop spectacular 'AI solutions' capable of transforming education. Instead, pursuing educational AI along more ethical lines requires considerable time and effort, and a considerable amount of deliberation, debate, dialogue and consensus building. All of this implies replacing ambitions of 'scaling-up' with a commitment to slowing-down. This book takes a great initial step in the right direction.

REFERENCES

Berendt, B. (2019). AI for the common good? *Paladyn, Journal of Behavioural Robotics*, *10*(1), 44-65.

Crawford, K. (2021). *Atlas of AI*. Yale University Press

Driver, J. (2005). Normative ethics. *The Oxford handbook of contemporary philosophy*. Oxford University Press (pp.31-62)

Galanos, V. (2019). Exploring expanding expertise. *Technology Analysis & Strategic Management, 31*(4), 421-432.

Green, B. (2018) *Data science as political action.* https://arxiv.org/pdf/1811.03435

Krafft, P., Young, M., Katell, M., Huang, K. and Bugingo, G. (2020). Defining AI in policy versus practice. in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 72-78).

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501-507.

Van Noorden, R. (2020). The ethical questions that haunt facial-recognition research. *Nature*, 20th November, www.nature.com/articles/d41586-020-03187-3

Verdegem, P. (2021). *AI for Everyone?* University of Westminster Press

Wagner, B. (2018). Ethics as an escape from regulation. In Bayamlioglu, E. *et al.* (eds) *Being profiled* (pp. 84-89). Amsterdam University Press.

Wajcman, J. (2019). The digital architecture of time management. *Science, Technology and Human Values*, *44*(2), 315-337.