AI & SOCIETY (2026) 41:135–149
https://doi.org/10.1007/s00146-025-02474-9

RESEARCH

AI preference prediction and policy making

James Edgar Lim[1] · Julian Savulescu[1,2]

Received: 14 March 2025 / Accepted: 26 June 2025 / Published online: 23 July 2025

Abstract
Democratic decision-making is difficult. Representatives often fail to represent the preferences of their constituents, and directly consulting members of the public can be costly. Inspired by these difficulties, several scholars have discussed the use of artificial intelligence (AI) models to support democratic decision-making. One such particular application is the use of AI to represent public policy preferences by predicting them. In this paper, we perform an analysis on the different ways AI models can be used to represent public policy preferences. We make distinctions between using AI as epistemic tools and for procedure; group and individual predictions; and predictions about preferences and inferences about values. We also describe how AI models can help policymakers screen policies for potential worries and objections, double-check any beliefs they have about the acceptability of their policies, and justify policy proposals. We also consider a number of worries about the use of AI in policymaking. We argue that these worries, while legitimate, can be mitigated or avoided in the way we have proposed the use of AI.

Keywords Artificial intelligence (AI) · Democracy · Augmented democracy · Preference prediction · Collective reflective equilibrium

1 Introduction

Good policymaking should take into account the preferences and values of the public.[1] But determining the preferences of a population is no easy task. In many liberal democracies (as well as semi-liberal ones), voters elect representatives to make policy based on the preferences and values of their constituents. But it is unclear how successful electoral mechanisms are at translating voter preferences into policy. Some studies have found that the preferences of representatives reflect the preferences of voters, and vice versa (Griffin and Newman 2005; Soontjens 2022a). But other scholars point out that the behavior of politicians and voters is often affected by party politics (Soontjens 2022b), social identity (Ågren et al. 2007; Jenke and Huettel 2020; Klar 2013), voter information (Nordin 2014), and the confidence of politicians (politicians less confident in re-election tend to reflect voter preferences) (Soontjens and Sevenans 2022). Instead of a representative system, nations and communities can use more direct methods for policymaking. For example, they can use referendums, extensive consultation, or public deliberation. However, these methods are resource intensive and time consuming. They are therefore impractical to use at the national or state level for all decisions. Furthermore, it can be demanding for members of the community to participate

✉ Julian Savulescu
jsavules@nus.edu.sg

James Edgar Lim
jameslim@nus.edu.sg

[1] Centre for Biomedical Ethics, Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[2] •Oxford Uehiro Centre for Practical Ethics, University of Oxford, Oxford, UK

[1] There are some theorists who point out that policymaking should sometimes override voter preferences. E.g., see (Jackson 2023; Sunstein 1991). This view is compatible with our claims in this paper, as we do not claim that policymaker should always act according to the preferences of the public.

Instead, we assume that it is generally better for a community when policymakers have a more accurate and detailed understanding of the preferences of the community. Even when voter preferences are misinformed, biased, or bad in some other way, it can be valuable for policymakers to be aware of these preferences, so they can anticipate pushback, respond to pushback, engage in more productive discourse, etc. We think that this is a plausible assumption, although we acknowledge that competing considerations could often make it undesirable for policymakers to have too much information. For example, if the policymaker deliberately intends to manipulate the public for their own gain, she may be able to abuse the information she has.

in frequent political decision-making, and as a result, decisions may be disproportionately representative of people with more time and resources on hand. Finally, when very technical or novel issues are on the line, such as regulation concerning novel technologies, people may not have enough information to have any preferences at all. Thus, even under (generous and) relatively ideal conditions where all members of society behave in good faith, there is no easy way to ensure that policies represent the preferences and values of the public.

What if we could find out the preferences of a group of people without asking them? Enter artificial intelligence (AI). For our purposes, AI refers to any computational system used for problem-solving, decision-making, data analysis, and other complex tasks. These include generative AI like Large Language Models (LLMs) and predictive algorithms. In particular, algorithms have been used (with varying levels of accuracy) to predict the preferences of individuals in a multitude of domains, such as in entertainment (Maroto-Gómez et al. 2023), marketing (Cai 2001), travel (Topal and Uçar 2019), healthcare (Biller-Andorno and Biller 2019; Earp et al. 2024; Ferrario et al. 2023), and online lodging systems (Ma et al. 2021), among many others. It is thus plausible that algorithms can be developed and used to predict the political and moral preferences of either individuals or groups of people (Lazar 2024). Infamously, the 2018 Cambridge Analytica scandal exposed the company's use of the social media data of 50 million Facebook users to build voter profiles, which were sold to political campaigns (Confessore 2018). Advances in AI within the last seven years could greatly increase the potential of such voter profiles—for both good and evil. The Cambridge Analytica scandal reminds us of how some groups can use data to violate the privacy of individuals and gain illegitimate power and influence over our electoral systems. But it also demonstrates the potential for new ways in which policymakers can understand the preferences and values of their electorate without organizing costly consultations and referendums. And as long as policymakers abide by norms regarding privacy and transparency, it is at least possible that algorithms can be used for good.

In this paper, we aim to bring further clarity and precision to the growing discussions on how AI can be used to improve outcomes for democratic communities. To this end, we perform an analysis on the distinctions between different ways in which AI models can be used to play a role in policymaking and democratic decision-making. We describe different sorts of predictions AI models can make regarding people's preferences, and explain the different roles they can play in policymaking. We discuss how AI predictions can be epistemically helpful for policymakers, thus helping them be more responsive to the preferences of their constituents, thus enhancing the quality of public deliberation and lawmaking.

To be clear, we do not suggest that the use of AI models replace voting or existing mechanisms of democratic governance; voting and democratic decision-making are still of vital importance. Nor do we want to minimize the risks of using AI models to make these kinds of predictions. Discussing these risks is incredibly important, but a full assessment of these risks is outside the scope of this paper. Finally, while this paper is related to many discussions in political theory, it does not provide a full political theoretic justification or repudiation of the use of AI in policymaking. Such a project would be very worthwhile, but it is outside the scope of this paper.

Here is an outline of this paper. In Part II, we provide an overview of some distinctions in the ways that AI models can be used to predict people's preferences. We discuss the differences between using AI as epistemic tools and for procedure; group and individual predictions; and predictions about preferences and inferences about values. In Part III, we make proposals regarding how AI models can be used to help people and policymakers gain valuable information about preferences, for the purposes of policymaking. These involve primarily epistemic uses, where policymakers use AI to gain information about public preferences. They include using AI to screen and double-check policies, as well as using AI as part of "collective reflective equilibrium in practice" (CREP), a method for forming justified proposals in public spaces (Savulescu et al. 2021). Finally, in Part IV, we discuss some worries about the use of AI models in policymaking, and argue that while these worries are indeed important considerations, there are not any decisive reasons to rule out the use of AI predictive tools altogether.

2 How can AI models be used?

Researchers have explored a number of different ways AI models—specifically Large Language Models (LLMs) can be used to aid democratic decision-making (this paper will discuss AI models generally, although a large amount of empirical research is on LLMs and generative AI specifically). These include:

- Summarization: presenting a set of views to policymakers in a digestible report (Small et al. 2023).
- Aggregation of summarized content to find consensus (Devine et al. 2023; Huang et al. 2024; Konya et al. 2023; Lazar & Manuali 2024; Porsdam Mann et al. 2023).
- "Representation": predicting and being "proxy agents acting on a principal's behalf in a democratic process" (Lazar & Manuali 2024, p. 5).[2]

[2] Also see (Bakker et al. 2022).

- Facilitation of discussion (Edelman 2023; Landmore 2023; Mendoza 2023; Small et al. 2023).

In this paper, we are primarily interested in the use of AI models to predict people's policy preferences. As we shall see, this use of AI is related to summarization, aggregation, and representation, although these categories are not very important for our purposes. Researchers have achieved some success in developing models which can make these kinds of predictions. For example, Gudino-Rosero et al. (2024) trained an LLM to predict the individual political choices and the aggregate preferences of a group of participants (by inviting participants to provide some demographic details and by asking participants to choose between pairs of proposals). In their study, Gudino-Rosero et al. found that an LLM trained on a person's preferences was more accurate at predicting preferences than a "bundle rule"—which predicts that a person will prefer policies which match the political ideology of the participant (Gudiño-Rosero et al. 2024).

2.1 The value of prediction: epistemic vs participatory

For our purposes, predictions about policy preferences can have two uses: epistemic and participatory. An AI model is used primarily for epistemic reasons if the primary aim of the user is to gain some information about the preferences held by a group of people. For example, a policymaker who wants to know how her constituents feel about a lower speed limit might use an LLM to predict how her constituents feel about the policy. In addition, AI models could also be used to make predictions about how constituents might feel about issues they do not know about yet, such as issues regarding new technologies. These predictions may be based on the values held by a group, or their preferences regarding other issues. We shall discuss such predictions in more detail later. In comparison, an AI model is used primarily for participatory reasons if it is used as a proxy for some procedural requirements of policymaking in democratic societies. For example, we might imagine a system where, instead of voting, citizens delegate their choice to an AI model, particularly a personalized model (Porsdam Mann et al. 2023), which casts their "votes" for them.[3] Such a use case of AI may have an epistemic element, because these AI votes do provide policymakers with information about the preferences of the citizens. But because AI models take over the role of voting, the use of AI in this case is also participatory. Other participatory use cases may be AI models which "represent" citizens in deliberative democratic forums, to write formal petitions to parliament on behalf of citizens, or to replace elected representatives in some legislative functions (for example, AI models which participate in hearings or committee meetings).

It is easier to see how the use of AI as a purely epistemic tool can offer benefits for democratic communities. A number of studies have found that politicians are often not very good at predicting the policy preferences of their voters (Broockman & Skovron 2013; Hedlund & Friesema 1972; Walgrave et al. 2023) with politicians often either projecting their views onto their constituents (Sevenans et al. 2023), and politicians believing their constituents to have more conservative views (Broockman & Skovron 2018). In particular, Hedlund and Friesema (1972, p. 741) note that politicians are particularly inaccurate when it comes to issues which are not "highly charged" (and thus not discussed heavily in media). This current state of affairs is clearly not ideal for voters (or policymakers), whose preferences may be misunderstood. AI models can be useful in determining the preferences and values held by a group of people when it would otherwise be difficult or infeasible to solicit their actual opinions. These include situations where the cost of soliciting opinions is high, in times of urgency and emergency, or when an issue is new, unfamiliar, or too technical for most members of the public. AI models can also be used by policymakers to conduct quick and rough surveys of the views probably held by their constituents, which can inform them whether and how to collect further information.

We are less bullish on participatory uses of AI in policymaking, particularly when AI models are used in place of existing democratic processes. We cannot perform a comprehensive analysis of all the procedural use cases of AI here, but we will provide a quick note on our general concern. Democratic processes do not just provide lawmakers with information about what citizens want. Free and fair democratic processes, from voting to deliberative forums, express the equality of citizens, allow citizens to exercise some power and make demands of politicians, provide lawmakers with authority and legitimacy, facilitate the exercise of civic virtue, enable the transformation of preferences through dialog, and give citizens reasons to accept outcomes when their will is not enacted (Christiano and Bajaj 2024). If AI models replace the role humans play in democratic processes, communities may end up foregoing some of the value associated with democracy. For example, citizens cannot develop civic virtue simply by delegating all their voting choices to AI models (Lazar and Manuali 2024).

Unfortunately, the distinction between the epistemic and participatory uses of an AI tool is not always a neat one. Sometimes, a tool used for epistemic purposes may affect the procedures with which policies are considered, even if

[3] Interestingly, this sort of "democratic" system is depicted in the video game *Helldivers 2,* where "Utilizing computer aided voting software, citizens are asked to answer several questions, and the outcome of their vote is decided upon by the computer." (Kain 2024) The tone of the game is distinctly dystopian.

the user does not intend to use the tool in such a way. For example, consider a policymaker who uses an AI tool to generate a list of policies based on what people are likely to prefer. Initially, this AI may seem like an epistemic tool, as its primary function is to show a policymaker which policies their constituents will be most likely to support. But the tool actually has some agenda-setting power. It can determine which policies are "on the table" in the first place. And if the tool is biased—perhaps in a way that ignores the preferences of a minority group—the policymaker may never consider the ideas favored by that group.

In practice, it may be difficult to compartmentalize the use of AI tools, such that their function is *purely* epistemic. AI tools can have unintended consequences on the process of democratic decision-making at many stages of the process. For example, depending on what information an AI tool presents, and how it presents the information, it can cause subconscious bias in human decision-makers. Perhaps more worryingly, the use of powerful AI tools can also affect our practices of holding policymakers responsible for the decisions they make. These tools may facilitate "agency laundering": the act of "obfuscating one's moral responsibility by enlisting a technology or process to take some action and letting it forestall others from demanding an account for bad outcomes that result" (Rubel et al. 2019, p. 1018). In our case, policymakers may be tempted to evade responsibility for their decisions by saying that they were following the directions of some (reliable) AI model. In many real-life situations, AI models will not be some neutral tool which merely gives us information. They can affect the way we consider ideas, discuss them, and how we respond to the way others implement those ideas.

However, a couple of things can be said in defense of the use of AI in some cases. First, there are some precautions that can be taken to minimize the unwanted roles AI tools may play in decision-making. Norms and regulations can be put in place which require policymakers to (continue to) consult their constituents in some way, in addition to any use of AI. This could ensure that those policymakers will consider options which are raised by people but not by the AI. Or again, norms and regulations can go some way to ensuring that policymakers are held responsible for the decisions they make, and cannot instead blame an AI tool.

Of course, these precautions may not be perfect. But this brings us to a second point: other conventional tools can also play a significant role in shaping the process of democratic decision-making. And sometimes, the role these tools play may have undesirable side effects. For example, a survey can also have some agenda-setting power, if conducted near the beginning of a decision-making process. A policymaker may look at the results of the survey and fail to consider any options that were not represented in the survey. If the survey fails to capture the views of some minority group, then their ideas will not be considered by the policymaker. Similarly, focus groups or expert consultations can also have an agenda-setting power. Thus, while it may be true that any use of an AI tool may affect democratic decision-making procedures in an adverse way, this is not necessarily a decisive reason to reject the use of the tool. It is instead important to weigh the risks of a tool against the potential benefits as well as the alternatives. As far as we can tell, it is not obvious that the risks of some specific uses of AI constitute decisive reasons to reject its use.

Since we think that the epistemic uses of AI offer more advantages while creating fewer ethical concerns than the participatory uses, our analysis in subsequent sections will focus primarily on the epistemic uses of AI.

2.2 Group predictions vs individual predictions

AI predictions can be further distinguished according to whose preferences are predicted. There are group-level predictions, and individual-level predictions. As we shall claim, both are potentially useful for policymakers interested in implementing the preferences held by their constituents. An AI model which combines the two may be the best use of AI for these purposes. A group-level prediction is a prediction about what a group of people would prefer, aggregated and summarized via some procedure. For example, Large Language Models (LLMs) are neural networks trained on massive amounts of human language. With this training, they create probabilistic models for how individual words and sequences of words relate to each other. As a result, they learn how to finish incomplete sentences and how to respond to prompts, effectively generating sentences and responses to questions (*What Is a Large Language Model (LLM)?*, n.d.). In one study, Bakker et al. (2022) developed a model which made statements about policy questions (like "should we lower the speed limit on roads") and found that the model outperformed human counterparts at making statements that human participants agreed with.

For illustrative purposes, we asked ChatGPT to make predictions for results if a survey was done on Singaporean opinions about the moral acceptability of in-vitro fertilization (IVF), embryonic selection based on polygenic risk scores (ESPS) and gene editing.[4] We compared those results to an actual study conducted by Haining et al. (2025), which

[4] An example of a prompt we used was, "Suppose a survey was done in Singapore about moral acceptability of embryonic selection based on polygenic risk scores (whereby prospective parents can select an embryo during preimplantation genetic testing as part of the in vitro fertilization). Suppose the poll had the options "ESPS is morally acceptable", "ESPS is not morally acceptable", "ESPS is not a moral issue", and "not sure". Please make a prediction for which option a majority of Singaporeans would choose.".

was unpublished at the time of drafting this part of this paper (hence, ChatGPT would not have access to results of the survey) (Haining et al. 2025) Table 1.

ChatGPT was successful in some general predictions (that Singaporeans would be more accepting of these technologies than not). But it was less successful in making fine-grained predictions. Perhaps, slightly more impressively, ChatGPT was also somewhat accurate in making predictions which compared the attitudes of different ethnic groups within Singapore toward these reproductive technologies (which it disambiguated into treatments for therapeutic and enhancement reasons) (Table 2).

Again, there is some agreement with empirical results, with Haining et al.'s (2025) finding that Chinese Singaporeans were more willing to use both embryo selection and gene editing than Indian Singaporeans. This is a more important result, because most people are intimately familiar with one, or maybe a handful of cultures. Policymakers and pollsters are no exception. A tool which can predict the varied preferences of multiple demographic groups can thus do something that most humans are unable to do without the help of consulting those groups. And as a result, such a tool can plug gaps in knowledge that many people may have.

Again, the results of this particular investigation were for illustrative purposes. More systematic research should be conducted to determine if general, off-the-shelf, LLMs can make robustly accurate predictions about preferences. We suspect that they are only useful for basic outlines of the preferences of their constituents. This is where individualized predictions come in. Algorithms which make individualized predictions are not new, and have been used in search engines, online shopping tools, and music services like Spotify, among other things. Recently, researchers have explored the use of personalized language models and determined that such models made more accurate predictions than non-personalized language models (Woźniak et al. 2024), and have explored the applications of such models like in healthcare (Earp et al. 2024). Personalized language models are refined using data about specific individuals, such as their writing, social media posts, Google searches, etc. For example, Porsdam Mann et al. (2023) have shown that LLMs can generate detailed and high-quality academic writing based on a researcher's writing which makes plausible predictions about what the researcher would say about an issue.

When it comes to moral and policy preferences, there could be personalized LLMs refined through a process in which individuals voluntarily answer purpose-designed surveys and questionnaires. For example, Kim et al. (2018) developed a model to predict the moral preferences of individuals regarding Moral Machine cases (where respondents are asked to choose whether an AV should swerve in a particular direction or not, either harming one hypothetical group of moral subjects or another. Moral subjects varied according to gender, age, species, employment status, criminal status, etc.). They obtained data from 10000 respondents to the moral machine experiment (Awad et al. 2018), including the responses of the respondents to 13 cases. Their model was trained on the first eight responses for a given respondent, and was able to mostly predict the next five responses made by the respondent (Kim et al. 2018).

One particular area where individualized preference prediction has received quite a lot of attention is in predicting the preferences of incapacitated patients using demographic

Table 1 A comparison of ChatGPT's predictions and Haining et al.'s (2025) survey results

Scenario	ChatGPT's predictions			Haining et. al.'s survey results[a]		
	Morally acceptable	Not morally acceptable	Not a moral issue	Morally acceptable	Not morally acceptable	Not a moral issue
IVF	~65–75%	~5–10%	~10–20%	48%	8%	36%
ESPS	~40–55%	~15–30%	~10–20%	40%	20%	28%
Gene-Editing	~50–65%	~15–25%	~10–15%	32%	31%	22%

[a]The survey also included the option "Not sure". In our prompt to ChatGPT, we also asked ChatGPT to predict responses which indicated "Not sure". We have left the option out in favor of a neater presentation, because it does not significantly affect our investigation here

Table 2 ChatGPT's predictions for survey results among Chinese Singaporean and Indian Singaporean attitudes to reproductive procedures

Scenario	Chinese Singaporeans (morally acceptable)	Indian Singaporeans (morally acceptable)
Therapeutic embryo selection	~55–65%	~45–55%
Enhancement embryo selection	~25–45%	~15–35%
Therapeutic gene editing	~65–75%	~50–60%
Enhancement gene editing	~25–40%	~15–30%

data and a patient's medical history (Ferrario et al. 2023). While there may be a lot of ethical issues that arise from the use of such a tool, there is evidence that machine-learning algorithms can match or improve on the performance of alternatives (like consulting next-of-kin) in making accurate predictions (Nagendran et al. 2020; Rid & Wendler 2010). Combining the methods used by Kim et al. and these patient preference predictors, we can imagine individualized models developed using an individual's response to survey questions, their published writing, their online interactions (with informed consent), prior voting patterns (provided voluntarily), and demographic data. From the existing research on individualized algorithms and LLMs, there are good reasons to think that such a model will be quite successful in making predictions about a person's policy preferences.

Both uses of AI—for group-level predictions and predictions about individual preferences—are potentially useful for identifying the preferences of a group of voters. On the one hand, because policymakers are presumably interested in satisfying the interests and preferences of groups of people, group-level predictions offer more straightforward information.

That said, predictions about individual preferences (or small groups) can still be useful where policies are particularly relevant to small groups, like minorities. In particular, there tends to be less data about the preferences of small rural communities and indigenous groups. Current LLMs are good at representing the "centre of mass" of "high-frequency observations", but are not as good at representing potentially important but "low-frequency" content (Lazar 2024). This pattern of representing only high-frequency observations has also been called "generative monoculture" (Wu et al. 2024). That is to say, they can plausibly represent the general views held by many members of a group, but not necessarily as well more nuanced views or views held by minorities. A good AI model would be able to inform policymakers what the most common preferences are *and* what specific individuals or minority groups would prefer. This may involve designing such models, such that they can make multiple predictions about how different salient groups or representative individuals within society might respond to policies, and present those predictions to a policymaker, so that the policymaker can deliberate about the policies based on this information.

We think that the best approach to using AI will combine the best of both worlds. We envision a model comprised of two parts. First, a series of personalized AI models. Each of these models is trained from a common base model to predict the policy preferences of a single individual. In principle, there could be a model for every member of a constituency, although this may be infeasible. In practice, it may be enough for models to be trained on a representative group (the group could either be a cross-sectional group representing society at large, or if relevant to the aims, a more focused group representing the views of a minority group). Ideally, these personalized models will be trained using as much information as developers can ethically attain. If attaining the relevant information is infeasible, the personalized models can be trained on a custom-made survey answered by individual participants. The second part of our envisioned model is an LLM which aggregates and summarizes the results when a question is posed to the first set of models. The final result is a model which could outperform general models like ChatGPT in predicting the preferences of a group, while also providing a policymaker with an overall view of what their constituents might prefer.

2.3 Preference predictions and value predictions

Another distinction worth making is between making particular predictions and inferences about the values or principles underlying the preferences of a group or an individual. When a model makes particular predictions, it makes a prediction about the preferences held by an individual or group regarding specific questions. This can be as simple as asking a model a yes/no question, like "Would a majority of Singaporeans support a policy legalizing moral enhancement?" On the other hand, an algorithmic model might also be used to articulate the underlying principles and values held by a group of people—perhaps even ones that they may not be aware of. Using data gathered from public statements, surveys, and studies, a model might be able to "discover" or infer the underlying moral considerations for the beliefs held by people, and relative weight that people attach to various moral considerations (Kim et al. 2018). For example, consider again the Moral Machine experiment, which had participants make choices on which groups ought to be saved (in a series of choices with two options each). A series of choices can reveal certain values which a respondent considers important, like overall utility, moral desert, or respect for the elderly (Awad et al. 2018). Analyzing the vast amount of information about the choices made by people is no small feat, but AI models are able to draw some inferences about people's values from a large pool of data.

A model which makes particular predictions has one obvious advantage over a model which makes inferences about values—it provides direct feedback to policymakers about what their constituents probably want. So why should policymakers want a model which interprets and articulates underlying values and principles? In part, policymakers may be better equipped to explain, justify, and engage in dialog about their decisions with their constituents if they have access to information about the values and principles held by those constituents. Thus, AI models which provide information about the values of a group of people can help policymakers live up to the ideals of deliberative democracy,

which places importance on "equal and inclusive public deliberation where reasons are offered in support of alternatives, and where citizens can have their preferences informed and possibly transformed through discussion with others" (Benson, forthcoming, p. 3).

In addition, it is sometimes difficult to explain to people the precise details regarding different policy contexts and policy alternatives. Thus, when a policymaker presents people information about their predicted policy preferences, it may be difficult for people to assess such information. For example, suppose that a policymaker has a choice between two policies, X and Y, which differ only in some very technical legal details. If a policymaker told her constituents, "We are going to enact X policy over Y policy, because that is what the AI model says you will prefer", her constituents may have no way of assessing these statements, unless they have access to technical expertise. By explaining how policy X fits the values of her constituents, they may have more context with which to assess the policies. A similar point may be made about policies regarding new or complex technologies. By articulating how one policy may be more consistent with a society's values, a policymaker will be more able to engage with their constituents, and conversely, constituents will be more well equipped to engage with policy.

3 Implementation

Having covered some distinctions in how policymakers can use AI models, we want to consider some specific examples for how policymakers can use AI models to improve outcomes for their constituents. As we noted earlier, we believe that the most useful and most ethically justifiable use of AI are epistemic uses. Therefore, this section will focus on epistemic use cases (with the caveat that using AI for primarily epistemic reasons can sometimes have participatory effects). What we discuss here is not exhaustive, but rather should be read as an illustrative list of use cases for AI preference prediction in policymaking.

3.1 Pre-proposal screening and double-checking

Policymakers can use AI models to "screen" policy proposals prior to tabling them for official discussion. When a policymaker wants to propose a policy, she might run the proposal through an AI model trained to provide predictions on how different groups within her constituency will react. The information she obtains can tell her whether the policy is ready to be proposed publicly, whether it needs some changes, whether further consultation and workshopping is required, or whether it should be shelved altogether. If the AI model indicates that some groups may have concerns about a proposal, the policymaker can take this as a clear sign that she needs to dedicate more resources to working with those groups to assuage their concerns. The AI model thus helps the policymaker to target her efforts at engagement more effectively and efficiently, in turn making it easier for her to meet the needs of her constituents.

In a very similar vein, policymakers can also use AI models to double-check information they have—or think they have. For example, a policymaker may believe that her constituents favor a specific policy based on personal interactions and focus groups. Instead of putting together a policy based on that information alone, she may also consult an AI model which predicts how the wider society may respond to the policy. This sort of double-checking is useful in at least two ways. First, it can either confirm information that the policymaker has or contradict the information, in which case it can motivate the policymaker to investigate the matter further before making any proposals. Second, it can alert the policymaker to gaps in her current sources of information, perhaps by alerting her to the views of groups that were not represented in focus groups and discussions.

Of course, AI models may have their own shortcomings. They may not be entirely accurate, offer vague predictions, and may nonetheless fail to represent all groups in society. However, the performance of an AI model should be compared to how policymakers otherwise gain information about their constituents. Standardly, politicians rely on a number of sources, including direct contact with citizens, traditional media, their confidants, social movements and interest groups, and social media, among other sources (Walgrave and Soontjens 2023). But these sources have their limitations. Sources like the media, confidants, and interest groups are often biased and unrepresentative. And furthermore, some groups are more politically engaged than others, and politicians (like most people) tend to have more contact with select groups of people than others. As a result, the information that politicians have access to may be unrepresentative of their entire constituency. Politicians, just like ordinary people, often experience "filter bubbles" which affect their perception of the world (Bozdag and van den Hoven 2015). Thus, it is important for policymakers to have access to other sources of information regarding the preferences of their constituents, especially groups of people less frequently represented in traditional sources of information. We do not believe that AI models should replace traditional sources of information; it is still very important for policymakers to engage with their constituents in real life. But the use of AI models can help policymakers obtain information which would otherwise be difficult or costly to obtain, which can then be used to inform further steps.

3.2 Collective reflective equilibrium

Reflective equilibrium is a prominent method of moral philosophy, which involves analyzing both our considered moral judgments about particular cases and our deeper moral principles and values. We adjust our moral judgments and principles when cases and principles come into conflict with each other, with the aim of achieving coherence between our judgments and principles. Originally outlined by John Rawls, reflective equilibrium requires an ethicist to use their own judgments, or the judgments of competent judges, as key data points in the method (Knight 2025). Recently, a number of philosophers have suggested that this methodology can be extended to groups of people as *Collective Reflective Equilibrium,* which is not just a method of moral philosophy, but a method for justified policymaking. In Collective Reflective Equilibrium, policymakers consider the judgments of groups of people, including policymakers, ethicists, lawyers, and relevant medical and scientific experts, and lay public and citizens (Savulescu et al. 2021). These judgments are held up against plausible moral theories, principles, and values, and revised until a coherent set of policy conclusions are reached.

Collective Reflective Equilibrium is valuable in two distinct ways. First, it is epistemically valuable. Moral philosophers often rely on their own intuitions regarding cases, like the Trolley Problem, to arrive at conclusions about principles and theories. These intuitions can be affected by bias, emotion, and idiosyncratic features of their own reasoning (Greene 2016). By surveying groups of people and obtaining their judgments, ethicists can gain information about the moral intuitions of others, possibly revealing information about which intuitions and judgments are more or less robust. Of course, groups of people are often subject to biases and faulty reasoning, which can lead to bad judgments. But part of Collective Reflective Equilibrium involves "laundering" survey data by excluding judgments which are clearly the products of bias or contradict important moral principles (Savulescu et al. 2021). Nevertheless, information about the considered judgments of large groups of people can give us information about which policies are more likely to cohere with an equilibrium of judgments and values. Second, Collective Reflective Equilibrium can help justify policy. As Savulescu et al. (2021, p. 656) point out, "In liberal democracies, the legitimacy of a policy depends on whether the public have a role in shaping it". Thus, the considered judgments of a constituency can confer justification on relevant policies.

What role can AI models play in Collective Reflective Equilibrium? One obvious answer is that AI models can predict the judgments people might have about some issue, or make inferences about people's underlying values. These judgments and values can be directly factored into the Collective Reflective Equilibrium process, and used in the same way actual judgments and values would be used in Reflective Equilibrium. This sort of use has epistemic value, especially when it is difficult or costly to obtain the actual judgments of people in surveys or consultations. Moreover, vast numbers are accessible to AI which potentially increases the breadth and depth of prediction. But it seems unlikely that this sort of use, by itself, can justify policy. After all, policies in democratic societies gain legitimacy from the actual actions performed by its members, like their votes, not predictions of how they would vote.

Another role AI models can play in Collective Reflective Equilibrium is in the laundering process. As noted earlier, the judgments of participants should meet certain criteria to "count" as considered judgments. For example, we should not revise our moral principles and values, so that they cohere with clearly racist judgments. AI models can help to launder judgments in a number of ways. For example, a model could generate sets of judgments which are most consistent with a set of pre-established underlying values. A secondary AI model might then be used to compare these generated judgments with actual judgments, filtering out judgments which are most inconsistent with the generated judgments.

In addition, AI models can help compare people's revealed preferences to their stated preferences. These two sets of preferences are not always the same. For example, while a majority of women say they will not terminate a pregnancy if they received a diagnosis of Down's Syndrome, most women do in fact terminate their pregnancies following such a diagnosis (Savulescu et al. 2021). In Collective Reflective Equilibrium, we may therefore have to make decisions between revealed and stated preferences, should they be different. As we noted earlier, algorithms like those used in the Moral Machine experiment can be used to make inferences about the values underlying a set of decisions. In turn, these inferences about values can compared to a set of stated values—perhaps obtained through a survey. At this point, a human decision-maker will have to make a choice about whether to prioritize the revealed preferences of the relevant group or their stated preferences. In some cases, the revealed preferences of a person may be closer to their actual preferences than their stated preferences, and may thus be more likely to provide justification for policy. In other cases, the revealed preferences of a person may constitute some implicit biases we should reject (for example, a person raised in a racist community may make choices which reflect some implicit racism, even though they may disavow the racism). Based on what the decision-maker knows about the context of the preferences made, they will have to make some judgment about whether a group's revealed preferences or stated preferences best represents the considered judgments held by the group as a whole. Once this decision is made, an AI

model may once again be used to help filter out preferences which are deemed to be not representative or inconsistent with the underlying values and principles held by the group.

The use of AI in Collective Reflective Equilibrium can be taken further with the integration of direct democratic participation, which allows AI models and users to adapt to the changing preferences of groups. We envision a process that involves an AI model which undergoes constant reinforcement and updating with human feedback, and a representative citizen's jury which interacts with the model over time. This process has a number of potential benefits. First, it ensures that the AI model is constantly updated and improved to better match the preferences of the represented. Second, it gives people more chances to participate in a democratic process, which in turn has a number of potential benefits, like facilitating the development of civic virtue. Third, it is an iterative process where judgments made by an AI model may be presented to human participants, who make their own judgments which are fed back to the model, and so on. This bears some similarity to the method of reflective equilibrium outlined by Rawls, which involves going back-and-forth between particular judgments and broader principles (Rawls 1971). This iterative method allows theorists to take into account new information which may lead them to revise our judgments. Similarly, an iterative process of presenting AI judgments to human judges, asking them to revise their own judgments in response, and feeding the judgments back to the AI model could result in decisions which are ultimately more coherent and responsive to new information.

"Enlightened Preference" Prediction

In the book, *Debating Democracy: Do We Need More or Less,* Brennan describes an alternative to democratic voting systems which he calls "Enlightened Preference Voting". Roughly, in this system, citizens are asked to state their preferences on election day. In addition, they are asked for their demographic information and are quizzed on basic political information. After this, "The government electoral commission then uses the data to estimate, via predetermined methods, what the public *would have wanted* if it were demographically identical but had gotten a perfect score on the knowledge test" (Brennan 2021, p. 99). Essentially, Brennan's proposal involves making a projection of what people would choose if they were more well informed on political matters—in other words, their "enlightened preferences".

We do not share Brennan's revisionary goal of providing an alternative to democracy. Nevertheless, there is something to be gained from knowing how some groups of people would choose if they had access to more information, time, or experience. Voters are often affected by misinformation or lack of information. For example, Brennan points out that many voters are misinformed about some verifiable facts, like the number of immigrants within a country (Brennan 2021). Or consider again issues which are highly technical or involve new technologies, where people may form opinions based on incorrect or misleading information.[5] This misinformation can lead people to make choices against the interests of themselves and the community—choices they would not have made had they been better informed. As a result, knowing how people would choose, were they better informed, can help policymakers determine which policies are both in the interests of the voters and consistent with their underlying values.

There are reasons to believe that AI tools can make plausible predictions about people's enlightened preferences. One simple method of making such predictions involves determining the preferences most educated and well-informed members of each demographic group. In principle, human agents can make these predictions, but machine-learning tools have the ability to find associations between many more data points (about each person's demographic characteristics), allowing for finer-grained and more accurate predictions. A more complicated, but perhaps more justifiable method, involves asking people questions about their underlying values and priorities, and using this, together with demographic data, to project what a person with their set of values would prefer.

While we do not favor the replacement of actual voting, as Brennan does, these AI-generated enlightened preferences can be used a number of ways. For one, they can be used in Collective Reflective Equilibrium to help launder people's views. By comparing people's enlightened preferences to their expressed preferences, we can determine which expressed preferences are clearly inconsistent with enlightened preferences, which could help weed out preferences which are inconsistent with empirical facts. Another possible use of AI-generated enlightened preferences is in a more deliberative system. We can imagine, for instance, citizen's juries where citizens are invited to state and discuss their preferences. At some point during deliberations, they may be presented with information about their enlightened preferences, and invited to deliberate further on policy with this new set of insights. There can even be some back-and-forth here, where new discussions are held after participants are shown their enlightened preferences, which are in turn fed into the AI tool to generate a new model, producing potentially new results. Finally, enlightened preferences may simply be used as a data point by policymakers, helping

[5] For example, see (Levin 2008), who points out that opinion polls on reproductive technologies fail to capture the fact that large portions of the (American) public are not familiar with those technologies.

them make decisions which are representative of the values held by their constituents.

4 Challenges

We shall now consider challenges to the use of AI in policy making. One theme of many worries is that the use of AI to represent people simply does not work. For example, it is unclear that it is possible "to design LLM representatives that reliably predict and track a principal's judgments" (Lazar and Manuali 2024, p. 11). Related problems include a propensity to make things up and being excessively "lossy" (Lazar and Manuali 2024, p. 3). While work on AI preference prediction is promising, it is still in its infancy. And limited work has been done on predicting the political or moral preferences of individuals. Predicting the preferences of groups may be a bit easier, but AI models tend to create output which is representative of high-frequency content, meaning that minority views and nuanced views tend to be left out (Lazar and Manuali 2024).

However, given the rate of research and improvement in this field, it is prudent to treat objections along this theme as tentative. The reliability and accuracy of AI preference prediction is likely to improve significantly within a matter of years. In fact, some recent developments like the Delphi experiment (Jiang et al. 2025) show that AI systems can be trained to make moral judgments to a relatively impressive degree, achieving an accuracy of 92.8% in making moral judgments (measured against the Norm Bank, a textbook of crowdsourced moral judgments). Furthermore, when we assess the reliability of LLM representatives, we should compare them to human representatives and politicians. As we noted at the beginning of this paper, human representatives have at best a mixed record at representing their constituents. And in representative democracies, a majority or plurality is typically sufficient to elect a representative, which means that the representative may not represent the preferences of a minority within their electoral districts. Thus, given the right use cases, even somewhat unreliable AI representatives may still offer some potential improvements over the status quo.

Lazar and Manuali (2024, p. 9) point out (plausibly) that "we should expect a higher standard from a computational means of democratic information processing than from humans" (Lazar and Manuali 2024). One reason (among a few, although this seems the most important) they provide for this claim is that humans can be held accountable for their errors, while AI models cannot (Lazar and Manuali 2024). This seems correct, but users of AI models can be held responsible for bad outcomes if they are found to have used insufficiently reliable AI models, or if they used AI models without the right precautions. Furthermore, institutions can adopt "strict liability" norms and regulations for the use of AI, such that (under the right conditions) AI users will be expected to take responsibility and account for bad outcomes even when the AI functions as expected. Finally, Lazar and Manuali's claim is compatible with the conclusion that we should use AI models to support policymaking, as long as those models outperform their human counterparts by a sufficiently large margin. We need more evidence to determine whether AI models meet this standard. But because human representatives have some serious flaws, there are good reasons to suspect that AI models can meet the standard sometimes.

A second theme of potential objections and worries focuses on the fundamental flaws which we know LLMs and generative AI tend to have. These include a tendency to "hallucinate" or make things up, (Zhang et al. 2023) introducing "subtle but significant" changes in meaning, (Lazar and Manuali 2024, p. 9) tendencies to overrepresent or only represent high-frequency content, (Lazar and Manuali 2024) and replicating biases (Liu et al. 2022). Finally, people worry about the opacity of many AI models.

Let us start with the problem of hallucination, and the introduction of changes in meaning. Because the role of a model here is to make predictions rather than represent existing data, it cannot hallucinate (any prediction involves making things up to some extent). Rather the question is how closely the predictions match the choices people would make. To some extent, this sort of accuracy is something that can be tested (by comparing the model's predictions to actual polls) and improved over time. Similarly, because AI models are not used to represent existing data, no issue arises regarding changes in meaning.

Next, the problem of bias. It is worth disambiguating between two kinds of bias. One kind of bias is that which arises from the replication of human bias from training data. For example, if a group of people have biased views, then any LLM trained on their stated views will also be biased. The fact that AI models replicate biases and overrepresent high-frequency content is a potential risk, but the risks can be mitigated. If user is aware that a given output may replicate biases or may be unrepresentative, she need not be misinformed by the model. In fact, it can often be useful for policymakers to be made aware of sectarian opinions or common biases held by members of their constituencies. Knowledge about such opinions and policies may give policymakers the tools to navigate around certain issues more deftly, or engage their constituents in terms they will more readily understand and accept. As an analogy, surveys can and often do represent the biases held by a population; this does not mean that surveys are useless. It is instead important to design AI models in a way such that they explicitly inform their users that their predictions can replicate biases by amplifying commonly made views and claims.

Another kind of bias that can occur in machine-learning involves other issues in the way data are collected and generated. This kind of bias can include "specification bias"—"bias in the choices and specifications of what constitutes the input and output in a learning task", "sampling bias", which occurs when there is disproportionate representation from segments of a population, and "annotator bias", which refers to biases that may be transferred from human annotators to a machine-learning model (Hellström et al. 2020). These kinds of biases are potential issues, because an AI tool can misrepresent the preferences of a group of people. Nevertheless, the use of AI tools cannot be ruled out at this theoretical stage. It is possible to "audit" an AI tool or algorithm to assess how biased it is (Bandy 2021; Metaxa et al. 2021). And even if an AI tool is not entirely accurate, it may still outperform or match other means of obtaining information about preferences. This includes tools like surveys, which can also be biased. In practice, it may be best for a policymaker to use multiple tools, including both AI and traditional tools, to understand the preferences of their constituents, as a diverse set of sources may provide them with the most accurate picture.

Let us now discuss the issue of opacity: the property of being difficult or impossible for users or operators to understand. In machine-learning, one major concern is opacity that arises because of a mismatch between "mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation" (Burrell 2016, p. 2). When an AI tool is opaque in such a way, even the developers of the tool may not fully understand how it turns specific inputs into outputs. Opacity in machine-learning creates a few concerns. For one, opacity means that an AI tool may have some biases which we are unaware of, or produce results for reasons that the represented user does not have. This can lead to the AI tool making predictions that are simply incorrect (Dowding and Taylor 2024). This concern is in principle mitigable; as noted earlier, we can audit an algorithm "from the outside" to determine how accurate or biased it is. This can be done, for example, by testing the results provided by the algorithm against a known data set.

Another set of concerns about opacity in machine learning has to do with justification. Many believe that in order for an agent to justify their decisions, they must be able to explain how they arrived at those decisions. Because we do not know how opaque AI tools translate inputs to outputs, a policymaker who uses such a tool cannot justify decisions based on those tools. Relatedly, we might worry that a policymaker who does not justify their decisions cannot be held accountable, or that their decisions will be illegitimate (Lazar 2023). This issue of opacity is much larger than we can do justice to here. That said, while we think these are legitimate concerns, they are not necessarily reasons to reject the use of opaque AI tools completely. There is a difference between why a decision was made, and the justification for that decision. A justification refers to our objective reasons to believe or accept a decision. Sometimes, knowing how a decision was made can give us reasons to accept it. But even when we do not know why an AI made a decision, we can sometimes still determine if the decision is justified (Muralidharan et al. 2024). For example, suppose that a policymaker uses an LLM which tells her that her constituents are likely to support policy X over policy Y. Even if she does not know how the LLM works, there are still ways for her to obtain information about whether this conclusion is justified. She might check the historical accuracy of the LLM, consult surveys which corroborate the conclusion, or rely on her knowledge about which policy is more consistent with the values of her constituents. Together, these sources of information can give her reasons to believe (or disbelieve) the conclusion. In turn, she can relay this information to her constituents, giving them reasons to accept the conclusion or not.[6]

A third theme of worries are about how the use of AI models in democratic decision-making can result in undermining the "participatory reasons to value democracy" (Lazar and Manuali 2024, p. 11). That is to say, there are a number of goods that societies can gain when people actually participate in democratic decision-making. These include the cultivation of democratic virtue and character (Christiano and Bajaj 2024), conferral legitimacy or authority on representatives, public justification (Christiano and Bajaj 2024), the expression of equality (Singer 1973), and the transformative effect of participating in political dialogs (engaging with others can help us recognize their values and give us reasons to change our own preferences) (Lazar and Manuali 2024). The thought is that when people participate in democratic processes like voting, their societies benefit in ways that they would not otherwise because of these goods. As a result, any attempt to replace these processes or automate them will result in the loss of something distinctly valuable.

We agree that AI models should not be used in place of existing democratic procedures like voting. Policymakers and representatives should continue to consult their constituents in dialog when making important policy proposals. Such consultation is particularly important with respect to minority groups and groups with minority opinions, because their voices are unlikely to be represented by a majority vote. Instead, what we have discussed is the use of AI models,

[6] And conversely, constituents can hold policymakers to account if they fail to provide adequate justification—for example, if they rely on an opaque AI tool without providing any other justification for their decisions.

Table 3 A summary of acceptable uses of AI, and some constraints and limitations

Acceptable uses of AI	Constraints and limitations
Epistemic uses—i.e., policymakers using AI to learn more about the preferences of their constituents	Representative uses—AI should not be used to replace democratic procedures like voting
Pre-proposal screening	Biases—users should be aware that AI systems can replicate biases held by groups of people
Double-checking	Replacing engagement—in using AI, users should not ignore the need to engage their constituents
Collective reflective equilibrium—using AI to generate data points	
Collective Reflective Equilibrium—view laundering	
Enlightened preference prediction—helping users understand what people would want under ideal conditions	

including LLMs, to remediate existing shortcomings in the way policymakers currently seek information regarding the preferences of their constituents. It can be costly or infeasible to get information about certain groups of people, like rural communities. It may also not be possible or feasible to more directly obtain large numbers of preferences and analyze them, say in an emergency or crisis. Or it may be difficult to ask people for their preferences with respect to new or highly technical issues, like new genetic editing technologies. These limitations can be exacerbated by biases held by the policymakers, overconfidence, false beliefs, and sheer laziness. While AI models will not automatically make these limitations disappear, they can make it easier for policymakers to get an initial sense of how their constituents and particular groups may feel about a policy or state of affairs. In turn, the policymakers may be better equipped to engage with those groups in person, or be alerted to issues and considerations they may not have thought of in the first place.

This sort of use of AI models—for epistemic purposes—comes with some risk. If policymakers have easy access to information about certain groups via AI models, they may see less of a need to engage with those groups in person. Given pressures to cut costs and reduce workloads, politicians may be motivated to replace actual engagement with people with the use of AI models. This is not the outcome we want. Nevertheless, there is some natural pressure against such an outcome occurring. People want to raise their opinions and concerns, and want to be heard. Even when most people do not participate actively in town halls or council meetings, we look for other ways to express our opinions, like on social media. Thus, it seems somewhat unlikely that people will simply accept a status quo where policymakers use AI models at the exclusion of engagement with real people. Furthermore, there are measures that societies can take to reduce the risk of policymakers replacing engagement with using AI models. Government institutions can introduce and maintain formal requirements for the passing of policies and regulations. For example, the Government of Canada has codified a "duty to consult" Indigenous groups when it considers conduct which "might adversely impact potential or established Aboriginal or treaty rights" (Canada 2012). While such duties do not guarantee that policymakers act in ways which respect the claims of such groups, they can at least ensure that policymakers sit down with representatives of those groups and listen to them. This could assuage concerns that access to AI preference predictors might cause policymakers to forego any engagement altogether.

One final worry we want to address is the possibility that policymakers become overly sensitive to the preferences of voters. Some theorists hold that policymakers should act as "trustees" on behalf of the wider community, doing their best to promote the interests of the community (Dovi 2018). Policymakers may sometimes have to act against their preferences of their voters if doing so is in the interests of the wider community. This is especially important if there are reasons to believe that members of a constituency hold poorly informed or biased views. If these theorists are correct, we would want policymakers to have some discretion to implement what they believe is best, and we do not want them to be oversensitive to the views of their constituents. That said, there are ways to use AI prediction tools in ways which do not merely present preferences to policymakers. As we have discussed, AI tools can be used to articulate the values underlying people's views, generate preferences which are most consistent with those underlying values, and to launder preferences and views which are based on misinformation or bias. Thus, the risk of policymakers becoming oversensitive to voter preferences is a function of how they use the AI tools. Merely using any AI tool will not necessarily result in this sort of oversensitivity (Table 3).

5 Conclusion

There are many ways in which AI models and LLMs can influence policy, policymakers, and democratic procedure. Some of these ways are benign and beneficial, and some of them potentially harmful. In this paper, we have discussed one particular way in which AI models can be used to augment policymaking in democratic societies. They can be used by policymakers to gain a better understanding of the preferences and values held by their constituents, thus

enabling them to be better representatives. We have also clarified that this use of AI models should not take the place of the ways in which people participate in democratic decision-making, like in voting or other consultations. While AI models are far from perfect, we believe that they can improve on the current status quo where representatives do not seem to understand the wants and needs of their voters.

Acknowledgements The authors would like to thank Sinead Prince for comments on an earlier version of this paper, as well as Casey Haining for contributing ideas for the paper. The authors thank the anonymous referees also for providing comments. In addition, Seth Lazar and Lorenzo Manuali, for a related discussion at a workshop, which contributed ideas instrumental to the writing of this paper.

Author contributions J.E.L. wrote the main manuscript text, with J.S. contributing substantially to both arguments and writing. All authors reviewed the manuscript.

Funding This work was funded by NRF/AISG Governance, AISG3-GV-2023-012, AISG3-GV-2023-012, NUS/Start up, NUHSRO/2022/078/Startup/13, NUHSRO/2022/078/Startup/13, Wellcome Discovery, under Grant Nos. 226801/Z/22/Z and 226801/Z/22/Z.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest JS is a Bioethics Committee consultant for Bayer. JS is a Bioethics Advisor to the Hevolution Foundation.

References

Ågren H, Dahlberg M, Mörk E (2007) Do politicians' preferences correspond to those of the voters? An investigation of political representation. Public Choice 130(1):137–162. https://doi.org/10.1007/s11127-006-9077-1

Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon J-F, Rahwan I (2018) The moral machine experiment. Nature 563(7729):59–64. https://doi.org/10.1038/s41586-018-0637-6

Bakker MA, Chadwick MJ, Sheahan HR, Tessler MH, Campbell-Gillingham L, Balaguer J, McAleese N, Glaese A, Aslanides J, Botvinick MM, Summerfield C (2022) Fine-tuning language models to find agreement among humans with diverse preferences. Adv Neural Inform Processing Syst 35:38176–38189

Bandy J (2021) Problematic machine behavior: a systematic literature review of algorithm audits. Proceed Acm Human-Comput Interaction 5(1):1–34

Benson J (forthcoming). Is fake news a threat to deliberative democracy? Partisanship, Inattentiveness, and Deliberative Capacities. Social Theory and Practice.

Biller-Andorno N, Biller A (2019) Algorithm-aided prediction of patient preferences—An ethics sneak peek. N Engl J Med 381(15):1480–1485. https://doi.org/10.1056/NEJMms1904869

Bozdag E, van den Hoven J (2015) Breaking the filter bubble: democracy and design. Ethics Inf Technol 17(4):249–265. https://doi.org/10.1007/s10676-015-9380-y

Brennan, J. (2021). Alternatives to Democracy. In J. Brennan & H. Landemore (Eds.), *Debating Democracy: Do We Need More or Less?* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780197540817.003.0005

Broockman D. E., & Skovron C. (2013) What Politicians Believe About Their Constituents: Asymmetric Misperceptions and Prospects for Constituency Control. https://www.vanderbilt.edu/csdi/miller-stokes/08_MillerStokes_BroockmanSkovron.pdf. Accessed June 2025

Broockman DE, Skovron C (2018) Bias in perceptions of public opinion among political elites. Am Political Sci Review 112(3):542–563. https://doi.org/10.1017/S0003055418000011

Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data Soc 3(1):2053951715622512. https://doi.org/10.1177/2053951715622512

Cai C. (2001) Personal preference prediction [University of Guelph]. https://hdl.handle.net/10214/20181

Canada G. of C. C.-I. R. and N. A. (2012). Government of Canada and the duty to consult [Administrative page]. Government of Canada. https://www.rcaanc-cirnac.gc.ca/eng/1331832510888/1609421255810. Accessed 8 July 2025

Christiano T., & Bajaj S. (2024) Democracy. In E. N. Zalta & U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy (Summer 2024). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2024/entries/democracy/. Accessed 8 July 2025

Confessore, N. (2018, April 4). Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. The New York Times. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html. Accessed 8 July 2025

Devine F., Krasodomski-Jones A., Miller C., Shu Y. L., Cui J.-W. 'Peter' C., Marnette B., & Wilkinson R. (2023) Recursive Public: Piloting Connected Democratic Engagement with AI Governance. Recursive Republic. https://vtaiwan-openai-2023.vercel.app/Report_%20Recursive%20Public.pdf. Accessed 8 July 2025

Dovi S. (2018) Political Representation. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/political-representation/. Accessed 8 July 2025

Dowding K, Taylor BR (2024) Algorithmic decision-making, agency costs, and institution-based trust. Philosophy Technol 37(2):1–22. https://doi.org/10.1007/s13347-024-00757-5

Earp BD, Porsdam Mann S, Allen J, Salloch S, Suren V, Jongsma K, Braun M, Wilkinson D, Sinnott-Armstrong W, Rid A, Wendler D, Savulescu J (2024) A personalized patient preference predictor for substituted judgments in healthcare: technically feasible and ethically desirable. Am J Bioeth 24(7):13–26. https://doi.org/10.1080/15265161.2023.2296402

Edelman, J. (2023). Democratic Fine-Tuning. https://www.lesswrong.com/posts/ncb2ycEB3ymNqzs93/democratic-fine-tuning. Accessed 8 July 2025

Ferrario A, Gloeckler S, Biller-Andorno N (2023) Ethics of the algorithmic prediction of goal of care preferences: From theory to practice. J Med Ethics 49(3):165–174. https://doi.org/10.1136/jme-2022-108371

Greene JD (2016) Solving the Trolley Problem. In: Buckwalter W, Sytsma J (eds) Blackwell Companion to Experimental Philosophy. Blackwell, pp 173–189

Griffin JD, Newman B (2005) Are Voters Better Represented? The Journal of Politics 67(4):1206–1227. https://doi.org/10.1111/j.1468-2508.2005.00357.x

Gudiño-Rosero J., Grandi U., & Hidalgo C. A. (2024) Large Language Models (LLMs) as Agents for Augmented Democracy. Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences. https://doi.org/10.1098/rsta

Haining CM, Toh HJ, Savulescu J, Schaefer GO (2025) Singaporean attitudes to cognitive enhancement: a cross-sectional survey. J Med Ethics. https://doi.org/10.1136/jme-2024-110490

Hedlund RD, Friesema HP (1972) Representatives' perceptions of constituency opinion. J Politics 34(3):730–752. https://doi.org/10.2307/2129280

Hellström T., Dignum V., & Bensch S. (2020) Bias in Machine Learning—What is it Good for? (No. arXiv:2004.00686). arXiv. https://doi.org/10.48550/arXiv.2004.00686

Huang S., Siddarth D., Lovitt L., Liao T. I., Durmus E., Tamkin A. & Ganguli D. (2024) Collective Constitutional AI: Aligning a Language Model with Public Input. The 2024 ACM Conference on Fairness, Accountability, and Transparency, 1395–1417. https://doi.org/10.1145/3630106.3658979

Jackson K (2023) Administration as democratic trustee representation. Leg Theory 29(4):314–348. https://doi.org/10.1017/S1352325223000204

Jenke L, Huettel SA (2020) Voter preferences reflect a competition between policy and identity. Front Psychol 11:566020

Jiang L, Hwang JD, Bhagavatula C, Bras RL, Liang JT, Levine S, Dodge J, Sakaguchi K, Forbes M, Hessel J, Borchardt J, Sorensen T, Gabriel S, Tsvetkov Y, Etzioni O, Sap M, Rini R, Choi Y (2025) Investigating machine moral judgement through the Delphi experiment. Nature Mach Intell 7(1):145–160. https://doi.org/10.1038/s42256-024-00969-6

Kain E. (2024). Do People Still Not Realize That "Helldivers 2" Is A Parody Of Authoritarianism? [Substack newsletter]. Diabolical. https://diabolical.substack.com/p/do-people-still-not-realize-that. Accessed 8 July 2025

Kim R., Kleiman-Weiner M., Abeliuk A., Awad E., Dsouza S., Tenenbaum J., & Rahwan, I. (2018) A Computational Model of Commonsense Moral Decision Making (No. arXiv:1801.04346). arXiv. https://doi.org/10.48550/arXiv.1801.04346

Klar S (2013) The influence of competing identity primes on political preferences. J Politics 75(4):1108–1124. https://doi.org/10.1017/S0022381613000698

Knight C. (2025) Reflective Equilibrium. In E. N. Zalta & U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy (Spring 2025). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2025/entries/reflective-equilibrium/

Konya A., Schirch L., Irwin C., & Ovadya A. (2023) Democratic Policy Development using Collective Dialogues and AI (No. arXiv:2311.02242). arXiv. https://doi.org/10.48550/arXiv.2311.02242

Landmore H. (2023) Fostering More Inclusive Democracy with AI by Landemore. IMF. https://www.imf.org/en/Publications/fandd/issues/2023/12/POV-Fostering-more-inclusive-democracy-with-AI-Landemore

Lazar S., & Manuali L. (2024) Can LLMs advance democratic values? (No. arXiv:2410.08418). arXiv. https://doi.org/10.48550/arXiv.2410.08418

Lazar S. (2023) Legitimacy, Authority, and Democratic Duties of Explanation (No. arXiv:2208.08628). arXiv. https://doi.org/10.48550/arXiv.2208.08628

Lazar S. (2024) Can we really trust AI to channel the public's voice for ministers? The Guardian. https://www.theguardian.com/commentisfree/2024/apr/25/ai-public-voice-ministers-large-language-model-chatgpt

Levin Y (2008) Public opinion and the embryo debates. The New Atlantis 20:47–62

Liu R, Jia C, Wei J, Xu G, Vosoughi S (2022) Quantifying and alleviating political bias in language models. Artif Intell 304:103654. https://doi.org/10.1016/j.artint.2021.103654

Ma Y, Sun H, Chen Y, Zhang J, Xu Y, Wang X, Hui P (2021) DeepPredict: a zone preference prediction system for online lodging platforms. J Soc Comput 2(1):52–70

Maroto-Gómez M, Castro-González Á, Castillo JC, Malfaz M, Salichs MÁ (2023) An adaptive decision-making system supported on user preference predictions for human–robot interactive communication. User Model User-Adap Inter 33(2):359–403. https://doi.org/10.1007/s11257-022-09321-2

Mendoza G. B. (2023) Can we use AI to enrich democratic consultations? RAPPLER. https://www.rappler.com/technology/features/generative-ai-use-enriching-democratic-consultations/

Metaxa D, Park JS, Robertson RE, Karahalios K, Wilson C, Hancock J, Sandvig C (2021) Auditing algorithms: understanding algorithmic systems from the outside in. Foundations Trends ® Human-Comput Interaction 14(4):272–344

Muralidharan A, Savulescu J, Schaefer GO (2024) AI and the need for justification (to the patient). Ethics Inf Technol 26(1):16. https://doi.org/10.1007/s10676-024-09754-w

Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M (2020) Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ 368:m689. https://doi.org/10.1136/bmj.m689

Nordin M (2014) Do voters vote in line with their policy preferences?—The role of information. Cesifo Econom Studies 60(4):681–721. https://doi.org/10.1093/cesifo/ifu012

Porsdam Mann S, Earp BD, Møller N, Vynn S, Savulescu J (2023) AUTOGEN: a personalized large language model for academic enhancement—Ethics and proof of principle. Am J Bioeth 23(10):28–41. https://doi.org/10.1080/15265161.2023.2233356

Rawls J. (1971) A Theory of Justice: Original Edition. Harvard University Press. https://doi.org/10.2307/j.ctvjf9z6v

Rid A, Wendler D (2010) Can we improve treatment decision-making for incapacitated patients? Hastings Cent Rep 40(5):36–45. https://doi.org/10.1353/hcr.2010.0001

Rubel A, Castro C, Pham A (2019) Agency laundering and information technologies. Ethical Theory Moral Pract 22(4):1017–1041. https://doi.org/10.1007/s10677-019-10030-w

Savulescu J, Gyngell C, Kahane G (2021) Collective reflective equilibrium in practice (CREP) and controversial novel technologies. Bioethics 35(7):652–663. https://doi.org/10.1111/bioe.12869

Sevenans J, Walgrave S, Jansen A, Soontjens K, Bailer S, Brack N, Breunig C, Helfer L, Loewen P, Pilet J-B, Sheffer L, Varone F, Vliegenthart R (2023) Projection in politicians' perceptions of public opinion. Polit Psychol 44(6):1259–1279. https://doi.org/10.1111/pops.12900

Singer P (1973) *Democracy and disobedience* (First Edition). Clarendon Press

Small C. T., Vendrov I., Durmus E., Homaei H., Barry E., Cornebise J., Suzman T., Ganguli, D., & Megill, C. (2023) Opportunities and Risks of LLMs for Scalable Deliberation with Polis (No. arXiv:2306.11932). arXiv. https://doi.org/10.48550/arXiv.2306.11932

Soontjens K (2022a) Do politicians anticipate voter control? A comparative study of representatives' accountability beliefs. Eur J Polit Res 61(3):699–717. https://doi.org/10.1111/1475-6765.12481

Soontjens K (2022b) Inside the party's mind: Why and how parties are strategically unresponsive to their voters' preferences. Acta Politica 57(4):731–752. https://doi.org/10.1057/s41269-021-00220-9

Soontjens K, Sevenans J (2022) Electoral incentives make politicians respond to voter preferences: Evidence from a survey experiment with members of Parliament in Belgium. Soc Sci Q 103(5):1125–1139. https://doi.org/10.1111/ssqu.13186

Sunstein CR (1991) Preferences and Politics. Philos Public Aff 20(1):3–34

Topal İ, Uçar MK (2019) Hybrid artificial intelligence based automatic determination of travel preferences of chinese tourists. IEEE Access 7:162530–162548

Walgrave S, Soontjens K (2023) How politicians learn about public opinion. Res Politics 10(3):20531680231200692. https://doi.org/10.1177/20531680231200692

Walgrave S, Jansen A, Sevenans J, Soontjens K, Pilet J-B, Brack N, Varone F, Helfer L, Vliegenthart R, van der Meer T, Breunig C, Bailer S, Sheffer L, Loewen PJ (2023) Inaccurate politicians: elected representatives' estimations of public opinion in four countries. J Politics 85(1):209–222. https://doi.org/10.1086/722042

What is a large language model (LLM)*?* (n.d.). Retrieved September 6, 2024, from https://www.cloudflare.com/learning/ai/what-is-large-language-model/

Woźniak S., Koptyra B., Janz A., Kazienko P., & Kocoń J. (2024) Personalized Large Language Models. arXiv.Org. https://arxiv.org/abs/2402.09269v1. Accessed 8 July 2025

Wu F., Black E., & Chandrasekaran V. (2024) Generative Monoculture in Large Language Models. arXiv.Org. https://arxiv.org/abs/2407.02209v1. Accessed 8 July 2025

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models (No. arXiv:2309.01219). arXiv. https://doi.org/10.48550/arXiv.2309.01219. Accessed 8 July 2025