Against the Double Standard Argument in AI Ethics

Scott Hill

(This is a draft. Please cite the final version which is forthcoming in *Philosophy & Technology*.)

Introduction

In an important and widely cited paper, Zerilli, Knott, Maclaurin, and Gavaghan (2019) (hereafter ZKMG) argue that the concern that opaque AI should be more transparent is mistaken. ZKMG's argument is this:

The Double Standard Argument

(1) Opaque AI decision makers are at least as transparent as human decision makers.
(2) If opaque AI decision makers are at least as transparent as human decision makers, then the concern that opaque AI should be more transparent is mistaken.
(3) Therefore, the concern that opaque AI should be more transparent is mistaken.

The motivation for premise (1) is that human decision makers are very opaque. Humans can cite reasons for their decisions after they make those decisions. But often those reasons do not match the actual process by which the human arrived at that decision. They are a rationalization of those decisions after the fact. Intuition and hunches, on which many human decisions are based, are determined by opaque processes. ZKMG show that this is no less opaque than the ways in which the forms of AI in question are opaque.

The motivation for (2) is this: We are fine with human decision makers given their limited transparency. Human decision makers are opaque. But the ways in which they are transparent are good enough to employ them as decision makers. Opaque AI is, or at least can easily be, crafted so that it is at least as transparent as humans. Every dimension along which a human is transparent, Opaque AI can be transparent in that way too. If it is good enough in the case of humans, then it is good enough in the case of AI.

Against the Double Standard Argument

My target is premise (2). I think there is a relevant difference between human decision makers and opaque AI decision makers. It is not that AI fails to be transparent in a way in which humans are transparent. It is instead that AI fails to be opaque in the way that humans are opaque.

Consider figures such as Bostrom (2014) and Muehlhauser (2013) who think AI is an existential threat. No human decision maker will choose to turn the universe into paper clips. But, such figures argue, opaque AI might. The reason it might do so is not that it fails to be transparent in a way that humans are transparent. The reason is instead that in its opaque processes it will find some strange way of fulfilling its instructions that would cause it to choose such an option. But as opaque as human decision makers are, there is nothing in their opaque processes that would cause them to make such a choice.

AI may be able to cite plausible post hoc reasons in its justification of decisions arrived at by processes that are opaque to us. And that may not be any worse with respect to transparency than what we get from humans. But that is cold comfort to the theorists concerned that AI opaque processes are different from human opaque processes in such a way that the decisions arrived at by such processes may well lead to the destruction of humanity. Indeed, the ability of opaque AI to be 'transparent' in this way, and to offer explanations of its decisions that are post hoc and plausible seeming, is one of the features of AI that troubles such theorists. The real decision making processes will float free from concerns about human welfare. And

the ability of AI to be 'transparent' in the way humans are will disguise those aspects of the processes that would otherwise alarm humans.

Of course, there is room to be skeptical that AI really poses this sort of danger. Maybe Bostrom and Muehlhauser are mistaken. But merely pointing out that AI is transparent in the ways humans are doesn't do anything to engage with the concerns of people who are calling for AI to be more transparent. Humans having the degree of transparency they do is fine because the opaque processes that generate human decisions will not result in the decision to turn the universe into paper clips. But the very same degree of transparency in AI, if people like Bostrom and Muehlhauser are correct, likely will. Higher transparency is called for in AI because the opaque processes of AI are more dangerous than the opaque processes of humans.

ZKMG cite Muehlhauser (2013) as a figure who is concerned about ensuring that AI is transparent. And they note that even he recognizes that human decision processes are opaque. They quote him as saying:

> We can observe its inputs (light, sound, etc.), its outputs (behavior), and some of its transfer characteristics (swinging a bat at someone's eyes often results in ducking or blocking behavior), but we don't know very much about how the brain works. We've begun to develop an algorithmic understanding of some of its functions (especially vision), but only barely.

Notice that, although the processes involved in the example ZMKG use to illustrate the way in which human decisions are opaque, such processes are stable, uniform across humans, and safe. Such processes will not result in a decision that leads to the destruction of humanity. Muehlhauser shares[1] Bostrom's concerns[2] about AI posing an existential threat. And he thinks making AI transparent is one way to help mitigate that threat. So to address the concerns of figures like Muehlhauser, it is not sufficient to simply show that opaque AI can offer post hoc rationalizations for decisions rendered by opaque processes just as humans can. It is instead necessary to show that the opaque processes in such AI are not dangerous in the way that such figures fear.

Furthermore, not all concerns about AI transparency are so high flying. There are concerns that are more down to earth as well. Consider people who think AI threatens to exacerbate current discrimination. The worry is that opaque AI generates new ways of promoting racism, sexism, and other forms of discrimination as well as new ways to hide such discrimination. Humans will continue to be racist in predictable ways. And that will be due in part to opaque processes that guide how humans make decisions. But AI will create new ways of being racist. And it will create new ways to hide that racism. And that will be due to the opaque processes that guide how AI makes decisions. AI may be just as transparent as humans in citing reasons for a decision. But in its opacity it hides new ways of supporting and maintaining prejudice that are not present in human opacity. Addressing concerns about racial bias in opaque artificial intelligence, ZKMG (2019, p. 673) say:

> While we do not deny these forms of bias, again we are not convinced that they are unique to AI…. Hence, once again the existence of machine bias on its own cannot justify the imposition of a higher standard of transparency for AI. Standards for machines taking the form of mandated software upgrades and maintenance procedures would be analogous to mandatory continuing education programs for professionals and would probably solve the clinical diagnostics problem which Mittelstadt et al. (2016) also cited. A consistent standard of transparency across the board is therefore possible in principle and seems reasonable in the circumstances.

---

[1] https://lukemuehlhauser.com/replies-to-people-who-argue-against-worrying-about-long-term-ai-safety-risks-today/

[2] https://lukemuehlhauser.com/treacherous-turns-in-the-wild/

Insofar as the worry about opaque AI is just that it will be racist in the same way humans are racist goes, this is a powerful reply. But the deeper worry is that AI will take our racism and run with it. It will invent new and more deeply hidden ways of being racist. And merely pointing out that the opaque processes in humans result in racist decisions does not address this deeper worry. As Di Bello and Gong (forthcoming, p. 3) put it:

> One might conjecture that, if distortions in the data and injustices in society were eliminated, predictive algorithms should no longer be cause for concern. But this conjecture would be premature. Predictive algorithms can still be the target of what we might call an inner critique. This inner critique stems from a number of theorems in the computer science literature about the impossibility of algorithmic fairness.

As Milano and Prunkl (forthcoming, p. 1) put it:

> We show how algorithmic profiling can give rise to epistemic injustice through the depletion of epistemic resources that are needed to interpret and evaluate certain experiences. By doing so, we not only demonstrate how the philosophical conceptual framework of epistemic injustice can help pinpoint potential, systematic harms from algorithmic profiling, but we also identify a novel source of hermeneutical injustice.

So the worry is that opaque AI will generate new forms of unfairness. And that worry is not addressed by pointing out that opaque processes in humans are unfair too.

Conclusion

The way in which human decision makers are opaque is stable and uniform across humans. The way in which AI decision makers are opaque is unstable and diverse. The opaque decision making processes in humans evolved over a long time, in similar environments, and are relatively fixed across humans. Opaque AI evolves rapidly, in many different environments constituted by very different data sets, and is programmed in different ways with different goals. If there was one kind of opaque AI, if it evolved as slowly as us, if we had observed it over a long time and its opaque processes were as predictable as humans, then perhaps there would be no relevant difference between AI and humans decision makers. If there were new and diverse kinds of humans that evolved as fast as AI and that evolved in as many different environments with as many different starting instructions as AI and whose opaque processes might be as indifferent to human concerns as AI's may be, then we should not entrust our decisions to such humans. In that case there would be no relevant difference either. So there is no double standard. The requirement for transparent AI is not generated by the fact that opaque AI is not human. What matters is the degree to which the opaque processes of a class of decision makers are stable, uniform, and safe. The degree to which such processes have these features in humans is higher than the degree to which such processes have these features in opaque AI. Therefore, AI should be held to a higher standard of transparency than humans.

References

Bostrom, N (2014) *Superintelligence: Paths, Dangers, Strategies* Oxford University Press

Di Bello, M and Gong, R (forthcoming) Informational Richness and its Impact on Algorithmic Fairness *Philosophical Studies*

Milano, S and Prunkl, C (forthcoming) Algorithmic Profiling as a Source of Hermeneutical Injustice *Philosophical Studies*
Muehlhauser, L (2013) Transparency in Safety-Critical Systems. Intelligence.org August 15, 2013. Available at: https://intelligence.org/2013/08/25/transparency-in-safety-critical-systems/.

Zerilli, J; Knott, A; Maclaurin, J; Gavaghan, C (2019) Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy and Technology* 32: 661-683

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and material: Not applicable

Funding: Not applicable