The Root of Algocratic Illegitimacy

Mikhail Volkov
London School of Economics

PLEASE CITE THE PUBLISHED VERSION:
10.1007/s13347-025-00879-4

Abstract

Would a political system where the governance was overseen by an algorithmic system be legitimate? The intuitive answer seems to be no. This paper considers the philosophical effort to justify this intuition that argues for algocracy, a rule by algorithms, being illegitimate. Taking as the paradigmatic example the anti-algocratic argument from Danaher that attempts to ground algocratic illegitimacy in the opacity of algocratic decision-making, it is argued that the argument oversimplfies the matters. Opacity can delegitimise - but not simpliciter. It delegitimises because of the presence of certain downstream violations of obligations and rights of the public that result from the opaque governance. Algocratic decision-makers, however, seem not to be subjects to these normative constraints in the relevant sense. The paper therefore argues that the standards of legitimacy that have been deployed for or against different kinds of human governance do not apply to the algorithmic decision-making systems with quite the same force. New avenues for rooting the illegitimacy of algocratic decision-makers have to be developed.

1 Introduction

What about a world where algorithms based on machine learning and large swathes of data become not simply important or indispensable for the functioning of society but the main decision-makers and executives in charge of how the society functions? This possibility is one that is being taken more and more seriously in the literature across philosophy, social and computer sciences. The first manifestations as well as impending problems of such a society are being carefully considered. However, the classic question of political philosophy as to whether such a system can ever be legitimate have not been assessed as closely – probably because the answer 'no' seems so obvious and intuitive. The most notable philosophical justification of this intuitive position is given by Danaher whose argument for algocratic illegitimacy proceeds by analogy with arguments for epistocratic illegitimacy (2016). Roughly, for decision-making to be legitimate, it must be justifiable to the public in acceptable terms. With epistocracy this was argued to not obtain because decision-making by epistocrats would be opaque (incomprehensible and uncontestable); simultaneously, we may reasonably suspect the epistocratic elite of common cognitive biases and therefore find their rule inherently unjustifiable. Danaher argues the same is true for rule by algorithms: their decision-making is opaque and therefore unjustifiable.

The argument is often used to justify the undesirability of the algorithmic governance, and its conclusions are generally accepted (e.g. Henin and Le Métayer 2021; Holm 2023).[1] The philosophical merits of the attempt have not been considered extensively – which is not a gap one can allow in such a context. Political legitimacy is the fundament of frictionless cooperation between the decision-making system and the public, and one of the lenses through which to assess the system's desirability. Simultaneously, there has been a growing awareness that algocracy might be a rather imminent prospect and not at all a fanciful speculation (König 2019; Danaher 2020). Hence, if algocracy is to be a realistic form of government and a near concern, the question of its potential legitimacy or lack thereof needs to come under closer scrutiny. In particular, the prima facie plausible intuition about its illegitimacy should not go uncontested. A critical examination of the question is exactly what the present work attempts to contribute with results that will suggest a need for a much more meticulous debate about legitimacy and AI.

Indeed, I do not think Danaher's argument is quite correct; resorting to old and tried strategies in political philosophy might not be the most promising way to delegitimise a rule by algorithms. Although an algocratic

[1] A notable exception is (Sætra 2020), where an objection to algocracy from the positions of procedural legitimacy is considered and resisted. This is, however, a different route to arguing for illegitimacy from opacity which is what we are concerned with here.

system may well be opaque, it is unclear that presenting opaqueness as the source of algocratic illegitimacy packs the same punch as similarly structured epistocratic arguments. My goal will therefore be to show that it is harder to run the same argument for illegitimacy from opacity against a rule by algorithms than it was against a rule by the educated.

The paper is structured as follows. Section 2 outlines Danaher's argument and makes its scope of application precise. Section 3.1 unpacks why exactly legitimacy of decision-making is rooted in its non-opacity. For Danaher's argument to work, algocratic governance must in fact be illegitimate and in a sense that does not collapse into epistocratic illegitimacy. There must be something illegitimate about the governance by algorithms specifically – that is, something must be wrong with a society where everyone is on equal cognitive footing versus the non-human decision-makers. Section 3.2 further argues the following thesis. Once the relation between legitimacy and opacity is understood as in epistocratic arguments, it is generally harder to locate such illegitimacy in algocratic governance. Danaher's argument from analogy with epistocracy seems therefore weakened compared to its original target. This is because, as non-human agents, algorithms do not have a responsibility to respect our rights – and securing respect for rights is at the root of why non-opacity is necessary for legitimacy. Section 3.3 surveys the immediate objections to my argument. Section 4 concludes with mentioning alternative possibilities for delegitimising algocracy.

2 Danaher's Argument for Algocratic Threat

By algocracy one means a 'governance system...organised and structured on the basis of computer-programmed algorithms' (2016, p. 247). Algorithms mentioned here are primarily those responsible for collection, organization and processing of large amounts data; whether humans are in-, on- or out-of-the-loop is not crucial for the argument (ibid., pp. 247-248). In particular, Danaher envisions complicated and interconnected environments, whereby one algorithm's working is built upon workings of other algorithms. An example would be an algorithmic decision-making system designating those at more risk of committing a crime on the basis of data collected by surveillance. On the scale Danaher seems to have in mind, we may add a political algorithmic decision-maker on top that recommends police funding depending on the amount of crime perceived by the first algorithm, and so on. Each algorithm may be relatively easy to follow but with increasing computational complexity and growth of described environments, humans may well be prevented from fully comprehending the decision-making of the entire system. The opportunity for meaningful human participation will be correspondingly limited.

Danaher's upshot then rests on two observations. First: as mentioned, the relevant algorithmic decision-making tends to be opaque due to its computational complexity and is likely to get even more opaque in the future. Second: opacity may serve as a delegitimising feature of a decision-making system. The latter point is explicitly borrowed from anti-epistocratic arguments. A crucial premise of such arguments is that a legitimate decision-making system must be justifiable to those bound by its decisions (Estlund 2003, 2008). However, one may expect opacity in epistocratic decision-making that prevents successful justification of decisions to the public. This is due to different concerns that this work will delve into later but as an example, educated elites may not be able to articulate their line of reasoning well. Alternatively, epistocrats may have shared biases that cannot be justified to the public because it does not share the biases or might even be disfavoured by them. The *threat* of algocracy ensues for the following reason: if we favour algorithms as decision-makers for their superior epistemic qualities as advocates of epistocracy do the educated, we end up with a similarly illegitimate form of governance (Danaher 2016, pp. 251-252). It seems that if one buys into the plausible anti-epistocratic arguments, the illegitimacy of algocracy follows by much the same route.

Before considering whether algocratic illegitimacy can be rooted in opacity analogously to epistocratic illegitimacy, let us clarify some components of Danaher's argument. Firstly, what exactly is meant by opacity here? It refers specifically to the opacity of algorithmic decision-making, so let us consider three senses in which it can be deemed opaque (Burrell 2016):

1. Intentional opacity. This can come either in the form of a purposefully complicated algorithm design or concealing the algorithm design from the public altogether. An online-shopping firm withholding information about its user data collection is an example of such opacity.

2. Opacity due to technical illiteracy. Even if the workings of the algorithm are laid bare, not everyone can read code. Therefore, to some portion of the public any algorithm will be incomprehensible.

3. Opacity due to the algorithm's scale and complexity. This refers to systems opaque even to their designers – 'black boxes'.

Before settling on the sense of opacity most relevant to the issue in question, recall that our clarification of Danaher's argument must not collapse it into an argument for illegitimacy of some form of human governance. If the algorithmic illetigimacy can be traced back to human decision-makers, the illegitimacy is not of an interesting sort – we know humans are able to come up with bad forms of government. The badness of algocracy would then be no different from illegitimacy of human decision-makers which (a) does not seem to be faithful to how politically and morally unique a rule by algorithmic decision-makers promises to be, and (b) renders anti-algocratic arguments mostly trivial consequences of deligitimizing arguments against different forms of human governance.[2]

In light of this, which sense should we adopt? Adopting the first definition is a false move: then, the opacity and hence illegitimacy of decision-

[2] One can object that this is happening in Danaher's argument: the cause of epistocratic illegitimacy is cited against algocracy. But this is not the same as tracing the illegitimacy of algorithmic decision-makers to the illegitimacy of its human creators. If successful, Danaher's argument shows that algocracy is illegitimate due to the same reason as epistocracy, but not because of it.

making rest not with the algorithms but rather with their designers and managers. The deceitful nature of human decision-makers is ultimately responsible for the opacity of the entire system. A system that only contains this kind of opacity is further not algocratic in Danaher's sense insofar as humans can fully participate in and control the algorithmic decision-making. Furthermore, the second conception also seems to be disqualified. The issue here is that while such algorithms are not subject to efficient public scrutiny, they are no worse in terms of opacity than some existing decision-makers whose epistemic expertise we rely on and consider legitimate. For instance, the majority of people understandably lack the expertise to identify why a jury ruled in favour of a certain court decision because for any given case, we are unlikely to be acquainted with the same amount of evidence or the criteria for the verdict as the jury. However, we generally treat the coercive action that ensues from their judgement as legitimate. Moreover, the more epistemic expertise the members of the jury possess and hence the more sophisticated the decision-making is, the more legitimate their decision. Opacity due to technical illiteracy does not induce illegitimacy and is further not unique to algorithmic decision-making. Indeed, much of today's governmental decision-making proceeds in this manner: most of us are not well-versed in constitutional law, environmental predictions, secret intelligence, etc., and yet we generally accept actions there as legitimate when they are in line with the expertise.

Therefore, we are concerned with the third sense: opacity that places all humans on roughly the same cognitive footing versus algorithmic decision-makers. Note that this already weakens the scope of Danaher's argument as he claims the threat would stand 'even if it were possible to deconstruct and understand the system as a whole' (2016, p. 255); we must be concerned with the case where neither governments nor their subjects understand the algorithmic expertise (Chomanski 2022).

Lastly, for ensuing illegitimacy to rest solely with algocracy, the use and threat of opaque techniques must emerge despite human volition. This could (and does) obtain, for instance, when deployed algorithms have hidden biases that designers are unaware of. Alternatively, recall that on the scale of an environment of many interacting algorithms, we could have the situation whereby the environment's interactive nature is very complex, each algorithm separately is functioning as intended and yet their interaction generates unexpected emergent effects that are not equivalent to those of any given algorithm. Then, the decision-making of the resultant system emerges despite any particular designer's volition. As a toy example, one can imagine a shopping app learning to more frequently prompt its users to buy gift cards, which leads to these users being branded a fraud risk by their banking app. We may think that both algorithms have performed their own function well yet their interaction has yielded an unsavoury outcome. More speculatively, we can have in mind a widespread deployment of algorithms of the self-modifying kind that, despite initially passing all reasonable standards of human assessment, may still self-modify in an unexpected direction. This includes AI algorithms that can autonomously alter their architecture to improve performance or even learn to perform new tasks (e.g. Sheng and Padmanabhan 2023).[3]

While all of the above would instantiate relevant cases of algocratic decision-making, human decision-makers are strictly speaking at the start of the causal chain in all of them. Despite this, one can maintain that when the algorithmic environment is sufficiently complex or when the decision-making of the algorithms necessarily includes some unpredictability, blaming humans for the emergent effects is unwarranted. The complexity of the algorithmic environment from which a negative effect might emerges will very likely impede blame attribution to any concrete set of humans behind the environment's individual independent components.

In contrast, there will be cases where the undesirable effect of algorith-

[3] For further philosophical and formal discussion of predictability and safety of self-modifying systems, see, e.g., (Schmidhuber 2006; Bostrom 2014).

mic decision can be straightforwardly attributed to humans. For instance, if human decision-makers decide to deploy a predictably opaque and unwieldy algorithmic decision-maker due to lack of care, the illegitimacy of all consequent decision-making lies first and foremost with said human decision-makers and the system that placed them in power. Similarly, algorithms must have decision-making power with at least some tangible consequence. That is, we are unconcerned with situations whereby an opaque algorithm merely recommends an action and a human decision-maker decides whether to proceed with the recommendation or veto it (so-called 'oracle' systems (Armstrong, Sandberg, and Bostrom 2012)). Note that even if opaque algorithms are given tangible decision-making powers, not all instances where they take potentially dangerous actions make them the locus of the (il)legitimacy debate. For instance, suppose an algorithm used by a car manufacturer repeatedly leads to avoidable car accidents. Then, despite the opacity of the algorithm and its tangible decision-making, we would still think the responsibility rests chiefly with the car manufacturer, the marketers of the car and other humans behind the product. This is because the situation described would mean either that the humans behind the algorithm committed a predictable wrong by deploying the algorithm in the first place or that they are actively committing a wrong by keeping the algorithm in use after it has proven dangerous. In these cases, if the algorithm's functioning is dangerously opaque, the problem would reduce to humans forming their decisions upon epistemically dubious grounds (opaque computational recommendations, insufficient warrant for deploying or keeping the algorithm) and therefore to *their* illegitimacy as decision-makers. We are interested in cases where the possibility of this reduction is absent and whether we can proclaim these illegitimate.

3 Algocracy and Illegitimacy

Having sketched the anti-algocratic argument and the relevant kind of algocracy, consider more closely the concept of legitimacy and why exactly opacity has been seen as a threat to legitimacy.

3.1 Illegitimacy Reduced. On a very influential interpretation in political philosophy, legitimacy is necessary for a political decision-making body to justify its use of power asymmetry against the public (Beetham 1991; Rawls 2001). In the sense relevant to us, this power asymmetry is more justified if it is agreeable and known to those bound by its consequences. This direction of grounding legitimacy in public reason is pursued by, for instance, Rawls (2001) and Pettit (2012). But what is so valuable about non-opacity of decision-making that *it* is a legitimising component? For instance, what is less legitimate about a situation where I make decisions as a lone dictator than one where I make them in consultancy with the public? Those directly engaged with this issue answer that decision-making being contestable by and open to the public prevents abuse of power and the ability of the decision-makers to coerce the public arbitrarily (Larmore 1999; Binns 2017; Pettit 2012, Ch. 3). The latter, in turn, is wrong because it mistreats the public: using power to involuntarily coerce people violates respect for persons (Larmore 2002) or individual freedom (Pettit 2012, p. 147) or some other individual right.

Although agreeable, the wrong-making features of such violations by decision-makers need unpacking. There are at least two senses in which infringing upon others' rights in this manner is wrong and at least one of these senses does not translate to the algocratic case. On the one hand, a decision-maker abusing a citizen's right to, say, freedom is bad intrinsically (1): regardless of the context and the decision-maker's moral status, it is irreducibly bad that the citizen's freedom was lost because freedom is *a self-authenticating right* (Benn and Lazar 2022). For example, we may say that villagers losing their rightful property to a flood is bad. Perhaps the badness is not moral and no one is to blame but the situation is nevertheless undesirable. On the other hand, we may claim that the violation of the citizen's freedom is bad because the decision-maker violates their own moral responsibility (2): as a moral agent, they are obligated to not violate rights of others. So, for instance, a dictator putting a political opponent in jail is wrong not only because the opponent has a right to freedom but also because the dictator *qua* moral agent has an obligation to not violate others' freedom. A further wrong-making feature seems to be introduced if the violation is self-serving (2*). That is, if the decision-maker consciously violated the right of a person to advance their own ends.

Taking intermediate stock:

- Decision-making is legitimate if its use of power asymmetry is justifiable.
- Use of power asymmetry by decision-makers is justifiable if it avoids (1) and (2/2*) (i.e., avoids loss of rights of those bound by decision-making).
- Avoiding (1) and (2/2*) necessitates limits to the extent decision-makers can coerce those bound by their decision-making.
- These limits can be secured if those bound have a say in the decision-making binding them – that is, if the decision-making is subject to public reason.
- Decision-making cannot be subject to public reason if it is opaque.

This explicates why opacity is a delegitimising feature.

Coming back to Danaher's analogy, this connection of illegitimacy and opacity agrees with epistocratic arguments. Most notably, Estlund (2003) argues that epistocratic decision-making cannot be comprehensively justified to those bound by it and therefore is illegitimate – but this is not all. It cannot be comprehensively justified precisely *because* it potentially violates the rights of the public or the obligations of the ruling class: '[epistocratic decision-makers] either have unusual [self-serving] motives,... or their perfectly ordinary motives serve their statistically abnormal interests' (ibid., p. 66). The central objection to epistocracy Estlund lays down, the demographic objection, centres around it not being beyond reasonable doubt that the epistocrats possess discriminatory epistemic attitudes that, when translated into practice, produce a negative effect on the public. This corresponds to the wrong-making features (2) and (2*): the decision-makers may malignantly leverage the power asymmetry to benefit themselves or do so due to a shared unconscious bias – either way the decision-making proceeds at the expense of rights of those bound by it. Thus, the opacity of epistocratic (or any other) decision-making is not wrong intrinsically but rather because of its downstream effect on the public's rights. Crying opacity is not a delegitimising spell: we do not require public contestability of the jury's decisions nor of any given secret service decision, unless we have independent reason to suspect the decision-making of wrong-making qualities. Opacity only delegitimises in cases where we can reasonably suspect malign downstream effects to ensue from it. Accordingly, responses to Estlund make use of this consideration and contend that there is no proven such trait of the educated that could violate rights of the public in this manner (Mulligan 2015).

What follows from Danaher staking his argument on the analogy with epistocracy is that algorithmic opacity must be delegitimising for similar reasons which is to say it must lead to violations of rights in the senses (1) and (2/2*). While Danaher convincingly shows that political decision-making by algorithms may well be opaque, he does not do enough to show that it is the right-violating kind of opacity. That is, that algocracy entails (1) and (2/2*) in more robust ways than legitimate forms of government like democracy. The burden of proof seems to me to be on the anti-algocrats but I will attempt to preemptively argue that, in fact, it is not true that algorithmic decision-making is somehow especially at risk of producing said violations. In particular, algorithms as decision-makers do not seem to be subject to (2/2*) at all and may fare better with respect to avoiding (1) than some legitimate forms of government. This jeopardizes whether algocracy can be argued illegitimate on the same grounds that epistocracy can and on the grounds of its opacity more generally.

3.2 Algorithms and Moral Responsibility. Let us now consider whether algocratic systems can be expected to entail violations (1) and (2/2*) and therefore to possess the delegitimising kind of opacity.

Start with the thesis that to legitimately wield power, the decision-maker must avoid (2/2*) and uphold their moral responsibility to not violate the rights of others, especially for own gain. It is clear that to be able to lose legitimacy for the reason outlined, the moral status of the decision-maker must be mensurable. For an entity to possess moral status, it must have some degree of autonomy and agency. The flood sweeping the village is not blameworthy for the consequences because it is not a moral agent. Is a computational decision-maker a moral agent or is it rather akin to the flood? While the entire debate cannot be summarized here, the predominant position in the literature is that even more advanced computational systems, let alone existing machine learning algorithms, will not be autonomous agents (Bostrom and Yudkowsky 2014; Hakli and Mäkelä 2019). Arguments for computational decision-makers qualifying as moral agents exist (e.g., Sullins 2006), but these tend to refer to AI systems with decidedly speculative properties.

The general sentiment about the moral responsibility of a machine learning algorithm being at best very constrained stems from the overwhelming human influence on its eventual decisions. Algorithmic recommendations/actions depend on the training sample fed to it by programmers; the algorithm itself is human-conceived and human-implemented; etc. To see why this may create a problem for the moral status, consider a general example. There is person A who has depended on person B for any and all information they received. A is then asked to perform some public decision-making with hefty consequences. Surely, if person A's decision-making turns out harmful, they are only partially (if at all) to blame because they are not fully autonomous (Mele 1995); they were simply incapable of reflecting on the decision-making in a manner that would not be completely swamped and swayed by B's inputs.

Our opponent might be unconvinced. A natural counter here is to point out that regular humans are like person A as our beliefs are also swamped by exogenous as well as evolutionary influences. This is valid, we may have had little control over many formative influences and inputs that have been fed to us. However, one has to remember that it is not only the etiology of beliefs that defines our agency but also the character of our choices. It is not only where our beliefs come from but what we can in principle do with them that settles the agency question. Thus, should lack of autonomy on the basis of exogenous influences be unconvincing to the proponent of algorithmic morality, we should point them to the implausibility of algorithmic agency. In particular, the relevant condition necessary for ascribing agency and a fortiori moral responsibility is that choices be up to the agent and not due to a deterministic input-output mechanism or random choice (List 2023; Mele 2005). Algorithmic choices are a combination of the latter two, even if the underlying working is not always accessible to us due to the often-cited complexity and opacity. One could wage the familiar line of attack here also, stating that human decision-making mechanisms are of the deterministic sort. But this stronger version of the already mentioned critique most naturally supports a view where humans and complex machines possess equal and (close to) non-existent level of moral responsibility, not one on which they are both responsible.[4]

[4] I highly doubt a concept as tied to moral rights and obligations as legitimacy would survive an accommodation of this view, which is why I allow myself to bracket it from this discussion.

Hence, machine learning algorithms cannot be attached full degree (or, more likely, any) moral responsibility due to the human role in their design, functioning and deployment. This consideration obtains even for stronger forms of algorithmic decision-makers: the history of information acquisition and functionality of realistic computational systems will be too dependent on conscious human effort. Yet more plausible than absence of responsibility, algorithms lack desire. Even if their means of obtaining some ends are obscure to humans, the ends themselves are human-programmed; algorithms do not come to want to classify and predict. Therefore, even if at the most extreme interpretation an algorithm is blameworthy for violating a human right, it is not additionally blameworthy for doing so consciously and purposefully.

Thus, algocratic decision-makers are probably not moral agents, lack moral responsibility and desires in the sense in which these features apply to humans. However, if there is no moral responsibility upon algorithms to not violate the rights of others, then they cannot violate (2/2*) as this requirement necessitates an antecedently fixed moral responsibility to not violate rights. A decision-maker does not lose in legitimacy due to violating constraints that are inapplicable to it. Therefore, we cannot delegitimise an algorithmic decision-making by appealing to (2/2*) (i.e., to the moral responsibility of decision-makers). Thus, in an important sense, algorithmic decision-making does not lose legitimacy where human decision-making can.

What about (1)? Maybe we could argue for illegitimacy from the fact that although an algocratic decision-maker cannot be held morally responsible for losses of rights, it multiplies them. This would obtain if algorithmic decision-making could be expected to generate more harm to self-authenticating rights than the existing political decision-makers. We could then draw something of a benchmark on the loss of rights a system can endure before it is declared illegitimate, with select democratic political systems sitting above the benchmark and the algocratic system below it. However, this miserable state of the public's rights under algocracy does not seem to be a very likely one, as Danaher admits himself (2016, pp. 255-257). There are reasons why algocracy may ultimately *promote* human rights: improved and extended algorithmic decision-making in hospitals upholds the right to life, law-enforcement algorithms assisting taxation can help public infrastructure, algorithms designed for crime prevention can, if rid of biases, promote safety, and so forth. These are examples conceived on the already existing scale of algorithmic decision-making. With advances in big data collection and processing, we may expect potential positive effects of algocracy to be amplified and exceed current human decision-makers in what concerns securing human rights. For example, human decision-makers are inherently biased, whereas some argue algorithms do not have to be (Zarsky 2012), which could secure right to equal treatment if unbiased algorithms are implemented globally. This is opposed to existing rights deprivations even in democratic decision-making systems like those related to migration (Digidiki and Bhabha 2020) or foreign interventions (Sanyal 2009). The claim that algocracy will be a hindrance to the realization of human rights is not at all granted.

It is therefore possible that algocracy will not result in multiplying downstream effects (1) and (2/2*). The undesirability of algocracy cannot be traced to its violating (2/2*) because there is no moral duty upon algorithms to avoid (2/2*) in the first place. Pragmatically, however, it may even be more effective in upholding human rights and so unlikely to multiply their losses. Therefore, we cannot argue algocracy to be illegitimate on account of it resulting in (1) and (2/2*). This blocks one of the premises in our reduction of legitimacy to non-opacity.

3.3 Objections. My argument here is essentially that the requirement of public reason can be waived depending on the moral status of the decision-maker and the preservation of the rights of the public. This may well be unsavoury seeing how much of a commonplace the requirement of public participation in politics is (Rawls 2001; Habermas 1995; Nagel 1991). The critic might consequently be pushed to hold the right to transparent political decision-making to be a standalone self-authenticating right, even if the decision-making system in question violates virtually no other self-authenticating rights. That is, it is a right that cannot be erased by the contingencies of the decision-making like its success or the specifics of the decision-makers; we must always be able to peek under the hood of how we are governed and to understand the process.

It is hard to argue for waiving a requirement against an objection that simply brands said requirement a self-authenticating and thus indispensable right. I will nevertheless try to argue that conceiving of the right to non-opaque governance as necessary does not quite rhyme with the original purpose of the public reason requirement. Recall that the latter says that decision-making should be justified to all bound by it *beyond reasonable rejection*:

> [O]ur exercise of political power is fully proper only when it is exercised in accordance with a constitution the essentials of which all citizens...may reasonably be expected to endorse[.] (Rawls 2005, p. 137)

But once we assume that all other rights except the right to a transparent look into the decision-making are reliably secured by stable and benign computational decision-making, there is little prospect for justified doubt about such a system. This is because the main ground for reasonable rejection of a political system is that it might violate rights. This forms the basis for the requirement of public reason and the right to a scrutinizable political system – these procedures are necessary to make sure that no rights are violated. But once the system is in place that violates no self-authenticating rights, the requirement for justification before public reason must be satisfied automatically. It is simply unreasonable to complain about a system that robustly preserves the fundamental rights: to life, to freedom of speech and assembly, etc. Thus, any system guaranteeing these protections must be beyond reasonable reproach of the public. And if the main function of the right to non-opaque governance is to cement the exercise of public reason, within such a system, it must also lose some ground. In this sense, the right to transparency may be better conceived of not as self-authenticating but rather as contingent on a reasonable suspicion that the decision-makers are suboptimal, malign, unfair, etc. Once the suspicion becomes unreasonable, the right wanes accordingly.

Granted, it is conceivable that algocracy is an existential risk that will squash human rights left and right if enabled. I have not argued in any detail that the opposite of this prospect is a guarantee. I have only asked the reader to grant me that it is a consistency - that the opposite is live possibility that algocracy brings exactly the opposite. Then, the sceptic would be forced to showcase their reasonable doubt in light of a system that is better with respect to human rights than anything humans ever executed themselves. If such benign algorithmic governance comes to pass, one can still have suspicions regarding whether the decision-makers are acting with the right intentions or are not self-serving (e.g., are not actually primarily guided by keeping themselves in power). But this is exactly the issue of algorithms not being subject to (2/2*). Even if we may have such suspicions, they are not grounds for assigning illegitimacy because even if that is the case, the decision-makers in question are, strictly speaking, violating *none* of their obligations to us. Whereas we could have objected to the epistocrats at the top on the basis of them possibly having self-serving biases even if they do not directly harm the public, we cannot object to algorithmic governance on the same basis. Algorithms are simply not blameworthy for not being completely moral.

The opponent's final push may be to hold that it is always reasonable to suspect the decision-making system of possible malign actions and therefore there is always the right to public oversight. This de facto makes the right to transparency a necessary presence in our society. This is a promising response which, I think, ultimately depends on how good we can expect a realistic algocracy to be. Here, I assumed that it might be such that violations of rights are minimized and the governance becomes more impartial and fairer than it is now. One could argue this is not a possibility. I think such a view is fully valid and, if one is of such an opinion, my argument becomes of little relevance. But I think it at least useful to entertain the case where algocracy turns out to be great with respect to securing our rights simply because our intuitions about its illegitimacy would be similarly strong even then. In that case, they become hard to justify.

4 Conclusion

Let us briefly recap the critique I provided above. Danaher's preferred source of delegitimising algocracy is opacity and, if not simpliciter, opacity delegitimises because of the potential for violations of moral rights and obligations. In particular, a decision-making system was deemed unjustifiable if it entailed (1) the self-authenticating rights of the public being violated and/or (2/2*) decision-makers violating their obligations to us (possibly for selfish reasons). Building a case to the effect that algorithms do so is tricky. It is not a foregone conclusion that they will cause more violations of self-authenticating rights than current legitimate systems in addition to potentially lessening said violations, which may exonerate algocracy from (1). Further, algorithms do not have any moral obligations to not commit violations of rights, which always exonerates them from (2/2*). While this line of argument does not demonstrate conclusively that algocracy is not illegitimate, it makes us doubt whether algocratic illegitimacy is to be rooted in the same factors as epistocratic illegitimacy, as Danaher wants.

Understandably, there remains a strong intuition that opaque algocratic governance, whether good or bad, is undesirable. If we want to justify this undesirability in the illegitimacy of this form of governance, we can search for avenues different from those supplied to us by anti-epistocratic arguments. We may develop the idea of decision-making algorithms as moral agents to run the argument from opacity more successfully or hope for a greater extent of sufficiently interpretable algorithms to avoid the opacity from the start. The latter project does not seem as fruitless as Danaher portrays. Alternatively, we could argue that moral agency is a necessary precondition for legitimacy for any political decision maker. That is, one simply has to be an uncontroversial subject to obligations and rights as humans are and is illegitimate otherwise.

Another possible avenue to argue opaque AI decision-making illegitimate is more elaborate. Even though we might think AI algorithms are not moral agents, there is still a case for attaching *legal* personhood to them (Kurki 2019, Ch. 6). This would result in the possibility of ascribing claim-rights and duties to the relevant system *qua* a legal person or to facilitating a liability transfer procedure for the case the system malfunctions, as it is done for corporations in many industries (List 2021). In the former case, we could elaborate on the idea of the AI decision-maker being deligimatised due to not living up to its duties *qua* a legal person. In the latter case, the responsibility gap between the AI decision-maker and humans would be closed: for any wrongdoing of the algorithmic decision-makers, *some* humans would carry the full responsibility by design. Then, the perceived illegitimacy of the decision-making could always be traced back to the illegitimacy of the human decision-makers behind the algorithm implementation.[5]

For now, however, opacity on its own does not seem to be a universal delegitimiser for algorithmic decision-making. If effective and fair algorithmic systems of the opaque type come to be the main decision-makers in our lives, it may be hard to blame the ensuing illegitimate use of the

[5] In this case, there would still be the question regarding the status of algocratic systems that emerge without a well-worked out liability protocol in place.

power-asymmetry on anyone but humans themselves.

Statements and Declarations

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript. The authors have no relevant financial or non-financial interests to disclose.

References

Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. 2012. "Thinking Inside the Box: Controlling and Using an Oracle Ai." *Minds and Machines* 22 (4): 299–324. https://doi.org/10.1007/s11023-012-9282-2.

Beetham, David. 1991. *The Legitimation of Power.* Houndmills, Basingstoke, Hampshire ; New York, Ny: Palgrave Macmillan. ISBN: 9780230279728.

Benn, Claire, and Seth Lazar. 2022. "What's Wrong with Automated Influence." *Canadian Journal of Philosophy* 52 (1): 125–148. https://doi.org/10.1017/can.2021.23.

Binns, Reuben. 2017. "Algorithmic Accountability and Public Reason." *Philosophy & Technology* 31, no. 4 (May): 543–556. https://doi.org/https://doi.org/10.1007/s13347-017-0263-5.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press, Cop. ISBN: 9780199678112.

Bostrom, Nick, and Eliezer Yudkowsky. 2014. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence,* edited by Keith Frankish and William M.Editors Ramsey, 316–334. Cambridge University Press. https://doi.org/10.1017/CBO9781139046855.020.

Burrell, Jenna. 2016. "How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3, no. 1 (January): 1–12. https://doi.org/https://doi.org/10.1177/2053951715622512.

Chomanski, Bartek. 2022. "Legitimacy and Automated Decisions: the Moral Limits of Algocracy." *Ethics and Information Technology* 24, no. 3 (August). https://doi.org/https://doi.org/10.1007/s10676-022-09647-w.

Danaher, John. 2016. "The Threat of Algocracy: Reality, Resistance and Accommodation." *Philosophy & Technology* 29, no. 3 (January): 245–268. https://doi.org/https://doi.org/10.1007/s13347-015-0211-1.

———. 2020. "Freedom in an Age of Algocracy." In *The Oxford Handbook of Philosophy of Technology,* edited by Shannon Vallor. Oxford University Press, Usa.

Digidiki, Vasileia, and Jacqueline Bhabha. 2020. "EU Migration Pact Fails to Address Human Rights Concerns in Lesvos, Greece." *Health and Human Rights* 22, no. 2 (December): 291–296. EU%20Migration%20Pact%20Fails%20to%20Address%20Human%20Rights%20Concerns%20in%20Lesvos,%20Greece.

Estlund, David. 2003. “Why Not Epistocracy?” In *Desire, Identity and Existence: Essays in honor of T. M. Penner,* edited by Naomi Reshotko, 53–69. Academic Printing / Publishing.

———. 2008. *Democratic Authority: A Philosophical Framework.* Princeton University Press.

Habermas, Jürgen. 1995. “Reconciliation Through the Public Use of Reason.” *Journal of Philosophy* 92 (3): 109–131. https://doi.org/https://doi.org/10.5840/jphil199592335.

Hakli, Raul, and Pekka Mäkelä. 2019. “Moral Responsibility of Robots and Hybrid Agents.” *The Monist* 102, no. 2 (March): 259–275. https://doi.org/https://doi.org/10.1093/monist/onz009. https://academic.oup.com/monist/article/102/2/259/5374583.

Henin, Clément, and Daniel Le Métayer. 2021. “Beyond Explainability: Justifiability and Contestability of Algorithmic Decision Systems.” *AI & Society* 37 (July). https://doi.org/https://doi.org/10.1007/s00146-021-01251-8.

Holm, Sune. 2023. “Algorithmic Legitimacy in Clinical Decision-Making.” *Ethics and Information Technology* 25 (3): 1–10. https://doi.org/10.1007/s10676-023-09709-7.

König, Pascal D. 2019. “Dissecting the Algorithmic Leviathan: on the Socio-Political Anatomy of Algorithmic Governance.” *Philosophy & Technology* 33 (July). https://doi.org/https://doi.org/10.1007/s13347-019-00363-w.

Kurki, Visa AJ. 2019. *A Theory of Legal Personhood.* Oxford University Press, August. ISBN: 9780198844037. https://doi.org/https://doi.org/10.1093/oso/9780198844037.001.0001.

Larmore, Charles. 1999. “The Moral Basis of Political Liberalism.” *The Journal of Philosophy* 96 (12): 599–625. ISSN: 0022362X, accessed February 19, 2024. http://www.jstor.org/stable/2564695.

———. 2002. “Public Reason.” In *The Cambridge Companion to Rawls,* edited by SamuelEditor Freeman, 368–393. Cambridge Companions to Philosophy. Cambridge University Press. https://doi.org/10.1017/CCOL0521651670.011.

List, Christian. 2021. “Group Agency and Artificial Intelligence.” *Philosophy & Technology* 34, no. 4 (August): 1213–1242. https://doi.org/https://doi.org/10.1007/s13347-021-00454-7.

———. 2023. “Agential Possibilities.” *Possibility Studies & Society* 1, no. 4 (September): 461–470. https://doi.org/https://doi.org/10.1177/27538699231200093.

Mele, Alfred R. 1995. *Autonomous Agents: From Self-control to Autonomy.* New York ; Oxford: Oxford University Press. ISBN: 9780195150438.

———. 2005. “Libertarianism, Luck, and Control.” *Pacific Philosophical Quarterly* 86, no. 3 (September): 381–407. https://doi.org/https://doi.org/10.1111/j.1468-0114.2005.00233.x.

Mulligan, Thomas. 2015. “On the Compatibility of Epistocracy and Public Reason.” *Social Theory and Practice* 41 (3): 458–476. https://doi.org/10.5840/soctheorpract201541324.

Nagel, Thomas. 1991. *Equality and Partiality.* Oxford University Press, November. ISBN: 0195069676.

Pettit, Philip. 2012. *On the People's Terms: a Republican Theory and Model of Democracy.* Cambridge ; New York: Cambridge University Press. ISBN: 9781107005112.

Rawls, John. 2001. *Justice as Fairness: a Restatement.* Edited by Erin Kelly. Cambridge, Ma: Bleknap Press of Havard University Press. ISBN: 9780674005112.

———. 2005. *Political Liberalism: Expanded Edition.* New York Columbia University Press. ISBN: 9780231527538.

Sætra, Henrik Skaug. 2020. "A Shallow Defence of a Technocracy of Artificial intelligence: Examining the Political Harms of Algorithmic Governance in the Domain of Government." *Technology in Society* 62 (August): 101283. https://doi.org/https://doi.org/10.1016/j.techsoc.2020.101283. https://www.sciencedirect.com/science/article/pii/S0160791X19305925.

Sanyal, Sagar. 2009. "Us Military and Covert Action and Global Justice." *International Journal of Applied Philosophy* 23 (2): 213–234. https://doi.org/10.5840/ijap200923217.

Schmidhuber, Juergen. 2006. *Goedel Machines: Self-Referential Universal Problem Solvers Making Provably Optimal Self-Improvements.* arXiv: cs/0309048 [cs.LO]. https://arxiv.org/abs/cs/0309048.

Sheng, Alex, and Shankar Padmanabhan. 2023. *Self-Programming Artificial Intelligence Using Code-Generating Language Models.* arXiv: 2205.00167 [cs.AI]. https://arxiv.org/abs/2205.00167.

Sullins, John P. 2006. "When is a Robot a Moral Agent." *International Review of Information Ethics* 6 (12): 23–30.

Zarsky, Tal Z. 2012. "Automated Prediction." *Communications of the ACM* 55, no. 9 (September): 33–35. https://doi.org/https://doi.org/10.1145/2330667.2330678.