AI Rule and a Fundamental Objection to Epistocracy[1]

Forthcoming in *AI & Society*

PENULTIMATE VERSION, PLEASE ASK PERMISSION TO CITE

Abstract: Epistocracy is rule by whoever is more likely to make correct decisions. AI epistocracy is rule by an artificial intelligence that is more likely to make correct decisions than any humans, individually or collectively. I argue that although various objections have been raised against epistocracy, the most popular do not apply to epistocracy organized around AI rule. I use this result to show that epistocracy is fundamentally flawed because none of its forms provide adequate opportunity for peoples (as opposed to individuals) to develop a record of meaningful moral achievement. This Collective Moral Achievement Objection provides a novel reason to value democracy. It also provides guidance on how we ought to incorporate digital technologies into politics, regardless of how proficient these technologies may become at identifying correct decisions.

Keywords: epistocracy, democracy, collective self-determination, moral achievement

[1] For generous conversation or support while writing this article, I'd like to thank Christian Barry, Nicholas Carroll, Shalom Chalson, Alan Hajek, Brian Hedden, Seth Lazar, Jonathan Quong, Pamela Robinson, Nicholas Southwood, Nick Willis, Shang Long Yeo, and audience members at the ANU's Machine Intelligence and Normative Theory Lab and Thursday Seminar.

How should we engineer digital technologies for use in politics? There are at least two options. The first involves increasingly supplanting human decision-makers with AI systems that develop and assess policies.[2] The second involves increasingly supporting human decision makers with tools that make political participation more inclusive, efficient, and secure, such as deliberative platforms, consensus finders, and blockchain voting.[3]

Which direction we ought to take depends on the answer to a philosophical question, namely, What kind of government is best? Here I consider the two kinds of government most closely aligned with these extremes: epistocracy and democracy. Democracy is a kind of government where citizens in general rule. What we may call the *Democratic Thesis* is the claim

[2] The final end of pursuing this direction – replacing human rulers with AI – is a long way off but still worth taking seriously. AI policy recommendation is experimentally feasible. (Zheng et al. 2022) The current trend is to continue developing products in this direction. Technology companies typically promote AI systems as delivering maximally accurate predictions by supplanting human decision makers. (Wang et al. 2023) And well-funded organizations, like the consulting firm Deloitte, advocate for providing government services with AI. (Eggers et al. 2021) Finally, some may consider developing AI capable of ruling beneficial or inevitable and so worth promoting now to ensure the best outcomes. As several of OpenAI's board members declared in 2023: "… we believe it would be unintuitively risky and difficult to stop the creation of superintelligence. Because the upsides are so tremendous, the cost to build it decreases each year, the number of actors building it is rapidly increasing, and it's inherently part of the technological path we are on, stopping it would require something like a global surveillance regime, and even that isn't guaranteed to work. So we have to get it right." (Altman et al., 2023)

[3] Examples of these tools include, respectively, Barcelona's Dicidim system (Barcelona's Digital Democracy 2021), the consensus finder Polis (Input Crowd, Output Meaning 2023), and the various systems catalogued by Jafar et al. (2021: 12-14).

that citizens in general ought to rule.[4] Epistocracy is a kind of government where those who are more likely to make correct decisions rule. What we may call the *Epistocratic Thesis* is the claim that those who are more likely to make correct decisions ought to rule.[5]

If the Epistocratic Thesis is true, there are various things we ought to do. We ought to determine whether it's possible to build epistocratic AI systems that are more likely than humans to make correct decisions. If this proves to be the case, we ought to shift our focus away from developing technology to support democracy. We ought instead work toward making our political arrangements like the first extreme by increasingly placing ourselves under epistocratic AI rule.

But *is* the Epistocratic Thesis true? I say no. This is because of what I call the Collective Moral Achievement Objection. Democratic theorists have long claimed democracy is valuable partly because it gives individuals opportunities for self-development.[6] This is no embarrassment to epistocracy, however, since epistocracies plausibly give individuals adequate opportunity for self-development as well.[7] I claim a different reason to value democracy is that it gives not just individuals but some important kinds of collectives opportunities for self-development.

[4] 'Ought' in both theses denotes not a mere suggestion, but a moral obligation.

[5] For explicit endorsements of the Epistocratic Thesis, see Brennan (2016), Gibbons (2021), Jeffrey (2018), and Jones (2020).

[6] Macpherson (1977: 2), Held (2006: 91-92), Schneirov and Fernandez (2013: 10-11), Dahl and Shapiro (2015: 55-56).

[7] Brennan (2016: 102; 106-109).

Specifically, democracy gives *peoples* adequate opportunity to develop records of meaningful moral achievement.

It's valuable for individuals to belong to peoples with the opportunity for moral achievement – so valuable that it's permissible to adopt governments that provide this opportunity, even if they involve rule by those who are substantially less likely to make correct decisions. Since democracy provides this opportunity and epistocracy doesn't, it's permissible to adopt democracy instead of epistocracy. This Collective Moral Achievement Objection falls short of establishing the Democratic Thesis because no one is obligated to strive for moral achievement. Nevertheless, it shows the Epistocratic Thesis is false.

But *will* it be possible to build epistocratic AI systems? I say it doesn't matter for the importance of the Collective Moral Achievement Objection. Although AI rule exists on the further horizons of current technological possibility, it promises to be an ideal form of epistocracy that avoids the most popular objections against the view. Idealizing a view often helps reveal what makes it problematic.[8] In this case, thinking about AI rule reveals the Collective Moral Achievement Objection to be a fundamental problem for epistocracy in that it applies to any form this kind of government may take. This objection thereby provides guidance about how we should arrange our institutions now.

We need this guidance because, as democratic theorists often note, democracy is both fragile and rare in the broad scope of history.[9] With AI systems possessing expertise surpassing

[8] Burgess (2024: 132).

[9] Held (2006: 1), Schneirov and Fernandez (2013: 1), Dahl and Shapiro (2015: 44).

their creators becoming increasingly common, the instrumental benefits of these systems may encourage many to join epistocrats in welcoming, and even working toward, a regression to the historical, non-democratic mean. Democracy could suffer even if AI rule never emerges. The fundamental objection I offer may not be the strongest reason to resist this regression in a given context, but it is still beneficial. Fundamental objections are less contingent and so forestall debate about whether the presuppositions of non-fundamental objections hold.[10] The Collective Moral Achievement Objection for its part shows we have good reason to resist epistocracy, even if its rulers – be they humans, AI systems, or something else – can guarantee their decisions are correct.

To establish these claims, I start by presenting the basic argument epistocrats offer for their view and the current strongest objections against it. I then show epistocratic AI rule avoids these objections and so constitutes an ideal form of epistocracy. At the same time, these objections suggest arguing against epistocracy by finding a surplus value only non-epistocratic governments produce that is important enough to justify choosing them over epistocracy. Using thought experiments and evidence from psychology, I argue this value is collective moral achievement. I then explain why democracies adequately produce this value and epistocracies can't. I conclude with broader considerations about what this result implies for integrating technology and politics.

[10] My view is that concerns about equality and the demographic objection provide the strongest reasons to reject epistocracy in current societies. But, as I argue below, the prospect of epistocratic AI rule makes these concerns contingent on technological development.

1. Arguing for and against Epistocracy

Despite the aforementioned historical trends, most governments currently claim to be democracies, and The Democratic Thesis likely strikes many readers as a platitude needing little defense. We must consider why the Epistocratic Thesis might be true instead.

Epistocrats appeal to instrumentalist considerations. They observe that different kinds of government produce different moral outcomes, where 'outcomes' are assumed distinct from the political processes that bring them about.[11] Residents of Hobbesian anarchies enjoy less goods like health, security, and happiness than residents of dictatorships, and even less of these goods than residents of liberal democracies. Epistocrats claim we ought to adopt whatever government likely produces better outcomes. Epistocracy satisfies this description because it puts those who are more likely to make correct decisions in change.

We can express these considerations more carefully as what we may call the *Basic Argument for Epistocracy*:[12]

(P1) Adopting epistocracy is feasible.

(P2) When choosing among feasible governments, we ought to adopt whichever likely produces better moral outcomes.

(P3) The moral outcomes epistocracy produces are likely better than governments where those who are less likely to make correct decisions rule.

[11] For discussion of the process-outcome distinction, see Estlund (2008: 65), Brennan (2016: 10-14, 138, 140, 182), Jeffrey (2018: 421), and Jones (2020: 19, 111).

[12] Brennan (2016: 10-16), Jeffrey (2018: 418-419), Jones (2020: 111), and Gibbons (2021: 192) endorse expressions of this argument.

(C) We ought to adopt epistocracy.

This argument requires some clarification. First, let 'better moral outcomes' denote whatever outcomes the reader takes to be morally better.[13] Second, let 'to make correct decisions' denote to select whichever options are true or morally better, according to the kind of decision at hand.

These clarifications simplify the tasks of raising objections to the Basic Argument and assessing whether AI rule avoids them. They ensure governments ruled by those who are more likely to make correct decisions by definition will be more likely to produce better moral outcomes than governments where such individuals don't rule. The third premise, then, is unassailable. What about the others? They fare less well, as the following most popular objections demonstrate.

Consider the first premise. Even if it would be great to put those who are more likely to make correct decisions in charge, we still have to find these individuals. This presumably requires testing individuals for political competence and allowing them to contribute to political decisions only if they pass. There is reason to doubt such tests are feasible.

Making a correct decision requires both identifying the correct option and being motivated to select it. Proponents of the so-called demographic objection argue that epistocrats have not shown we can reliably test the former quality.[14] In the presence of confounding factors,

[13] There will of course be reasonable disagreement about which outcomes are morally better. If the reader considers the impermissibility of imposing outcomes against such disagreement a side-constraint on political action, let 'better moral outcomes' denote outcomes that are better according to values everyone can reasonably accept. This set is not empty: health, security, happiness, and many other things are surely within it.

[14] See Estlund (2008: 215) and Ingham and Wiens (2021) for statements of this objection.

good performance on political competence tests may negatively correlate with recognizing which options are correct in many situations.[15] Suppose economics departments are wrongly biased against unionization. Their graduates will likely pass competence tests about economics but make incorrect decisions about unionization. Although confounders like these are ubiquitous, epistocrats have not explained how to avoid them. Tests for good motivation have their own problems. To name one: acts that initially correlate positively with good motivation, such as donating to charity, will become uninformative as individuals start performing them just to be deemed competent to make political decisions.[16]

Consider instead the second premise. The strict distinction epistocrats draw between outcomes and processes threatens its plausibility because political processes can be morally valuable independently of the values of their outcomes. By focusing exclusively on the latter, the second premise requires us to disregard any value processes may have for choosing a government.

Yet turning this observation into an objection requires overcoming what we may call the *Surplus Value Challenge*: it's permissible to adopt a government that produces worse outcomes only if its decision-processes produce a compensating surplus value. In the present context, the challenge is to find a value that decision-processes in non-epistocratic governments produce that compensates any deficiency of their outcomes.

[15] Ingham and Wiens (2021: 325).

[16] Kogelmann (2022: 8).

We can try to meet this challenge by considering how epistocratic decision-processes negatively impact individuals. These impacts can be direct or indirect. Epistocracies typically improve outcomes by disenfranchising some individuals to some extent (namely, those deemed incompetent). So, a strong argument for the former might appear to be that epistocracies directly disempower individuals from controlling outcomes affecting their lives. Yet this appearance is misleading: other governments offer a trifling gain in individual control over outcomes.[17] Being disenfranchised from a group of a few individuals significantly decreases your control over the group's decisions; being disenfranchised from a group of millions does not. Given the assumption that outcomes in epistocracies are likely better, you stand a greater chance of being negatively impacted by outcomes in non-epistocratic governments.

For this reason, it's better to consider whether disenfranchisement has a direct, negative impact on individuals other than through reducing their control over outcomes. The most promising thought along this line is that disenfranchisement negatively impacts equality. Even if epistocracy disenfranchises no one completely, it still creates a distinction between those who are and are not competent to fully participate in political decisions. This distinction undermines the horizontal equality among citizens because it invites competent individuals to look down on incompetent ones. It also erodes the vertical equality between citizens and rulers because competent individuals will have more power and so dominate the incompetent. Democracies do better on this score insofar as they enfranchise everyone equally on the basis of criteria like age that do not require discrimination between levels of political competence. It can be justified to

[17] Brennan (2016: 31, 80, 86).

adopt democracy instead of epistocracy, then, as long as the value of equality compensates deficiencies in democratic outcomes.

Next, epistocratic decision-processes could impact individuals indirectly. The difficulty with arguing along this line is finding a viable candidate for indirect impact. Epistocrats readily acknowledge that although epistocracy marginally disempowers individuals, it significantly disempowers collectives by disenfranchising their members.[18] Epistocracy may not significantly disempower any individual black voter, for instance, but if epistocracy disenfranchises enough black voters, they will be significantly disempowered as a group. Accordingly, a potential response to the Surplus Value Challenge – one I aim to develop – is that epistocracy has an indirect, negative impact on individuals because it undermines the ability of some collectives individuals care about to control political decisions.

There is good reason to think this is a promising way to argue against epistocracy. Collective self-determination is widely considered valuable, so much so that the first article of the UN's International Covenant on Civil and Political Rights proclaims it a right of all peoples.[19] But collective self-determination as typically described is too general to support an adequate objection. Epistocrats can reply that collectives in epistocracies are free to self-determine in lots of ways, such as by choosing their own leisure activities, foods to cherish, languages, artistic traditions, holidays, architectural styles, and so forth.[20] What they are not free to do (at least not

[18] Brennan (2016: 76, 98, 110).

[19] For influential and recent philosophical discussions of this value, see Margalit and Raz (1990), Stilz (2016), and Lovett and Zuehl (2022).

[20] Margalit and Raz (1990: 443-444).

without their members being competent) is fully participate in morally important decisions. The literature on collective self-determination provides no clear reason why this particular restriction has a significant, negative impact on individuals, especially if making it likely results in better moral outcomes for the same collectives being restricted.[21]

Finally, as an alternative to objecting to the first and second premises, one may adopt an accommodating response to the Basic Argument. Epistocracy, as I have defined it, is neutral about who is more likely to make correct decisions. Accepting the Basic Argument need not compel one to change one's view if the rulers in one's preferred government and those who are more likely to make correct decisions are identical. This is the view of the so-called epistemic democrats, who argue that citizens in properly arranged democracies are sufficiently likely to make correct decisions that epistocratic considerations do not require adopting an alternative government.[22]

The accommodating response is nevertheless open to a significant contingency worry.[23] It makes accepting one's preferred government depend on our knowledge of how different kinds of government perform, which is constantly changing. If epistocrats ever showed that some non-

[21] Lovett and Zuehl (2022: 495) in fact explicitly claim that collective self-determination – or, in their terminology, democratic autonomy – is compatible with epistocracy. They argue protecting equality requires not accepting epistocracy.

[22] See Goodin and Spiekermann (2018) and Estlund (2008) for examples of such views. Note that these authors break from epistocrats in that they see epistemic qualities of democracy as merely *pro tanto* reason to prefer this government as opposed to the ground of its authority.

[23] Landemore (2013: 50-52).

democratic government significantly outperforms democracy, then epistemic democrats committed to the Basic Argument would have to abandon democracy in its favor.

Epistocrats can and have mounted further replies to most of these objections. I will pass over them and instead mount a comprehensive response on their behalf by considering the possibility of a new and ideal kind of epistocracy: epistocratic AI rule. In the next section, I show AI rule realizes the contingency worry about the accommodating response. I also show AI rule avoids the objections just surveyed – including responses to the Surplus Value Challenge – as well as or better than other kinds of epistocratic governments.

2. The Challenge of Epistocratic AI Rule

We can understand AI rule by comparing it to human rule. Whereas human rule consists in humans deciding which laws and policies to impose on a population, AI rule consists in AI systems deciding which laws and policies to impose. AI rule does not presuppose AI systems also force obedience to these laws and policies; this could be done by humans in the police, military, etc. habitually deferring to the decisions of an AI legislator instead of a human one. What makes either type of rule epistocratic – at least aspirationally – is that it is intentionally arranged to put whoever is more likely to make correct decisions in charge.

We may suppose an epistocratic AI ruler could be engineered with the same techniques commonly used to engineer the most advanced AI systems.[24] This involves providing a learning algorithm training data consisting of input-output pairs from which it develops a rule for

[24] See Huyen (2022: 3-8) for an overview.

generating future output in response to novel input. In the case of an AI used for medical diagnosis, the training data might be input-output pairs of symptoms and disease labels from which the learning algorithm develops a rule to predict which disease a patient has based on her symptoms. In the case of an AI ruler, the training data might be an input of metrics about various social situations paired with descriptions of laws or policies implemented in response to them.[25] From these, the learning algorithm can develop a rule to predict which laws or policies to adopt in response to new social situations.

Making such an AI ruler epistocratic requires good training data.[26] If the labels knee pain and conjunctivitis are frequently paired when training an AI for medical diagnosis, it will predict that patients with knee pain have conjunctivitis, even though the symptom and disease are unrelated. Frequently pairing mild economic inflation with racial segregation will result in an AI recommending racial segregation in response to mild economic inflation. In general, an AI ruler cannot become likely to make correct decisions without being provided a large number of social situations associated with laws or policies that are correct to adopt in response to them.

At this point, one might doubt that epistocratic AI rule could be substantially better than epistocratic human rule. First and foremost, during the AI's construction, the need for humans to select correct training data to make the AI ruler epistocratic threatens to reproduce the problems with epistocratic human rule canvassed in the previous section. The technical nature of AI systems introduces further issues. An AI ruler could be hacked or compromised by malicious

[25] Lovett and Zuehl (2022: 469-470).

[26] Huyen (2022: 81).

system administrators. An AI ruler could fail due to technical problems while making high-stakes, time sensitive decisions about national security, disaster response, or public health, with catastrophic results. The opaque, complex nature of an AI ruler could make its particular decisions or operation as a whole unacceptable to the public.[27] An AI ruler could also fail to be scalable. That is, it might make accurate decisions about constrained, localized problems (such as the best traffic laws for a city) but be unable to make accurate decisions about issues involving multiple interacting systems (such as the best policy to reduce pollution for a nation).

These are important concerns. Yet as reasons to prefer one government to another, they face the same problem as the accommodating response to the Basic Argument: they are open to a contingency worry. All of them are conceivably solvable engineering problems as opposed to necessary shortcomings. Scalability issues could be addressed by developing faster processors and standardizing data collection so that the AI ruler could reliably integrate larger amounts of information. The AI ruler's source code could be made publicly available, and advances in explainable AI could make the system just as if not more transparent in how it makes decisions than human rulers.[28] Steps could be taken to make the AI ruler as secure and reliable as the computer systems people depend on every day to perform banking and market transactions, which would also have catastrophic results if they failed.

---

[27] Danaher (2016: 254-255).

[28] See Linardatos et al. (2021) for a review of methods for engineering explainable AI. Coglianese and Lehr (2019) and Wischmeyer (2020: 79, 94-97) provide optimistic accounts of how AI systems can be sufficiently transparent to meet rule of law constraints.

Addressing the need for humans to select correct training data to construct an epistocratic AI is more difficult. I revisit this topic Section 4, where I consider how human involvement in training may instead help the case for epistocracy. For now, note that reproducing the problems of human epistocratic rule is conceivably avoidable through what we may call *democratic bootstrapping*.[29] This is a hypothetical process by which some countries become obligated to adopt epistocratic AI rule through the democratic development of this government in other countries. The upshot is that this process does not require countries to identify and discriminate between competent and incompetent citizens to be carried out. It goes as follows.

Suppose Country[0] democratically decides to develop an AI ruler that can make increasingly important and complex political decisions. Country[0] also uses democratic methods to build the AI by having its citizens vote on which pairs of social situations and laws or policies are 'correct' during its training. Once the AI starts deciding what laws and policies to implement, the training incorporates an assessment phase in which citizens vote retrospectively on which decisions were 'correct'. These results can further sensitize the AI to the most relevant social metrics for producing legislation. Given enough training, the AI system can operate autonomously.

A system developed in this way can plausibly become more reliable than humans in identifying correct legislative options. As epistemic democrats have argued, sufficiently large, well-organized electorates can track the truth better than small groups of experts.[30] These

[29] 'Bootstrapping' alludes to the computer science term whereby a smaller process is used to initiate increasingly larger ones.

[30] Goodin and Spiekermann (2018).

electorates stand a better chance at identifying correct laws or policies retrospectively, after observing the results of their implementation. Combining these features with the advantages of AI systems offers additional benefits. AI systems have better memories than democratic electorates, whose members constantly change through birth and death. AI systems can better focus on details that are most relevant to assessing laws and policies. Members of democratic electorates, by contrast, often rely on amorphous heuristics, like asking 'How well is my life generally going now?', to make the same assessments.[31] AI systems have more processing power than democratic electorates and could potentially assess thousands of policies simultaneously. One can think of additional advantages.

Consequently, once Country[0] democratically establishes its AI ruler, it's reasonable to imagine life for its citizens substantially improving along metrics of publicly agreed upon goods like wealth, health, security, and happiness. Suppose Country[0] offers to export the system on the basis of this success, and Country[1] through Country[n] each democratically agree. This exportation process needn't happen all at once but could be done piecemeal within differing government sectors: a country might adopt AI rule over traffic, then healthcare, then education, and eventually integrate them into a unified, ruling system. Along the way, these sector specific systems can be further democratically refined to fit different social contexts until the unified system requires increasingly less human oversight. As AI rule is successively rolled out in these countries, each similarly demonstrates improvement in their citizens' quality of life.

[31] Achen and Bartels (2017: 138, 142-145).

At some point, there will be a country – call it Country[n + 1] – whose government sectors are not significantly different from a mixture of those found in Country[0] through Country[n]. Country[n + 1] will consequently face epistocratic pressure to export the systems already refined in these other countries and assemble them into an AI ruler as a merely technical endeavor. There will be significant inductive evidence that Country[n + 1]'s adopting AI rule will produce what its own citizens consider substantially better moral outcomes compared to its current, human-led government.[32] This evidence furthermore will be strong enough that while the same transition was merely permissible for Country[0] through Country[n], for Country[n + 1] it will be obligatory.

A little thought shows the most serious objections from the previous section do not apply to an AI ruler constructed by democratic bootstrapping. The AI has gained a superior ability to track the truth through democratic means that do not require identifying experts. This preserves horizontal equality because it avoids distinguishing citizens on the basis of competence. Although the citizens of Country[n + 1] have no input into the AI's training, this is just a matter of luck – they could have trained the AI if it were made available to them earlier in the bootstrapping process, and citizens of other countries have no reason to see them as less competent at identifying correct outputs.

Finally, vertical equality is preserved because the citizens of Country[n + 1] are not dominated by the citizens who trained the system. This is by no means true for AI systems in general. As Jonne Maas has recently argued in this journal, a large number of developers of an AI system can have sufficient power to dominate the system's end-users if they influence how the

[32] I owe inspiration for the inductive portion of this argument to Pamela Robinson.

system behaves and the end-users depend on the system to fulfil their goals.[33] Developers of a chatbot that end-users must interact with to gain life-saving medicine have power over those end-users; they can determine whether the users are disrespected by controlling whether the chatbot asks disrespectful questions, for instance. But the epistocratic AI ruler is not like this chatbot. The citizens of Country[n + 1] are obligated to adopt AI rule only if the AI will be a better ruler than themselves and the citizens have sufficient evidence this is the case. The chatbot's end-users are faced with a different choice: subject themselves to its behavior or die.

We can see, then, that concerns introduced by the technological aspects of AI rule are avoidable and that an AI ruler can be constructed without falling prey to the previous section's objections. Let's therefore assume it's possible to make an AI ruler that is epistocratic not only aspirationally but in fact. To what extent do these objections apply to this ruler post-construction? It turns out they are avoided just as well if not better than under epistocratic human rule.

Consider the objections to the first premise. Epistocratic AI rule avoids the demographic objection because it replaces the difficult task of finding a confounder-free proxy for the ability to identify correct options with the simpler one of directly assessing the AI's decisions. Finding epistocratic rulers by testing populations for competence is unnecessary because we build a ruler – the AI system itself. The task of assessing the AI's performance is by no means straightforward. But since we must always consider whether our rulers make correct decisions, it will be one we are left with in any case. There's furthermore good reason to think this task will be easier with

[33] (2023: 1496)

an AI system. An AI can be fine-tuned to perform well on specific kinds of decisions and thoroughly examined for accuracy in ways that would be too invasive for any human. There would also be no difficulties with testing an AI ruler for appropriate moral motivation. Unlike humans, an AI would have no impulse to lie or deceptively conform its behavior to anyone's expectations.[34]

Next, consider the responses to the Surplus Value Challenge. Epistocratic AI rule avoids concerns about preserving equality. Consider the vertical equality between the AI ruler and citizens. Someone cannot be unequal unless there is a person she is unequal to. Such a person simply does not exist under AI rule.[35] All are equally required to follow the decisions of the AI ruler, and we have no good reason to believe the AI must count as a person. Although the AI's decisions involve the complex task of running a state, this is insufficient for personhood. AI systems like AlphaFold perform complex tasks like predicting protein structures better than any human, but we do not for that fact consider them persons. Next, consider concerns about preserving horizontal equality. Since we are assuming the AI ruler is epistocratic, its decisions are more likely correct than those of any human or humans, including decisions impacting equality between citizens. When choosing between governments, we consequently have a better chance at preserving equality under epistocratic AI rule.

Finally, we come to concerns about preserving collective self-determination. This objection remains too general. Citizens are no less free to collectively self-determine under

[34] Burgess (2022: 104; 2024: 136-137) makes a similar point.

[35] Jayaram and Sparks (2022: 205-206) and Lovett and Zuehl (2022: 469-470) also endorse this point.

epistocratic AI rule than under epistocratic human rule, which is to say both governments permit groups to control morally arbitrary decisions.

Overall, epistocratic AI rule avoids many of the objections to epistocratic human rule and performs at least as well with respect to others. An AI ruler therefore constitutes an ideal epistocracy, at least assuming the as-yet contingent issues with its feasibility can be addressed. For the sake of finding a fundamental objection, I suppose they can.

3. Objecting to the Basic Argument

How, then, can we object to such an ideal epistocracy? As mentioned earlier, we can argue against epistocracy by answering the Surplus Value Challenge: we need to find a value only non-epistocratic governments adequately produce that is important enough to justify choosing them over epistocracy. In this section, I begin to answer this challenge by motivating a new value: collective moral achievement. Specifically, it's valuable for individuals to belong to peoples that have adequate opportunity to develop records of meaningful moral achievement.

As I'll explain shortly, collective moral achievement is valuable for many of the same reasons we value individual moral achievement. I'll also suggest why democracy can provide this value while epistocratic AI rule can't. Then, in the next section, I'll fully answer the Surplus Value Challenge by showing that not only AI rule but all forms of epistocracy fail to adequately produce this value.

We can start to appreciate the value of collective moral achievement with a thought experiment. Suppose an AI system were always available to make *any* decision, not just those relevant to politics. Suppose, furthermore, you knew the system was substantially more likely to

produce correct decisions than yourself. If you ought to use whichever decision-process likely produces better moral outcomes, then you ought to always defer to this system when you face a moral choice. For instance, when deciding between going camping with your friends or gathering toys for homeless children, you ought to rely on the AI to tell you what to do.

Maybe it's permissible for you to always defer to the AI. But you're also surely permitted to make many morally important decisions yourself, even if doing so runs a significantly higher risk of producing worse moral outcomes.[36] This is at least because you have a strong interest in having the opportunity to develop a record of meaningful moral achievement. Doing what you are told can be an achievement if sufficiently difficult, so the above scenario does provide you some opportunity for moral achievement.[37] Yet it is not meaningful if what you can morally achieve is exhausted by your complying with someone else's decisions. Permitting you a few achievements to call your own here and there also inadequately satisfies your interests. What you require instead is the opportunity to develop an open-ended record of meaningful moral achievements, one you can display to yourself and others and develop by making new achievements as you see fit.

Of course, the Basic Argument is about governments, not individuals. This thought experiment nevertheless suggests a way to show this argument's second premise is false. We can do so by establishing the following Key Claim: it is permissible not only for individuals to have

[36] This judgment remains even discounting merely self-affecting decisions, as well as decisions involving individuals like partners, friends, and family who often want us to make up our own minds about how to treat them.

[37] Bradford (2016: 797).

adequate opportunity to develop a record of meaningful moral achievement but also the peoples to which they belong.

'A people' for my purposes refers to any set of individuals than can form a collective agent and whose members share socially salient characteristics.[38] Examples of peoples include residents of geographic regions, such as the European people; members of political states and nations, such as the Japanese people; members of religious groups, such as the Muslim people; and even members of special interest groups, such as Mothers Against Drunk Driving.

We should accept the Key Claim. Since the peoples individuals belong to form a significant part of their social identities, many of the reasons it is permissible for individuals to develop a record of moral achievement are reasons it is permissible for peoples to develop such a record.[39] We can sort the interests grounding these permissions into those that are intrinsic and those that are extrinsic.

Consider the former. As the thought experiment of the always-available-AI demonstrates, individuals have an intrinsic interest in having the opportunity to develop a record of moral achievement. But individuals also have an intrinsic interest in the peoples they belong to having the opportunity to develop a record of moral achievement. Judgements of cases likewise support this claim. Consider two scenarios: Bump and Buy.[40] In Bump, an ice cream truck hits a bump in the road next to a playground. This causes a couple tubs of ice cream to fall off, which are then happily consumed by children. In Buy, members of your community pool their money to buy the

[38] See Applbaum (2019: 120-121) for a similar definition.

[39] On the relation between group membership and identity, see Davis et al. (2019: 256).

[40] I owe this example to Nick Willis.

same amount of ice cream to be just as happily consumed by just as many children. Assuming all else remains equal, which is better? The answer is clearly Buy. This is difficult to explain unless the collective moral achievements of your community are valuable for their own sake.

There are yet more reasons to accept that individuals have intrinsic interests in peoples' moral achievement. For one, individuals often make substantial sacrifices to contribute to collective moral achievements even when others are likely to bring about the same outcomes. Think, for instance, of the many individuals who forego substantial amounts of income to work for charities and non-profits, even in crowded markets where similar organizations duplicate services. This is difficult to explain unless they value their particular groups alleviating problems. For another, moral achievements often have a prominent place in peoples' collective narratives. Think, for instance, of the narrative of progressing freedom American liberals tell about themselves, which involves achievements like ending slavery, enfranchising women, ending segregation, and establishing a legal right to gay marriage.[41]

Next, consider extrinsic interests. Moral motivation provides one reason it's important for individuals to have adequate opportunity to develop a record of meaningful moral achievement. Our moral achievements are central to our sense of identity, which provides an important source of moral motivation: we are motivated to do good things partly to maintain a consistent sense of ourselves as persons who achieve good things.[42] But the same plausibly goes for peoples. Since the peoples we belong to are part of our sense of identity, we are motivated

[41] Mayer (2014: 104).

[42] See Hardy and Carlo (2005: 235) and Schlenker et al. (2009).

to do good things to maintain a consistent sense of ourselves as members of peoples who achieve good things. Denying peoples adequate opportunity for moral achievement consequently denies us an important motive for moral behavior.

Another reason it's important for individuals to have adequate opportunity to develop a record of meaningful moral achievement is to control their social standing. Our moral achievements have a significant impact on how others treat and think about us. But so do the moral achievements of our peoples. Psychological studies suggest that perceived morality has a greater impact on our overall assessments of out-groups than their friendliness or competence.[43] A group's morality is also important for its integration with other groups: immigrant groups perceived as immoral are more strongly expected to adopt the culture of host groups, while host groups perceived as immoral tend to have their culture more strongly rejected by immigrant groups.[44] Denying peoples adequate opportunity for moral achievement accordingly denies them an important means of appropriately managing the judgements their members face from other groups. Because self-sorting into in-groups and out-groups is a general human tendency, we should expect that peoples must manage these judgments under any kind of government.

Given the above evidence, collective moral achievement is clearly valuable – so valuable that pursing it is worth risking some worse moral outcomes, just as with individual moral achievement. We should accordingly accept the Key Claim and so accept it is permissible for peoples to have adequate opportunity to develop a record of meaningful moral achievement.

[43] Brambilla et al. (2012: 160-161).

[44] Urbiola et al. (2021: 13).

But then we should reject the Basic Argument's second premise that when choosing between governments, we are obligated to adopt the one that likely produces morally better outcomes. A people cannot develop a record of moral achievement without making a sufficient number of morally important decisions for itself. However, a people that makes decisions by deferring to an AI does not make morally important decisions for itself. So, if we were really under the obligation just mentioned, and if epistocratic AI rule ever became feasible, this would have the absurd result that peoples shouldn't develop records of moral achievement. As we just saw, this is in fact permissible.

What, then, must a government be like to enable a people to develop a record of moral achievement? It must allow them to produce outcomes in a way that is compatible with what we may call the Principle of Collective Moral Achievement. This is the principle that a people develops a record of moral achievement only if its members freely control successive decisions over morally important outcomes. Without free control, a people cannot be responsible for outcomes so as to appropriately claim them as their achievements. And without making successive decisions, a people's record of moral achievement cannot be developed.

Democracy satisfies the Principle of Collective Moral Achievement. I follow various democratic theorists in considering democracy a form of government that protects the freedom of all adults to participate equally and effectively in a wide range of political processes.[45] Understood in this way, democracies enable peoples to develop records of moral achievement by giving their members free and equal control over morally important outcomes, whether

[45] Held (2006: 281-282), Schneirov and Fernandez (2013: 2, 4, 11), Dahl and Shapiro (2015: 37-41).

through establishing their right to vote, allowing them to petition and protest for social changes, or providing opportunities for them to act as interest groups that craft policy and persuade the public to ensure its legislation. Developing a record of meaningful moral achievement furthermore requires risking suboptimal outcomes. It would not be a meaningful achievement if the outcomes you brought about were all constrained beforehand to be morally good or neutral. Democracies do not require constraining options this way. Accordingly, we can point to democracy in responding to the Surplus Value Challenge: It is permissible to adopt democracy instead of AI epistocracy because democracy allows peoples to pursue moral achievements and AI epistocracy does not.

Note, however, that this does not establish a democratic right to do wrong.[46] It would be unacceptable for an individual to substantially risk depriving others of life, health, and to violate other basic rights in order to pursue moral achievement. It's similarly unacceptable for a people to substantially risk depriving others of life, health, and to violate other basic rights for the same end. Imagine a people demanding to coordinate disaster response or the provision of life-saving medical care instead of a much more qualified agent just because doing so successfully would be a great achievement. Such demands ought to be rejected.

This is only to say the importance of collective moral achievement, like that of all values, is limited. Respecting what importance collective moral achievement does have permits us to limit the encroachment of epistocratic governance into our collective decision-making. Since a vast number of peoples stand to claim the opportunity for moral achievement, and since it's

[46] See Øverland and Barry (2011) for discussion.

frequently difficult to control morally significant outcomes without making decisions in areas under the government's purview, we have good reason for these limits to be extremely broad. The importance of collective moral achievement accordingly provides not just an objection to the Basic Argument but the makings of a fundamental objection to epistocracy.

4. Objecting to the Epistocratic Thesis

Up to this point, I've established that collective moral achievement is valuable and that democracies can supply this value. But to fully answer the Surplus Value Challenge and show collective moral achievement provides a fundamental objection to epistocracy, I must address some likely responses from epistocrats. Epistocrats may reply that epistocracy in fact provides opportunities for collective moral achievement, whether by claiming that epistocratic AI rule adequately provides these opportunities or that human epistocracies provide them. We must rule out both possibilities.

Why, then, might one think epistocratic AI rule allows for collective moral achievement? One set of reasons comes from what it takes to construct an epistocratic AI. Recall that the outputs of AI systems are based on training data, which is ultimately supplied by humans. Epistocrats may reply that a people can develop a record of meaningful moral achievement through involvement in this process. That an AI ruler ends up legislating just laws, they may claim, is a meaningful moral achievement of the people on whose data it was trained.

A closer look at the ways to engineer AI systems shows this reply fails. The data for an AI ruler could be generated either from the revealed preferences of members of a collective or from the deliberate selection of these members. The collective in question could be either identical to

or different from the people the AI rules. This yields four general ways of engineering an AI ruler. In all of them, AI rule either violates the Principle of Collective Moral Achievement or fails to adequately satisfy the interests individuals have in their people's moral achievement.

To cover two cases, suppose the collective generating the training data is different from the people ruled. Then the AI system's output is not an achievement of the people regardless of whether the data is generated by revealed preferences or deliberate selection. This is because the people's members lack free control of the AI's decisions on account of their inability to influence its outputs. Even supposing the AI's decisions were an achievement, it would belong to the collective generating system's training data, not the people on whom the system is applied.

Suppose instead the people ruled is somehow generating the data. For a third case, suppose also that the people generates data for the AI by its members' revealed preferences. This involves the AI system inferring which outputs to consider correct from the members' behavior.[47] For instance, the AI might decide to increase income equality because it observes individuals have a salient tendency to move to areas with higher income equality. This approach has the benefit of avoiding the problem that individuals may not accurately express their preferences. Those who behave as if they value income inequality may be unsure or even strongly deny having this preference if asked.

Epistocratic AI rule is incompatible with the Principle of Collective Moral Achievement in this case, once more because the people inadequately controls the system's outputs. To control

[47] See Russell (2019: 190) for discussion of how AI systems can learn preferences from observing behaviour and Burgess (2022: 100) for a recent analysis of this technique in political contexts.

an outcome requires aiming at the outcome. It also requires that the outcome one aims at and the outcome one causes robustly covary.[48] The revealed preferences approach fails to satisfy this second condition. The outcomes a people aims at depend on its members' express preferences while the outcomes the AI system produces depend on its members' revealed preferences. Because both sorts of outcomes easily diverge, the outcomes produced by the AI are not the people's achievements, even if most of the time they happen to coincide. It seems absurd that a people with an express preference for income inequality (shown perhaps by their proclaiming as much, and even voting to increase inequality when given the chance) could appropriately claim legislation designed to produce income equality as their own achievement, let alone one they find meaningful.

For the fourth and final case, suppose the training data is generated through deliberate selection by a democratic electorate. Perhaps a people's members vote to consider laws banning animal cruelty correct, laws encouraging it incorrect, and so on. An AI could then be developed to craft legislation reflecting these preferences.

This approach might provide a people some opportunity for moral achievement, but not *meaningful* moral achievement. Training an AI by vote is similar to taking a poll on social media: responding is easy, there is no guarantee the collective result will lead to anything, and the vote may not even be made in response to actual circumstances. In fact, considering the amount of data typically required to train AI systems, individuals will likely have to vote on a large number of hypothetical laws and policies to ensure the AI ruler performs adequately. Successfully fighting

[48] I owe this point to Mikayla Kelley.

to push a law or policy though a human-run legislative process is much more difficult and engaging, sufficiently so to deserve being called a meaningful achievement.

Epistocrats may respond that complying with the AI's decisions can also be difficult. Doing what you are told can be an achievement – perhaps even more so when you have some influence over what you are told. But we should remember that the AI's decisions ultimately take the form of coercive orders. A people living under a democratically trained, epistocratic AI for this reason becomes akin to an individual who writes a list of actions which she hands off to a reliable assistant who later coerces her to perform similar ones. Even if this person wrote morally good actions on the list, it's justified to doubt whether she would have complied with the assistant's orders without being coerced, and so it's reasonable to reject that her compliance is a meaningful achievement.

At this point, epistocrats may claim there is another set of reasons epistocratic AI rule allows for collective moral achievement. These come from how an epistocratic AI could function post-construction. Paul Burgess has recently discussed the advantages of replacing human politicians with AI systems in representative democracy, where AIs would be elected to represent and craft laws for their human constituents.[49] Epistocrats might expand on this idea and argue that voting for epistocratic AI systems provides the best of both worlds: epistocratic advantage and sufficient participation for collective moral achievement.

However, this kind of government fails to avoid the Collective Moral Achievement Objection. One problem is that the choice over candidate AI representatives must be severely

[49] Burgess (2022: 101, 104, 107; 2024: 131).

restricted for this government to be epistocratic in the sense that it puts whoever is most likely to make correct decisions in charge. By an epistocrat's lights, if one candidate AI made significantly worse decisions than others, then a government denying that AI as an option for election would track correct decisions better than a government allowing the AI on the ballot. In short, all the candidate AIs must be as good as practically possible at making correct decisions for the government to be epistocratic. But then the control over options this government provides voters is too impoverished for meaningful moral achievement. Similarly, a person given only healthy food options can't claim being a healthy eater as an achievement.

Another problem is that elected AIs must represent the public as trustees rather than directed delegates for the government to be epistocratic.[50] If an AI acted as the latter kind of representative, it would always do the bidding of its constituents. There would be no epistocratic advantage in this arrangement unless the constituents were better at making correct decisions than the AI, in which case this government could hardly be called an AI epistocracy. If the AI were instead better at making correct decisions than its constituents, then by the epistocrat's lights it should act as a trustee and always do what it judges most beneficial for them. This would again leave the constituents, and the peoples they compose, impoverished control over outcomes. At most, the AI might consider its constituents' preferences when making decisions and edit them into what it judged more favorable directions. This is inadequate for collective moral achievement. Similarly, a person who is always given oatmeal smoothies or prunes when she asks

[50] See Pitkin (1972: 127, 134) for discussion of the trustee-delegate distinction in democratic theory.

for ‘something sweet’ can’t claim being a healthy eater as an achievement, especially if the person really wanted treats like donuts and soda.

We should conclude that AI epistocracy fails to provide adequate opportunity for collective moral achievement both during and after its construction. Could human epistocracy do any better?

Any epistocracy requires denying some individuals control over political decisions to improve outcomes. There are roughly two ways this can happen under human epistocratic rule. First, some of the people’s members can qualify for sole power, weighted power, or veto power over decisions by meeting some epistocratic standard. Second, some members can qualify for power after being randomly selected then trained according to an epistocratic standard.[51] Both cases again either are incompatible with the Principle of Collective Moral Achievement or inadequately satisfy individual interests in their peoples’ moral achievement.

This is most apparent when qualifying or randomly selected members are given sole power over political decisions. Since the people’s members are not generally free to control outcomes in these governments, whatever achievements these outcomes amount to are properly attributable to qualifying members alone.[52] This objection seemingly runs into trouble considering that, strictly speaking, at least some members of a people do not qualify to contribute

[51] Another form of epistocracy Jason Brennan discusses is what he calls government by simulated oracle (2016: 220-222). This involves using weighted scorings of voters’ policy preferences and political knowledge to predict what they would want if they were fully informed. Political decisions are then based on these predictions. My objections to giving competent voters weighted power or veto power apply to this form of epistocracy as well.

[52] Lovett and Zuehl (2022: 484-485).

to political decisions in almost any government, including democracy. But we rarely speak strictly anyway. The problem for epistocracy is that 'American Citizens who pass a voter competence test' is not a plausible disambiguation of who we refer to when we say 'The American people achieved a legislative victory', whereas 'American Citizens over 18' is.

Giving qualifying members weighted power over political decisions (e.g., by having their votes double-counted) or veto power might give more control to the people as a whole. However, these forms of epistocracy do not adequately ensure control over successive decisions. Adopting epistocratic governments is worthwhile only if they produce different outcomes than non-epistocratic governments a significant amount of the time. This entails, for the same reasons given above, that the people as a whole is not responsible for these decisions a significant amount of the time. This is incompatible with producing a pattern of outcomes that are clearly attributable to a people that its members can reflect on and modify with new achievements as they see fit.

Yet another problem with giving qualified members weighted power or veto power (one shared by epistemic training by random selection) is that these arrangements taint outcomes. All these governments constantly threaten that the achievements at which a people aims will be vetoed, outvoted, or unsupported after training by its own members. This incentivizes conforming to the preferences of these members, since what a people endeavors to do will more likely succeed the more it is favored by those deemed epistocratically competent. Consequently, whenever a people brings about an outcome, its members will have good reason to not consider the achievement fully their own, thereby significantly reducing its meaning. These governments will also create reasonable doubt among out-groups that the people's decisions are properly

attributable to it, thereby undercutting the potential benefits of collective achievement for social standing. One can imagine members of out-groups saying: 'The people achieved just outcomes only because different ones would have been vetoed', or 'They made the right decision only because of the epistocratic training'.

Overall, any substantive version of epistocracy, whether run by humans or AI, gives peoples inadequate opportunity to develop a record of meaningful moral achievement.[53] The basic reason why is that epistocracy provides no guarantee that free control over decisions will be equally distributed among a people's members. The Collective Moral Achievement Objection accordingly reveals a fundamental problem with epistocracy, one that is not contingent on political arrangements or technological developments.

5. Conclusion: The Permissible Ways to Incorporate AI in Politics

So how should we engineer digital technologies for use in politics? The above considerations show we are not obligated to supplant human decision-makers with AI systems, even if deference to latter guarantees substantially better moral outcomes. This is because we are at least permitted to preserve a sphere of human, democratic decision-making large enough to satisfy individual interests in collective moral achievement. Again, given the vast number of peoples with which individuals identify, this sphere must be very expansive, even though this risks worse moral outcomes overall.

[53] Except when epistemic democrats are right that properly arranged democracies qualify as epistocracies.

I do not claim, however, that we are morally required to support human decision-making in politics. This is because striving for collective moral achievement is itself not morally required. The members of many peoples may voluntarily forego its value because they see some good in living under full epistocratic management by AI rulers. I also do not claim we are permitted to expand the sphere of democratic, human decision-making as wide as possible. As discussed earlier, this is because it is impermissible to trade moral achievement – collective or individual – against substantially better protection of basic rights.

Epistocrats may eagerly note that this latter statement is compatible with the requirement to adopt a government where decisions risking substantial basic rights violations are made by experts (whether human or machine) and all others are made by democratic means. In short, the sphere of democratic, human-decision making ought to exist within an epistocratic shell of basic rights protections, perhaps managed by AI.

I acknowledge that this nested government avoids the Collective Moral Achievement Objection. But if this is the strongest reply epistocrats can make, their view is left denuded of the revisionary appearance that largely made it interesting in the first place. It is uncontroversial that experts ought to make decisions about maintaining critical infrastructure, responding to public health emergences, supervising national economies, and other issues where life, health, and maintaining basic living standards are clearly at stake – after all, experts already play this role. Their response is furthermore tempered by the fact that there is no expert consensus about which basic rights exist, as well as no expert consensus about how to appropriately form such a consensus. We will lack sufficient moral certainty to permissibly impose a single, overarching vision of basic rights protection for the foreseeable future.

As digital technology becomes increasingly integrated with politics, recognizing the value of collective moral achievement is crucial. Rather than supplanting human decision-makers, we can explore new ways to support them, possibly with AI systems designed for this purpose. Determining the boundaries between supporting and supplanting is a topic that deserves at least its own article. But respecting this boundary is essential if we want peoples to be free to strive for moral achievements of their own. I believe many of us do.

Conflict of Interest Statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data Availability Statement

I do not analyse or generate any datasets, because my work proceeds within a theoretical and mathematical approach.

Funding Statement

This research was funded by Seth Lazar's Australian Research Council Future Fellowship grant FT210100724.

References

Achen, Christopher H. and Larry M. Bartels (2017). *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton, NJ: Princeton University Press.

Altman, Sam, Greg Brockman, and Ilya Sutskever (2023, May 22). Governance of Superintelligence. *OpenAI Blog*. https://openai.com/blog/governance-of-superintelligence. Accessed 5 January 2024.

Applbaum, Arthur Isak (2019). *Legitimacy: The Right to Rule in a Wanton World*. Cambridge, MA: Harvard University Press.

Barcelona's Digital Democracy (2021). *City Quality Magazine*, 29 June, https://cityqualitymagazine.com/barcelonas-digital-democracy/. Accessed 10 October 2023.

Bradford, Gwen (2016). Achievement, Wellbeing, and Value. *Philosophy Compass* 11 (12): 795-803.

Brambilla, Marco, Simona Sacchi, Patrice Rusconi, Paolo Cherubini, and Vincent Y. Yzerbyt (2012). You Want to Give a Good Impression? Be Honest! Moral Traits Dominate Group Impression Formation. *British Journal of Social Psychology* 51 (1): 149-166.

Brennan, Jason (2016). *Against Democracy*. Princeton: Princeton University Press.

Burgess, Paul (2022). Algorithmic augmentation of democracy: considering whether technology can enhance the concepts of democracy and the rule of law through four hypotheticals. *AI & Society* 37: 97-112.

Burgess, Paul (2024). *AI and the Rule of Law: The Necessary Evolution of a Concept*. Oxford: Hart Publishing.

Coglianese, Cary and David Lehr (2019). Transparency and Algorithmic Governance. *Administrative Law Review* 71 (1): 1-56.

Dahl, Robert A. and Ian Shapiro (2020). *On Democracy*, 2nd ed. New Haven, CT: Yale University Press.

Danaher, John (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy and Technology* 29 (3): 245-268.

Davis, Jenny L., Tony P. Love, and Phoenicia Fares (2019). Collective Social Identity: Synthesizing Identity Theory and Social Identity Theory Using Digital Data. *Social Psychology Quarterly* 82 (3): 254-273.

Eggers, William D., David Schatsky, and Peter Viechnicki (2021). AI-Augmented Government: Using Cognitive Technologies to Redesign Public Sector Work. *Deloitte Insights* (Online). https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/artificial-intelligence-government.html. Accessed 13 October 2023.

Estlund, David M. (2008). *Democratic Authority: A Philosophical Framework*. Princeton University Press.

Gibbons, Adam F. (2021). Political Disagreement and Minimal Epistocracy. *Journal of Ethics and Social Philosophy* 19 (2): 192-201.

Goodin, Robert E. and Kai Spiekermann (2018). *An Epistemic Theory of Democracy*. Oxford, United Kingdom: Oxford University Press.

Hardy, Sam A. and Gustavo Carlo (2005). Identity as a Source of Moral Motivation. *Human Development* 48 (4): 232-256.

Held, David (2006). *Models of Democracy*, 3rd ed. Cambridge: Polity Press.

Huyen, Chip (2022). *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications.* Sebastopol, CA: O'Reilly Media, Inc.

Ingham, Sean and David Wiens (2021). Demographic Objections to Epistocracy: A Generalization. *Philosophy and Public Affairs* 49 (4):323-349.

Input Crowd, Output Meaning (2023). https://pol.is/home. Accessed 10 October 2023.

Jafar, Uzma, Mohd Juzaiddin Ab Aziz, and Zarina Shukur (2021). Blockchain for Electronic Voting System—Review and Open Research Challenges. *Sensors* 21 (17): 1-22.

Jayaram, Athmeya and Jacob Sparks (2022). Rule by Automation: How Automated Decision Systems Promote Freedom and Equality. *Moral Philosophy and Politics* 9 (2):201-218.

Jeffrey, Anne (2018). Limited Epistocracy and Political Inclusion. *Episteme* 15 (4): 412-432.

Jones, Garett (2020). 10% Less Democracy: Why You Should Trust Elites a Little More and the Masses a Little Less. Stanford, CA: Stanford University Press.

Kogelmann, Brian (2022). Finding the Epistocrats. *Episteme*: 1-16. <doi:10.1017/epi.2022.18>

Landemore, Hélène (2013). *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton, NJ: Princeton University Press.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23 (18): 1-45.

Lovett, Adam and Jake Zuehl (2022). The Possibility of Democratic Autonomy. *Philosophy and Public Affairs* 50 (4): 467-498.

Margalit, Avishai and Joseph Raz (1990). National self-determination. *Journal of Philosophy* 87 (9):439-461.

Maas, Jonne (2023). Machine learning and power relations. *AI & Society* 38: 1493-1500.

Mayer, F. W. (2014). *Narrative Politics: Stories and Collective Action*. New York, NY: Oxford University Press.

Macpherson, C. B. (1977). *The Life and Times of Liberal Democracy*. Oxford: Oxford University Press.

Øverland, Gerhard and Christian Barry (2011). Do Democratic Societies Have a Right to Do Wrong? *Journal of Social Philosophy* 42 (2): 111-131.

Pitkin, Hanna Fenichel (1972). *The Concept of Representation*. Berkeley, CA: University of California Press.

Russell, Stuart (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York, NY: Viking Press.

Schlenker, Barry R., Marisa L. Miller, and Ryan M. Johnson (2009). Moral identity, Integrity, and Personal Responsibility. In D. Narvaez and D. K. Laplsey (Eds.), *Personality, identity, and Character: Explorations in Moral Psychology*. New York, NY: Cambridge University Press (316–340).

Schneirov, Richard and Gaston A. Fernandez (2013). *Democracy As a Way of Life in America: A History*. New York, NY: Routledge.

Stilz, Anna (2016). The Value of Self-Determination. In David Sobel, Peter Vallentyne, and Steven Wall (eds.), *Oxford Studies in Political Philosophy, vol. 2*. New York, NY: Oxford University Press. pp. 98-127.

Urbiola, Ana, Lucía López-Rodríguez, María Sánchez-Castelló, Marisol Navas, and Isabel Cuadrado (2021). The Way We See Others in Intercultural Relations: The Role of Stereotypes in the Acculturation Preferences of Spanish and Moroccan-Origin Adolescents. *Frontiers in Psychology* 11 (1): 1-16.

Wang, Angelina, Sayash Kapoor, Solon Barocas, and Arvind Narayanan (2022). Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy. *FAccT Proceedings*: https://doi.org/10.1145/3593013.3594030

Wischmeyer, Thomas (2020). Artificial Intelligence and Transparency: Opening the Black Box. In: Wischmeyer T, Rademacher T (eds.) *Regulating Artificial Intelligence*. Springer International Publishing, 75–101.

Zheng, Stephan, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C Parkes, and Richard Socher (2022). The AI economist: Improving equality and productivity with AI-driven tax policies. *Science Advances* 8 (18): 1-17.