Algorithms, Bias, and the Importance of Agency

Alan Rubel[1], Clinton Castro[1], and Adam Pham[1]

[1] University of Wisconsin, Madison WI 53706, USA

Abstract. We argue that an essential element of understanding the moral salience of algorithmic systems requires an analysis of the relation between algorithms and agency. We outline six key ways in which issues of agency, autonomy, and respect for persons can conflict with algorithmic decision-making.

Keywords: Algorithms, Bias, Agency.

1 Algorithms and agency

The last few years have seen growing interest in the uses, misuses, and biases of automated, algorithmic information systems. One key area of inquiry concerns ways in which algorithms reflect various biases, for example in model choice, by reflecting existing social structures, and by reifying antecedent beliefs. The literature contains a number of arguments regarding how algorithms may cause harm, may discriminate, and may be inscrutable. There has been less scholarly focus on a different moral failing, namely algorithms' effects on agency. That is our focus here.

Consider the 2016 U.S. case of Wisconsin v. Loomis [1]. There, defendant Eric Loomis pleaded guilty to crimes related to a drive-by shooting. The trial judge ordered a presentence investigation report (or "PSI"), which in turn used a proprietary risk assessment tool called COMPAS. This tool is designed to make better decisions in allocating resources for supervision, and the company that developed it specifically warns against using it to make sentencing decisions. Nonetheless, the trial judge used the PSI and COMPAS report in his decision to sentence Loomis in the maximum range for the crimes to which he pled guilty.

Much of the literature about the use of algorithms recognizes that such uses of algorithms may discriminate by race, ethnicity, and gender and that because the algorithms are proprietary defendants cannot scrutinize their effects. But the Loomis case also presents a puzzle. It is plausible that, even though he received a lengthy prison sentence, Loomis was not harmed at all. That is, it is plausible that Loomis received exactly the sentence he would have received had the trial judge never ordered the PSI. Moreover, because Loomis is white, and the algorithm appears to disadvantage black defendants, he likely did not experience discrimination on the basis of race or ethnicity. Nonetheless, Loomis may have been wronged (but not harmed). And it is Loomis that was wronged, not (only) others who suffer discrimination from the use of algorithms. But how so?

The wrong consists not of discrimination or of excessive sentence (at least not on the basis of the algorithm), but of a procedural wrong. Loomis—like anyone facing the criminal justice system—has a claim to understand the mechanisms by which he is subjected to incarceration and to use that understanding to make his case. Denying him that understanding is not a harm in itself (though it may or may not result in a harm), but a failure of respect for him as a person. There are, of course, numerous calls for algorithmic transparency. However, absent an explanation for why transparency matters, the criticism is not well-grounded.

Our contention in this paper is that many algorithmic systems are similar to the Loomis case in that they engender wrongs that cannot be reduced to harms. More specifically, we will argue that a complete picture of the moral salience of algorithmic systems requires understanding algorithms as they relate to issues of agency, autonomy, and respect for persons.

2 Six conflicts between algorithms and agency

First, algorithmic systems may govern behavior or create rules that are not of the sort that any agent is capable of reasonably following. Cathy O'Neil provides the example of a school system using algorithms to evaluate (and fire) teachers, despite their model's inability to distinguish the effects of good (bad) teaching from background noise [2]. The moral upshot (in addition to failing to do anything good for the schools) is that teachers are evaluated according to criteria that no reasonable agent could agree to—and this is true even for those teachers not fired [3].

Of course it cannot be the case that *any* algorithm that causes harm—i.e., makes someone worse off than they would have other been—is unreasonable. After all, people can be made worse off for justifiable reasons. Genuinely ineffective teachers might be made worse off by not having their contracts renewed, and that could justified (assuming there are no other factors that demand renewal). What seems to matter is a combination of the *seriousness* of the harm (losing one's job is a very serious harm), the trustworthiness of the algorithm in the decision (in O'Neil's account, the algorithm appears quite unreliable), and whether one is able to control the outcome (a key problem in the teaching case is that background noise—outside of teachers' control) accounted for much of the evaluation. Having one's livelihood be determined on the basis of an unreliable system in which one cannot exercise substantial control is a state of affairs to which one cannot reasonably agree. Where people *do* agree, it may be evidence of deeper injustices yet.

Second, is the issue of epistemic agency. For a person to have a reasonable degree of agency requires that they know where they stand, regardless of whether they can take action. The basis for this claim is the idea that people are reasoning beings, who plan, their lives, and who think of themselves as members of a broader community. And where we stand in relation to other people and (more importantly) in relation to institutions that exercise power over us, *matters* to us. So, denying a person the ability to understand the reasons why they are treated as they are is a failure of respect for them as agents. This we can see in the *Loomis* case. The COMPAS algorithm is proprietary. Although Loomis (or anyone) can find out some basic information about the algorithm (e.g., its inputs), no one can access the algorithm itself. To the extent that it matters, then Loomis's lack of access prevents him from understanding the basis for how he is treated by the state of Wisconsin.

Third, algorithmic systems can leave individuals with no recourse to exercise what we will call "second-order power" (or appeals) over outcomes. That is, where algorithms are opaque, proprietary, or inscrutable, individuals cannot draw on their epistemic agency (described above) to take action and appeal decisions on the basis of reasons precisely because they are prevented from understanding the underlying reasons for the decision. In Loomis this problem appears as a failure of due process in a criminal case—Loomis cannot explain to the court why the COMPAS score is not (if it is not) appropriate for his case. But the issue arises in other contexts as well, for example in consumer scoring by credit agencies [4].

Fourth, algorithmic systems can fail to respect boundaries between persons. Algorithms can be used to make decisions about workers, such as when employers use algorithms to schedule employees in ways that frustrate their need for reasonable work hours, or about students, such as when advising systems nudge students towards majors based on anticipated success. By necessity, algorithms generalize about individuals, but doing so treats them as undifferentiated collections of work or credit hours. This treatment may fail to account for important aspects of their individual needs as persons for reasonable work hours or course schedules that are a good intellectual fit.

A fifth issue concerns not the agency of those who are affected by algorithms but by those who deploy them. Using algorithms to make decisions allows a person or persons to distance themselves from morally suspect actions by attributing the decision to the algorithm. The issue here is that invoking the complexity or automated nature of an algorithm to explain why the suspect action occurred allows a party to imply that the action is unintended and something for which they are not responsible. So, for example, in late 2017 the news organization ProPublica discovered that Facebook's system for tracking user interests and selling advertisements based on those interests allowed others to purchase ads targeting Facebook based on anti-Semitism [5]. Facebook's response was not to admit that there was a wrong or that they perpetrated a wrong. Rather, it was to point to the fact that the user categories were generated automatically, that the odious ones were only rarely used, and that "[w]e never intended or anticipated this functionality being used this way" [6]. This response serves to mask Facebook's agency in generating categories that can be misused by others using the Facebook platform.

Lastly, there's a bigger and deeper issue about the very nature of agency, autonomy, and individual liberty. In the recent philosophical literature on liberty and freedom, one important thread pushes back against notions of negative and positive liberty (roughly the freedom from constraint and the capability of acting to further one's interests, respectively) [7]. This view maintains that a person's freedom is a function of the quality of their agency, or that their desires, values, and interests are their own.

The recent literature on filter bubbles, fake news, and highly tailored advertising in social media suggests that algorithms are being extensively (and increasingly) used to manipulate persons' choice architectures to generate understandings, beliefs, and motivations that are not persons' own (in some cases, and to some extent). In other words, the concern about filter bubbles, fake news, and tailored advertising is not merely that bad consequences will result (perhaps so, perhaps not). Rather, it is that they diminish quality of agency and, hence, freedom properly understood.

To be clear, concerns about algorithms are many and varied. They include harm and they include discrimination. But we cannot fully understand the moral salience of algorithms without addressing their effects on agency, autonomy, and respect for persons.

3 Discussion

After the presentation of the longer paper based on this abstract at the BIAS workshop, participants raised a number of important points that are worth addressing here.

The first key question is whether criticizing the *Loomis* case on the grounds that an inscrutable algorithm played a role in the sentencing decision proves too much. After all, had the judge in the case simply issued a sentence for the charges to which Loomis pleaded guilty, the ultimate basis for the decision is also inscrutable. The judge can offer reasons—prior convictions, seriousness of charges and "read-in" charges, prior incidents of violation, but those are simply inputs. We likewise know the inputs to COMPAS. In both the algorithm case and in the judge-only case, there is at root an inscrutable process. In the COMPAS case it's the proprietary algorithm, in the judge case it is the psychology of the judge.

Our answer is two-fold. First, it is true that in some sense our argument is not unique to algorithmic decision systems. However, the fact that other inscrutable systems may conflict with agency, autonomy, and respect for persons neither diminishes the concern with respect to algorithms nor treats non-algorithmic systems differently. That is, if our arguments point to ways in which decisions by judges, or administrative agencies, or bureaucracies, are problematic then we should examine those systems more carefully with the concerns in mind.

Our second response is linked to our argument about agency laundering. Although (for example) a judge's psychology is opaque much as COMPAS is, there is at least one important difference between human and algorithmic decision-makers. Specifically, human decision-makers can be *culpable*. Machines, no matter how good their AI, can be causally effective in bringing about outcomes, but they cannot be morally responsible. That is, they are not agents. If COMPAS or another algorithm "gets it wrong" in some case or other, the moral responsibility for getting it wrong falls to the humans and groups of humans that developed and deployed the algorithm. The same is not true for a judge (or other human decision-maker). Had a judge come to a sentencing decision in the *Loomis* case without using COMPAS, she would be accountable to that decision. And the problem of algorithmic decisions is, in part, that their use can launder such exercises of agency. Hence, while inscrutability is a concern that is not unique to algorithms, there are key differences with respect to human decision-makers.

Another participant asked about what should done about the use of COMPAS and similar risk assessment systems being put to use in criminal justice contexts. Unfortunately, we do not have a comprehensive answer. But there are a couple of considerations worth thinking through. One is whether the arguments we make get some of their force from the severe penalties and high incarceration rates in the United States' criminal justice system. The unreasonableness (our first argument) of algorithmic decision systems, as we argue, is partly a function of their stakes. Where the stakes are higher, the ability of agents to reasonably abide a system's decisions diminishes. So, one possibility may be wrapped up in criminal justice reform. Another possibility is that use of such systems demands that agents using the systems recognize that the systems are tools, and only humans can make agential decisions.

A last comment addresses the nature and extent and agency laundering. For example, is the mere reliance on a system to make decisions enough to make it agency laundering [8]. We are currently developing a broader framework for understanding agency laundering, and our full account is beyond the scope of this extended abstract. However, our sense is that mere reliance or delegation is not enough. Rather, an agent has to ascribe some morally relevant quality to the algorithm (neutrality, accuracy, functionality) such that the agent's own role in a decision becomes obscured.

References

1. Wisconsin v. Loomis, 881 N.W.2d 749, 371 Wis. 2d 235 (Wis. 2016).
2. O'Neil, C., Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, pp. 3-10 . Crown, New York (2016).
3. As various political philosophers have noted, a key element of a liberal democratic society--and for institutions within such societies--is that members will only reasonably participate where they can abide fair terms of social cooperation, provided others do as well. See, e.g., Rawls, J., A Theory of Justice. Harvard University Press, Cambridge, MA (1971).
4. Pasquale, F. The Black Box Society: The Secret Algorithms That Control Money and Information, pp. 32-34. Harvard University Press, Cambridge, MA (2015).
5. Angwin, J., Carner, M., and Tobin, A. Facebook Enabled Advertisers to Reach 'Jew Haters'. ProPublica, Sept. 14, 2017. https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters, last accessed 2017/12/30.
6. Sanberg, S. Facebook post, September 20, 2017. https://www.facebook.com/sheryl/posts/10159255449515177, last accessed 2017/12/30.
7. Christman, J. Liberalism and Individual Positive Freedom. Ethics 101(2) (1991).
8. Thanks to Eric Meyers for this point.