The Overdemandingness of AI Ethics

Susan Dywer, University of Maryland (United States)
International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: Ethics, trust, ethical concepts

Extended abstract

Familiar normative theories like varieties of consequentialism and principle-based ethics have faced one or another kind of "overdemandingness" objection (Chappell, 2009; Murphy, 2000). On the one hand, maximizing good for the greatest number seems to require that we discount attending to legitimate personal interests (e.g., refraining from killing a sibling, even if her death would bring about great good for many others). On the other, strictly adhering to principles or moral rules stated in absolutist terms (e.g., 'Do not lie.') can require a too-narrow range of moral attention (e.g., being kind to my terminally ill grandfather is more important than telling him an uncomfortable truth). More generally, there is the interesting question of whether all normative theories ask us to be "moral saints", a state of being obviously inconsistent with the life of an ordinarily decent human being (Wolf, 1982). And finally, there is the Voluntarist Principle ('ought implies can'): if we are genuinely unable (physically, psychologically, or in some other way) to do some action A, then we cannot be morally obliged to do A (King, 2019).

In this paper, I argue that a host of current ethical demands on AI and its deployment violate the Voluntarist Principle: in short, it is not possible, e.g., to ensure the trustworthiness of deep learning-based decision systems, or to render the operations of deep-learning models transparent in any way that the vast majority of ordinary people would understand. Hence, there can be no moral obligation to do either of these things.

I'll suggest that part of what's going wrong with current AI ethics is that it appears to hold AI (here encompassing the models themselves and those who design and deploy them) to moral standards higher than those to which we hold each other in ordinary human life (see Bryson 2020). Some might suggest that this is as it should be, since AI is, in a variety of ways, more powerful or impactful than ordinary human agency. But if this is the case, why haul in ethical concepts, principles, or theories designed for less powerful human beings? There seems to be a mismatch here.

So, should we abandon the traditional ethical theories, principles, and concepts that were never designed (insofar as we can talk of design here) for artifacts like AI? And does this mean that there can be no AI ethics?

It seems to me there are two routes here. The first, which I think would be a *very* bad idea for AI, is to adopt the path taken over the last two decades in the field of business ethics. There we see the abandonment of traditional normative theory, precisely because it is seen as being too demanding. Better, the business ethicists say, to think in terms of "behavioral ethics" (Bazerman 2022). Let's not worry too much about intentions and complicated human motives and expect employees to deliberate about how they behave. Instead, let's focus on overt behavior and build compliance departments instead. All that matters is what we see on the surface. If the point of compliance departments is to reduce the likelihood of

a company facing federal fines, criminal complaints or costly civil suits, fine. But, please don't call this *ethics*. (This is the familiar worry about 'ethics washing' (Hao, 2019).)

The second route I'll discuss is that AI ethicists think in systematic ways about how existing ethical concepts might be extended and/or how new ones might be justifiably created. The aim here is to articulate and apply *adequate* ethical concepts in our evaluation of AI. Here is an example from a different arena. The debate about the moral permissibility of abortion has, for the most part, assumed that there is a tension between the rights or interests of two individuals: the pregnant woman has a right to decide what happens in her body, while the fetus has a right to life. It is obvious how unfruitful this approach has been. How might things have been different if more philosophers and others had taken seriously the fact that the 'relationship' between the pregnant woman and the fetus is utterly unique? They are not like two adult individuals with competing interests. In this sense, the individual rights approach is inadequate to address the issue (Little 1999).

Similarly, why think that the moralized notion of trust, so widely appealed to in AI ethics, is adequate to the task of addressing the risks and moral dangers that some critics see? Is there a notion richer than mere reliability, but weaker than trust, that would do a better job? And when it comes to all the questions about the moral status and moral agency of robots, why think we need to ask 'are they persons?'

References

Bazerman, M. 2022. *Complicity: How We Enable the Unethical and How to Stop*. Princeton: Princeton University Press.

Bryson, J. 2020. The Artificial Intelligence of the Ethics of Artificial Intelligence. In *The Oxford Handbook of Ethics of AI*. New York: Oxford University Press, pp. 3-25.

Chappell, T. 2009. *The problem of moral demandingness*. New York: Palgrave MacMillan. Hao, Karen, 2019. In 2020, let's stop AI-ethics washing and actually do something. *MIT Technology Review*, December 27.

King, A. 2019. *What We Ought and What We Can*. New York: Routledge. Little, M. 1999. Abortion, Intimacy, and the Duty to Gestate. *Ethical Theory and Moral Practice* 2(3): 295-312.

Murphy, L. 2000. *Moral Demands and Non-Ideal Theory.* New York: Oxford University Press. Wolf, S. 1982. Moral Saints. *Journal of Philosophy* 79(8): 419-439.