The Threat of Algocracy: Reality, Resistance and Accommodation

By John Danaher

Forthcoming in *Philosophy and Technology*

Abstract

One of the most noticeable trends in recent years has been the increasing reliance of public decision-making processes (bureaucratic, legislative and legal) on algorithms, i.e. computer programmed step-by-step instructions for taking a given set of inputs and producing an output. The question raised by this article is whether the rise of such algorithmic governance creates problems for the moral or political legitimacy of our public decision-making processes. Ignoring common concerns with data protection and privacy, it is argued that algorithm-driven decision-making does pose a significant threat to the legitimacy of such processes. Modeling my argument on Estlund's threat of epistocracy, I call this the 'threat of algocracy'. The article clarifies the nature of this threat, and addresses two possible solutions (named, respectively, "resistance" and "accommodation"). It is argued that neither solution is likely to be successful, at least not without risking many other things we value about social decision-making. The result is a somewhat pessimistic conclusion in which we confront the possibility that we are creating decision-making processes that constrain and limit opportunities for human participation.

Keywords: Algocracy; Epistocracy; Big Data; Data-Mining; Legitimacy; Human Enhancement

1. Introduction

We live in an age of algorithmic decision-making. There are algorithms trading stocks on Wall Street (Patterson 2013); algorithms determining who is the most likely to be guilty of tax evasion (Zarsky 2013); algorithms assisting in scientific discovery (Mayer-Schonberger and Cukier 2013); and algorithms helping us in dating and mating (Slater 2013). This is just a small sample: many more could be listed (Siegel 2013). With the ongoing data revolution, and the transition towards the so-called "Internet of Things" this trend can only be set to grow (Kitchin 2014a; Kellermeit & Obodovski 2013; Rifkin 2014).

The question raised by this article is whether the use of such algorithm-based decision-making in the public and political sphere is problematic. Suppose that the creation of new legislation, or the adjudication of a legal trial, or the implementation of a regulatory policy relies heavily on algorithmic assistance. Would the resulting outputs be morally problematic? As public decision-making processes that issue coercive rules and judgments, it is widely agreed that such processes should be morally and politically *legitimate* (Peter 2014). Could algorithm-based decision-making somehow undermine this legitimacy?

In this article, I argue that it could. Although many are concerned about the *hiddenness* of algorithmic decision-making, I argue that there is an equally (if not more) serious problem concerning its *opacity* (potential incomprehensibility to human reasoning). Using David Estlund's (1993; 2003; 2008) threat of epistocracy argument as my model, I argue that increasing reliance on algorithms gives rise to the *threat of algocracy* – a situation in which algorithm-based systems structure and constrain the opportunities for human participation in, and comprehension of, public decision-making. This is a significant threat, one that is difficult to accommodate or resist.

The article proceeds as follows. In section 2, I clarify the phenomenon of interest, explaining what is and is not meant by my use of the term 'algocracy'. In section 3, I make the case for the potential 'threat'. In section 4, I argue it is morally undesirable to resist the threat. And in section 5, I argue that it is hard to accommodate the threat in a satisfactory manner. I conclude on a somewhat pessimistic note,

highlighting the fact that we may be creating a governance structure that has instrumental and procedural virtues, but sacrifices human control and comprehension.

2. What is algocracy?

The term 'algocracy' has the potential to mislead. I use it in a precise manner here, building upon previous uses of the same term (Aneesh 2006; 2008), and linking it to the related concept of 'epistocracy' in political philosophy (Estlund 1993; 2003; 2008; and, particularly, Lippert-Rasmussen 2012).

Let me start by preempting and heading-off some potential misconceptions. I do not use the term to describe a system in which computers or artificial agents seize control of governmental decision-making bodies and then exercise power in way that serves their needs and interests. Something of that sort may be possible in the future, but I am here concerned with a more mundane (and extant) phenomenon.[1] Also, I do not mean for the term to carry pejorative connotations. The suffix 'cracy', when added to the end of related words like 'bureaucracy' and 'technocracy' often has such connotations, but this need not be the case. After all, the term 'democracy' has the same suffix and typically has positive (or at least neutral) connotations. I intend for the bare term 'algocracy' to have similarly neutral connotations. As will become clear below (section 3), I think that algocratic systems can have many positive qualities; it is just that they can also have negative qualities, some of which are identified by the argument to the threat of algocracy.

I use the term 'algocracy' to describe a particular kind of governance system, one which is organised and structured on the basis of computer-programmed algorithms.[2] To be more precise, I use it to describe 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

[1] The possibility of an intelligent AI controlling the world is explored at length in Bostrom 2014.
[2] I add 'computer programmed' here since algorithms are, in effect, recipes or step-by-step instructions for deriving outputs from a set of inputs. As such, algorithms do not need to be implemented by some
[2] I add 'computer programmed' here since algorithms are, in effect, recipes or step-by-step instructions for deriving outputs from a set of inputs. As such, algorithms do not need to be implemented by some computer architecture, but I limit interest to computer-programmed variants because the threat of algocracy is acutely linked to the data-revolution (Kitchin 2014a).

those systems interact with one another, the relevant data, and the broader community affected by those systems. This can be done by algorithms packaging and organizing the information in a particular way or even by algorithms forcing changes in the structure of the physical environment in which the humans operate (Kitchin and Dodge 2011). Such systems may be automated or semi-automated,[3] or may retain human supervision and input.

In using the term in this sense, I build upon pre-existing uses in the sociological literature. Aneesh (2006; 2009) for instance uses 'algocracy' in his analysis of labour migration to denote an organisational system that is distinct from a market or a bureaucracy. For Aneesh, a *market* is a system in which *prices* structure and constrain the ways in which humans act; a *bureaucracy* is a system in which *laws and regulations* structure and constrain the ways in which humans act; and an *algocracy* is a system in which *algorithms* structure and constrain the ways in which humans act. The boundaries between such systems are not precise: they often integrate with and overlap with one another. This is important in the present context because, as I understand them, algocratic decision-making systems can be integrated into pre-existing legal-bureaucratic decision-making systems.[4]

In adopting this definition, one can remain agnostic about the precise technological basis for an algocratic system. Nevertheless, I will give a more specific sense of the phenomenon with which I am concerned. I am particularly concerned about the growth in algocratic systems that are based on predictive or descriptive data-mining algorithms (Kitchin 2014a). I follow Zarsky (and others) in defining data-mining as: "the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data" (Zarsky 2011, 291). These patterns can be used descriptively — to explain or understand what has happened in the past — or predictively — to preempt or anticipate future behaviour. To give an example of the former, data-mining could be used to trawl through financial records to detect past instances of fraud. To give an example of the latter, data-mining could be used to predict, from historical datasets, which criminal is most likely to reoffend, or who is most likely to be a terrorist. One of the distinctive features of such data-mining systems

[3] Dormehl gives some striking illustrations of bureaucratic systems that are automated, e.g. the facial recognition algorithm system used to revoke driving licences in Massachusetts (Dormehl 2014, 157-58)
[4] There are also connections here with Lessig's work (1999 and 2006) on code as a type of regulatory architecture. Lessig is concerned primarily with who owns and controls that architecture; I am concerned with ways in which that architecture facilitates a lack of transparency in public decision-making.

in the modern era is that they rely on extremely large datasets ("Big Data"), collected from growing networks of data-monitoring technologies. Humans can be more or less involved in the data-mining process and in the kinds of decisions made on foot of the data-mining process. Humans can predetermine the patterns that data-mining algorithms search for ("subject-based searches"), or they can allow the algorithms to find the patterns ("pattern-based searches") (Zarsky 2011, 291-292); humans can review and scrutinise the recommendations made by algorithms, or they can essentially leave it up to the machines, acting as mere implementers of algorithm-based judgments. In some cases, the systems can be entirely automated.

The debate about military drones has generated some useful distinctions between types of robotic weapon system that is relevant to this topic. They are (Citron & Pasquale 2014):

> ***Human-in-the-loop weapons***: Robots can only select targets and deliver force with a human command.
>
> ***Human-on-the-loop weapons***: Robots can select targets and deliver force on their own, but there is human oversight and the possibility of human override.
>
> ***Human-out-of-the-loop weapons***: Robots act autonomously, selecting targets and delivering force without human oversight or override.

We are not concerned with robotic weapons here, of course, but the distinctions can be applied to any algocratic system. Take a tax law-enforcement system as an example. A human-in-the-loop version could rely on algorithms to select targets for auditing, but only if a human agent requests or demands this. Conversely, a human-on-the-loop system could work autonomously, constantly sorting through collected data, identifying important patterns, and automatically issuing recommendations, arrest warrants or even court summonses. These would ultimately be implemented by human agents who would choose whether or not to follow what the algorithm tells them. Human-out-of-the-loop systems would leave everything up to the machines.

One final conceptual distinction is needed before we can proceed to the argument proper. When considering the extent of human involvement in algocratic systems, we

need to be aware that some systems make this easier than others. There is a distinction between data-mining systems that are 'interpretable' and those that are 'non-interpretable' (Zarsky 2011 & 2013). The former are based on rationales and factors that can be interpreted and understood by human beings – in short that can be "reduced to a human language explanation" (Zarsky 2011, 293). Non-interpretable systems cannot be reduced to such explanations. They rely on factors that are too complex for humans to understand. If such systems were in place, even if humans were still "on" or "in" the loop, they may be ill-equipped to second-guess the algorithmic judgment. This is particularly important given the growth in the use of machine learning algorithms to find patterns and make predictions from data. Interpretability is a recognized problem in that field. I return to this in due course.

3. What is the threat of algocracy?

For the purposes of this discussion,[5] the recent growth in algocratic systems can be said to raise two moral and political concerns:

> **Hiddenness Concern**: This is the concern about the manner in which our data is collected and used by these systems. People are concerned that this is done in a covert and hidden manner, without the consent of those whose data it is.
>
> **Opacity Concern**: This is a concern about the intellectual and rational basis for these algocratic systems. There is a worry that these systems work in ways that are inaccessible or opaque to human reason and understanding.

The first of these concerns has given rise to a rich literature,[6] a contentious political debate[7] and a range of legal regulations and guidelines.[8] For example, in 2014 the European Court of Justice delivered a verdict striking down a European data retention directive.[9] The directive required telecoms operators to store data about their customers

[5] Debates about other systems, e.g. automated cars and weapon systems, can raise other moral and political issues.
[6] For an overview, see the *Stanford Law Review* symposium issue on Privacy and Big Data. Available at: http://www.stanfordlawreview.org/online/privacy-and-big-data (visited 10/4/14)
[7] The Edward Snowden controversy being, perhaps, the most conspicuous example of this.
[8] For example, the European Directive on this is Directive 95/46/EC
[9] Case C-293/12 (joined with Case C-594/12 *Digital Rights Ireland Ltd v. Minister for Communications, Marine and Natural Resources, and Ors* 8th April 2014

for up to two years. The court struck this down on the grounds that it "entail[ed] a wide-ranging and particularly serious interference with the fundamental rights to respect for private life and to the protection of personal data."[10] As is apparent from this statement, the normative grounding for the hiddenness concern lies in concepts of privacy and control over personal information.

The opacity concern is rather different and has generated less debate (although this is now beginning to change). The opacity concern has nothing to do with privacy and control over personal information, although it is testament to the overpowering nature of the hiddenness concern that those few theorists who have begun to discuss opacity often couch their analysis in terms of privacy or personal information (Morozov 2013; Crawford & Schultz 2014). The opacity concern has to do with our participation in political procedures, and how this participation is undermined by growing use of algocratic systems. The normative grounding for this concern is in concepts of political authority and legitimacy.[11] I will explain this normative grounding first before proceeding to defend an argument in relation to the opacity concern. This argument will constitute the *threat* of algocracy.

Legitimacy is the property that coercive public decision-making processes must possess if they are to rightfully exercise the requisite authority over our lives. There are many different accounts of what it is that makes a decision-making procedure legitimate (Peter 2014). Broadly speaking, there are three schools of thought. The *pure instrumentalists* think that a procedure gains legitimacy solely in virtue of its consequences: procedures are instruments that have normative aims (reducing crime, increasing well-being etc.), and the better they are at achieving those aims, the more legitimate they are.[12] Contrariwise, the *pure proceduralists* think that it is difficult to know what the ideal outcome is in advance. Hence, they tend to emphasise the need for our procedures to exhibit certain outcome-independent virtues (Peter 2008). For example, they might argue that our procedures should create ideal speech situations, allowing for those who are affected by them to comprehend what is going on, and to contribute to the decision-making process (Habermas 1990). It is also possible to adopt

[10] *Ibid*, para. 65
[11] There may also, of course, be a connection here with a more substantive conception of justice (Ceva 2012).
[12] I'm not sure that there are any pure instrumentalists, but those who endorse an epistemic theory of democracy certainly emphasise this virtue (Estlund 2008; List & Goodin 2001)

mixed or *pluralist* approaches to legitimacy, which focus on both the properties of the procedures and their outcomes.

I favour the mixed approach. There are two reasons for this. First, because I believe pure versions of instrumentalism and proceduralism can lead to odd conclusions about the legitimacy of a procedure.[13] If all you cared about was the outcome of a decision-making procedure, you might be able to justify an evidence-gathering process that included cruel or inhumane treatment of human witnesses, provided that such treatment facilitated a more accurate decision. Likewise, if all you cared about was the procedure itself, you might be able to justify a process which clearly led to a decision with bad consequences simply because it treated people with respect and allowed them some meaningful participation. Neither of these is intuitively appealing. Second, the concept of an 'outcome' or a 'procedure' is sufficiently fuzzy to allow for plenty of debate about what counts as being part of an outcome and part of a procedure. Is an evidence-gathering procedure that treats someone inhumanely but gathers accurate information warranted because of its outcomes? Or should the longer-term suffering of the person from whom the information is gathered be included in any assessment of those outcomes? The answer is not entirely clear, but instrumentalists might be inclined to favour the latter view since inhumane treatment feels like something that should undermine the legitimacy of a decision-making process. The advantage of the mixed approach is that does not need to concern itself with such debates. The treatment of the witness is relevant either way. This is important because in favouring the mixed approach one sometimes needs to assess decision-making processes in light of the trade offs between their instrumental and procedural virtues.

With the normative grounding clarified, the opacity problem can be articulated. It helps to do this by way of analogy with Estlund's threat of epistocracy argument (Estlund 2003 & 2008; Machin 2009; Lippert-Rasmussen 2012). Estlund's argument is that those who are enamoured with outcome-oriented approaches to legitimacy may be forced to endorse the legitimacy of epistocratic systems of governance. He points out that if we assume (plausibly) that legitimacy-conferring outcomes are more likely to be achieved by those with better epistemic abilities, then the following argument seems compelling:

[13] The oddness reflects arguments in the consequentialist/deontologist debate in ethics.

(1) There are procedure-independent outcomes against which the legitimacy of public decision-making procedures ought to be judged. (Cognitivist Thesis)

(2) In any given society, there will be a group of people with superior epistemic access to these procedure-independent outcomes. (Elitist Thesis)

(3) If there are people with superior epistemic access to these procedure independent outcomes, then procedures are more likely to be legitimate if those people are given sole or predominant decision-making authority.

(4) Therefore, in any given society, decision-making procedures are more likely to be legitimate if authority is concentrated in an epistemic elite. (Authority Thesis)

The argument depends on a normative claim (viz. outcomes confer legitimacy on decisions) and two factual claims. The first factual claim is that there is such a thing as an epistemic elite, a sub-group of the population with superior epistemic access to the legitimacy-conferring outcomes; the second is that handing over decision-making authority to this sub-group is likely to get us closer to those outcomes. There are ways in which we could critique these factual assumptions (Lippert-Rasmussen 2012). But I will not do so here since my goal is not to defend Estlund's argument but to develop a similar but different argument.

To do this, I need to consider in more detail what is meant by an 'epistocracy' and how it relates that of algocracy. Estlund defines the concept of epistocracy (2003, 53) by reference to sub-populations of human societies with *generally superior* epistemic capabilities. For him, this sub-population constitutes an epistemically elite group of citizens who, if the logic of the argument is to be followed, get to control and detain power across all public decision-making processes. Thus, when he talks about the threat of epistocracy he seems to be talking about a threat emanating from *stable group* of human agents who are allocated significant social power. But, as Lippert-Rasmussen (2012) points out, this conflation of epistocracy with a *stable* and *generally elite* sub-population is misleading; epistocracy is a broader concept than that. A sub-population could have superior epistemic access to legitimizing outcomes for *emergent* and *highly contingent* reasons. In other words, the individual members of the sub-population need

not have generally superior epistemic abilities. They may have superior access for a limited set of decisions, for a narrowly constrained period of time, or because the sub-population as a whole (and not any individual) emergently satisfies some set of conditions that enables them to have superior epistemic access.[14] The crucial point is that the sub-population is epistemically superior and is favoured for this reason. This epistemic favouring is what turns a democracy into an epistocracy.

If we adopt this broader definition of epistocracy, the threat alluded to by Estlund's argument would arise whenever we favour a sub-population of decision-makers for epistemic reasons. This broader definition is more in keeping with the concept of algocracy. An algocratic system is one organised on the basis of algorithms which structure and constrain the opportunities for human interaction with that system. One could imagine people favouring the implementation of such systems for epistemic reasons – in other words, because such systems are thought to have some privileged or superior epistemic access to legitimacy-conferring outcomes, when compared to a purely human alternative. Thus, when I talk about a *threat* of 'algocracy', I am talking about a threat that arises from this sort of epistemic favouring of algocratic systems.

The question, of course, is whether favouring such systems undermines legitimacy. Estlund thinks it does in the case of epistocractic systems. His argument is that such systems are problematic because they fail to satisfy important legitimacy conditions of general acceptability, reasonable rejectability, and publicity. In Estlund's model this means that the procedures must be justifiable to people in terms of reasons that are accessible and comprehensible to them (Estlund 2008; Machin 2009). 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 (Gaus 2010). And others equally insist that this requires procedures in which people can participate and deliberate (see discussions in Machin 2009; Habermas 1990; Besson and Marti 2006).

[14] A classic example would be if the sub-population satisfies the conditions for the Condorcet Jury Theorem or one of its extrapolations (e.g. List and Goodin, 2001).

The problem with epistocratic systems is that these non-opacity requirements may fail to be met. Relevant sub-populations may not have access to the rationales underlying their decisions (they may know more than they can tell);[15] they may have access but may not be able to make them comprehensible to the general population (Machin 2009 discusses the difficulties with this); or their epistemic superiority may be attributable to contingent or emergent factors that they themselves are unable to fully articulate. Initially we (or someone) may know that they satisfy relevant conditions of superiority, but over time we may lose sight of those conditions whilst still maintaining deference to that sub-population.[16]

My argument is that algocratic systems can likewise fail to meet the requirements of non-opacity. Indeed, the likelihood of non-opacity may be even higher in the case of algocratic systems. This is why it is meaningful to refer to a 'threat' of algocracy. The threat is one that can sneak up on us. We may initially favour algocratic governance systems for appropriate instrumental reasons, impressed by their greater speed, accuracy and insight (when compared to similar human systems), and we may be keen to take advantage of their impressive results. But in favouring them we may end up with systems that are increasingly opaque. Morozov expresses the point rather nicely:[17]

> *Thanks to smartphones or Google Glass, we can now be pinged whenever we are about to do something stupid, unhealthy or unsound. We wouldn't necessarily need to know why the action would be wrong: the system's algorithms do the moral calculus on their own. Citizens take on the role of information machines that feed the techno-bureaucratic complex with our data. And why wouldn't we, if we are promised slimmer waistlines, cleaner air, or longer (and safer) lives in return?*

We then become trapped, as Morozov puts it, in a web of "invisible barbed wire". We are convinced that the algorithmic control systems enhance our autonomy, increase our health and well-being, and improve social outcomes, but we don't have clear sense of

[15] This is a reference to the work of Michael Polanyi (1966).
[16] Estlund offers alternative arguments for thinking that epistocracies are politically problematic. These have to do with reasonable rejection on the grounds of suspicion of the epistemic elite. I ignore those arguments here since they tie into his conflation of epistocracy with rule by a stable group of generally superior human agents.
[17] Morozov (2013) - see the subsection entitled "Even programmes that seem innocuous can undermine democracy" for this quote.

how *exactly* they manage to do this. The result is social spaces that are opaque to human reason.[18]

A simple illustration might help to underscore this point.[19] In recent years, the online retailer Amazon has taken to stocking some of its large warehouses using a "chaotic storage algorithm" (Greenfield 2012; Bumbulsky 2013). For centuries, humans have stocked warehouses and similar storage facilities by following their own "algorithms". For example, they might stock them by grouping similar items together (books, DVDs, home furniture, appliances etc.) and then subdividing those groups along various lines (e.g. alphabetical order, sub-genre, type of furniture or appliance). The rationales behind these storage systems make sense, and are clearly understandable by ordinary human beings. Furthermore, the process of identifying items and fulfilling orders is one that humans can fully comprehend and participate in. The chaotic storage algorithm system is rather different. The system works by tagging every item that enters the warehouse with a barcode and then assigning it to a location in the warehouse based on available shelf-space. This is done by algorithm. The result is a system that is apparently far more efficient (less wasted product, faster turnover of stock), and in which very different products are located side-by-side on the shelves. When it comes time to fill an order, a human worker[20] must rely on an algorithm to plot a course through the warehouse for them to pick up the various items.

This creates a very interesting physical working environment. It is one in which humans are "on the loop", but whose organisation is determined by the algorithms and whose physical space cannot be navigated (by humans) without algorithmic assistance. There is consequently deference to the epistemic superiority of the algocratic system. Now, to be clear, the chaotic storage system is not *completely opaque* to human reason. It has an underlying purpose that can be followed by human beings (*viz.* assignment based on shelf-space leads to greater efficiency). That purpose is attractive, even appealing to the humans who create it. Who wouldn't want a more efficient storage system? The problem is that the actual mechanics of the algorithm are too complex for any one human to follow. A human could not keep track of the barcodes, nor the

[18] The society that worries Morozov is no imaginative dystopia. It is actively pursued by some: see Alex Pentland (2014)

[19] I take this illustration from the artist James Bridle who uses it in some of his talks. See http://shorttermmemoryloss.com/ for more.

[20] For the time being anyway. It is likely that, in the future, robot workers will take over such systems. Amazon already works with Kiva robots in some warehouses. See http://www.youtube.com/watch?v=3UxZDJ1HiPE (visited 1/3/15) for a video illustration.

available shelf space. They need to outsource all of this understanding to the machine. The result is that they start to imprison themselves in the "invisible barbed wire" mentioned by Morozov.

My argument for the threat of algocracy works on the belief that what is happening in Amazon warehouses can happen on a much larger and more invidious scale in public decision-making procedures. We could introduce and defer to more and more algocratic systems, starting with ones that are relatively easy to follow, but which morph into systems that are far more complex and outside the upper limits of human reason. It will be much more difficult to fall back on the need for participation and comprehension here because the scope for genuine human participation will be much more limited: the algorithms will be organizing and manipulating vast streams of data and will be grafted on top of an increasingly complex ecosystem of other algorithms. The result is that we end up with a set of decision-making procedures that are depleted of their legitimacy.

This can be summarised as a simple argument:

> (5) Legitimate decision-making procedures must allow for human participation in and comprehension of those decision-making procedures.
>
> (6) Increasing reliance on algocratic systems limits the scope for active human participation in and comprehension of decision-making procedures.
>
> (7) Therefore, reliance on algocratic systems is a threat to legitimate decision-making procedures.

There are two initial doubts one might have about this argument. First, we may wonder whether we can simply create algorithmic systems that allow for participation and comprehension (contrary to the claims of premise 6). And second, we may wonder whether this threat is really posed by the systems themselves, or the elite group of programmers and coders who design them.

The first of these doubts draws our attention to the nature of the participation and comprehension requirements. What level of understanding is needed in order for

legitimacy to be achieved? If, for example, we had a complicated tax-evasion monitoring algorithm, wouldn't it be enough for people to simply know that the system works by identifying those most likely to be tax evaders (just as the Amazon workers know, roughly, how the system works and its purpose)? Do they really need to know precisely which factors trigger the system? In other words, isn't a coarse-grained description of the rational basis for the system enough? No; this shouldn't be enough. If we are to respect the moral equality of individual citizens, we cannot legitimately exercise coercive authority over them in such a manner. It is not enough for them to simply know that the system is more likely to reach preferred outcomes; they must be able to scrutinise and critically engage with the factors that enable the system to do this. This doesn't mean that an extremely fine-grained understanding of the algocratic system is required, but we need more that just the general rationale.

But then we may ask: why can't we simply ensure that we create algorithmic systems that are more amenable to such understanding and participation? In principle this may be possible, but three factors combine to make it exceptionally difficult. The first is that many algorithmic systems are protected by secrecy laws, either because they are based on 'trade secrets' and associated commercial interests, or because they are used by government agencies and there are governmental interests in preventing people from gaming or hacking these systems (Pasquale 2015 discusses this issue at length). Laws of this sort could be dismantled and reconstructed to facilitate greater transparency, but the difficulty of doing so should not be underestimated given the powerful commercial and governmental interests at stake. The second factor is that modern data-mining systems increasingly rely on machine learning algorithms. This is partly due to the increase in the size of the datasets that must be mined for useful information. The unique thing about such algorithms is that humans do not have to pre-select or pre-determine the rules or principles the algorithms use to perform their tasks; instead the algorithms can be trained on large datasets to generate their own rules and principles. Famous examples include product-recommendation algorithms and IBM's Watson. The problem is that the interpretability of the outputs of such algorithms is a significant and recognised problem in the field of machine learning.[21] The algorithms are often not able to tell programmers exactly why they produce the outputs they do. People are working on more interpretable methods, but there seem to tradeoffs involved

[21] For example, neural network models are widely recognized as having an interpretability problem. See, for example, the discussion in Miner et al 2014, 249.

in making such systems more interpretable (Vellido, Guerrero and Lisboa 2012; Lisboa 2013; Otte 2013; Chase Lipton 2015; Zeng, Ustun and Rudin 2015).[22] Finally, compounding these two problems, there is the fact algorithms are not singular phenomena. Any new algorithm is likely to be grafted on top of others, collectively authored by teams of coders using pre-existing coded architectures, and then woven into increasingly complex algorithmic ecosystems (Seaver 2013; Kitchin 2014a & 2014b). It is the interaction between all the members of this algorithmic ecosystem that produces the useful output, not the operation of the single new algorithm. But when you have such a complex ecosystem, the scope for individual participation and understanding is further limited. Even if it were possible for an individual to deconstruct and understand the system as a whole, it would be an extremely time-consuming and labourious process.[23] A lack of opacity is consequently likely.

This leads to the second worry.[24] Isn't it true to say that in the case of any algocratic system there is a set of human elites behind it? Thus, the threat is not posed by deference to the systems themselves, but rather to the elites that programme and engineer them. This looks right at a first glance. In the case of something like the Amazon chaotic storage algorithm, there is a group of algorithm designers and company management who use their preferred ideology to create an algocratic system that structures their warehouses and constrains their workers. A similar process would surely be followed in other domains: politicians (or other public authorities) would present project 'specs' to computer programmers, who would then use their superior epistemic abilities to create an algocratic system that implements the relevant ideological aim ('efficiency', "crime reduction", "well-being enhancement" or whatever). But (a) as just mentioned, there as ways in which such systems could go beyond the comprehension of even these elites; and (b) even if true, this should provide us with no real solace as ceding political authority to such a group is also procedurally problematic. It reduces the threat of algocracy to the threat of epistocracy. I return to this point in section 5, below.

[22] It is also worth noting that 'interpretability', for many working in this field, seems to mean 'interpretability by appropriately trained peers'. This would be insufficient for political purposes.
[23] I would like to thank an anonymous reviewer for encouraging further discussion of this issue.
[24] I am indebted to DI for pressing me on this point. This reduction would raise similar kinds of concerns to those animating Lessig in his classic works on the topic (1999 & 2006).

4. Should we resist the threat?

If the threat is real, should we do something to resist it and to protect the legitimacy of our political system? This certainly seems to be the view of some commentators. Evgeny Morozov (2013), for example, urges us to *politicise the problem* and *sabotage the system* in order to protect our democratic values. Some people may be attracted to this model of political resistance but there are two reasons to question it. First, it is not clear that resistance of this sort would be practically achievable across the full spectrum of public decision-making processes. Second, and probably more importantly, it is not clear that resistance of this sort is *morally preferable*: there is a moral case to be made for the use of algocratic systems both on instrumentalist and proceduralist grounds. There is consequently a tradeoff of values involved that may render accommodation more appropriate than resistance.

The practicality of resistance is not my major concern here, but if we assume that resistance requires us to block and dismantle algocratic systems, then there are two hurdles that are worth noting. The first is simply the increasing *ubiquity* of the relevant technologies, in particular the data-monitoring technologies that feed algocratic systems. The second is the increasing *hiddenness* of those technologies. Ubiquity and hiddenness might look like uncomfortable bedfellows, but the ubiquitous presence of data-monitoring and mining technologies often leads them to hide in plain sight. We all now make use of technologies with data-mining potential on a daily basis, we do so because they are essential to how we live and work, but this can often desensitise or blind us to the algocratic possibilities. We know the systems there, but we are not fully cognisant of their uses and effects. This trend is only likely to increase as monitoring technologies become smaller, more efficient, and more ubiquitous (Brin 1997).

More important than this, however, is the overarching desirability of algocratic systems. The 'threat' of algocracy challenges such systems because of their likely opacity, but this is just one moral mark against them. It needs to be weighed alongside other marks (such as the impact on privacy) and alongside other benefits. It is important not to ignore the benefits. There are often powerful instrumental benefits associated with the construction and use of algocratic systems (Mayer-Schonberger and Cukier 2013). We are collecting and drawing together ever-larger datasets, and algocratic technologies give us some hope of leveraging those datasets to good effect. This is true

for both social authorities and for the public at large. To give a simple example, smart electricity grids, which rely heavily on data-monitoring and mining technologies, can help to boost the effectiveness and efficiency of renewable energy sources (Rifkin 2014, Ch 5). This is highly desirable in an era of climate change and energy insecurity. Amazon's chaotic storage algorithms — whatever you might think of the company itself and its wider work practices — do help to reduce waste and inefficiencies and increase profitability. And even self-monitoring and self-tracking apps, like the ones we use on our phones everyday, can help to improve individual productivity, health and well-being, primarily by helping us with goal setting, self-experimentation and habit formation.[25]

The same is true when we consider the public sphere. To give an example, tax evasion is a major problem: 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's 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 (Bishop & Trout 2002; Meehl 1996). 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.

And it is not all about outcomes either. There may be procedural benefits to algocratic systems too. Zarsky makes the case for this (2011 & 2012). He argues that one major procedural deficiency with human-based decision-making systems is their susceptibility to implicit bias. Consider the profiling debate in relation to anti-terrorism and crime-prevention as an example. One concern with profiling is that it can arbitrarily target and discriminate against certain racial and ethnic minorities. That is something that we could do without. If people are going to be targeted by such measures, they need to be targeted on legitimate grounds (*i.e.* because they are genuinely more likely to be terrorists or to commit crimes). The problem is that, because of implicit biases, the

[25] A stark example of this is the Pavlok, a technology which uses basic principles of psychological conditioning to encourage behavioural change. See http://pavlok.com - note how the website promises to 'break bad habits in five days'.

human authorities may not be able to do this. Automated algocratic systems could be constructed in such a way as to not be prone to the same implicit biases. As such, they may be procedurally preferable to human-based systsms. As Zarsky puts it:

> *[A]utomation introduces a surprising benefit. By limiting the role of human discretion and intuition and relying upon computer-driven decisions this process protects minorities and other weaker groups.*
>
> (Zarsky 2012, 35)

Indeed, Zarsky goes even further and suggests that one explanation for the unease towards algocratic systems might be the preference of the privileged majority for systems that place the burdens on minorities (Zarsky 2012, 35). Thus, the privileged would prefer a profiling system administered by humans, because they could rely on those humans being biased in their favour. They could not rely on the automated system doing the same.

I do not wish to endorse Zarsky's argument here. There are, as he and others have noted (Citron & Pasquale 2014), reasons for thinking that automated systems could replicate the biases of humans. Algorithm construction is a *translation process* (Kitchin 2014b): a problem or task must be converted into a set of step-by-step instructions which must in turn be translated into computer code. There is plenty of space in this translation process for implicit or even explicit biases to play a role. But if we are conscientious about this possibility, we may be able to filter out or reduce the potential for bias. In this sense, Zarsky's argument points us in an interesting direction. It suggests that in addition to securing better outcomes, algocratic systems could be procedurally fairer to those affected by them. Thus, in assessing how to respond to the threat of algocracy, we will need to balance the loss in comprehension and participation against the potential gains in outcomes and procedural fairness. The fact that such a complex weighting exercise may need to be undertaken should give us some reason to reject resistance as a solution to the threat. Perhaps, instead, we should try to keep the algocratic systems and preserve participation in some other way?

5. Can we accommodate the threat?

In this section, I look at four *accommodating* solutions to the threat of algocracy. Each of these solutions tries to keep humans in the decision-making loop, and preserve their ability to participate in that loop. This would protect against the problem of opacity whilst still allowing us to reap the benefits of the algocratic systems. The solutions move from the relatively mundane case of insisting upon human review to the more outlandish possibilities of human-machine integration. I argue that, in each case, it is difficult see how the solution could accommodate the threat by itself, though in various combinations they may suffice.

5.1 - Insist upon human review of algorithms

This is a solution that straddles the boundary between resistance and accommodation. It tries to avoid the threat of algocracy by keeping humans on the loop, and allowing them some substantial review and/or override power.

A version of this solution is already part of the law in the European Union. According to Article 15 of the European Directive 95/46/EC (the *Data Protection Directive*), there must be human review of any automated data-processing system that could have a substantial impact on an individual's life. The official wording is as follows (emphasis added):

> *15.1 - Member States shall grant the right to every person not to be subject to a decision which produces legal effects concerning him or significantly affects him and which is based solely on automated processing of data intended to evaluate certain personal aspects relating to him, such as his performance at work, creditworthiness, reliability, conduct, etc.*

The Directive does, however, allow for certain exceptions to this rule. Specifically, it allows for people to voluntarily contract themselves out of this right, and for governments to override it so long as other measures are taken for protecting the individual's "legitimate interests".[26]

[26] Directive 95/46/EC, Art. 15.3

Similar solutions have been proposed in recent papers from Citron and Pasquale (2014; and Citron 2012), and Crawford and Schultz (2014). Both sets of authors are concerned with the impact of automated prediction on due process rights. Citron and Pasquale focus in particular on algorithms used for assessing creditworthiness. They argue that it is essential for the due process rights of those who might be unfairly stigmatised by such assessments to be protected. To this end, they call for some regulatory oversight of the algorithms, as well as public transparency in how the algorithmic systems work. They acknowledge the possibility that such transparency will allow individuals to "game the system", but they dismiss this on the grounds that there is no credible evidence in favour of it. Crawford and Schultz cast their net more broadly than Citron and Pasquale, covering many different uses of predictive algorithms. Nevertheless, their proposed solution is similar. They call for a system of "procedural data due process" rights. This would consist of three elements: (i) *notice* - i.e. subjects are made aware when they have been targeted by an algocratic system; (ii) *opportunity for a fair hearing* - i.e. subjects are allowed to review the evidence used against them and the algorithmic logic applied; and (iii) *judicial review by an impartial adjudicator* - i.e. subjects are allowed to appeal algocratic decisions to an impartial adjudicator, such as a court of law.

Can human review balance the potential benefits of algocratic systems with the concerns about participation and comprehension? There are at least two reasons for thinking that it cannot. The first is that the nature of the underlying technology may be such that the possibility for human review is blocked. This is particularly true if it relies on non-interpretable data-mining processes or if algorithms only make sense when understood in relation to the broader ecosystem of algorithms in which they operate. Of course, one could perhaps insist (legally) that only interpretable processes be used, but as mentioned previously this may reduce the instrumental gains and encounter resistance from the corporate or governmental interests that are advanced by such processes. In addition to this, it may be very difficult to implement such a solution "after-the-fact", *i.e.* after the opaque processes have already come to dominate in a particular domain.

The second reason for doubting the reviewability solution is that, to the extent that algocratic systems could be made to rely on interpretable processes, the likely effect would be to replace the threat of algocracy with the threat of epistocracy. It is highly

unlikely that any particular citizen would have the background knowledge and expertise to review, engage and understand the algorithmic processes by themselves. They would have to rely on some human epistemic elites to distill and convey the necessary information to them. Likewise, courts charged with judicially reviewing algocratic decisions would also have to rely on epistemic elites to inform them about how those decisions work. It is highly unlikely that any of these actors would have the confidence to fully challenge or engage with what these elites would tell them. The result would be a new epistemic elite taking over our public decision-making processes, much to the chagrin of political theorists like Estlund. This is not something to relish.

Some may respond to this by arguing that deference to such elites is already part and parcel of our public decision-making processes. No one human being is capable of understanding all the rationales and reasons for the decisions that affect them. They often require the assistance of experts to package and translate difficult ideas, and to make decisions on their behalf. This is certainly a feature of the status quo, but it is not clear that it is something to be cherished or preserved. It may represent a current compromise in the tradeoff between instrumental and procedural virtues, but if we could overcome it, we probably should. Also, the viability of this tradeoff may be undermined if the asymmetry between the epistemic elites and ordinary citizens is exaggerated or accentuated by data-mining technologies. This is something for which a number of recent papers argue by highlighting the emerging and increasing 'big data divide', arising out of disparities in the ability to leverage the benefits of data-mining systems between creators/controllers and the citizens who are affected (Andrejevic 2014; Mittelstadt and Floridi 2015).

A more satisfying accommodating solution would try to directly empower those who are affected by algocratic decisions, thereby obviating the need for human epistemic elites. This is what the three remaining solutions attempt to do.

5.2 - Epistemic Enhancement of human beings

The first of these direct-empowerment solutions is suggested to us by a recent paper from Danaher (2013). The paper tries to connect the debate about the use of human enhancement technologies with the debate about legitimate decision-making. It argues that enhancement technologies could be used to improve the instrumental and

procedural legitimacy of public decision-making processes. Could this argument be co-opted to address the threat of algocracy?

We'll need to understand the argument a bit better before we can answer that question. In brief outline, the argument relies on the concept "Epistemic Enhancement", which is defined as "any biomedical intervention intended to improve or add to the capacities humans use to acquire knowledge, both theoretical and practical/moral" (Danaher 2013, 88). This covers a wide range of potential technologies, from drugs that manipulate and enhance cognition and affect, to neural stimulators or implants that do the same. The claim is that as long as those technologies "allow those who participate [in public decision-making processes] to process more information, dampen distorting emotions, remember more facts and so on" (Danaher 2013, 99-100) they could be used to enhance procedural legitimacy. In particular, they could be used to protect against the possibility of an epistemic elite taking over, provided that the technologies are made available to all. In the process of defending this view, Danaher responds directly to charges that a social demand for epistemic enhancement would be coercive or autonomy-undermining.

The proposal is attractive for two reasons. The first is that it focuses directly on improving the cognitive and affective capacities of ordinary people who may be affected by public decision-making processes. In doing so, it tries to offer an (admittedly partial) antidote to the problems of increasing decisional complexity and incomprehensibility. This could help to address the threat of algocracy, provided that the epistemic enhancement is of the right kind. The second attraction of the proposal is that by including both moral and theoretical reasoning within the domain of enhancement, it offers some hope of balancing the procedural benefits of algocracy against its costs. Suppose, for example, that Zarsky is right that implicit bias is a serious problem when humans are kept on the loop and allowed to override algorithmic decisions. In that case, epistemic enhancement could be directed at neutralising the problem of implicit bias, while at the same time increasing cognitive ability, thereby allowing for human participation without undermining procedural benefits of the algocratic system.

Despite these attractions, the proposal is ultimately unpersuasive. To see why, we need to draw some conceptual distinctions. As Nicholas Agar points out (2013), there is

an important distinction between what we might call "modest" and "radical" forms of enhancement. Modest enhancement is that which is intended to enhance us up to, or slightly beyond, the current extremes of human performance and ability. Radical enhancement is that which tries to transform us into posthumans, into beings with capacities and abilities that exceed what is currently possible for humans (Kurzweil 2006). It's not entirely clear whether Danaher intends his concept of epistemic enhancement to cover both modest and radical forms of enhancement. His definition speaks of improving or *adding to* current human capacities, which suggests it might be both. Nevertheless, for the time being, I shall assume that it only refers to modest forms of enhancement. I'll return to the possibility of radical enhancement toward the end of this article.

Working with that assumption, I think it becomes obvious why Danaher's proposal would not resolve the threat of algocracy. Algocratic processes are unlikely to be human-like, particularly when they involve datasets with billions of components, when the matching and sorting processes are non-interpretable, and when they are integrated into complex algorithmic ecosystems. In short**: without external constraints** algocratic systems are likely to rely on processes and capacities that are radically beyond what is possible for human beings to understand. Thus, even if enhancement technologies enabled more humans to reach the extremes of human ability, they would do nothing to address the threat of *algocracy*. At most they might level the playing field between different groups of human beings. This might address the threat of epistocracy, but not that of algocracy.

This, however, throws open an intriguing possibility. If epistemic enhancements could be used to stave off the threat of epistocracy, and if the human reviewability solution discussed above could succeed in legally restricting algocratic systems to those that are understandable by epistemically elite human beings, then we might have a way in which to accommodate the threat. For then, we could have the advantages of the algorithms, and avoid replacing the threat of algocracy with the threat of epistocracy.

This is certainly an intriguing possibility but we need to be realistic as well. It may not be possible to restrict algocratic systems to those that are understandable by human beings: the technical, economic, political and personal interests at stake may not allow for this. Furthermore, there are several obstacles that would need to be cleared in

order to implement a solution of this sort. First, we would actually need to have the requisite enhancement technologies. Much of the debate over human enhancement involves speculation about possible future technologies, not currently available ones. Current and proposed enhancement technologies have modest and incompletely understood effects. It may be that we cannot develop the requisite technologies in time to address the threat of algocracy. Second, even if we did have such technologies, they would need to be made widely available, not just restricted to wealthy elites who can afford them. Third, availability by itself would not address the problem since the technologies would not magically imbue us with the requisite knowledge and understanding. There would need to be a wide-ranging public education programme on the nature of the various algocratic systems as well (see Machin 2009 on the difficulties of public education). It may be possible to clear these obstacles, but we shouldn't be tricked into thinking it will be easy.

But perhaps there are easier ways? Ones which rely on more immediately available technologies? The final two solutions consider this possibility.

5.3 - Embrace sousveillance technologies

This "solution" probably wouldn't bear mentioning except for the fact that some have actually suggested it in response to concerns about algocracy,[27] and, more importantly, because it sets up the more interesting final solution. "Sousveillance" is a twist on the term "surveillance". Where the latter term means to watch "from above" (i.e. from a position of authority) the former means to watch "from below" (i.e. from the perspective of the ordinary citizen) (Mann 2013; Mann, Nolan & Wellman 2003).

Sousveillance advocates argue for a type of radical transparency (Brin 1997; Ali & Mann 2013). If the problem with big data algorithms is the constant monitoring and surveillance of our activities by economic and political elites, then the solution is to turn the surveillance technology back on those economic and political elites. Veillance technologies are, after all, widely available, and with the advent of Google Glass, and similar wearable monitoring devices, they are likely to become even more widely available. We can use the data captured by these devices to empower ourselves to hold those authorities to account. "Sunlight", "disinfectant" and other cliches abound.

[27] David Brin one of the chief proponents of sousveillance, has explicitly suggested this in response to Morozov's worries about the threat to democracy posed by algocratic control, see comments on Danaher 2014.

The father of the sousveillance movement — Steve Mann — has himself argued that the widespread use of sousveillance could correct for some of the legitimacy problems inherent in bureaucratic systems of control. He specifically argues that sousveillance can be used to correct for the asymmetries of information and understanding that are inherent in our transactions with bureaucratic institutions (like public authorities and courts). If we are unfairly targeted by such institutions, sousveillance technologies will allow us to share our story with the bureaucrats with "full documentary *evidence* rather than mere *testimony*" (Ali and Mann 2013, 250). This is empowering. And when coupled with freedom of information laws that give us access to the internal regulations and rules of bureaucratic institutions, Mann argues that sousveillance technologies provide a powerful recipe for restoring legitimacy.

To the extent that these bureaucratic systems are themselves reliant on algorithms, we might hope that sousveillance technologies could correct for the threat of algocracy too. But, of course, any such hope is forlorn. Contrary to what Mann seems to suggest, the mere possession of sousveillance technologies does not correct for epistemic asymmetries. The user of the technologies has to be able to understand the rational basis for the bureaucratic decisions, and they cannot do this with the veillance technology alone. If the rational basis for bureaucratic decisions was entirely determined by the human collection and processing of data, there might be a chance of correcting for the imbalance of power. Ordinary humans could then directly engage with and understand the reasoning process, and could use the sousveillance technologies to supply their own data and keep the bureaucrats honest in their dealings. But if the rational basis for the decisions is not determined by humans, but instead by complex ecosystems of algorithms, the situation is rather different. No amount of sousveillance could redress that imbalance.

The problem here is that sousveillance technologies, at least in their purest form, are mere data-collection devices. The comprehension and understanding of that data is up to their human users.[28] But this raises another possibility. What if every human being not only had their own veillance technologies but also had the assistance of their own data-mining and processing algorithms? In other words, what if each human being could

[28] Of course, there may be *some* processing whenever sousveillance technologies record digital and audio information, but that is not the kind of processing and sorting that would be made possible if humans had their own mining algorithms.

form a partnership or alliance with their own algorithms? Would that solve the problem? This is what the final solution suggests.

5.4 - Form individual partnerships with algorithms

I will discuss two possible forms that a partnering solution could take: the *non-integrative* form, which would involve individualised pairing with algorithmic systems that are not-integrated into human biology; and the *integrative* form, which would involve the integration of algorithmic systems into human biology. The former is realisable in the immediate future; the latter is much more fanciful and speculative. Both are doubtful solutions to the problem.

We shall start with the non-integrative form. This is just a slight modification of the sousveillance solution. Where the sousveillance advocate calls for everyone to have their own data-monitoring technologies in order to hold authorities to account, the advocate of non-integrative algorithmic partnerships simply adds to this the claim that everyone should have their own data-mining technology too. This is effectively like having your own private AI-assistant, who can help you to comprehend and understand the other algorithmic processes that affect your life. The idea is that this is empowering as you no longer need to defer to an epistemic elite in order to understand what is going on.

This kind of non-integrative partnership system is already being advocated by a number of economists and technologists (Cowen 2013; Brynjolfsson & McAfee 2011 & 2014). They focus on the problem of technological unemployment and argue that partnerships of this sort may be the only way in which human beings can maintain their employability in the coming era of artificial intelligence. The example of computer chess is often trotted out to illustrate the point (Cowen 2013; Brynjolfsson & McAfee 2014; Thompson 2013). Computers started to surpass elite humans in chess-playing ability in the late 1990s, but this did not render humans obsolete. In large part, this is because chess is just a game and a test of human, not machine ability; but it is also because humans started to pair-up with computers, forming human-computer chess-playing teams. This has had an interesting result. The best chess being played today is not being played by computers, nor by humans, but by these human-computer teams. It seems that by partnering-up with computers, humans have actually enhanced the quality

of their chess. The "Quantified Self" movement[29] provides another example of the benefits of such partnerships. Members of this movement advocate self-experimentation, and the use of individualised data-monitoring and processing technologies, in order to improve their self-understanding and enhance their performance. This has been done primarily in relation to personal health and fitness (Ferriss 2011), but it can encompass cognitive and emotional reasoning too (Thompson 2013).

So the basic idea is that by partnering up with algorithms, individual human beings can retain autonomy, enhance their cognitive powers and understanding, and this might just be enough to ensure their continued ability to meaningfully participate in algorithmic decision-making processes. Of course, this suggestion suffers from three major defects. First, it runs foul of the big data divide problem mentioned earlier (Andrejevic 2014; Mittelstadt and Floridi 2015). As several authors have pointed out, individual users are not well-placed to take advantage of the epistemic benefits of data-mining technologies. Those benefits accrue to those who can generate and control big datasets. This favours wealthy, large-scale concerns (companies, governments, universities), not individual citizens. Second, and adding to the first problem, most individual humans are unlikely to be able to design and create their own algorithmic partners. They would have to rely on others to do this for them, which would then simply bring us back to the problem of epistocracy. Third, it is not at all clear that this kind of non-integrative partnering system would ensure that humans can participate and engage with such processes. Again, the example of human-computer chess teams is instructive in this regard (Cowen 2013, Ch 5). The clear evidence from the past decade and half is that the top chess teams are not the ones in which the humans understand the game the best. Indeed, being a top-ranked individual chess player may actually be a disadvantage when partnering up with a computer. The top-ranked player is too inclined to second-guess the computer's judgment. It seems that greater deference to the computer's intelligence is needed in order to succeed (Cowen 2013, 82). But this suggests that human-computer partnerships might not resolve the threat of algocracy at all. Indeed, they might hasten it. If we all form individualised partnerships with

[29] See, generally, http://quantifiedself.com; Thompson (2013) also discusses the phenomenon. The story of Chris Dancy, a Denver-based IT executive who is known as the world's "most connected man", might also be instructive. Dancy wears up to 10 data-collection devices on his person every day, in addition to other non-wearable devices. He claims that this has greatly improved his life. See http://www.dw.de/worlds-most-connected-man-finds-better-life-through-data/a-17600597 for an interview with him (accessed 1/3/15).

algorithms, we might hasten our path to moral patiency, we would become recipients of the wisdom of our AI-assistants, not true agents involved in understanding and shaping our own destinies.

There is, however, a philosophical objection to this line of reasoning. Within the philosophy of mind, there is a school of thought that endorses the "extended mind thesis" (Clark & Chalmers 1998; Clark 2010). According to this thesis, our mental processes are naturally extended into our artifacts and technologies. The thesis derives support from the functionalist theory of mind, which holds that because mental states are determined by their position within a functional network there is no reason why such a network should be limited to what takes place inside the human brain. Thus, for example, my email folder could be viewed as an extension of my mental faculty of memory: it contains a record of conversations and exchanges I have had with family, friends and co-workers, and I frequently use it to assist my recall. In a similar vein, why couldn't our faculties of understanding and comprehension naturally extend into the algorithms with which we are partnered?

The extended mind thesis does hold out some hope for the defender of the non-integrative partnership solution. But there are two reasons for doubting its consolations. The first is that it is a controversial philosophical thesis and so an unpromising basis on which to rest a solution to serious social-political problem. The second is that even if the extended mind thesis provides a useful framework for explaining and understanding psychological processes, there can be further distinctions between those processes that affect its application to the threat of algocracy. I would suggest that legitimate participation in public decision-making requires conscious understanding of the rational basis for those decisions. There is nothing in the extended mind thesis to suggest that the external artifacts that form part of our "minds" deliver this kind of conscious understanding. When I rely on a calculator to perform some complex mathematical operation on my behalf, I do not consciously represent and understand that series of operations. There is no reason to think it would be any different when pairing up with other computerised processes.

That leaves us with the possibility of forming integrative partnerships with computers. The suggestion here is that instead of relying on external devices to assist our interactions with the world, we actually incorporate those devices into our

physiology, *i.e.* we turn our*selves* into bits of technology. This already happens, to some extent, with prosthetic devices that are integrated into human biology. To address the threat of algocracy, the integration would need to be at a cognitive level, allowing us, in a sense, to understand the world in the same way as the algorithm. The idea might be something along the lines of uploading our minds to a digital substrate or replacing our brains with a set of neural prosthetics.

As a solution to the threat of algocracy, the notion of integrative partnerships suffers from at least two defects. The first is that it is highly fanciful and speculative. Though the idea of digital copies and uploads is commonplace in science fiction, and beloved by transhumanists and techno-utopians, we are certainly a long way from realising such possibilities. And that's assuming that they are even conceptually coherent possibilities: some might argue that the mental could never really be replaced by an artificial analogue.

The second defect is rather more subtle and has to do with the possible effects of such integrative partnerships on the nature of human agency and on the kinds political organisation we value. The threat of algocracy is most acutely felt in a political system that is predicated upon liberal principles. After all, it is in such a system that the need to respect the individual's moral agency — to allow them to meaningfully participate in public decision-making — is an important concern. This concern in turn rests on certain core beliefs about what it means to be an autonomous moral agent. If an integrative partnership with technology is simply an attempt to preserve the human agent in an artificial form, then these values and concerns will still be relevant. The problem is that if the integrative partnership does nothing more than preserve the human agent, it is not clear that the threat of algocracy will be solved. For it is not clear that mere preservation would allow for comprehension and understanding of the algocratic systems. It may be that we need to integrate ourselves with those algocratic systems as well.[30] This might require our consciousnesses to be linked into the global internet of things, so that we can appreciate and understand the datastreams and mining processes that govern collective decision making. But, of course, everyone would have to do the same thing. It is not clear that the concept of the individual moral agent would survive such a

[30] This is the vision of transhumanists like Ray Kurzweil who seek to saturate the cosmos with our intelligence, i.e. to make everything in the universe an extension of and input into our cognitive processes (Kurzweil 2006, 29).

technological transformation (Lipschulz & Hester 2014). And so it is not clear that the threat of algocracy would be relevant in such a world.

After that flight of fancy, we must, alas, come back down to earth. I do not wish to completely disparage the notion that partnerships with technology could form part of a solution to the threat of algocracy. They certainly could. But there are difficulties here, both technological and philosophical.

6. Conclusion

This article has defended three major theses. First, it has argued that there is such a thing as the threat of algocracy. This is a threat to the legitimacy of public decision-making processes, which is posed by the opacity of certain algocratic governance systems. The threat is a real one, distinct from related concerns with privacy and ownership of data.

Second, it has argued that it may not be possible or desirable to resist the threat of algocracy, *i.e.* to simply stop relying on algocratic decision-making systems. The technologies that make algocracy possible are becoming less noticeable and more ubiquitous. And any costs they have in terms of opacity need to be weighed against their other instrumental and procedural benefits.

Third, it has argued that it is also difficult to accommodate the threat of algocracy, *i.e.* to find some way for humans to "stay on the loop" and meaningfully participate in the decision-making process, whilst retaining the benefits of the algocratic systems. Some accommodating solutions are naive and fanciful, others simply miss the mark, addressing the threat of epistocracy but not the threat of algocracy. In the end, the most viable solution may be some combination of *reviewability* and *enhancement* (which could encompass human-machine partnership, integrative or otherwise). The former might be able to legally limit the types of algocratic system that are used by insisting upon a right and possibility of human review; the latter might then be able to prevent this solution from simply collapsing into the threat of epistocracy.

But this conclusion is somewhat pessimistic. Although it may be relatively easy to restructure the legal system so as to insist on reviewability, the probability of

successfully creating and distributing appropriate enhancement technologies within the requisite timeframe is much more uncertain. Furthermore, the growth of algocratic systems combined with the ways in which such systems become woven into ever more complex algorithmic ecosystems, may be such as to push them beyond the control and understanding of their human creators. In that case, achieving individual epistemic elitism may no longer be enough. In short, we may be on the cusp of creating a governance system which severely constrains and limits the opportunities for human engagement, without any readily available solution. This may be necessary to achieve other instrumental or procedural gains, but we need to be sure we can live with the tradeoff.

Acknowledgments: The author would like to thank audiences at Exeter and Maynooth Universities, and two anonymous referees for feedback on earlier drafts of this paper.

Ethical Statement: The author declares no conflicts of interest. Research for this paper was not funded, nor did it involve any work involving human or animal subjects.

References

Agar, N. 2013. *Truly Human Enhancement.* Cambridge, MA: MIT Press.

Ali, MA and Mann, S. 2013. The Inevitability of the Transition from a Surveillance Society to a Veillance Society: Moral and Economic Grounding for Sousveillance. *IEEE International Symposium on Technology and Society ISTAS* 243-254 (available at http://wearcam.org/veillance/IEEE_ISTAS13_Veillance2_Ali_Mann.pdf accessed 31/7/14)

Aneesh, A. 2006. *Virtual Migration*. Duke University Press.

Aneesh, A. 2009. Global Labor: Algocratic Modes of Organization. *Sociological Theory* 27(4): 347-370

Andrejevic, M. 2014. The Big Data Divide. *International Journal of Communication* 8: 1673-1689.

Barrat, J. 2013. *Our Final Invention*. Thomas Dunne Books.

Besson, S. and Marti, JL. 2006. *Deliberative Democracy and its Discontents*. London: Ashgate.

Bishop, M. & Trout, JD. 2002. 50 Years of Successful Predictive Modeling Should be Enough: Lessons for Philosophy of Science. *Philosophy of Science: PSA 2000 Symposium Papers*, 2002 69 (supplement): S197-S208

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: OUP.

Brin, D. 1997. *The Transparent Society*. New York: Basic Books.

Brynjolfsson, E. and McAfee, A. 2011. *Race against the Machine*. Lexington, MA: Digital Frontiers Press.

Brynjolfsson, E. and McAfee, A. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a time of Brilliant Technologies*. New York: WW Norton.

Bumbulsky, J. 2013. Chaotic Storage Lessons. *Medium* (available at https://medium.com/tech-talk/e3b7de266476 - accessed 1/3/15.

Ceva, E. 2012. Beyond Legitimacy: Can Proceduralism Say Anything Relevant about Justice? *Critical Review of International Social and Political Philosophy* 15: 183

Chase Lipton, Z. 2015. The Myth of Model Interpretability, *KD Nuggets News* 15:n3 – available at http://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learning.html

Citron, D. 2010. Technological Due Process. *Washington University Law Review* 85: 1249

Citron, D. and Pasquale, F. 2014. The Scored Society: Due Process for Automated Predictions. *Washington Law Review* 86: 101.

Clark, A. 2010. *Supersizing the Mind*. Oxford: OUP.

Clark, A. and Chalmers, D. 1998. The Extended Mind. *Analysis* 58: 7-19

Cowen, T. 2013. *Average is Over: Powering America Beyond the Age of the Great Stagnation*. New York: Dutton.

Crawford, K. and Schultz, J. 2014. Big Data and Due Process: Towards a Framework to Redress Predictive Privacy Harms. *Boston College Law Review* 55: 93

Danaher, J. 2013. On the Need for Epistemic Enhancement: Democratic Legitimacy and the Enhancement Project. *Law, Innovation and Technology* 5(1): 85

Danaher, J. 2014. Rule by Algorithm? Big Data and the Threat of Algocracy. *Philosophical Disquisitions* – available at http://philosophicaldisquisitions.blogspot.ie/2014/01/rule-by-algorithm-big-data-and-threat.html (accessed 23/12/15)

Estlund, D. 1993. Making truth safe for democracy. In Copp, D., Hampton, J. and Roemer, J. (eds) *The Idea of Democracy*. Cambridge: Cambridge University Press

Estlund, D. 2003. Why not Epistocracy? In Naomi Reshotko (ed) *Desire, Identity, and Existence: Essays in Honour of T.M. Penner*. Academic Printing and Publishing.

Estlund, D. 2008. *Democratic Authority*. Princeton: Princeton University Press.

Gaus, G. 2010. *The Order of Public Reason*. Cambridge University Press.

Greenfield, R. 2012. Inside the Method to Amazon's Beautiful Warehouse Madness. *The Wire* (available at http://www.thewire.com/technology/2012/12/inside-method-amazons-beautiful-warehouse-madness/59563/ - accessed 1/3/15.

Grove, W. and Meehl, PE. 1996. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical statistical controversy. *Psychology, Public Policy and Law* 2: 293-323

Habermas, J. 1990. Discourse Ethics: Notes on a Program of Philosophical Justification. In *Moral Consciousness and Communicative Action*. Trans. Christian Lenhart and Shierry Weber Nicholson. Cambridge, MA: MIT Press.

Kellermeit, D. and Obodovski, D. 2013. *The Silent Intelligence: The Internet of Things*. DND Ventures LLC.

Kitchin, R. 2014a. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: Sage.

Kitchin, R. 2014b. Thinking critically about researching algorithms. *The Programmable City Working Paper 5* – available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2515786

Kitchin, R. and Dodge, M. 2011. *Code/Space: Software and Everyday Life*. Cambridge, MA: MIT Press.

Kurzweil, R. 2006. *The Singularity is Near*. London: Penguin Books.

Laudan, L. 2006. *Truth Error and Criminal Law: An Essay in Legal Epistemology*. Cambridge: CUP.

Lessig, L. 1999. *Code and other Laws of Cyberspace*. New York: Basic Books.

Lessig, L. 2006. *Code 2.0*. New York: Basic Books.

Lippert-Rasmussen, K. 2012. Estlund on Epistocracy: A Critique. *Res Publica* 18(3): 241-258

Lipschulz, R. and Hester, R. 2014. We are the Borg! Human Assimilation into Cellular Society. In Michael and Michael (eds). *Uberveillance and the Social Implications of Microchip Implantation*. IGI-Global.

Lisboa, P 2013. Interpretability in Machine Learning: Principles and Practice. In Masulli, F, Pasi, G and Yager, R (eds) *Fuzzy Logic and Applications* (Dordrecht: Springer, 2013)

List, C. and Goodin, R. 2001. Epistemic Democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy* 9: 277

Machin, D. 2009. The Irrelevance of Democracy to the Public Justification of Political Authority. *Res Publica* 15: 103

Mann, S. 2013. Veillance and Reciprocal Transparency: Surveillance versus Sousveillance, AR Glass, Lifeglogging, and Wearable Computing. Available at http://wearcam.org/veillance/veillance.pdf -- accessed 1/3/15.

Mann, S., Nolan, J. and Wellman, B. 2003. Sousveillance: Inventing and Using Wearable Computing Devices for Data Collection in Surveillance Environments. *Surveillance and Society* 3: 331-355

Mayer-Schonberger, V. and Cukier, K. 2013. *Big Data: A Revolution that Will Transform How we Live Work and Think*. John Murray.

Miner, L et al. 2014. *Practical Predictive Analytics and Decisioning-Systems for Medicine*. Academic Press.

Mittelstadt, B D, and Floridi, L. 2015. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Science and Engineering Ethics*. DOI: 10.1007/s11948-015-9652-2

Morozov, E. 2013. The Real Privacy Problem. *MIT Technology Review* (available at: http://www.technologyreview.com/featuredstory/520426/the-real-privacy-problem/ - accessed 1/3/15)

Otte, C. 2013. Safe and Interpretable Machine Learning: A Methodological Review. In Moewes, C and Nurnberger, A. (eds) *Computational Intelligence in Intelligent Data Analysis* (Dordrecht: Springer 2013).

Patterson, S. 2013. *Dark Pools:The Rise of AI Trading Machines and the Looming Threat to Wall Street.* Random House.

Pentland, A. 2014. *Social Physics*. London: Penguin Press.

Peter, F. 2008. Pure Epistemic Proceduralism. *Episteme* 5: 33

Peter, F. 2014. Political Legitimacy. In Edward N. Zalta (ed) *The Stanford Encyclopedia of Philosophy* Spring 2014 Edition -- available at http://plato.stanford.edu/archives/spr2014/entries/legitimacy/

Polanyi, M. 1966. *The Tacit Dimension*. New York: Doubleday.

Rifkin, J. 2014. *The Zero Marginal Cost Society: The Internet of Things, The Collaborative Commons and the Eclipse of Capitalism.* Palgrave MacMillan.

Seaver, N. 2013. Knowing Algorithms. In *Media in Transition 8*, Cambridge MA.

Siegel, E. 2013. *Predictive Analytics: the Power to Predict who will Click, Buy, Lie or Die*. John Wiley and Sons.

Slater, D. 2013. *Love in a time of Algorithms*. Current.

Thompson, C. 2013. *Smarter than you think: How technology is changing our minds for the better*. London: William Collins.

Vellido, A, Martín-Guerrero, J. and Lisboa, P. 2012. Making machine learning models interpretable. *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.

Zarsky, T. 2011. Governmental Data-Mining and Its Alternatives. *Penn State Law Review* 116: 285

Zarsky, T. 2012. Automated Predictions: Perception, Law and Policy. *Communications of the ACM* 15(9): 33-35

Zarsky, T. 2013. Transparent Prediction. *University of Illinois Law Review* 4: 1504

Zeng, J, Ustun, B and Rudin, C. 2015. Interpretable Classification Models for Recidivism Prediction. *MIT Working Paper*, available at http://arxiv.org/pdf/1503.07810v2.pdf