Algorithm Evaluation Without Autonomy

Scott Hill

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

Introduction

In *Algorithms & Autonomy*, Rubel, Castro, and Pham (hereafter RCP), argue that the concept of autonomy is especially central to understanding important moral problems about algorithms. I argue that although RCP are largely correct in their diagnosis of what is wrong with the algorithms they consider, those diagnoses can be appropriated by moral theories RCP see as in competition with their autonomy based theory. Most notably, proponents of consequentialism and virtue ethics can appropriate RCPs insights. And so there is no motivation for the controversial claims about autonomy they make. The most significant contribution of RCP, if I am right, is in their identification, presentation, and evaluation of concrete cases involving algorithms and not in the more controversial claims about theoretical ethics that RCP themselves see as central to what they are doing. This is good news for RCP and for the prospects of applying moral philosophy to algorithms because it ends up making the arguments of their book more ecumenical. It frees those arguments from being rooted to controversial moral views and shows that those arguments may be equally grounded in a diversity of views. In the end, different moral theories converge on the same result: The algorithms are bad for the reasons RCP maintain even if such reasons are universal rather than unique to RCPs theory.

The Point of the Book

It is important to get clear on the point of *Algorithms & Autonomy*. The main point is that autonomy is especially relevant for the moral evaluation of algorithms. As RCP (2021, p. 21) put it:

> The central claim of this book is that understanding the moral salience of algorithmic systems requires understanding how they bear upon the autonomy of persons….

They (2021, p. 19) also say:

> …. our view is that autonomy is an important value, and many moral concerns about algorithmic systems are best understood as, at bottom, issues of autonomy.

RCP (2021, p. .53) also suggest that competing theories be tested against theirs in terms of the ability of those theories to explain moral concerns about algorithms:

> And in any case, objecting to our larger project on the grounds that a different kind of social agreement principle better captures autonomy and social cooperation warrants an argument for why it would better explain concerns in the context of algorithmic systems.

In addition, RCP (2021, p. 39) maintain that if one does not accept something in the neighborhood of their views about autonomy and if instead one endorses consequentialism or virtue ethics, one will not find their arguments persuasive:

> So far, we have considered a few different conceptions of autonomy and offered the account we will use to ground the arguments in the remainder of the book…. Consequentialists and virtue ethicists (among others) might argue that other values are the proper measure of moral value. As important as

> those criticisms are, we won't offer a defense here. Rather, we will simply confirm that a rock-bottom assumption of this project is that autonomy is morally valuable, and it is an important enough (and rich enough) value that it can ground the arguments we offer throughout. If one disagrees with that assumption, this project probably won't be persuasive.

So RCP accept various controversial theories about autonomy and moral contractualism. They maintain that proponents of alternative theories will not accept the main arguments in their book. RCP also suggest that other theories be tested in terms of their ability to explain moral concerns about algorithms. Furthermore, RCP say that understanding the morality of algorithms requires their theoretical tools about autonomy. And finally, maintain that moral problems about algorithms are best understood, at bottom, as issues about autonomy.

Background: The Central Cases, the Reasonable Endorsement Test, and Autonomy
While RCP discuss many examples, the book is centrally framed around a few important cases[1]:

> *Loomis and COMPAS*: COMPAS takes as input information about a subject's criminal behavior, history, beliefs, and job skills. It gives as output assessments of pretrial release risk, general recidivism risk, and violent recidivism risk. Eric Loomis was charged with crimes related to an incident involving a shooting and a car theft. While he denied the allegations, Loomis pleaded guilty to two of the charges. COMPAS delivered the judgment that Loomis had a high pretrial release risk, general recidivism risk, and violent recidivism risk. The prosecution encouraged the court to sentence him partly on the basis of Loomis' COMPAS report. And the court referenced the report in its decision to give Loomis the maximum sentence, over a decade in prison, on the two charges to which he plead guilty.
>
> *Wagner, Braeuner, and TVASS*: TVASS takes as input information about student standardized test scores. It gives as output assessments of individual teachers. Many of the teachers evaluated by TVAAS did not teach subjects that were tested by TVASS. So while TVASS had as input information about student test scores in algebra and US history, for example, it did not have as input information about student test scores in physical education and art. In cases in which someone taught a subject not provided as input into TVAAS, such teachers were evaluated on the basis of the average performances of all students on all subjects at their schools. Teresa Wagner taught physical education and Jennifer Braeuner taught art. Wagner and Braeuner were, based on all other measures, consistently excellent teachers. However, one year the composite scores of the students in their schools dropped from the highest possible score to the worst possible score. As a result, Wagner's and Braeuner's individual evaluations dropped significantly. This kept Wagner from receiving a performance bonus and Braeuner from receiving tenure.

RCP diagnose these cases in terms of their moral theory.

> **Reasonable Endorsement Test:** An action is morally permissible only if it would be allowed by principles that each person subject to it could reasonably endorse.

[1] Here I discuss two of three cases that are central to their book. RCP's treatment of the third case is sufficiently like *Wagner, Braeuner, and TVASS* that I do not discuss it here. The third case is diagnosed by RCP in the same way as the other two cases.

Reasonable endorsement is analyzed by RCP in terms of autonomy. Autonomy is the capacity of self government. Certain background conditions in one's environment contribute to the degree to which a person can self govern and therefore is autonomous. RCP give a rich and detailed account of those background conditions. And they argue that the use of algorithms considered above is immoral because such use undermines the background conditions necessary for autonomy. RCP focus on four reasons that these algorithms undermine autonomy.

> *Reliability*: The algorithms need to make accurate predictions and evaluations.
>
> *Responsibility*: The algorithms should not make evaluations or predictions on the basis of things the subject being evaluated has no responsibility for.
>
> *Stakes*: High stakes exacerbate other respects in which algorithms might be problematic.
>
> *Relative Burden*: The algorithms must not impose a burden on one salient group relative to another.

These features of algorithms partly determine the background conditions required for autonomy. And therefore whether those subject to the algorithms could reasonably endorse those algorithms.

A Problem for the Reasonable Endorsement Test

RCP's theory is heavily influenced by Scanlon's (1998) and Parfit's (2011) versions of contractualism. However, there is a well known class of examples, usually applied in other contexts, that is a problem for these theories as well as RCP's theory. Examples of this form have been discussed by Brandt (1967), Copp (1995), Feldman (1978), Parfit (2011), Rosen (2009), Podgorski (2018), Lyons (1965), and Smith (2010). The key feature of such examples is that they are constructed in such a way that if everyone were to endorse a moral principle, something very bad would happen. Here I will introduce an instance of such an example that is relevant for our purposes:

> *AI Overseer*: After a nuclear war, only a small number of humans are alive. An AI oversees them all and takes care of them. It treats the humans well. But it is programmed to prize ideological diversity about morality. It detects which moral principles each person endorses. And it is programmed to punish intellectual consensus about morality. If a consensus in moral principles is reached, then the AI will torture and destroy all remaining humans. With this threat in place, it ensures that humans have diverse perspectives about morality and never arrive at a consensus about any moral principle. For if they were to arrive at such a consensus, the humans know the AI would torture and kill them all.

In this case, it seems like some actions would be morally permissible. It would be permissible, for instance, for one person to give another a hug. And yet, there are no moral principles that everyone could reasonably endorse according to which such an act is permitted. For if everyone were to endorse such a principle, the AI would torture and destroy humanity.

The point of examples like this is that, although what is right or wrong might correspond in many cases with what reasonable people would endorse, such counterfactual endorsement isn't a plausible account of the essence of morality. An action may be permissible even if, because of strange features of how the world happens to be, not everyone could reasonably endorse principles that permit it. Whether an act is permissible is one thing. Whether everyone could reasonably endorse principles permitting that act is another thing. There may be significant overlap in many actual cases. But the two should not be identified as one. Another way to put the point, even if the ability to reasonably endorse a moral principle is removed, morality remains.

RCP's Diagnoses of the Central Cases

Insofar as RCP's diagnosis of the algorithms in question depends on the truth of the Reasonable Endorsement Test, I think it will not succeed. Fortunately, I think RCP's diagnoses can be decoupled from the truth of that test. To understand how, it is important to look at RCP's diagnoses of the central cases. With respect to TVASS they (2021, p. 58) say:

> Wagner and Braeuner frame their case in terms of harms (losing a bonus, precluding tenure consideration, and so forth), but those harms matter only because they are wrongful. They are wrongful because TVAAS is an evaluation system that teachers could not reasonably endorse. Wagner and Braeuner's scores did not reliably track their performances nor did the scores reflect factors for which they were responsible, as the scores were based on the performance in subjects Wagner and Braeuner did not teach. And the stakes in the case are fairly high (there were financial repercussions for Wagner and job security for Braeuner). So, per our account, they were wronged.

With respect to COMPAS RCP (2021, p.60-5) say:

> researchers associated with Northpointe assessed COMPAS as being accurate in about 68 percent of cases. More important is that COMPAS incorporates numerous factors for which defendants are not responsible…. Further, the use of COMPAS in Loomis is high stakes. Incarceration is the harshest form of punishment that the state of Wisconsin can impose…. COMPAS… imposes a greater relative burden to Black defendants than to White defendants, it is one that at least some defendants cannot reasonably endorse.

The algorithms are not very accurate. They don't do a very good job of evaluating teacher performance or predicting recidivism. The algorithms judge people for things they have no responsibility for. They assess teachers on the basis of how students perform on tests unrelated to the subjects those teachers teach. They predict recidivism on the basis of whether the defendant is adopted. The stakes of output of the algorithm are high stakes. The outputs determine how long one goes to prison, whether one gets a promotion, and whether one can receive a loan. The algorithms impose a burden on one salient group relative to another. COMPAS makes false predictions about recidivism among black defendants at a higher rate than it makes false predictions about white defendants. These features of the algorithms damage the background conditions required for autonomy. So they cannot be reasonably endorsed by each person subject to the outputs of those algorithms. This is why such algorithms are immoral according to RCP.

The Convergence Approach to Applied Ethics

Consider an approach to applied ethics discussed in Brennan (2007), Temkin (2012), and Furey and Hill (2021). Quantum mechanics is good at predicting the behavior of very small objects but bad at predicting the behavior of very large objects. General relativity is good at predicting the behavior of very large objects but bad at predicting the behavior of very small objects. Neither theory gets all of the behavior of all objects right. Neither theory works in all contexts. But each theory is useful. Each theory works well in some contexts.

Now, imagine you are an astronomer and are asked by the public to predict the movement of a large celestial body. You use general relativity to make your prediction. You then report your prediction to the public. This seems fine. There is nothing wrong with your explanation. The fact that general relativity does not capture the whole of physical reality is not a problem. The fact that there is a class of object behavior, the behavior of very small objects, that general relativity gets wrong is no problem. The theory gets right what you were asked to talk about, large celestial bodies. And so it works perfectly for your purposes.

In the same way, what matters is not whether there are counterexamples to the Reasonable Endorsement Test. What matters is whether the test works well in the contexts to which it is applied.

There is another relevant issue. While there are contexts in which one moral theory works well and others work poorly, there are also contexts where all plausible moral theories agree. And in a context in which all plausible moral theories converge on a judgment and when commonsense matches that judgment, we can be confident that the relevant judgment is correct. I think RCP's diagnoses of the algorithms considered here is such a context.

Notice that all of the interesting work in the cases RCP diagnose is done by appeal to the four reasons and not by any claims concerning autonomy or reasonable endorsement. They discuss failures of reliability in evaluating people, evaluating them in terms of factors for which they are not responsible, evaluating them when the stakes are high, and evaluating them in a way that burdens some groups relative to others.

None of these reasons for thinking TVASS or COMPAS are immoral is employable only by RCPs notion of autonomy or the Reasonable Endorsement Test. They are all reasons that consequentialists and virtue theorists would adopt (Feldman (1978)). Consequentialism is the view that only consequences matter in evaluating actions. Virtue ethics is the view that expressions of virtue and cultivating virtue are central to evaluating actions.

Regarding TVASS: As RCP point out, it does not accurately evaluate teacher performance. Furthermore, TVASS produces judgments about teachers based on factors those teachers are not responsible for such as the performance of students on tests that have nothing to do with the content of what they teach. Finally, TVASS is used to make high stakes decisions about whether a teacher gets a promotion or tenure. A consequentialist will see all these reasons as relevant. Consequentialists care about the outcome of an algorithm's decisions. The consequentialist may maintain that making high stakes decisions on the basis of teacher evaluations that are unreliable and disconnected from the teachers responsibilities will lead to bad outcomes. So the consequentialist can appropriate these reasons. Moreover, making decisions about teachers in these ways does not exemplify virtue or cultivate virtue in the teachers. So the virtue ethicist can appropriate these reasons as well. Thus, it is unnecessary to appeal to autonomy or what one would reasonably endorse to employ RCP's diagnosis of this case.

Regarding COMPAS: Making decisions about whether someone goes to jail based on an algorithm that is only accurate in 68% of cases, that holds people responsible for what their parents and siblings do, and that imposes a greater burden on black people than it does on white people will have bad consequences. So the consequentialist can appropriate these reasons. And surely, making such decisions does not exemplify virtue or cultivate virtue in defendants. So the virtue theorist can appropriate these reasons as well.

Furthermore, commonsense vindicates RCP's diagnoses. Once we recognize that these algorithms have the features RCP identify, it seems clear that the algorithms are bad. Even without a theory in hand, we can tell via commonsense moral judgment that the algorithms are bad.

So here is an alternative to the Reasonable Endorsement Test that RCP could use.

> **Convergence Test**: If every plausible theory of normative ethics and commonsense intuitions about morality converge on the judgment that the use of an algorithm is impermissible, then the use of that algorithm is impermissible.

The considerations RCP raise show that the algorithms they consider fail to pass the Convergence Test. So, the use of those algorithms is impermissible.

References

Brandt, R. (1967). Some Merits of One Form of Rule-Utilitarianism. In University of Colorado Studies in Philosophy, pp. 1–22. University of Colorado.

Brennan, J (2007). The Best Moral Theory Ever: The Merits and Methodology of Moral Theorizing. Dissertation. University of Arizona

Furey, H & Hill, Scott (2021) MIT's Moral Machine Experiment is a Psychological Roadblock to Self Driving Cars *AI & Ethics* 1, 151-155

Lyons, D. (1965). Forms and Limits of Utilitarianism. Oxford: Oxford University Press.

Parfit, D. (2011). On What Matters, Volume 1. Oxford: Oxford University Press.

Podgorski, A. (2018). Wouldn't it be Nice? Moral Rules and Distant Worlds. Nous 52 (2), 279–294.

Scanlon, T. M. (1998). What We Owe to Each Other. Cambridge, MA: Harvard University Press.

Smith, H. (2010). Measuring the Consequences of Rules. Utilitas 22, 413– 33.

Temkin, L. (2011). Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning. Oxford University Press, Oxford