Ethics and Information Technology (2022) 24:34
https://doi.org/10.1007/s10676-022-09647-w

ORIGINAL PAPER

Legitimacy and automated decisions: the moral limits of algocracy

Bartek Chomanski[1]

Published online: 8 August 2022

Abstract
With the advent of automated decision-making, governments have increasingly begun to rely on artificially intelligent algorithms to inform policy decisions across a range of domains of government interest and influence. The practice has not gone unnoticed among philosophers, worried about "algocracy" (rule by algorithm), and its ethical and political impacts. One of the chief issues of ethical and political significance raised by algocratic governance, so the argument goes, is the lack of transparency of algorithms.

One of the best-known examples of philosophical analyses of algocracy is John Danaher's "The threat of algocracy" (2016), arguing that government by algorithm undermines political legitimacy. In this paper, I will treat Danaher's argument as a springboard for raising additional questions about the connections between algocracy, comprehensibility, and legitimacy, especially in light of empirical results about what we can expect the voters and policymakers to know.

The paper has the following structure: in Sect. 2, I introduce the basics of Danaher's argument regarding algocracy. In Sect. 3 I argue that the algocratic threat to legitimacy has troubling implications for social justice. In Sect. 4, I argue that, nevertheless, there seem to be good reasons for governments to rely on algorithmic decision support systems. Lastly, I try to resolve the apparent tension between the findings of the two preceding Sections.

Keywords Algocracy · Algorithmic governance · AI ethics · Legitimacy

Introduction

Governments claim the unique right to enforce their decisions by violence or threat thereof. Governments' decisions are consequential: they can deprive citizens of liberty, property, and, at the limit, life. This is a fearsome power, and its every aspect deserves close philosophical scrutiny. It is thus no accident that governments' use of emerging technologies to drive their decisions has become a topic of interest in recent years.

With the advent of automated decision-making, governments have increasingly begun to rely on artificially intelligent algorithms to inform policy decisions across a range of domains of government interest and influence, from immigration control, through crime prevention, to welfare provision. The practice has not gone unnoticed among philosophers, worried about "algocracy" (rule by algorithm), and its ethical and political impacts. One of the chief issues of ethical and political significance raised by algocratic governance, so the argument goes, is the lack of transparency of algorithms.

One of the best-known examples of philosophical analyses of algocracy is John Danaher's "The threat of algocracy" (2016), arguing that government by algorithm undermines political legitimacy. In this paper, I will treat Danaher's argument as a springboard for raising additional questions about the connections between algocracy and legitimacy, especially in light of empirical results about what we can expect the voters and policymakers to know.

The paper has the following structure: in Sect. 2, I introduce the basics of Danaher's argument regarding algocracy. In Sect. 3 I argue that the algocratic threat to legitimacy has troubling implications for social justice. In Sect. 4, I argue that, nevertheless, there seem to be good reasons for governments to rely on algorithmic decision support systems. Lastly, I try to resolve the apparent tension between the findings of the two preceding Sections.

✉ Bartek Chomanski
b.chomanski@gmail.com

[1] Department of Philosophy, Adam Mickiewicz University, Poznan, Poland

Algocracy and opacity

Should governments make decisions on the basis of algorithms the mechanism of which neither they nor their subjects understand? This is the problem of algocracy.

In what follows, I'll adopt Danaher's definition of algocracy throughout this article:

> a system in which algorithms are used to collect, collate and organise the data upon which decisions are typically made and to assist in how that data is processed and communicated through the relevant governance system. In doing so, the algorithms structure and constrain the ways in which humans within those systems interact with one another, the relevant data and the broader community affected by those systems (2016).

I will use the term "algorithmic decision-making" to denote any decision-making relying on algorithmic outputs as a consideration in making a decision, and the term "algorithmic decision support systems" to denote systems composed of algorithmic decision-making models, their designers and their end-users (ultimate decision-makers).

In his paper, Danaher considers the impact of algocracy on the legitimacy of government decisions. In so doing, he seems to entertain the following argument:

(1) Algocracy involves algorithms.
(2) Algorithms are opaque.
(3) Opacity prevents comprehensibility and informed participation.
(4) In the absence of comprehensibility and informed participation, there can be no legitimacy for governments' decisions (call it the "Strong comprehensibility condition on legitimacy").
(5) Therefore, opacity undermines legitimacy.
(6) Therefore, algocracy undermines legitimacy.

Premise (1) is true by definition. Premise (2) draws its support from the "black box" nature of many algorithms (I'll speak more about it later). Premise (3) also seems true by definition, while Premise (4) appears to be a broadly endorsed principle in democratic theory (see especially Estlund (2008)). Danaher offers the following rationale for it:

> In Estlund's model, … the procedures must be justifiable to people in terms of reasons that are accessible and comprehensible to them… This requires non-opacity: The rationales underlying the mechanics of the procedure must not be opaque to those who are affected by those procedures. In appealing to non-opacity conditions, he is not alone. Many theories of political legitimacy insist that decision-making procedures must be rationally acceptable to those who are affected by them (2016) [references omitted].

Then, (5) follows from (2), (3) and (4), and (6) follows from (1) and (5). The argument is certainly a strong one – however, it does not appear to be entirely endorsed by Danaher himself. Given the nature of the premises, the only one that can be rejected (or, more plausibly, relaxed) is Premise (4). In this light, a more promising version of this argument could rely on a weaker comprehensibility condition on legitimacy, perhaps along the lines of:

Premise (4'): The absence of comprehensibility and informed participation is a prima facie reason to think the legitimacy of governments' decisions is diminished.[1]

Rephrasing it thus does three things: first, it makes the criterion more plausible; second, it merely places the justificatory burden on the governments planning to institute some form of algocratic governance, rather than outright denying legitimacy to the results of such governance; third, it makes explicit the idea that legitimacy is a matter of degree (just like comprehensibility and informed participation). Consequently, just as decisions could be more or less understandable, they can also be more or less legitimate. (This is not to say that the degree of legitimacy is always proportional to the degree of comprehensibility; other factors could of course be at play.)

Returning to premise (2), it is worth pointing out that the opacity of algorithmic procedures can come in a variety of flavors. In this paper, I adopt Jenna Burrell's (2016) tripartite distinction between different kinds of opacity. Burrell distinguishes opacity due to secrecy, opacity due to technological illiteracy, and opacity due to the inherent black-box nature of the algorithms. In fact, I will put opacity as secrecy to a side (see Fink (2018) for some concerns about the secrecy and disclosure of algorithms used by public institutions), and focus primarily on the other two types.

Burrell defines opacity due to technological illiteracy as "stemming from the current state of affairs where writing (and reading) code is a specialist skill" (2016) and black-box opacity as arising "from the mismatch between mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation" (2016).

While the former kind of opacity is relatively straightforward, one may need further elaboration on the latter. Burrell obliges by offering the following gloss:

[1] Accordingly, the conclusion of the amended argument would be something to the effect that (6') Prima facie, under algocracy, the legitimacy of governments' decisions is diminished.

> When a computer learns and consequently builds its own representation of a classification decision, it does so without regard for human comprehension. Machine optimizations based on training data do not naturally accord with human semantic explanations. … the workings of machine learning algorithms can escape full understanding and interpretation by humans, even for those with specialized training, even for computer scientists (2016).

Consequently, opacity due to illiteracy is in principle possible to overcome by learning the relevant skills (“writing (and reading) code”). Opacity due to the black box nature of some algorithms, in contrast, may make these algorithms forever inscrutable.

Algocracy and social justice

Suppose it’s true that democratic legitimacy is prima facie undermined by the of lack informed participation from the subjects of the decision (where informed participation requires comprehensibility). Suppose an algocratic decision is defended on the basis that it is, in principle, understandable. That is, it is neither shrouded in legally enforced secrecy, nor supported by an indecipherable black-box algorithm. In other words, the operations of the automated decision-support system could be comprehended, were one to expend a certain amount of resources and effort on acquiring the relevant specialist skills. Of course, opacity in this sense is essentially a matter of degree – the more skillful an individual, the better they will be at comprehending the algorithmic procedures and the more recondite the algorithmic operations, the more resources need to be spent unearthing its inner workings.

Comprehensibility, a necessary condition for informed democratic participation, is thus at least formally met as long as the algorithm is not kept secret and its operations do not exceed “human-scale reasoning.” That is, in such circumstances, there’s a way for the decision subject to understand the processes that led to the actual outcome. However, in-principle comprehensibility need not be enough to assure any meaningful understanding on the part of the decision subjects, and consequently, seems to fail to meet the strong comprehensibility condition on legitimacy.

Comprehensibility in the real world

Consider the following scenario to get us started:

> *Consultation*
>
> Suppose that one of the provisions of the Springfield city charter states that any binding decision of the city council must be preceded by a public consultation. Today is the day of just such a consultation meeting about proposed changes to one of the city’s laws. Anyone, regardless of race, disability, gender, or any other characteristic, can participate, and the announcement about the meeting is widely circulated. However, the meeting is held in a remote location only accessible by helicopter. After the consultation is concluded, the city passes the changes in law, claiming that its doing so is in accordance with the city charter.

There are some problems with the council’s claim. For starters, Springfieldians would need to sacrifice a substantial amount of time and resources to be able to attend the consultation. If there were economically marginalized individuals in Springfield, they would be more adversely affected by the city’s choice of location than the economically advantaged. This is, of course, because someone poor, like Cletus, would be less likely to afford helicopter rides to the meeting than someone rich, like Mr. Burns. Setting this aside, however, it still seems wrong to condition democratic participation on the possession of a certain amount of resources. The city’s policy would remain wrong even if everyone in Springfield had the same level of income and wealth (except of course if everyone in the city could easily afford a helicopter).

What *Consultation* demonstrates, I take it, is that *having to bear substantial costs as a condition on democratic participation is* prima facie *unjustified*. The city council would have to provide powerful reasons indeed for its choice of the meeting location (perhaps if the policy proposal was exclusively concerned with taxes on helicopter owners, it would be permissible to hold the meeting there; or perhaps if that was the only way to keep the meeting participants safe). This insight is applicable to algocratic governance just as much as it is to the governance of Springfield. Acquiring technical skills that enable people to understand algorithms is costly (maybe not as costly as helicopters, but probably more expensive than helicopter rides). Moreover, it also requires that the prospective learner be in fact willing to learn.

Putting it differently, there are at least two necessary conditions on skill acquisition. One is the presence of an incentive to do so, the other is the possession of resources. People won’t learn unless they want to and can afford it.

Empirical research on voter behavior strongly suggests that voters will generally avoid acquiring knowledge required to make an informed choice about policy matters (see Somin (2015) for a primer). The researchers tend to take this fact to reflect rational choices on the voters’ part. Given the very small expected benefit from voting well and given the substantial amount of costly study required to become informed enough to vote well, it is generally rational to remain ignorant. What's true of acquiring knowledge in general will also be true when it comes to acquiring the particular knowledge of how algorithms work. If individual votes (or other forms of political participation) are unlikely to influence algocratic procedures and outcomes, it is not irrational to decide against learning about algorithms in the first place.

As a corollary, we should also expect low level of voter willingness to acquire information if the relevant decisions are unlikely to affect the majority of them. E.g. suppose that algorithmic decision support systems are being used in determining the likelihood of child abuse, as described by Eubanks (2017). If most voters believe themselves to be extremely unlikely to ever come into contact with the part of the state that Eubanks describes, their incentive to understand its algorithmic procedures is likewise diminished. Similarly, if most voters expect not to come into contact with the systems administering pre-trial detention, they may be unwilling to become informed about the algorithms it uses.

Thus, we have two kinds of likely circumstances in which voters would be disincentivized from becoming informed about the operations of algorithmic decision support systems. And it's not merely theoretical speculation. Rational ignorance has already been demonstrated to explain citizen behavior in other circumstances where knowledge of information technologies (IT) is required for community participation. In a fairly recent attempt to promote a more participatory model of city planning, some jurisdictions have decided to rely on IT-based geographic information systems (GIS). However, this initiative failed to increase citizen participation, likely due to rational ignorance. As Krek (2005) summarizes "For most citizens the personal benefit of getting involved in planning activities and learning how to use a public participatory GIS application is usually low and the cost of participation high. Therefore, they rather decide to ignore the possibility of participation."

It is reasonable to expect this mechanism to apply to algocratic institutions, given the expected costs and benefits to individual voters of becoming informed about algorithms.

Opacity and disparities

Nevertheless, even if voters do have an incentive to be informed about their government's algorithmic decision-making practices, and specifically about the inner workings of the relevant algorithms, obstacles remain. As I already mentioned, skill acquisition (just as helicopter acquisition) is costly, and not everyone will be in a position to devote the required amount of resources to learning about algorithms. It is therefore likely that participatory engagement in public matters aided by algorithmic decision-support systems will only be possible for those at or above a certain level of income or wealth. This would remain true even for people likely to be significantly affected by algocratic decisions.

Given these constraints, should the governments have the right to impose their algorithmic decisions on their subjects? Suppose the outcome of such a decision has the potential to substantially affect the wellbeing of the decision subject, but that most citizens are extremely unlikely to be subjected to these decisions. At the same time, the cost one must bear to acquire the ability to comprehend the algorithm's operations is also substantial.

As *Consultation* demonstrates, the government cannot claim comprehensibility merely because it is in principle possible to understand its decision, were one to have access to a substantial amount of resources. This is because it effectively bars those decision subjects who *lack* access to such resources from informed democratic participation. If those very same people end up being primarily subjected to the governments' algorithmic decision-making, the legitimacy of such decisions is thrown into serious doubt. Suppose, for instance, that Springfield's consultation concerns some poverty relief-related policy. Those able to attend the consultation and those most affected by the policy changes are probably completely different groups of people. Consequently, we'd be hard pressed to agree with the city's claim that its new policy meets the provisions of the charter.

This could have significant implications for social justice, given the actual wealth and income disparities among different groups of decision subjects. Members of marginalized communities, when they become targets of algorithmic decision-making by governments, would tend to have a more difficult time engaging in participatory democratic processes than members of dominant groups. This is because members of marginalized communities would be more likely to lack the resources needed to acquire the relevant skills necessary for comprehending the nature of the algorithmic influences on the governments' decisions. At the very least, members of such communities would be more likely disadvantaged in access to these skills relative to the dominant groups.[2] This would undermine their ability to participate in the public decision-making processes, and, consequently, undermine the legitimacy of decisions affecting them.

As a result, governments need to have a reasonable expectation that acquiring information about, and understanding of, its decision processes is not excessively onerous before claiming that their decisions are comprehensible. In other

[2] Given that marginalized communities are in fact (sometimes intentionally) *targeted* by algorithm-driven policies, this makes the problem more acute.

words, government cannot easily claim legitimacy while at the same time imposing high (and unequally spread) costs on informed participation.

One way of addressing this problem is by ensuring that marginalized communities do have both the incentive and the resources to learn about algorithmic policymaking, while the remainder of the population lacks at least one of the two. This could be done, for example, by education programs directed at such communities. If these programs worked, then we'd have a pool composed of a minority of informed citizens and a majority of uninformed citizens, potentially engaged in democratic participation in various forms. On some models, such an unequal distribution of relevant knowledge is no problem for reaching informed democratic outcomes. If the majority's votes are randomly distributed, the minority's informed opinion will prevail.

Unfortunately, matters are more complex here than initially appears. For starters, it is not obvious that education programs will in fact equip people with the relevant skills. Schooling's empirical record of achieving its pedagogical goals of providing students with knowledge is less than stellar (Caplan, 2018). Even if that were overcome, though, it may still be the case that the uninformed opinions that shape voting behavior are not random (and thus don't cancel each other out). It is an empirical question whether the uninformed majority holds relevant *systematic* biases that could override the informed minority's democratic decision in any particular case (Caplan, 2007). We thus cannot *a priori* assume that changes to education policies will improve matters.

Comprehension without transparency

Sandra Wachter and colleagues (2018) have recently suggested that counterfactual explanations ('if you were earning 2x more, you'd have been approved for a loan') can meet something like the comprehensibility criterion without opening the black box. The idea is that using techniques such as "Adverse Perturbation," and applying them even to the state-of-the-art black-box models, allows the researchers to generate human-understandable counterfactual explanations. Applying such techniques doesn't require "opening the black box," thus ensuring comprehension without the need for understanding the specific parameters of the model.

However, this need not help us in our situation. At least *some* technical skill is required to understand the Adverse Perturbation techniques themselves. After all, such awareness seems necessary to determine whether the counterfactual explanations themselves are accurate and trustworthy. Wachter et al.'s solution seems merely to shift the need for comprehension by one level, rather than offering a genuine way out of the algocratic predicament.

Opacity and the decision-maker

It thus looks like governments have at least a presumptive duty not to engage in algorithmic decision-making unless the populations most likely affected by such decisions are in a position to comprehend their basis. If comprehension can only be acquired at a substantial cost, and especially if such costs are imposed primarily on already disadvantaged communities, the claim to legitimacy is seriously undermined. The higher the costs and the more significant the consequences, the less can the government claim to have the right to make its decisions via algocratic means.

The preceding section showed, then, that there are normative consequences to the *public's* limited understanding of algorithmic processes. In this section, I examine the claim that normative consequences also flow from the limited understanding that the government *agents* (understood as broadly as possible to include lawmakers, regulators, and law enforcement) have of the algorithmic processes they may rely on in their decisions.

Let's start with some stories.

> *Ideal Queen*
>
> Queen Oona is sitting in judgment over you. You're accused of a crime which carries a substantial prison penalty. The evidence has now been presented and the queen must consider it all and render her judgment. The queen carefully ponders the evidence, judiciously weighing each bit according to her best judgment. After many laborious hours have passed, the queen decides that the evidence indicates there's a very high chance that you're guilty beyond reasonable doubt. Accordingly, she finds you guilty and passes the sentence. She also clearly and accurately lays out her reasons for this judgment, in a way that you and any other interested party can understand.

The queen seems to be acting justly in the above scenario. You'd have no grounds for complaining about her conduct.

Consider now a more realistic version of the queen.

> *Non-ideal queen*
>
> Queen Oona is sitting in judgment over you. You're accused of a crime which carries a substantial prison penalty. The evidence has now been presented and the queen must consider it all and render her judgment. The queen looks at the evidence to determine your guilt, but her thought process isn't entirely rational. There are factors that influence her thinking which have nothing to do with the evidence's quality. For example, the queen is a bit more likely to issue a

> harsher judgment when she's hungry[3], or when her favorite polo team unexpectedly lost the recent game[4], or when the defendant is unattractive[5]. Secondly, her thinking is prone to irrational biases that affect her decisions in potentially objectionable ways[6]. Lastly, justice is not the only goal at which the queen aims in issuing her verdicts. She also wants to be well-liked by and retain the favor of the nobles and other important people in the kingdom, and build her own and her family's wealth.[7]

Non-Ideal Queen is, indeed, quite far from ideal, though it seems to correspond better to reality (at least when real-world professional judges are concerned – and, though less work has been done on the cognitive shortcomings of experts and policymakers, what evidence there is does seem to suggest that they too suffer from similar biases in their thinking and decision-making[8]). Submitting to her judgment doesn't look especially appealing. Nevertheless, it seems, we have little choice but to tolerate this state of affairs if we want any justice to be done (what's the alternative? Abandon the monarchy?). At least, I will assume so for the remainder of this paper.

Improving the non-ideal?

Consider now the following scenario.

> *Advisor*
> As in *Non-Ideal Queen*, except this time Oona is advised in her deliberations by the court sage; the sage is an expert on judicial matters – including the determination of the likelihood of guilt – but she never explains her reasoning to anyone – all she does is look at the same evidence as Oona and declare her view on whether the defendant is guilty or not, plus how confident she is in her judgment. Consequently, Oona does not (indeed cannot) understand the way in which the sage thinks and arrives at her judgments. Nevertheless, the final decision is always Oona's alone. Trustworthy independent auditors have determined that when Oona issues a judgment with the sage's advice, she is more accurate than if she were to deliberate on her own.

It seems that, even though *Advisor* retains all of Oona's faults, the situation is a clear improvement over *Non-Ideal Queen*. Moreover, *Advisor* offers a clear analogy to automated decision-making systems with human beings "in the loop" of the decision-making process who are given the final say over what to do, while taking into account the opaque algorithms' predictions. Risk-assessment tools used in sentencing and pre-trial detention decisions (see in general Christin et al., (2015), and Cadigan & Lowenkamp (2011) for a discussion of pre-trial detention decisions specifically) are probably the best known and most controversial examples of the use of such systems in something like an advisory capacity.

According to research (see, e.g. Green & Chen (2019); Grgić-Hlača et al., (2019)), decision-making systems which involve machine advice to human beings are more accurate than those where human beings (including experts) make their decisions unaided.[9] Consequently, if our choice is between *Non-Ideal Queen* and *Advisor*, then it seems clear that we should welcome the sage's introduction.[10]

Consider now a further development in Oona's story.

> *Know-how*
> Suppose that, as her experience issuing judgments with the sage's advice grows, Oona acquires a degree of inarticulable judicial know-how,[11] so that her own accuracy improves markedly. Now, given that her

[3] Danziger et al., (2011) for evidence that judicial decisions tend to be more lenient just after food breaks, and get harsher the more time passed from the most recent food break.

[4] See Eren & Mocan (2018) for evidence that juvenile court judges' sentences are harsher after their local football team unexpectedly loses a game.

[5] See Stewart (1980)and Downs & Lyons (1991) for evidence that the defendant's attractiveness influences sentencing length.

[6] See Englich et al., (2006) and a review by Peer & Gamliel (2013) for evidence of judges' thinking being influenced by cognitive shortcuts and biases. Also note that all the research cited in footnotes 3–6 concerns legal professionals (especially judges), not the general population.

[7] See e.g. Lemieux (2004) for a brief introduction to the way of thinking about public officials as self-interested utility maximizers rather than as exclusively concerned with the pursuit of the common good.

[8] On experts, see Cassidy & Buede (2009) and, in general, Koppl (2018). On policymakers' cognitive and motivational errors, see in general Cairney & Kwiatkowski (2017), and Houghton (2008) and Yetiv (2013) for research on biases in the context of foreign policy decision-making.

[9] Nevertheless, there is some controversy about such claims. In a well-publicized article, Dressel & Farid (2018) have found that the infamous COMPAS recidivism prediction algorithm is no more accurate in its predictions than a random sample of non-experts. However, though the result has been replicated in subsequent work by Lin Zhiyuan et al., (2020), the latter authors have also found that changing aspects of the experimental setup reintroduced the machine advantage over humans, and that such new setups were importantly similar to what one can expect in real-world scenarios.

[10] I set aside here the widely discussed issue of fairness of such decisions and assume that the increase in accuracy of judgments is not traded off against more biased decisions.

[11] Some legal scholars embrace this type of legal skill: "Law is not all reasoning and analysis-it is also emotion and judgment and intuition and rhetoric. *It includes knowledge that cannot always be explained, but that is no less valid for that* [emphasis added]" (Gewirtz, 1995).

> know-how is inarticulable, Oona sometimes cannot explain how she arrived at her decisions. Perhaps certain judgments just "feel right" to her, and, more often than not, they turn out to be correct. Indeed, in time, she attains – exclusively through the acquisition of the know-how – a significantly greater accuracy than what she has in *Advisor*.

Know-How seems like a clear improvement over *Advisor*. That is to say, I submit, that we are more willing to accept as legitimate Oona's decision in *Know-How* than we would be in *Advisor*. We should welcome Oona's acquisition of greater judicial skill.

Now look at yet another scenario.

> *Defer*
> The very same trustworthy independent auditors we met a few cases earlier also discover that if the sage were to make all decisions herself, she would have been more accurate than she is when in tandem with Oona (in other words, arranging the decision-makers in the ascending order of accuracy of the final decisions, we have: Oona alone, then Oona with the advisor, then the advisor herself). When Oona finds out about the auditors' report, she decides to defer to the sage in any situation where the pair's judgments diverge. This improves her accuracy to that of the sage herself.

Assume that the *Know-How* Oona's accuracy matches but does not exceed the *Defer* Oona's accuracy. If we thought that *Know-How* is an improvement over *Advisor*, we should think the same about *Defer* as they do not differ in other respects (specifically, in both cases Oona is unable to provide reasons for some of her decisions). This suggests that we are willing to accept marginal improvements in accuracy despite losses in transparency. Consequently, we should accept not just situations where algorithmic decision-support systems aid human decision-makers, but perhaps also those where human decision-makers defer to machine judgments when the two diverge, provided we can be reasonably sure that the machine's accuracy is superior to the unaided human judgment.

Some evidence suggests that, in very many cases, we can have such assurances. For example, the already quoted Green & Chen (2019) and Grgić-Hlača et al., (2019), as well as a sprawling literature review by Garb & Wood (2019) find that the accuracy of statistical prediction tools (including machine learning algorithms) is greater than that of professionals on their own *and* greater than that of professionals aided by these support tools.

Transparency and the real world

The problem remains, however. As we gain accuracy, we *are* losing transparency. We no longer have access to reasons which guide each decisions, unlike in cases where Oona goes through the process of reasoning herself and makes it explicit to the interested parties.

Or do we?

Recall that the ability to report her reasoning clearly and accurately was an attribute of the *ideal* Oona. In more realistic circumstances, there is some doubt about whether we can expect the same from real decision-makers. For example, Jane Goodman-Delahunty and Siegfried Ludwig Sporer (2010) have reviewed a number of studies on judicial decision-making and found, startlingly, that "[t]hese studies demonstrate that reliance on the articulated reasons for sentencing decisions can be misleading as *these reasons are not reliable indicators of the considerations and factors that influence judicial sentences* [emphasis added]." For example, one of the reviewed studies[12] found "that the judges' descriptions of factors that motivated their decisions rarely matched their performance. In fact, most judges relied on a simple strategy or rule to reach decisions"; in another one,[13] researchers discovered that.

> [a]lthough a complex set of factors to provide individualized justice were mentioned as considerations, and judges inevitably specified that they considered the offender's community ties, analyses of the predictors of their decisions revealed that in fact they applied a simple decision-making strategy and followed the prosecutors' recommendations. In other words, the prosecutors' recommendations predicted bail decisions in the observed cases (Goodman-Delahunty & Sporer, 2010).

This result, as Goodman-Delahunty and Sporer report, was replicated outside the original setting.[14]

Consequently, it is not at all clear whether we do have access to the real reasons behind judicial decisions in the real world. In light of the research that Goodman-Delahunty and Sporer review, it looks like real-world human decision-makers (even highly respected professionals) seem frequently unaware (at best) of the real mechanisms[15] behind

[12] See Ebbesen & Konečni (1981) for details.

[13] See Ebbesen & Konečni (1975) for details.

[14] See Raine & Willson (1995) for details.

[15] Moreover, as Schwartzman (2008) catalogs, surprisingly many legal scholars have advocated for the view that judges ought sometimes to conceal the real reasons for their decisions from the public, the position sometimes explicitly justified by appeal to maintaining the judiciary's legitimacy (e.g. (Idleman, 1994)). However, if such

their own decisions. The comparison between typical human transparency and typical machine opacity appears to compare the ideal with the non-ideal. There seems to be much less human transparency in the real world than in our *Ideal Queen* scenario.[16]

On the whole, then, it looks like the lack of transparency stemming from the use of algorithms in governmental decision-making does not present itself as a formidable obstacle to legitimacy, especially when compared to a viable alternative.[17]

This conclusion appears to agree with Danaher's considered view on algocracy which allows for its permissibility under some conditions:

> A failure to collect sufficient tax undermines many valuable public services. Government revenue agencies (particularly in the wake of the Great Recession) are often understaffed and under-resourced. What is more, the individual humans within those agencies are not always capable of exploiting and seeing connections between different pools of financial data. Algorithms can help. They can mine the relevant data pools for useful patterns, do so tirelessly and efficiently and make recommendations for audits. This could be a great boon for tax collection. The benefits are not hypothetical either. It has already been proven that algorithmic systems are better at making predictions than human experts in certain fields …. Thus, in many instances, it may turn out to be true that if we want to achieve *better outcomes*, we would be well-advised to defer to an algocratic system (2016, references omitted, emphasis added).

Danaher's view expressed in the passage above is, I take it that at least in some cases, outcomes alone should be a deciding factor in whether algocratic governance is implemented. The conclusions of this Section converge: under some circumstances, procedural scruples notwithstanding, algocracy could be morally acceptable because of the goods it can deliver.

Conclusions

Given all I have said so far, one question is bound to suggest itself: what are we to make of the apparent tension between the conclusions of Sect. 4 and those of Sect. 3? After all, Sect. 4 concludes that algorithmically aiding governmental decisions seems acceptable, given the kinds of people we can expect to occupy government roles. On the other hand, imposing such decisions on a population that collectively is unlikely to understand the mechanisms behind them appears problematic, as Sect. 3 concludes.

One resolution of this tension is that there's an implicit ideal vs. non-ideal comparison in Sect. 3 that I have not made clear. I was pointing out informational deficits of real-world voters in the context of algorithmic opacity without clearly setting up a real-world contrast. Of course, in the ideal world, a large degree of comprehensibility and informed participation can be expected of the citizens. But real-world research on voter ignorance and behavior – as cited in this paper – seems to suggest that such conditions are rarely met, regardless of problems of algocracy. Perhaps, then, real-world voters have little to lose from algocratic governance – whether government decisions are made with or without algorithms, they are in fact imposed on populations largely (rationally) ignorant of the mechanisms behind them.

As a result, it seems that, when thinking about algocracy, we should, similarly to what Danaher suggests in the passage cited above, look towards whether it can improve social outcomes instead.

Acknowledgements I am grateful to Daan Kolkman and Anthony Skelton for valuable comments about the manuscript.

Funding n/a.

Availability of Data and Material n/a

Code Availability n/a.

Declarations

Conflicts of Interest/Competing Interests n/a

References

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 2053951715622512. Retrieved from https://doi.org/10.1177/2053951715622512. doi:10.1177/2053951715622512

Cadigan, T. P., & Lowenkamp, C. T. (2011). Implementing risk assessment in the federal pretrial services system. *Federal Probation*, 75(2), 30–38

Cairney, P., & Kwiatkowski, R. (2017). How to communicate effectively with policymakers: combine insights from psychology and

arguments were sound, they could also apply to algorithmic decision-making, where real reasons for some decision could remain obscured and insincere reasons provided instead.

[16] A similar point about a "double standard" with regards to transparency has been proposed by Zerilli et al., (2019). See also Robbins (2019), and some arguments in Zarsky (2013) for different types of skepticism about the transparency ideal.

[17] Interestingly, this conclusion suggests that when real-world policymakers and enforcers insist on transparency to the detriment of other objectives (as some interviewed by Veale et al., (2018) do), they aren't necessarily doing the right thing.

policy studies. *Palgrave Communications, 3*(1), 37. Retrieved from https://doi.org/10.1057/s41599-017-0046-8. doi:10.1057/s41599-017-0046-8

Caplan, B. D. (2007). *The myth of the rational voter: why democracies choose bad policies*. Princeton: Princeton University Press

Caplan, B. D. (2018). *The case against education: why the education system is a waste of time and money*. Princeton, New Jersey: Princeton University Press

Cassidy, M. F., & Buede, D. (2009). Does the accuracy of expert judgment comply with common sense. *Management Decision, 47*(3), 454–469. Retrieved from https://doi.org/10.1108/00251740910946714. doi:10.1108/00251740910946714

Christin, A., Rosenblat, A., & Boyd, D. (2015). Courts and predictive algorithms. *Data & civil rights: A new era of policing and justice*, 1–13. Retrieved from https://datasociety.net/wp-content/uploads/2015/10/Courts_and_Predictive_Algorithms.pdf

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology, 29*(3), 245–268. Retrieved from https://doi.org/10.1007/s13347-015-0211-1. doi:10.1007/s13347-015-0211-1

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences, 108*(17), 6889–6892

Downs, A. C., & Lyons, P. M. (1991). Natural Observations of the Links between Attractiveness and Initial Legal Judgments. *Personality and Social Psychology Bulletin, 17*(5), 541–547. Retrieved from https://doi.org/10.1177/0146167291175009. doi:10.1177/0146167291175009

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), eaao5580. Retrieved from https://www.science.org/doi/abs/10.1126/sciadvhttps://doi.org/10.1126/sciadv.aao5580

Ebbesen, E. B., & Konečni, V. J. (1975). Decision making and information integration in the courts: The setting of bail. *Journal of Personality and Social Psychology*, 32(5), 805

Ebbesen, E. B., & Konečni, V. J. (1981). The process of sentencing adult felons. In B. D. Sales (Ed.), *The trial process* (pp. 413–458). Springer

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, 32(2), 188–200

Eren, O., & Mocan, N. (2018). Emotional judges and unlucky juveniles. *American Economic Journal: Applied Economics*, 10(3), 171–205

Estlund, D. M. (2008). *Democratic authority: a philosophical framework*. Princeton, N.J.: Princeton University Press

Eubanks, V. (2017). *Automating inequality: how high-tech tools profile, police, and punish the poor* (First Edition). New York, NY: St. Martin's Press

Fink, K. (2018). Opening the government's black boxes: freedom of information and algorithmic accountability. *Information, Communication & Society*, 21(10), 1453–1471

Garb, H. N., & Wood, J. M. (2019). Methodological advances in statistical prediction. *Psychological assessment*, 31(12), 1456

Gewirtz, P. (1995). On 'I Know It When I See It'. *Yale Law Journal, 105*(4), 1023–1048. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/ylr105&i=1057

Goodman-Delahunty, J., & Sporer, S. L. (2010). Unconscious influences in sentencing decisions: a research review of psychological sources of disparity. *Australian Journal of Forensic Sciences, 42*(1), 19–36. Retrieved from https://doi.org/10.1080/00450610903391440. doi:10.1080/00450610903391440

Green, B., & Chen, Y. (2019). *Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments.* Paper presented at the Proceedings of the conference on fairness, accountability, and transparency

Grgić-Hlača, N., Engel, C., & Gummadi, K. P. (2019). Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1–25

Houghton, D. P. (2008). Invading and occupying Iraq: Some insights from political psychology. *Peace and Conflict*, 14(2), 169–192

Idleman, S. C. (1994). Prudential Theory of Judicial Candor. *Texas Law Review, 73*(6), 1307–1418. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/tlr73&i=1325

Koppl, R. (2018). *Expert failure* (1 Edition. ed.). New York: Cambridge University Press

Krek, A. (2005). *Rational ignorance of the citizens in public participatory planning.* Paper presented at the 10th symposium on Information-and communication technologies (ICT) in urban planning and spatial development and impacts of ICT on physical space, CORP

Lemieux, P. (2004). The public choice revolution. *Regulation*, 27, 22

Lin Zhiyuan, J., Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances, 6*(7), eaaz0652. Retrieved from https://doi.org/10.1126/sciadv.aaz0652. doi:10.1126/sciadv.aaz0652

Peer, E., & Gamliel, E. (2013). Heuristics and Biases in Judicial Decisions. *Court Review, 49*(2), 114–119. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/ctrev49&i=114

Raine, J. W., & Willson, M. J. (1995). *Conditional Bail Or Bail with Conditions?: The Use and Effectiveness of Bail Conditions*. Institute of Local Government Studies, the University of Birmingham

Robbins, S. (2019). A misdirected principle with a catch: explicability for AI. *Minds and Machines*, 29(4), 495–514

Schwartzman, M. (2008). Judicial sincerity.Virginia Law Review,987–1027

Somin, I. (2015). Rational ignorance. *Routledge international handbook of ignorance studies*, 274–281

Stewart, J. E. (1980). Defendant's Attractiveness as a Factor in the Outcome of Criminal Trials: An Observational Study 1. *Journal of Applied Social Psychology*, 10(4), 348–361

Veale, M., Van Kleek, M., & Binns, R. (2018). *Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making.* Paper presented at the Proceedings of the 2018 chi conference on human factors in computing systems

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the Gdpr. *Harvard Journal of Law & Technology*, 31(2), 841

Yetiv, S. A. (2013). *National security through a cockeyed lens: How cognitive bias impacts US foreign policy*. JHU Press

Zarsky, T. Z. (2013). Transparent Predictions. *University of Illinois Law Review, 2013*(4), 1503–1570. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/unilllr2013&i=1537

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: is there a double standard? *Philosophy & Technology*, 32(4), 661–683

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.