Generative Social Choice

Sara Fish[1], **Paul Gölz**[2], **David C. Parkes**[1], **Ariel D. Procaccia**[1], **Gili Rusak**[1], **Itai Shapira**[1], and **Manuel Wüthrich**[1]

[1]Harvard University [2]UC Berkeley and Cornell University

Abstract

The mathematical study of voting, *social choice theory*, has traditionally only been applicable to choices among a few predetermined alternatives, but not to open-ended decisions such as collectively selecting a textual statement. We introduce *generative social choice*, a design methodology for open-ended democratic processes that combines the rigor of social choice theory with the capability of large language models to generate text and extrapolate preferences. Our framework divides the design of AI-augmented democratic processes into two components: first, proving that the process satisfies representation guarantees when given access to oracle queries; second, empirically validating that these queries can be approximately implemented using a large language model. We apply this framework to the problem of summarizing free-form opinions into a proportionally representative slate of opinion statements; specifically, we develop a democratic process with representation guarantees and use this process to represent the opinions of participants in a survey about chatbot personalization. In a trial with 100 representative US residents, we find that 93 out of 100 participants feel "mostly" or "perfectly" represented by the slate of five statements we extracted.

Significance Statement

Presently, there is a vigorous drive to integrate artificial intelligence (AI) into democratic participation tools. This integration presents great opportunities like allowing participants to express their preferences in natural language, and higher scalability due to a reduced need for human facilitation. But the integration of AI also raises concerns about the legitimacy of such democratic processes, which our design framework addresses in two ways: First, the AI's evaluation can be broken down into testing its performance on precisely defined subtasks and can rely on mathematical guarantees for the process overall. Second, the normative desirability of an AI-enhanced democratic process can be analyzed using the techniques of social choice theory. Thus, our work helps unlock the potential of AI-driven democratic innovation.

1 Introduction

Voting is a key way in which groups — be they national electorates, members of a legislature, or members of a board — make common decisions. The theoretical foundation of voting is provided by the field of *social choice theory*, which studies mathematical guarantees in the context of how different voting rules aggregate individual preferences into a collective decision. The typical social choice setting involves a *small, predetermined set of alternatives* (e.g., the candidates in an election), over which voters specify their preferences and from which the voting rule selects an outcome.

Many pressing policy questions, however, are too nuanced to fit this neat template of choosing between a few alternatives. The need for open-ended forms of democratic input is demonstrated, for example, by the increased use of *deliberative minipublics* [13, 28], which provide policy recommendations to governments on complex issues such as climate change [40] and electoral reform [14].

Similarly nuanced questions arise around the alignment of artificial intelligence (AI) with societal interests; in this context, Meta [9] and OpenAI [12] have been experimenting with democratic processes that seek public input to open-ended questions such as "how far [...] personalization of AI assistants like ChatGPT to align with a user's tastes and preferences should go?" [41]. Though deliberation can address such open-ended questions, it lacks two key strengths of voting: scalability [e.g., 16] and guarantees on its outcomes.

To address these shortcomings, we introduce a new paradigm for the design of democratic processes: *generative social choice*. It fuses the rigor of social choice theory with the flexibility and power of generative AI, in particular large language models (LLMs), to reach collective answers to open-ended questions in a scalable and principled way.

1.1 How LLMs Address the Limitations of Classical Social Choice

In our view, there are two fundamental obstacles to applying classical social choice to open-ended questions, both of which can be overcome by LLMs.

- *Unforeseen Alternatives.* In classical social choice, the set of alternatives is explicitly specified and static. Take the 2016 Brexit referendum, for example, in which the alternatives were either to maintain the status quo or make a clean break with the European Union. Since intermediate options were not specified, they could not be selected by voters, even if they might have enjoyed a much larger degree of support. Even in participatory budgeting [8], the set of alternatives is limited to the budget-feasible subsets of *previously proposed* projects.
  By contrast, LLMs have the capability of *generating* alternatives that were not initially anticipated but find common ground between agents. In principle, the possible outcomes of an LLM-augmented democratic process may span the universe of all relevant outcomes for the problem at hand, e.g., all possible bills or statements.
- *Extrapolating Preferences.* In classical social choice theory, voters specify their preferences in a rigid format. Typically, agents evaluate each alternative independently, or, if the alternatives form a combinatorial domain,[1] a voting rule might assume that preferences have a restricted parametric shape and only elicit its parameters. Clearly, this approach does not suffice if a democratic process may produce alternatives that were not previously anticipated, and therefore not elicited: to even know which alternatives would be promising to generate, the process must be able to extrapolate agents' preferences.
  LLMs can address this problem as they enable participants to *implicitly* specify their preferences by expressing their opinions, values, or criteria in natural language. The LLM can act as a proxy for the participant, predicting their preferences over any alternative, whether foreseen or newly generated.

1.2 A Framework for Generative Social Choice

It is clear, at this point, what LLMs can contribute to social choice. LLMs and social choice theory make an odd couple, however, because social choice focuses on rigorous guarantees whereas LLMs are notoriously impervious to theoretical analysis. We propose a framework for generative social choice that addresses this difficulty by breaking the design of democratic processes into two interacting components.

[1] This is the case, for example, in multi-winner elections or participatory budgeting.

- *First component: Guarantees with perfect queries.* Assume that the LLM is an oracle that can precisely answer certain types of queries, which may involve generating new alternatives in an optimal way or predicting agents' preferences. Once appropriate queries have been identified, the task is to design algorithms that, when given access to an oracle for these queries, provide social choice guarantees.
- *Second component: Empirical validation of queries.* Assuming the LLM to be a perfect oracle is helpful for guiding the design of a democratic process, but of course not an accurate reflection of reality. In the second component, the task is to implement the proposed queries using calls to an LLM, and to empirically validate how well these implementations match the queries.

Naturally, the two components interact: The theory identifies queries that are useful for social choice and should hence be validated empirically. Conversely, experiments show which queries can be answered accurately in practice, raising the question of which guarantees algorithms relying on these queries might provide.

A key benefit of this framework is that theoretical results derived in it are future-proof: as LLMs continue to rapidly improve, they will only grow more reliable in answering queries, making the LLM-based aggregation methods ever more powerful.

1.3 Our Results: A Case Study in Generative Social Choice

In addition to introducing the framework presented above, we demonstrate it in one particular setting: summarizing a large body of free-form opinions into a slate of few statements, in a representative manner. In this setting, participants share free-form opinions about a given policy issue on an online platform such as *Polis* [35] or *Remesh*,[2] or as part of a qualitative survey. Then, a voting rule selects a slate of k statements that is proportionally representative of the diversity and relative prevalence of viewpoints among the participant population.

The setting of statement selection was formalized by Halpern et al. [17] in the language of multi-winner approval elections: If we think of statements as candidates, and of an agreement between participants and statements as binary approval votes, the slate should satisfy axioms for proportional representation from this literature such as *justified representation* (JR). In our work, we allow cardinal (rather than just binary) levels of participant–statement agreement. Furthermore, we introduce a novel strengthening of JR, *balanced justified representation (BJR)*, which we believe to be particularly well suited for our statement-selection application and of independent interest.

Whereas previous summarization systems can only select a slate among users' statements, our process can *generate new statements*, which might find new common ground between participants and allow for more representative slates. Our process takes as input each user's interactions on the platform as a description of their preferences. The process then employs an LLM to *(1)* translate these descriptions into participants' utilities for any new statements (*discriminative* queries, in the language of machine learning), and *(2)* generate statements that maximize the utility of a subset of participants, based on their descriptions (*generative* queries).

Following our framework's first component, we show that, with access to polynomially many of these queries, a democratic process resembling *Greedy Approval Voting* [1] guarantees BJR. Crucially, this guarantee holds not just relative to a set of predetermined statements but to the space of all possible statements (Section 3.1).

A potential issue with this process is that, through the generative query, it calls the LLM with a prompt whose length scales linearly in the number of participants. This is problematic since LLMs can only handle input of bounded length. We show that, unless one makes assumptions

[2] https://www.remesh.ai/

on the structure of preferences, this problem of linear-size queries is unavoidable for any process guaranteeing BJR with subexponentially many queries (Section 3.2). If, however, the space of statements and preferences is structured, specifically, if it has finite VC dimension [37], democratic processes based on sampling can guarantee BJR (with high probability) using a polynomial number of queries whose length is independent of the number of participants (Section 3.3).

In Section 4, we present a practical, LLM-based implementation of discriminative and generative queries. Empirical validation shows that the proposed implementation of the discriminative query accurately extrapolates agents' preferences to unseen statements. Further, we show that the proposed implementation of the generative query consistently produces high-agreement statements by leveraging the complementarity of different LLM-based generation methods.

Equipped with these query implementations, we then deploy the full democratic process in Section 5. As part of OpenAI's grant program for *Democratic Inputs to AI* [12], we pilot our process to study US residents' opinions on the extent to which chatbots should be personalized. We elicit free-text opinions about this topic from a sample of 100 participants and distill them into a representative slate of five statements, using our LLM-enhanced democratic process. To validate that these statements faithfully represent the population, we conduct a second survey with a fresh sample of 100 US residents. After matching the five statements to equal-sized blocs of participants, 93% of participants indicate that their assigned statement in our slate captures their opinion on chatbot personalization "mostly" (18%) or "perfectly" (75%). To support future research on online participation, we made the participants' full responses available as a public data set.[3]

1.4 Related Work

In a recent position paper that is independent of our work, Small et al. [36] discuss the opportunities and risks of LLMs in the context of Polis. The opportunities they identify include topic modeling, summarization, moderation, comment routing, identifying consensus, and vote prediction. Most relevant to us are their experiments for the vote prediction task, which are closely related to our implementation and evaluation of discriminative queries. In the future, our democratic process as a whole could serve in the summarization role envisioned by Small et al. [36], for which they do not propose specific algorithms and perform no systematic experiments.

Our discriminative queries using LLMs are also related to work by Konya et al. [19], who integrate an LLM with a latent factor model to predict preferences. More broadly, the paradigm of *virtual democracy* facilitates automated decisions on ethical dilemmas by learning the preferences of stakeholders and, at runtime, predicting their preferences over the current alternatives and aggregating the predicted preferences; example papers, which employ classical machine-learning algorithms, apply the paradigm to domains such as autonomous vehicles [27], food rescue [23], and kidney exchange [15]. These papers all aim to predict preferences on a fixed set of alternatives — they do not generate new alternatives.

A source of inspiration for our work is the paper of Bakker et al. [2]. They fine-tune an LLM to generate a single consensus statement for a specific group of people, based on written opinions and ratings of candidate statements. Reward models are trained to capture individual preferences, and the acceptability of a statement for the group is measured through a social welfare function. One difference from our work is that we do not attempt to find a single statement that builds consensus across the entire group — we instead allow for multiple statements representing distinct opinions. A more fundamental difference is that we view our experiments as an instance of a broader framework that allows for a systematic investigation of the types of queries an LLM can perform and the theoretical guarantees they provide.

[3] https://github.com/generative-social-choice/chatbot_personalization_data/

Finally, we build on the rich literature on justified representation in approval-based committee elections [1, 22]. As we have already mentioned, Halpern et al. [17] also study representation axioms from this literature in a statement-selection context. The key technical challenge in their work is that they only have access to partial approval votes. The learning-theoretic approach they adopt, as well as a later refinement by Brill and Peters [7], bears technical similarity to the algorithm we propose for obtaining representation with size-constrained generative queries. All previous papers in this literature assume a non-generative setting with a fixed set of alternatives.

2 Model

Let N be a set of n *agents*, and let $\mathcal{U}$ denote the *universe of* (well-formed, on-topic) *statements*, which may be finite or infinite. Each agent $i \in N$ has a *utility function* $u_i : \mathcal{U} \to \mathbb{R}$ that maps statements to utilities. Whereas our positive results apply for arbitrary real-valued utility functions, our impossibilities will even hold in the restricted setting of *approval utilities*, where utilities are 0 or 1, which much of the prior work [22] has focused on. An *instance* of the statement-selection problem consists of N, $\mathcal{U}$, $\{u_i\}_{i \in N}$, and a *slate size* $k \in \mathbb{N}_{\geq 1}$.

A *democratic process* is an algorithm that, when run on an instance, returns a *slate*, i.e., a multiset consisting of k statements from the universe.[4] Crucially, this algorithm receives only N and k in its input, but not $\mathcal{U}$ or the u_i, which it must instead access through queries as we describe below.

For convenience, we denote the rth largest element in a finite set X of real numbers (for $1 \leq r \leq |X|$) by $\max_{(r)}(X)$. To deal with edge cases, we set $\max_{(0)}(X) \coloneqq \infty$ for all sets X.

2.1 Queries

Since the democratic process does not receive the statements and preferences in its input, it instead accesses them indirectly through *queries*. The democratic processes we develop make use of two query types:

Discriminative Queries. Discriminative queries extrapolate an agent's utility function to unseen statements. For an agent i and statement α, $\textsc{Disc}(i, \alpha)$ returns $u_i(\alpha)$.

Generative Queries. For a set of agents S of size at most t and an integer $0 \leq r \leq |S|$, t-$\textsc{Gen}(S, r)$ returns the statement in $\mathcal{U}$ that maximizes the r-highest utility among the members of S. Formally, the query returns

$$\operatorname*{argmax}_{\alpha \in \mathcal{U}} \max\nolimits_{(r)}(\{u_i(\alpha) \mid i \in S\}), \tag{1}$$

breaking ties arbitrarily.

Intuitively, the generative query's parameter r interpolates between finding a lowest common denominator (t-$\textsc{Gen}(S, |S|)$ maximizes the minimum utility over S) and finding a statement that precisely matches a narrow coalition in S (e.g., t-$\textsc{Gen}(S, 1)$ gives some agent maximum utility, but

[4] Allowing a slate to contain the same statements multiple times avoids technical problems with the edge case where generative queries return the same statement, in which case no query-based algorithm would be able to procure k distinct statements. We also believe this choice to be suitable for our application domain, where representing multiple segments of the population by identical statements might sometimes be appropriate, for example if all agents in these segments have identical preferences. For ease of exposition, we will slightly abuse notation and treat slates as if they were sets; this essentially amounts to assuming that different generative queries do not return exactly the same statement.

may be unpopular among the remaining agents). For convenience, we will simply write $\textsc{Gen}(\cdot,\cdot)$ to refer to generative queries without a size limit or to talk generally about generative queries with different size constraints t.

2.2 Representation Axiom

The aim of our democratic processes is to produce a slate of statements W that is representative of the agent population. If agents have approval utilities, statement selection reduces to the classic setting of multi-winner approval voting. Therefore, our target axiom is inspired by the family of *justified representation* axioms [1] in this literature:

Definition 1. A slate W satisfies *balanced justified representation (BJR)* if there is a function $\omega : N \to W$ matching agents to statements in a balanced way,[5] for which there is no coalition $S \subseteq N$, statement $\alpha \in \mathcal{U}$, and threshold $\vartheta \in \mathbb{R}$ such that *(i)* $|S| \geq n/k$, *(ii)* $u_i(\alpha) \geq \vartheta$ for all $i \in S$, and *(iii)* $u_i(\omega(i)) < \vartheta$ for all $i \in S$.[6]

In words, if there is a coalition of agents that is *(i)* large enough to "deserve" a statement on the slate by proportionality and *(ii)* has cohesive preferences (i.e., there is a statement for which all these agents have utility at least ϑ), then *(iii)* the coalition must not be "ignored", in the sense that at least one member must be assigned to a statement with utility at least ϑ.

Our notion of BJR strengthens the classical axiom of justified representation, and is logically incomparable to several other axioms in the social choice literature. We prove these relationships and motivate the need for a new axiom in Appendix A. Throughout this paper, we will aim to build democratic processes that satisfy BJR, even when the universe of statements is very large and can only be navigated through queries.

3 First Component: Guarantees with Perfect Queries

In this section, we instantiate the first component of the generative social choice framework. We defer all proofs to Appendix B.

3.1 Unconstrained Queries

We begin by constructing a democratic process that guarantees BJR in polynomial time. This algorithm uses queries of type $\textsc{Disc}(\cdot,\cdot)$ and n-$\textsc{Gen}(\cdot,\cdot)$, i.e., generative queries without constraints on the number of input agents. The democratic process we propose, shown in Process 1, can either be seen as a generalization of *Greedy Approval Voting* [1], or as a variant of the *Greedy Monroe Rule* [34] that selects statements based on an egalitarian rather than utilitarian criterion. Our democratic process iteratively constructs a slate, adding statements one at a time. In each iteration, it identifies a set T of n/k (up to rounding) remaining agents and a statement α such that $\min_{i \in T} u_i(\alpha)$ is maximized. It then adds α to the slate, removes the agents T (who are now satisfied), and repeats. Our proof in Appendix B that this process satisfies BJR follows in structure the argument by Aziz et al. [1] that Greedy Approval Voting satisfies JR.

Theorem 2. *Process 1 satisfies balanced justified representation in polynomial time in n and k, using queries of types n-$\textsc{Gen}(\cdot,\cdot)$ and $\textsc{Disc}(\cdot,\cdot)$.*

[5] That is, each statement on the slate is matched to $\lfloor n/k \rfloor$ or $\lceil n/k \rceil$ agents.

[6] This axiom can also be defined in a setting where slates are sets of statements, rather than multisets. In this case, the statements α are restricted to lie in $\mathcal{U} \setminus W$, to make the axiom satisfiable. This axiom can be satisfied by a variant of Process 1, in which the choice of statements in each iteration is restricted to statements that have not previously been selected.

Process 1: Democratic Process for Balanced Justified Representation

Inputs: agents N, slate size k

$\bar{r} \leftarrow n\frac{1}{k}$

$S \leftarrow N$

$W \leftarrow \emptyset$

for $j = 1, 2, \ldots, k$ **do**

  $\alpha \leftarrow \textsc{Gen}(S, \lceil \bar{r} \rceil)$

  $W \leftarrow W \cup \{\alpha\}$

  $r \leftarrow \begin{cases} \lceil \bar{r} \rceil & \text{if } j \leq n - k \cdot \lfloor \bar{r} \rfloor \\ \lfloor \bar{r} \rfloor & \text{else} \end{cases}$

  $T \leftarrow$ the r agents in S with largest $\textsc{Disc}(\cdot, \alpha)$

  $S \leftarrow S \setminus T$

end

return W

3.2 Size-Constrained Generative Queries

So far, our generative queries could generate optimal statements even if the queried set S of agents was as large as n. When implementing a generative query using an LLM, however, the prompt to the LLM must include, for each agent in S, enough information to extrapolate the agent's preference across the universe of statements. Since this information can easily take hundreds of tokens per agent in S, the context windows of current LLMs (GPT-4 and PaLM: 8K tokens, LLaMA: 4K tokens) limits the size of S. Even recent models with extended context windows (GPT-4-32k: 32K tokens, GPT-4-turbo: 128K tokens, Claude 2: 100K tokens) struggle to effectively use the entirety of their context window [24]. As a result, democratic processes might for now be restricted to generative queries for moderate $|S|$. Therefore, we investigate in this section whether democratic processes can still ensure BJR when generative queries are limited to sets of agents of some size t that is substantially smaller than n. Immediately, we see that, if the query size t is even just slightly smaller than n/k, representation cannot be attained:

Proposition 3. *No democratic process can guarantee balanced justified representation with arbitrarily many $\frac{n}{k}(1 - \frac{1}{k})$-$\textsc{Gen}(\cdot, \cdot)$ and $\textsc{Disc}(\cdot, \cdot)$ queries. This impossibility even holds in the subsetting of approval utilities and for the weaker axiom of justified representation.*

Conceptually, the proof of this theorem and the subsequent impossibility theorem are based on the idea of *overshadowing*. Specifically, we construct instances that have few "popular" statements and many "unpopular" statements with lower support. For a given set S of at most t agents, our instances will ensure that some unpopular statement will be at least as well liked *within* S as any popular statement. Thus, all generative queries might return unpopular statements, and we design the instance such that no slate composed entirely of unpopular statements is representative. In Appendix B, we apply this idea in a straightforward way to prove Proposition 3.

On the face of it, slightly larger size-constrained generative queries seem promising for achieving BJR, since there is a democratic process that achieves BJR with queries of size $t = \lceil n/k \rceil$. Indeed, observe that, for any S and r,

$$\textsc{Gen}(S, r) = \operatorname*{argmax}_{\alpha \in \mathcal{U}} \max\nolimits_{(r)}(\{u_i(\alpha) \mid i \in S\}) = \operatorname*{argmax}_{\alpha \in \mathcal{U}} \max_{\substack{S' \subseteq S \\ |S'| = r}} \max\nolimits_{(r)}(\{u_i(\alpha) \mid i \in S'\})$$

$$= \operatorname*{argmax}_{\alpha \in \left\{ \textsc{Gen}(S', r) \,\middle|\, S' \subseteq S, |S'| = r \right\}} \max\nolimits_{(r)}(\{u_i(\alpha) \mid i \in S\}),$$

which shows that any call to $\textsc{Gen}(S, r)$ can be simulated by (exponentially many) r-$\textsc{Gen}(\cdot,\cdot)$ queries and discriminative queries. By applying this simulation to Process 1, in which all generative queries satisfy $r \leq \lceil n/k \rceil$, Theorem 2 immediately implies that BJR can be implemented by $\lceil n/k \rceil$-$\textsc{Gen}(\cdot,\cdot)$ queries, though the time complexity of the modified process is obviously prohibitive.

Proposition 4. *There exists a democratic process that satisfies balanced justified representation using (exponentially many) queries of type $\lceil n/k \rceil$-$\textsc{Gen}(\cdot,\cdot)$ and $\textsc{Disc}(\cdot,\cdot)$.*

Unfortunately, the exponential running time of this naïve democratic process turns out to be unavoidable, even if the generative queries can have linear size in n. Our proof must necessarily be more complicated than our previous impossibility in Proposition 3, in which we constructed an explicit instance on which *any* democratic process with t-bounded generative queries had to violate representation. A more sophisticated proof is necessary since, for any instance, there exists a democratic process that satisfies BJR in polynomial time with $\lceil n/k \rceil$-$\textsc{Gen}(\cdot,\cdot)$ queries on this instance; namely, a variant of the algorithm from Proposition 4 that guesses the right subset S' and returns the corresponding statement $\textsc{Gen}(S, r)$. We prove our impossibility (in Appendix B) by showing that, for any fixed polynomial-time algorithm, *there exists* an instance on which this algorithm violates BJR, through an application of the probabilistic method.

Theorem 5. *No democratic process can guarantee balanced justified representation with any number of $\textsc{Disc}(\cdot,\cdot)$ queries and fewer than $\frac{2}{k}\, e^{n/(12k)}$ queries of type $\frac{n}{8}$-$\textsc{Gen}(\cdot,\cdot)$. This holds even for the subsetting of approval utilities and the weaker axiom of justified representation. As a corollary, if $k \in O(n^{0.99})$, then any democratic process guaranteeing BJR with $\frac{n}{8}$-$\textsc{Gen}(\cdot,\cdot)$ and $\textsc{Disc}(\cdot,\cdot)$ queries has exponential running time.*

3.3 Structured Preference Settings

While the last section's lower bounds are potentially worrisome, a silver lining is that the instances we used to prove them were contrived. Our impossibility proofs were constructed by drowning popular statements in an overwhelming number of relatively unpopular statements: for any set of agents (of a given size), there was a statement that was well liked by only these agents and not by any other agent. Since statements and preferences in the real world presumably have some structure, it seems highly implausible that such an abundance of orthogonal statements would exist for real-world populations. Note that, by "structure" we are not referring to any fixed geometry of alternatives (in contrast to, say, spatial models of voting). Instead, we only require that preferences do not have infinite "complexity".

To formally define this complexity, we introduce the notion of a *statement space* $(\mathcal{U}, \mathcal{F})$, which consists of a universe of statements $\mathcal{U}$ and a set of possible utility functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{U}}$. A statement-selection instance belongs to $(\mathcal{U}, \mathcal{F})$ if its universe of statements is $\mathcal{U}$ and if each agent i's utility function u_i appears in $\mathcal{F}$.

To measure the complexity of a statement space, we borrow a fundamental complexity notion from learning theory, the *VC dimension* [37]. We extend the definition of VC dimension to statement spaces in a natural way: The VC dimension of $(\mathcal{U}, \mathcal{F})$ is the largest $d \in \mathbb{N}$, for which there exist $u_1, u_2, \ldots, u_d \in \mathcal{F}$ such that, for any index set $\mathcal{I} \subseteq \{1, \ldots, d\}$, there is a statement $\alpha \in \mathcal{U}$ and threshold $\vartheta \in \mathbb{R}$ such that $u_i(\alpha) \geq \vartheta$ for all $i \in \mathcal{I}$ and $u_i(\alpha) < \vartheta$ for all $i \notin \mathcal{I}$. If no largest integer d exists, the VC dimension is infinite. In other words, d is the size of the largest set of participants, such that for any subset of participants there is a statement that has a utility above some threshold for this subset and none of the agents outside of this subset.

Process 2: Democratic Process for BJR with Size-Constrained Queries.
(differences with Process 1 are highlighted in color)

Inputs: agents N, slate size k, VC dimension d, error probability δ
$n_x \leftarrow O\left(k^4\left(d + \log(k/\delta)\right)\right)$
$\epsilon \leftarrow \frac{1}{4k^2}$
$\bar{r}_x \leftarrow n_x\left(\frac{1}{k} - \epsilon\right)$
$\bar{r} \leftarrow n\left(\frac{1}{k} - 2\epsilon\right)$
$S \leftarrow N$
$W \leftarrow \emptyset$
for $j = 1, 2, \ldots, k$ **do**
  $X \leftarrow$ draw n_x agents from N without replacement
  $Y \leftarrow X \cap S$
  $\alpha \leftarrow \begin{cases} \textsc{Gen}(Y, \lceil \bar{r}_x \rceil) & \text{if } |Y| \geq \bar{r}_x \\ \text{some arbitrary } \alpha \in \mathcal{U} & \text{else} \end{cases}$
  $W \leftarrow W \cup \{\alpha\}$
  $r \leftarrow \begin{cases} \lceil \bar{r} \rceil & \text{if } j \leq n - k\lfloor \bar{r} \rfloor \\ \lfloor \bar{r} \rfloor & \text{else} \end{cases}$
  $T \leftarrow$ the r agents in S with largest $\textsc{Disc}(\cdot, \alpha)$
  $S \leftarrow S \setminus T$
end
return W

This notion of VC dimension of a statement space $(\mathcal{U}, \mathcal{F})$ is identical to the classic, learning-theoretic VC dimension of a hypothesis set $\mathcal{H}$, constructed as follows. We define a family of functions $h_{\alpha,\vartheta}$ that map the utility functions $u \in \mathcal{F}$ to binary labels as follows:

$$h_{\alpha,\vartheta}(u) := \begin{cases} 1 & \text{if } u(\alpha) \geq \vartheta \\ 0 & \text{else} \end{cases}$$

That is, $h_{\alpha,\vartheta}(u)$ indicates whether an agent with utility function u assigns a utility of ϑ or larger to a statement α. Then, the VC dimension d of a statement space $(\mathcal{U}, \mathcal{F})$ is identical to the classic, learning-theoretic VC dimension of the hypothesis set $\mathcal{H} := \{h_{\alpha,\vartheta} \,|\, \alpha \in \mathcal{U}, \vartheta \in \mathbb{R}\}$, consisting of binary classifiers over $\mathcal{F}$.

It seems unlikely that d would be huge in real-world settings, as it would imply, for instance (assuming a one-dimensional simplification), that we could find a statement such that people that lie at opposite sides of the space of opinions all support that statement, while people that lie in the middle disagree with it. If, hence, the VC dimension of the statement space is finite in realistic settings, we can obtain BJR even with size-constrained generative queries, as formalized by the following theorem.

Theorem 6. *Let d be the VC dimension of the statement space and $\delta > 0$ the maximum admissible error probability. Then, Process 2 runs in polynomial time in n, k (independent of d) and satisfies BJR with probability at least $1 - \delta$ using $\textsc{Disc}(\cdot,\cdot)$ and t-$\textsc{Gen}(\cdot,\cdot)$ queries for $t \in O\big(k^4(d + \log \frac{k}{\delta})\big)$.*

The proof of this theorem can be found in Appendix B. The process that achieves this result, Process 2, is an adaptation of Process 1. The key difference (Process 2) is that here we run $\textsc{Gen}(Y, \cdot)$ on a random subset $Y \subseteq N$ of the agents. Importantly, the size of this subset does not grow with the total number of agents n.

To illustrate the power of this theorem with a simple example, suppose that opinions on a discussion topic vary along three dimensions, say socially conservative vs. liberal, fiscally conservative vs. liberal, and religious vs. secular. Suppose furthermore that agents and statements can be represented as points in this three-dimensional space, such that the utility $u_i(\alpha)$ is a (strictly monotonically decreasing) function of the Euclidean distance between the agent i and statement α in this space. Then, the hypothesis set $\mathcal{H}$ (as introduced above) of this statement space is just the set of all spheres in $\mathbb{R}^3$, which is well known to have VC dimension $d = 4$ [e.g., 5, p. 122]. Hence, Process 2 produces BJR slates (up to a failure probability below 10^{-6}) using t-GEN$(\cdot, \cdot)$ queries with $t \leq \text{const} \cdot k^4 \left(4 + \log(k) + 6 \cdot \log(10)\right)$. If n is large, this t is much smaller than the lower bounds on t that are implied by Proposition 3 and Theorem 5 when we assume an unstructured statement space.

Importantly, Theorem 6 extends to far more complicated preference structures, and it does not require the structure to be known, but only (an upper bound on) the VC dimension. If, for example, the set of statements $\mathcal{U}$ consists of all sequences of w many words in English (which has below 10^6 words), a naive upper bound on the VC dimension of the statement space is $d \leq \log_2(|\mathcal{U}|) \leq w \log_2(10^6)$. Thus, $t \leq \text{const} \cdot k^4 \left(w + \log(k)\right)$ suffices to virtually guarantee BJR.

In summary, despite the negative worst-case results from Section 3.2, it is highly likely that relevant statement spaces in reality have enough structure to allow for a BJR guarantee with high probability and a relatively small number of queries, which is *independent of the number of agents* n. This means that we can scale the democratic process to any number of participants, say to a national audience, even when using an LLM with bounded context window size.

4 Second Component: Empirical Validation of Queries

We established in the previous section that, with access to *perfect* generative and discriminative queries, we can guarantee BJR. In this section, we describe how we implement these queries as subprocedures interfacing with an LLM, and we empirically study how well our implementations approximate the idealized queries.

Evaluation Data. To evaluate the query implementations, we use the data collected in our pilot study on chatbot personalization, which we discuss in detail in Section 5 and Appendix C. The dataset consists of survey responses by a representative sample of 100 US residents. Each participant extensively describes their views on chatbot personalization in free-form responses to multiple questions. Furthermore, each participant rates six example statements. Each statement consists of a concrete rule for chatbot personalization, a brief justification for the rule's importance, and an example illustrating the rule.[7] We elicited these ratings by asking participants "to what extent does this statement capture your full opinion regarding chatbot personalization?" Participants were then asked to choose a rating on a 5-point scale (with the levels "not at all" (0), "poorly" (1), "somewhat" (2), "mostly" (3), and "perfectly" (4)) and to give a short free-text response to explain their rating. We equate ratings with utilities, e.g. an agent i rating a statement α with "mostly" means that $u_i(\alpha) = 3$. Note, however, that the choice of numerical values is largely inconsequential given that Process 1 is invariant to monotone transformations of the rating scale.

[7] For examples of such statements, see Appendix C.2.

4.1 Discriminative Queries

Our implementation of the discriminative query $\textsc{Disc}(i, \alpha)$ takes as input agent i's survey responses and the statement α, and returns a prediction of the rating $u_i(\alpha)$. We implement these queries with a single call to GPT-4's base model, which has not been fine-tuned using *Reinforcement Learning from Human Feedback (RLHF)* [29].[8]

The prompt text is constructed as follows: It starts with the participant's free-form survey responses. Then, as few-shot examples, we list the example statements shown to the participant, each followed by the participant's rating and free-text explanations. At the end of the prompt, we append the statement α, such that, given the previous few-shot examples, the natural next token would be the agent's rating for α (see Appendix D for more details on the prompt). Hence, we can interpret GPT-4's prediction of the next token as an estimate of $u_i(\alpha)$. Since the GPT-4 base model allows for access to token probabilities, we can construct a probability distribution capturing the model's uncertainty about $u_i(\alpha)$. In our implementation of $\textsc{Disc}(i, \alpha)$ we return the expected $u_i(\alpha)$ (between 0 and 4), which will be used in our algorithms. An important advantage of using the expected (rating rather than, say, the mode) is that this virtually eliminates the possibility of ties when Process 1 chooses which agents to remove from consideration.

To evaluate this implementation of the discriminative query, we study how well it predicts a participant's rating of an example statement when the other five example statements are included in the prompt. Figure 1 displays the result of this analysis, for all 100 participants and all 6 choices of held-out statement (hence a total of 600 datapoints). Specifically, both subfigures compare the actual agreement rating given by the participant (row) with the predicted ratings (column) produced by GPT-4. The subfigures differ in that Figure 1a shows the average distributions, whereas Figure 1b gives a histogram when these distributions are collapsed to their expected values.

The pronounced diagonal in Figure 1a indicates that the generated rating distributions concentrate around the true rating. Predictions are typically within one step of the true rating, and there is no clear bias.

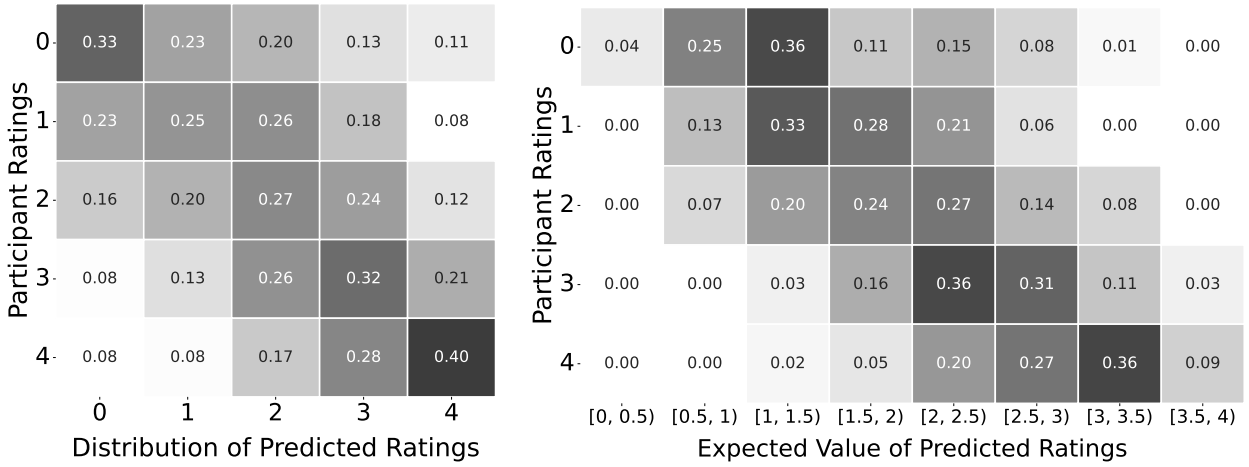
Since our implementation of the discriminative query returns only the expected value of these distributions, Figure 1b relates more directly to the performance of our democratic process. Again, there appears to be a clear linear relationship between true ratings and predictions, even if there is some visible bias towards intermediate ratings. This bias is not surprising, since, for example, a statement with a ground-truth rating of 4 ("perfectly") can be under- but not overestimated. Fortunately, since Process 1 is unaffected by monotone transformations of the rating scale, all we need for our democratic process to work is that there be a monotone relation between the true rating and the output of $\textsc{Disc}(\cdot, \cdot)$, which appears to be the case.

These results suggest that our LLM-based implementation of the discriminative query successfully extrapolates participants' preferences to new statements. This implementation provides a good approximation to the idealized $\textsc{Disc}(\cdot, \cdot)$ and we will hence use it in our democratic process.

4.2 Generative Queries

Passing the free-form responses of all 100 participants to GPT-4 would exceed the context window of 32K tokens of the GPT-4 version we are using. Nevertheless, we are able to circumvent this limitation by summarizing each agent's free-form responses more succinctly using the LLM, and

[8] We primarily use the base model, rather than the RLHF model since, at the time of our experiments, OpenAI did not provide access to log-probabilities for its RLHF model. Moreover, Santurkar et al. [31] found that RLHF models are more biased towards the opinions of certain demographics than corresponding base models, which might cause discriminative queries implemented using RLHF models to systematically skew towards certain viewpoints.



(a) (b)

Figure 1: Confusion matrix of discriminative queries, with normalized row-sums. Both figures represent the same 600 predicted rating distributions (100 participants times 6 choices of held-out statement). These 600 predicted distributions are partitioned according to the associated ground-truth participant rating, each partition corresponding to a row. In Figure 1a each row contains the average predicted distribution (i.e. the average of the distributions with the ground-truth rating corresponding to the row). In Figure 1b, each row is a histogram of expected ratings (that is, we take the expected rating for each distribution).

passing these condensed representations to the LLM for generation. As a result, LLM queries scale to the full size of our group of participants, which is why our evaluation uses Process 1 rather than the sampling-based Process 2.

We initially implemented the generative query with a single LLM call, but found a series of challenges that persuaded us to adopt a multi-prompt design instead. A first challenge with the single-prompt approach was a lack of stability, in the sense that we found the LLM's responses to be sensitive to details of the prompt text, such as wording and the order of the agent descriptions. A second challenge was that the LLM did not seem sufficiently responsive to the parameter r in our t-GEN(S, r) queries, although the task differs substantially depending on whether the statement should represent a small subgroup at a high minimum utility (small r) or a large subgroup at a moderate minimum utility (larger r).[9] A final challenge was that, when calling the prompt for large sets S of agents, the generated statements tended to be milquetoast, likely due to the LLM attempting to satisfy everyone, rather than cohesive subgroups.

To avoid these drawbacks, we implement our generative query through an *ensemble*: we generate a *pool* consisting of several candidate statements by applying the LLM prompt (see Appendix D for details on the prompt) to different subsets of agents in S. We then use discriminative queries to estimate agents' utilities for each statement in the pool and return the one that maximizes the objective of the idealized generative query, see Eq. (1).[10] For our pilot experiment, our ensemble contains the following statement sources:

- We initialize the pool with four statements generated by clustering participants using a k-means

[9] This lack of responsiveness persisted despite chain-of-thought prompting [39].

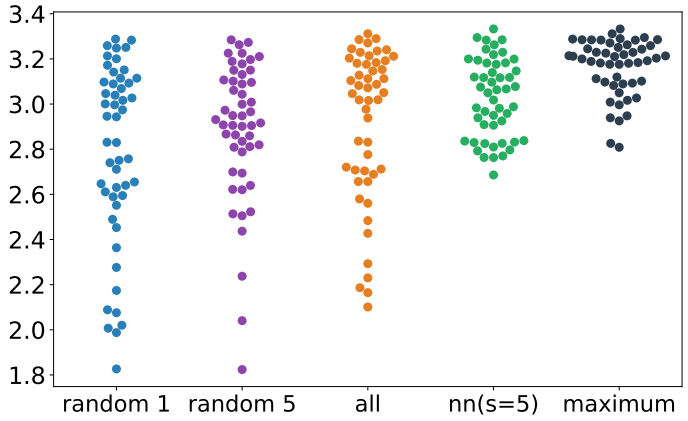
[10] We maintain all statements from previous calls to the generative query in the pool since this might identify good statements without requiring any additional LLM calls.

heuristic, and applying the generation prompt to each cluster.[11]

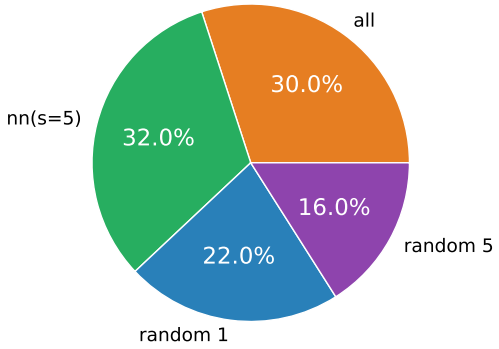
- For each generative query, we call the LLM prompt twice on all agents in S; once with a generation temperature of 0 and once with a temperature of 1.
- For each generative query, we additionally call the LLM prompt on a set of five agents, randomly selected without replacement from S.
- Furthermore, for each generative query, we compute three statements by applying the LLM prompt to different sets of agents produced by a nearest-neighbor heuristic.[12]

Evaluating the generative queries in a quantitative way is difficult for several reasons. First, since the 100 participants never see the statements generated based on their responses, we do not know their real ratings for these statements and have to rely on the discriminative queries as a proxy. Second, since the optimization over all possible statements in the idealized generative query (see Eq. (1)) cannot be computed in practice, we lack a ground truth for how far our implementation is from the ideal query.

Therefore, we restrict our evaluation to a more modest goal: comparing the types of statement generation sources in our ensemble and showing that they are complementary, i.e., that we benefit from choosing an ensemble approach rather than any single source. For this experiment, we randomly draw 40 out of the 100 agents and attempt to find a statement that maximizes the 20th-highest rating.[13] We then generate one statement from each of four sources (three of which feature in our ensemble), by applying the LLM prompt once to all 40 agents ("all"), once to a random subset of five agents ("random 5"), once to a group of six agents generated by the nearest-neighbor heuristic ("nn($s = 5$)", see Footnote 12 for details), and once, as a point of reference, to a single random agent ("random 1"). We run this experiment 50 times and show the results in Figure 2a.



(a) Distribution of the 20th-highest utility obtained by the statements from different sources.



(b) Percentage of experiments in which each statement source obtained a higher 20th-highest utility than all others.

Figure 2: Evaluation of the 20th-highest utility obtained by different generation sources in our experiments. Each of the 50 datapoints corresponds to a random sample of 40 out of the 100 agents.

[11] Since none of these statements were selected for the slate, we omit a detailed description.

[12] Specifically, we select a random agent i, and use the discriminative query to order the other agents in terms of how much they agree with agent i's free-form opinions. We then select the s most aligned agents for some fixed number s, and apply the LLM prompt to the resulting cluster of $s + 1$ agents. Three sets are produced with $s = 5$, $s = 10$ and $s = 15$. The three clusters are produced with $s = 5$; $s = 10$; and $s = 5$ except only a random subset of 20 agents is considered.

[13] This simulates what is required of the generative query in the fourth round of running Process 1 with $n = 100, k = 5$.

The main difference across the generation sources is their robustness, i.e., how often they yield statements whose 20th-highest utility is below 2.8. If we were to generate statements based on a single random agent ("random 1"), such low-utility statements would be frequently chosen, which is to be expected given that the randomly chosen agent's opinions need not align with the remaining agents. When we instead apply the LLM prompt to a random set of five agents ("random 5"), unpopular statements become less frequent. Including all agents in the prompt ("all") further increases the chance of very good statements, and decreases the incidence of very bad outliers. Still, this kind of generation does not entirely dominate the "random 5" generation source since it produces statements with mediocre highest-20 ratings (between 2 and 2.8) more frequently than the random-5 approach. The nearest-neighbors heuristic clearly outperforms the other three generation approaches. Indeed, all statements it produces have a 20-th highest utility above 2.7, which demonstrates the promise of applying the LLM prompt to subsets of agents that are chosen to be aligned in their opinions, rather than on random subsets or all agents.

Though the nearest-neighbor generation yields better statements than any other generation source in isolation, this does not mean that these other sources become redundant. Indeed, the last entry ("maximum") of Figure 2a shows that taking the best out of all four generation sources further reduces the lower tail of 20th-highest utilities, and that only this approach manages to generate very popular statements (rating above 3.2) most of the time. Figure 2b shows that the best-out-of-four statement is only chosen as the nearest-neighbor statement in about a third of our experiments. This implies that, even though the other three approaches individually lack robustness, it is rare that all three fail on the same instance. Our ensemble approach to statement generation makes use of this complementarity between generation sources.

There would be much to learn by extending this experiment to larger groups of remaining agents, different values of r, and more generation sources. In particular, it would be very interesting to study whether including multiple copies of the same generation source pays off or not. Unfortunately, the financial cost of running these experiments is currently limiting our analysis. This cost is mainly due to the large number of GPT-4 calls made for the discriminative queries, which on their own cost around $500 for the experiment in Figure 2.[14] OpenAI's recent announcement of GPT-4-turbo, a GPT-4 variant priced at about a third of the one we used, makes us hopeful that the cost of such experiments will soon decrease.

5 Pilot on Chatbot Personalization

We piloted our democratic process as part of the OpenAI "Democratic Inputs to AI" grant program [12], using our method to study public opinion on chatbot personalization. We ran surveys studying this topic on November 1 and 2, 2023 and generated a slate of five statements representing public opinion. To obtain actionable guidelines for the development of chatbots, we adopt a statement format that consists of the rule that participants judge most important for chatbot personalization, a brief justification for the rule's importance, and an illustrating example.

[14] The experiment's 50 random seeds, 40 agents per seed, and 4 statement to be evaluated per agent result in $50 \cdot 40 \cdot 4 = 8000$ discriminative queries. Since each discriminative query has a length of approximately 2000 tokens, at a current cost of $0.03 per thousand tokens for the GPT-4 base model, the cost for the experiment's discriminative queries alone is about $500.

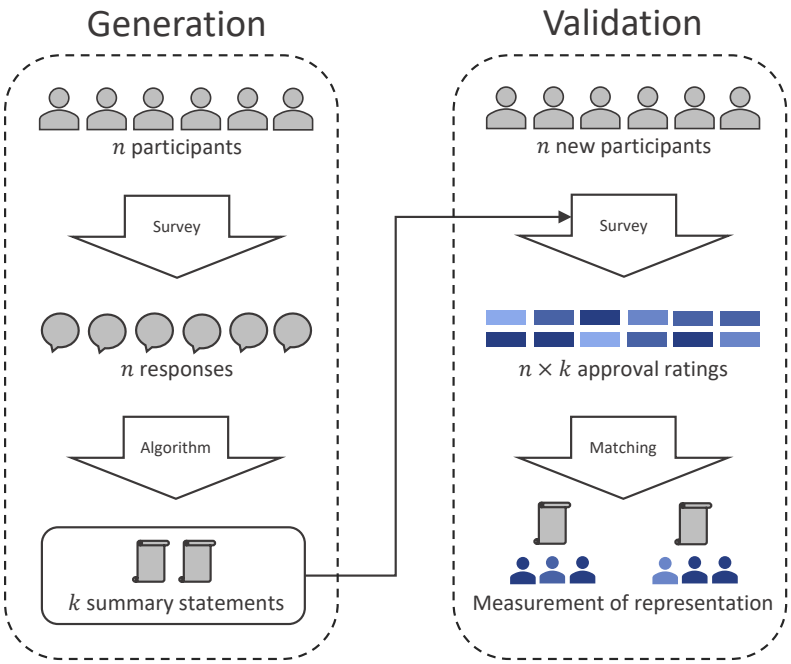


Figure 3: Overview over the pilot run of our process: In the first stage ("generation", left), we survey $n = 100$ participants. We then feed their responses into Process 1 to generate a slate of $k = 5$ statements. In the second stage ("validation", right), we validate these statements by asking a fresh sample of $n = 100$ participants to rate the five statements. Based on these ratings, we match participants optimally to statements, such that each statement represents an equal number of participants.

5.1 Pilot Description

We illustrate the setup of the pilot in Figure 3. We first recruit 100 participants through the online platform Prolific[15], which we refer to as the *generation sample*. Our sample consists of US residents, stratified with respect to age, gender, and race.[16] We ask these participants to complete a survey on chatbot personalization.[17] To introduce participants to the topic of chatbot personalization, we first show them background information and ask them whether a chatbot should personalize its answer in each of three example scenarios. Then, we asked participants to describe their stance on chatbot personalization, by answering four questions in free-text form that ask about the trade-offs of personalization, the rules participants would like to see imposed on chatbot personalization, as well as arguments for and against their proposed rules.

We also ask participants to rate their agreement with six example statements, which we generated with a single call to GPT-4 and without knowledge of participant responses. These ratings are given on the five-level scale described at the beginning of Section 4. Both the initial scenarios and statements are shown to participants in random order.

Based on participant responses, we extract a slate of five representative statements using Process 1. To evaluate this slate, we then launch a second survey with a new set of 100 stratified participants, the *validation sample*, to evaluate the slate's statements (see Figure 3, on the right). In this validation survey, after showing participants the same introductory information about

[15] https://www.prolific.com/

[16] For more details on the demographic composition of the sample, see Appendix C.1.

[17] See Appendix E.1 for the verbatim survey questions.

chatbots, we ask them to rate the five statements on the slate (using the same question format as at the end of the generation survey). For reproducibility, and to support future research on online participation, we made participants' full responses publicly available at https://github.com/generative-social-choice/chatbot_personalization_data/.

5.2 Results

Due to space limitations, we defer the slate of five statements we generated to Appendix C.2. No statement on the slate is categorically opposed to personalization but each statement expresses restrictions on personalization that major groups of US residents believe should be respected. We understand the following three points to be the main themes of the slate:

Privacy and data security: Four out of five statements stress the importance of privacy and of preventing chatbot data from being used in other contexts.

User control: The same four statements state that users should have control over which of their data are stored and used for personalization.

Truthfulness: The remaining statement's primary concern is that chatbots should never provide inaccurate or misleading information.

A striking feature of the slate is the high level of agreement between statements: Indeed, four of the five statements each express a concern about *privacy and data security* while recommending *user control* as a guardrail on personalization. Both the high level of agreement between participants, and the popularity of these two themes came as a surprise to us. The slate's fifth statement stresses that chatbot personalization should not go so far as to compromise the chatbot's *truthfulness*.

In Appendix C.2, we document which of the ensemble's generation sources each statement originated with, examine differences between the statements, and argue that these statements do not seem derivative of our priming scenarios. In Appendix C.3, a qualitative analysis of all responses in the generation sample confirms that concerns about *privacy and data security* and *user control* highlighted by our slate were frequently brought up by participants (72 of 100 participants mentioned one or both of these themes)[18] and that an example statement about user control received very favorable approval ratings, indicating its popularity.

In the remainder of this section, we demonstrate the representativeness of the slate using the validation sample, a fresh sample of 100 US residents. This validation is crucial since these participants actually rated the statements of our slate, which allows us to reason about their preferences without extrapolation. Furthermore, this validation sample received no introductory materials other than a brief explanation of chatbot personalization, which reduces the risk of priming participants in favor of or against certain statements.

According to the ideal of proportional representation, each statement in our generated slate should represent 20% of the US population as accurately as possible. Following this principle, we match the participants of our validation sample to the statements of our slate such that each statement has 20 participants matched to it and the sum of participants' rating levels for their assignment is maximized, i.e., such that the balanced assignment maximizes the representation objective of Monroe [26]. We then study the ratings of participants for their assigned statements.

As can be seen in Figure 4, 75% of the participants say that their assigned statement "perfectly" captures their full opinion on chatbot personalization, and an additional 18% of participants say it "mostly" captures their full opinion. Only 7% of participants feel only "somewhat" represented or less. Hence, the vast majority of participant opinions are represented accurately by our slate.

[18]The manual labeling of responses with themes is also contained in our data repository.

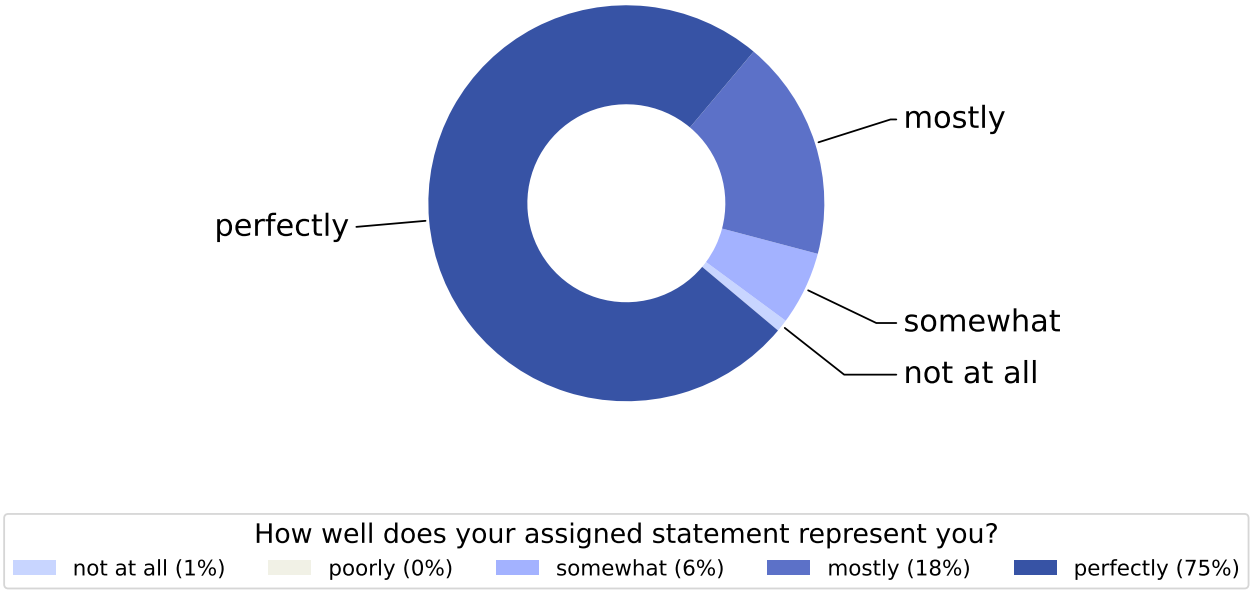


Figure 4: Ratings of participants from the validation survey for their *assigned* statement.

Remarkably, none of the 100 agents have a higher rating for a statement other than their assigned statement, which means that the requirement to assign *an equal number* of agents to each statement is not a binding constraint. This is a good sign for our claim of proportional representation, which could be in question if, say, many agents would rather be matched to the statement about truthfulness than their current assignment. It also shows that, should the slate violate BJR, this violation would have to be based on an entirely different kind of statement. Moreover, since such a violation would have to strictly increase the utility of all 20 members of the deviating coalition, it would have to unite most of the 25 agents who are not yet "perfectly" represented and would have to "perfectly" represent all coalition members who are already "mostly" represented. While we cannot entirely rule out such a BJR violation, this narrow path makes the existence of a violation seem unlikely.

Naturally, we must closely inspect the seven agents who feel relatively badly represented by their assigned statement, since their responses might reveal viewpoints missing from our slate. Though the free-text explanations given with the ratings are generally short, they allow us to understand what the seven participants dislike about the selected statements. While certain themes occur repeatedly among these seven participants,[19] their reasons for feeling relatively unrepresented are eclectic. Since proportionality axioms like BJR only guarantee representation to large, cohesive groups, these responses also give us no reason to doubt the representativeness of our slate.

Having established that the slate of statements represents the population well, an interesting question is how distinct the preferences of different groups are. Does our balanced matching identify distinct opinion clusters, or would participants in one group be just as happy with another group's statement? To answer this question, we consider the distribution of ratings across statements for each group, shown in Figure 5. Comparing the different plots, it is clear that the preferences of different groups substantially differ. In particular, each group has a clear preference for its assigned

[19]For instance, four of these participants do not believe that chatbot companies can be trusted to not collect data despite their customers' privacy choices or to keep collected data safe; and three express that the advantages of including all available data outweighs potential privacy risks. At least three of the participants doubt that chatbots can meaningfully identify truth or should be relied on as truthful.

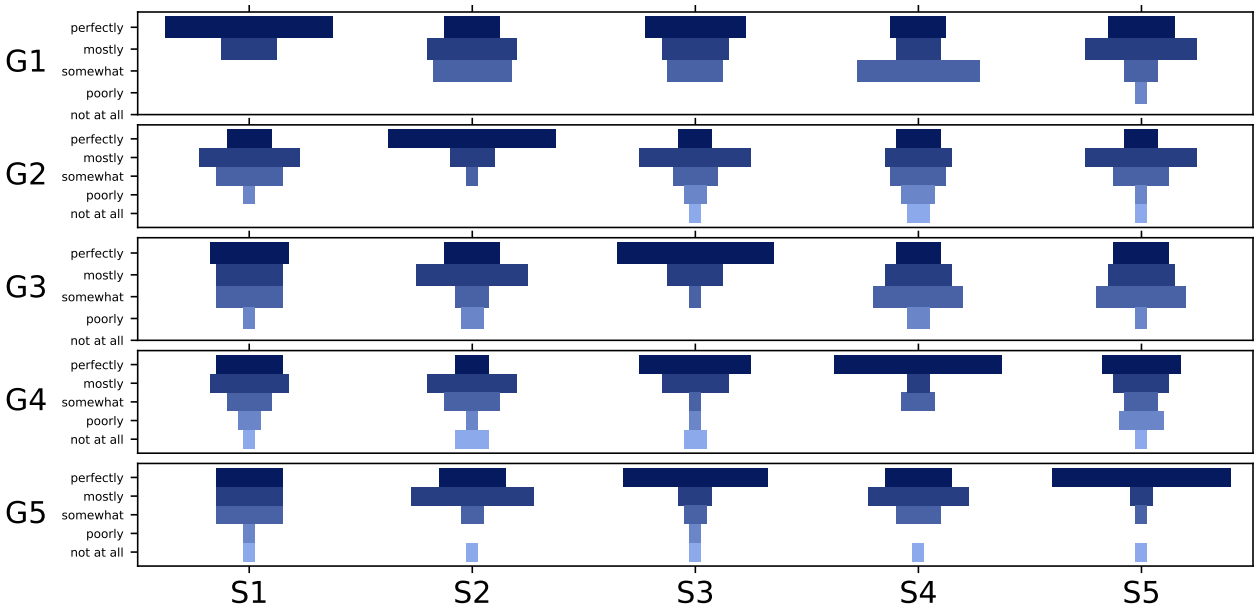


Figure 5: Agreement of participants in different groups with each of the statements. Each row corresponds to a group; for example, G1 represents the 20 participants assigned to statement S1. For each group, we plot the frequencies of rating levels given by members of this group to statements S1 through S5. S4 denotes the statement about truthfulness.

statement over the other statements (in Figure 5, see the distributions on the diagonal, from top left to bottom right).

Taken together, Figures 4 and 5 indicate that there is heterogeneity in opinions across the population and that our slate accurately represents this heterogeneity.

6 Discussion

As a result of the increase in power, availability, and steerability of LLMs, we are currently witnessing an explosion of creative prototypes for participative processes with generative-AI components [e.g., 10, 20, 25, 33]. This expansion of the capabilities of participation is thrilling, but — as these prototypes continue to proliferate and eventually turn into deployed practices — we ought to critically interrogate the legitimacy of these processes on two fronts.

The first line of questioning has already received broad attention [e.g., 36]: *can the AI building blocks in the process be trusted?* Taking our process as an example, we have started answering this question by measuring the average accuracy of our LLM queries, by overcoming an observed lack of robustness through the ensemble implementation of our generative query, and by piloting the process in practice. Before our process is ready for high-stake deployments, though, it must yet be hardened against malicious participant input (e.g., prompt injections [38] meant to unduly sway generative queries), and the effect of biases against groups of people [4, 21] and viewpoints [e.g. 18] in the LLM must be studied and counteracted. The thorniest question is whether participants themselves will trust the LLMs, for which our best suggestion is to grant participants recourse in the case of errors,[20] though this sacrifices some scalability for greater legitimacy.

[20]Say, if a discriminative query misjudged their preference, or if they can suggest a more popular statement than what a generative query produced.

We want to raise a no less important question: *is the process around the AI components democratic?* Granted, an AI participation process solicits input from all participants, and might even treat participants symmetrically. But that property alone (*neutrality*) is an utterly unimpressive benchmark for a voting rule. Instead, voting rules with AI elements, like those without, should argue their case based on social choice axioms that ensure, for example, the rule's responsiveness, efficiency, and fairness.

At its heart, generative social choice articulates a vision of what it means for an AI-enhanced voting rule to be democratic. By showing the required ingredients — the axioms targeted by the rule, necessary conditions on the behavior of the LLM, and evidence that the LLM meets these conditions — a voting rule can assuage the above two threats to legitimacy, while tapping into the possibilities enabled by generative AI.

Acknowledgments

We thank Nika Haghtalab and Abhishek Shetty for helpful pointers on how to apply sampling bounds to sampling without replacement. This work was partially supported by OpenAI through the "Democratic Inputs to AI" program and by the Office of Naval Research under grant N00014-20-1-2488. Manuel Wüthrich was partially funded by the Swiss National Science Foundation (SNSF). Paul Gölz was supported by the National Science Foundation under Grant No. DMS-1928930 and by the Alfred P. Sloan Foundation under grant G-2021-16778 while in residence at the Simons Laufer Mathematical Sciences Institute (formerly MSRI) in Berkeley, California, during the Fall 2023 semester.

References

[1] H. Aziz, M. Brill, V. Conitzer, E. Elkind, R. Freeman, and T. Walsh. 2017. Justified Representation in Approval-Based Committee Voting. *Social Choice and Welfare* 42, 2 (2017), 461–485.

[2] M. Bakker, M. Chadwick, H. Sheahan, M. Tessler, L. Campbell-Gillingham, J. Balaguer, N. McAleese, A. Glaese, J. Aslanides, M. Botvinick, and C. Summerfield. 2022. Fine-Tuning Language Models to Find Agreement Among Humans With Diverse Preferences. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*.

[3] P. L. Bartlett and S. Mendelson. 2002. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research* 3 (2002), 463–482.

[4] C. Basta, M. R. Costa-Jussà, and N. Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*.

[5] A. Blum, J. E. Hopcroft, and R. Kannan. 2020. *Foundations of Data Science*. Cambridge University Press.

[6] M. Brill, P. Gölz, D. Peters, U. Schmidt-Kraepelin, and K. Wilker. 2022. Approval-Based Apportionment. *Mathematical Programming* (2022).

[7] M. Brill and J. Peters. 2023. Robust and Verifiable Proportionality Axioms for Multiwinner Voting. In *Proceedings of the 14th ACM Conference on Economics and Computation (EC)*.

[8] Y. Cabannes. 2004. Participatory Budgeting: A Significant Contribution to Participatory Democracy. *Environment and Urbanization* 16, 1 (2004), 27–46.

[9] N. Clegg. 2023. Bringing People Together to Inform Decision-Making on Generative AI. Blog post. https://about.fb.com/news/2023/06/generative-ai-community-forum/

[10] F. Devine, A. Krasodomski-Jones, C. Miller, S. Y. Lin, J.-W. Cui, B. Marnette, and R. Wilkinson. 2023. Recursive Public. Report. https://vtaiwan-openai-2023.vercel.app/Report_%20Recursive%20Public.pdf

[11] R. El-Yaniv and D. Pechyony. 2009. Transductive Rademacher Complexity and Its Applications. *Journal of Artificial Intelligence Research* 35 (2009), 193–234.

[12] T. Eloundou and T. Lee. 2024. Democratic Inputs to AI Grant Program: Lessons Learned and Implementation Plans. Blog post. https://openai.com/blog/democratic-inputs-to-ai-grant-program-update

[13] B. Flanigan, P. Gölz, A. Gupta, B. Hennig, and A. D. Procaccia. 2021. Fair Algorithms for Selecting Citizens' Assemblies. *Nature* 596 (2021), 548–552.

[14] P. Fournier (Ed.). 2011. *When Citizens Decide: Lessons from Citizen Assemblies on Electoral Reform*. Oxford University Press, New York.

[15] R. Freedman, J. Schaich Borg, W. Sinnott-Armstrong, J. P. Dickerson, and V. Conitzer. 2020. Adapting a Kidney Exchange Algorithm to Align with Human Values. *Artificial Intelligence* 283 (2020).

[16] R. E. Goodin. 2000. Democratic Deliberation Within. *Philosophy & Public Affairs* 29, 1 (2000), 81–109.

[17] D. Halpern, G. Kehne, A. D. Procaccia, J. Tucker-Foltz, and M. Wüthrich. 2023. Representation With Incomplete Votes. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*.

[18] J. Hartmann, J. Schwenzow, and M. Witte. 2023. The Political Ideology of Conversational AI: Converging Evidence on ChatGPT's pro-Environmental, Left-Libertarian Orientation. arXiv:2301.01768.

[19] A. Konya, Y. L. Qiu, M. P. Varga, and A. Ovadya. 2022. Elicitation Inference Optimization for Multi-Principal-Agent Alignment. Manuscript.

[20] A. Konya, L. Schirch, C. Irwin, and A. Ovadya. 2023. Democratic Policy Development Using Collective Dialogues and AI. arXiv:2311.02242.

[21] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*. 166–172.

[22] M. Lackner and P. Skowron. 2023. *Multi-Winner Voting with Approval Preferences*. Springer.

[23] M. K. Lee, D. Kusbit, A. Kahng, J. T. Kim, X. Yuan, A. Chan, R. Noothigattu, D. See, S. Lee, C.-A. Psomas, and A. D. Procaccia. 2019. WeBuildAI: Participatory Framework for Fair and Efficient Algorithmic Governance. In *Proceedings of the 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW) (article 181)*.

[24] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. 2023. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172.

[25] B. Marnette and C. McKenzie. 2023. Talk to the City: an open-source AI tool for scaling deliberation. Blog post.

[26] B. L. Monroe. 1995. Fully Proportional Representation. *American Political Science Review* 89, 4 (1995), 925–940.

[27] R. Noothigattu, S. S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. D. Procaccia. 2018. A Voting-Based System for Ethical Decision Making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*. 1587–1594.

[28] Organisation for Economic Co-operation and Development. 2020. *Innovative Citizen Participation and New Democratic Institutions: Catching the Deliberative Wave*. OECD. https://doi.org/10.1787/339306da-en

[29] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[30] D. Peters, G. Pierczynski, and P. Skowron. 2021. Proportional Participatory Budgeting with Additive Utilities. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*. 12726–12737.

[31] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. 2023. Whose Opinions Do Language Models Reflect?. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. 29971–30004.

[32] S. Shalev-Shwartz and S. Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

[33] E. Shaotran, I. Pesok, S. Jones, and E. Liu. 2023. Aligned: A Platform-based Process for Alignment. arXiv:2311.08706.

[34] P. Skowron, P. Faliszewski, and A. Slinko. 2015. Achieving Fully Proportional Representation: Approximability Results. *Artificial Intelligence* 222 (2015), 67–103.

[35] C. Small, M. Bjorkegren, T. Erkkilä, L. Shaw, and C. Megill. 2021. Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. *Revista De Pensament I Anàlisi* 26, 2 (2021).

[36] C. T. Small, I. Vendrov, E. Durmus, H. Homaei, E. Barry, J. Cornebise, T. Suzman, D. Ganguli, and C. Megill. 2023. Opportunities and Risks of LLMs for Scalable Deliberation with Polis. arXiv:2306.11932.

[37] V. N. Vapnik. 1998. *Statistical Learning Theory*. Wiley.

[38] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

[39] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*.

[40] R. Willis, N. Curato, and G. Smith. 2022. Deliberative Democracy and the Climate Crisis. *WIREs Climate Change* 13, 2 (March 2022), e759.

[41] W. Zaremba, A. Dhar, L. Ahmad, T. Eloundou, S. Santurkar, S. Agarwal, and J. Leung. 2023. Democratic Inputs to AI. Blog post. https://openai.com/blog/democratic-inputs-to-ai

Appendix

A Relationship Between BJR and Other Justified Representation Axioms

Our notion of BJR is closely related to several axioms in the social choice literature.[21] Suppose for the time being that we were to relax BJR by not requiring the matching of agents to statements to be balanced, in which case each agent would be matched to their most preferred statement without loss of generality. In the subsetting of approval utilities, this relaxed axiom coincides with the *justified representation (JR)* axiom of Aziz et al. [1]. For our setting of general cardinal utilities, the relaxed axiom is implied by *extended justified representation (EJR)* and *full justified representation (FJR)* as defined by Peters et al. [30].

Table 1: Utility matrix of first example instance, with $k = n = 3$.

	α	α'	β	β'
u_1	1	1	0	0
u_2	1	1	0	0
u_3	0	0	1	1

Table 2: Utility matrix of second example instance, with $k = n = 2$.

	α_1	α_2	β	β'
u_1	3	0	2	2
u_2	0	3	2	2

The need for a new, balanced-matching-based notion of justified representation is best explained using two simple examples. The first example, given in Table 1, is standard: $k = 3$ statements must be selected, two thirds of the agents (specifically, agents 1 and 2) approve statements α, α', and the remaining third of the agents (agent 3) approves statements β, β'. As has been frequently observed [e.g., 1, Example 3], JR (and thus the relaxation of BJR with unbalanced matchings) is satisfied by the slate $\{\alpha, \beta, \beta'\}$. This is problematic since this slate is patently unproportional: it represents two thirds of the population by one third of the slate, and vice versa.

JR cannot rule out this form of unproportionality because each member of the two-thirds bloc is already represented by some statement they approve, and JR does not allow agents and coalitions to formulate any claims to representation beyond that point. Axioms like EJR and FJR allow coalitions to make stronger claims than JR by assuming that an agent (say, agent 1 in the previous example) may prefer to be represented by *multiple* statements rather than just one. Specifically, these axioms model an agent's utility as being the sum of their utilities for all statements on the slate.

Though this approach allows EJR and FJR to rule out the unproportional slates in the first example, it causes them to require slates on other instances that we find undesirable for the setting of statement selection, especially for non-approval utilities. Table 2 shows one such instance, in which two statements must be selected for two agents. Each agent $i \in \{1, 2\}$ has a statement α_i which is very specific to i and thus has a high utility for i but low utility for the other agent. In this instance, we believe that a slate consisting of these two statements would be a good choice since it represents the specificity of agents' preferences to the highest degree; indeed, only this slate satisfies BJR. EJR and FJR, by contrast, rule out these statements, since they prefer to represent both agents jointly by two less specific statements (namely, β, β') rather than each agent individually by

[21] Note that we defined slates as multisets, whereas these axioms typically define committees as sets. The discussion in this section is both valid if one translates the multi-winner axioms into the multiset setting, or by using the set variant of BJR described in Footnote 6.

a specific statement.[22]

Our axiom of BJR enforces more specificity on the second instance, while ruling out the unproportional slates on the first example instance. Instead of allowing a single agent to be represented by multiple statements, BJR's analysis of the shortcoming of JR in the first example is that too many agents were represented by a single statement on the slate. Philosophically, we see connections between our axiom and the notion of *fully proportional representation* of Monroe [26]: "voters should be segmented into equal-sized coalitions, each of which is assigned a representative, such that the preferences of voters are as closely as possible reflected by the representatives of their segment." In the remainder of this appendix, we show that BJR, other than implying JR, is incomparable to previously studied notions of justified representation, even in the setting of approval utilities.

Proposition 7. *Balanced justified representation (BJR) is incomparable with proportional justified representation (PJR), extended justified representation (EJR), full justified representation (FJR), and core stability. This incomparability holds even for approval utilities, and holds both in our setting where slates/committees are multisets*[23] *and in the classical setting where they are sets (using the adaptation of BJR in Footnote 6).*

Proof. We will show this incomparability in two steps: we first show that BJR implies none of the other axioms, and then that none of the axioms implies BJR.

BJR does not imply other axioms. Consider the instance with $n = 6$, $k = 4$, and the following utilities:

	α	α'	α^-	β	γ	δ
u_1	1	1	1	0	0	0
u_2	1	1	1	0	0	0
u_3	1	1	0	0	0	0
u_4	0	0	0	1	0	0
u_5	0	0	0	0	1	0
u_6	0	0	0	0	0	1

In this instance, the slate $\{\alpha^-, \beta, \gamma, \delta\}$ satisfies BJR since, if we assign agents 1 and 2 to α^-, agents 3 and 4 to β, agent 5 to γ, and agent 6 to δ, then only agent 3 is not already maximally satisfied. As a result, no potential deviating coalition can include the necessary $n/k = 3/2$ agents.

By contrast, this slate does not satisfy PJR because the coalition of agents 1, 2, and 3 is large enough to proportionally claim $\ell = 2$ statements, has two statements they all like in common (α, α'), but only one of the four statements on the slate is liked by any agent in this coalition.

Since EJR, FJR, and core stability imply PJR, none of them can be implied by BJR either.

Other axioms do not imply BJR. To prove this direction of the claim, consider the following instance with $n = 8$ agents and $k = 4$. The table below shows the agents' utilities for a subset of the statements:

[22] One might hope that EJR and FJR can be adapted to this perspective, by extending utilities to sets in a unit-demand rather than additive way. With this modification, however, they no longer rule out the unproportional slate in the first example instance.

[23] Brill et al. [6] give a formal embedding to translate existing justified representation axioms to the multiset setting ("party-approval elections", in their terminology). Whereas the existence of core stable committees is unresolved when committees are sets of alternatives, such committees are guaranteed to exist in the multiset setting [6].

	α	α'	β	β'
u_1	1	1	0	0
u_2, u_3, u_4	1	0	0	0
u_5	0	0	1	1
u_6, u_7, u_8	0	0	1	0

In addition, any pair of agents $\{i, j\}$ is associated with a statement $\gamma_{i,j}$, which exactly they approve.

In this instance, the slate $\{\alpha, \alpha', \beta, \beta'\}$ does not satisfy BJR. Indeed, since a balanced assignment assigns two agents to each statement of the slate, it holds for any such balanced assignment that some agent i assigned to α' and some agent j assigned to β' have 0 utility for their assigned statement. Since these two agents could deviate to the statement $\gamma_{i,j}$, BJR is violated.

By contrast, we will show that this slate satisfies core stability, and thus the weaker axioms of FJR, EJR, and PJR. Indeed, suppose that some non-empty coalition S along with a (multi)set T of at most $\frac{|S|}{n} \cdot k$ statements formed a core deviation. Suppose that S includes $0 \leq x \leq 2$ many among the agents $\{1, 5\}$. Since agents 1 and 5 have a utility of 2 for the candidate slate, they can only be part of a deviating coalition if the deviation T gives them utility at least 3. Analogously, since the other agents have a utilty of 1 for the candidate slate, they can only deviate if T gives them utility at least 2. If we define the *coalition welfare* cw as the sum of utilities, across the agents in S, for T, it follows that $cw \geq 3\,x + 2\,(|S| - x) = 2\,|S| + x$. Now, the average contribution of a statement in T to this objective is

$$\frac{cw}{|T|} \geq \frac{2\,|S| + x}{|S|\,k/n} \geq \frac{2\,n}{k} + x\,\frac{n}{|S|\,k} = 4 + x\,\underbrace{\frac{2}{|S|}}_{>0}. \tag{2}$$

Note that statements α and β are the only ones that can potentially contribute at least 4 to the coalitional welfare (since all other statements are approved by fewer than two agents), and they can also contribute only exactly an amount of 4, never more. Thus, it must be that $cw/|T|$ is equal to 4. This, in turn, implies that $x = 0$, i.e., that agents 1 and 5 are not in S, and that T consists only of the statements α and β (possibly with repetition). But now observe that, since agents 1 and 5 are not in the coalition, α and β cannot marginally contribute more than 3 to the coalition welfare, which contradicts Eq. (2) and thus shows that the slate satisfies core stability. □

B Deferred Proofs

Theorem 2. *Process 1 satisfies balanced justified representation in polynomial time in n and k, using queries of types n-GEN$(\cdot,\cdot)$ and DISC$(\cdot,\cdot)$.*

Proof. In this proof, we will use α_j, T_j to denote the values of α and T assigned in a given iteration $1 \leq j \leq k$. We construct the matching ω by, for each round $j = 1, \ldots, k$, mapping all agents that were removed from S in that round to the statement that was added to W in that round, i.e. for all $i \in T_j$ we have $\omega(i) = \alpha_j$. Clearly, this matching is balanced, since either $\lfloor n/k \rfloor$ or $\lceil n/k \rceil$ agents are removed in each round.

Now consider a coalition $S' \subseteq N$, a statement $\alpha' \in \mathcal{U}$, and a threshold $\vartheta \in \mathbb{R}$ such that $|S'| \geq n/k$ (and, by integrality, $|S'| \geq \lceil n/k \rceil$) and $u_i(\alpha') \geq \vartheta$ for all $i \in S'$. Once Process 1 terminates we have $S = \emptyset$, hence there must be an earliest iteration j where some agent $i' \in S'$ appeared in T_j. At the beginning of iteration j of the loop, it must thus still hold that $S' \subseteq S$. Note that

$$\max\nolimits_{(\lceil \vec{r} \rceil)}(\{u_i(\alpha') \mid i \in S\}) = \max\nolimits_{(\lceil n/k \rceil)}(\{u_i(\alpha') \mid i \in S\}) \geq \max\nolimits_{(|S'|)}(\{u_i(\alpha') \mid i \in S\})$$

$$\geq \max_{(|S'|)}(\{u_i(\alpha') \mid i \in S'\}) \geq \vartheta.$$

Thus, since $i' \in T_j$ and by the definition of the generative query (Eq. (1)), it must hold that

$$u_{i'}(\omega(i')) = u_{i'}(\alpha_j) \geq \max_{(\lceil \bar{r} \rceil)}(\{u_i(\alpha_j) \mid i \in S\}) \geq \vartheta.$$

We conclude that S', α', ϑ do not violate BJR. □

Proposition 3. *No democratic process can guarantee balanced justified representation with arbitrarily many $\frac{n}{k}(1-\frac{1}{k})$-GEN$(\cdot,\cdot)$ and DISC$(\cdot,\cdot)$ queries. This impossibility even holds in the subsetting of approval utilities and for the weaker axiom of justified representation.*

Proof. Set $t \coloneqq n/k\,(1-1/k)$. Let n be some multiple of k^2, so that t is an integer. Suppose that there is one "popular" statement α, which has utility 1 for all agents. Furthermore, for each set S of at most t agents, let there be an "unpopular" statement with utility 1 for S and 0 for all other agents. This unpopular statement is a valid answer for any query of the shape t-GEN$(S,\cdot)$, because the r-th largest utility among S for this statement is 1, the maximum possible utility of this instance. Thus, with the right tie breaking, one can implement all t-GEN$(\cdot,\cdot)$ queries to return unpopular statements, from which it follows that the process will have to return a slate W entirely of unpopular comments.

Since each unpopular statement has positive utility for at most t agents, at most $k \cdot t = n\,(1-1/k) = n - n/k$ agents receive positive utility from any statement in W. In other words, n/k agents have utility 0 for all statements in W, but have utility 1 for the popular statement α. This demonstrates a violation of (balanced) justified representation. □

Theorem 5. *No democratic process can guarantee balanced justified representation with any number of DISC$(\cdot,\cdot)$ queries and fewer than $\frac{2}{k}\,e^{n/(12k)}$ queries of type $\frac{n}{8}$-GEN$(\cdot,\cdot)$. This holds even for the subsetting of approval utilities and the weaker axiom of justified representation. As a corollary, if $k \in O(n^{0.99})$, then any democratic process guaranteeing BJR with $\frac{n}{8}$-GEN$(\cdot,\cdot)$ and DISC$(\cdot,\cdot)$ queries has exponential running time.*

Proof. Choose k to be an even integer and n as a multiple of 8, such that $t \coloneqq n/8$ is integer as well. Fix a process that makes fewer than $\frac{2}{k}\,e^{n/(12k)}$ many t-GEN$(\cdot,\cdot)$ and any number of discriminative queries. We will prove the claim using the probabilistic method: we will define a random instance and show that the process will fail BJR with positive probability, which means that there exists a deterministic instance where the process fails BJR. In fact, the random instances we construct will have approval utilities, and we will derive a contradiction to not just BJR, but also JR on this instance, to simultaneously prove the "this holds even…" part of the claim.

For given n, k, construct our instance as follows: Each set S of $\frac{n}{2k}$ many agents has infinitely many "unpopular" statements that have utility 1 for S and utility 0 for all other agents. Furthermore, each agent is uniformly and independently assigned a color in $\{1, 2, \ldots, k/2\}$, and all agents with the same color c have utility 1 for a "popular" statement β_c, which has utility 0 for everyone else. Since all utilities are 0 or 1, there will typically be many statements α that are tied in the definition of a generative query GEN(S, r) (Eq. (1)): if there exist statements that have utility 1 for at least r agents in S, any such statement may be returned; if no such statements exist, the query may return any arbitarary statement. To resolve this ambiguity, we assume that the generative query breaks ties in the "most favorable" way: the generative query will respond to GEN(S, r) with a statement that has utility 1 for as many agents in S as possible, and breaks remaining ties according to some canonical ordering of statements in which unpopular comments precede popular comments.

Consider the trajectory of the process on an instance with just the unpopular statements, i.e., where each t-GEN$(S, \cdot)$ query of the process is answered by a canonical unpopular statement that attains the maximum number $\min(|S|, \frac{n}{2\,k})$ of agents in S that have utility 1 for it.

Now, consider the random instance with unpopular and popular statements. We will show that, with positive probability, all t-GEN$(\cdot, \cdot)$ queries made by the process are still answered by their canonical unpopular statement, which means that the process will follow the same trajectory as above. This will be the case if, for each t-GEN$(S, \cdot)$ query made by the process and for each color c, at most $\frac{n}{2k}$ agents in S have color c, so that β_c will not be returned by the query. For a specific S and c, the probability of this event can be upper-bounded using Chernoff as

$$\begin{aligned}
&\mathbb{P}\left[\text{at least } \frac{n}{2\,k} \text{ agents in } S \text{ have color } c\right] \\
&= \mathbb{P}\left[\mathsf{Binomial}(n/8, 2/k) \geq 2 \cdot \frac{n}{4\,k}\right] \\
&\leq \exp\left[-\frac{n}{12\,k}\right].
\end{aligned}$$

By a union bound, it follows that, with positive probability, this event does not occur in any of the fewer than $\frac{2}{k}\, e^{n/(12k)}$ queries, for any of the $\frac{k}{2}$ colors. This implies that there is an instance in the support of our random instance on which the trajectory of the process remains the same as if there were no popular statements and where, in particular, the process must return a slate of unpopular statements.

Finally, we show that, when the process only returns unpopular statements, it must violate justified representation. (This always hold for our random instance, ex post.) Since each unpopular statements gives positive utility to at most $\frac{n}{2k}$ agents, no more than $\frac{n}{2}$ agents can be covered by the slate of k statements selected by the process. Therefore, there are at least $\frac{n}{2}$ uncovered agents, which are partitioned in some arbitrary manner across the $\frac{k}{2}$ many colors. By an averaging argument, there must be some color c with at least $\frac{n}{k}$ uncovered agents, which means that the process' output violates justified representation and BJR for β_c. □

Lemma 8 (Agnostic PAC learning for sampling without replacement)**.** *Let $\mathcal{H}$ be a hypothesis class, consisting of binary classifiers $h : \mathcal{X} \to \mathcal{Y}$, with $|\mathcal{Y}| = 2$, over some domain $\mathcal{X}$. Let $d < \infty$ denote the VC dimension of $\mathcal{H}$. For a given hypothesis $h \in \mathcal{H}$, denote its 0–1 loss on a nonempty finite set $S \subseteq \mathcal{X} \times \mathcal{Y}$ of labeled datapoints by $L_S(h) \coloneqq \sum_{(x,y)\in S} \mathbb{1}\{h(x) \neq y\}/|S|$.*

Let $D \subseteq \mathcal{X} \times \mathcal{Y}$ be a finite set of labeled datapoints. Consider a random process that chooses some number $m \leq |D|/2$ of labeled datapoints $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ from D uniformly and ***without replacement****, and denote by $\hat{h}$ the empirical risk minimizer* $\operatorname{argmin}_{h\in\mathcal{H}} L_S(h)$*. For any $0 < \epsilon < 1, 0 < \delta < 1$, this process will satisfy*

$$L_D(\hat{h}) \leq \min_{h\in\mathcal{H}} L_D(h) + \epsilon \tag{3}$$

and

$$|L_S(h) - L_D(h)| \leq \epsilon \quad \forall h \in \mathcal{H} \tag{4}$$

with probability at least $1 - \delta$, as long as

$$m \geq C \cdot \frac{d + \log 1/\delta}{\epsilon^2} \tag{5}$$

for some absolute constant C.

Proof. If $|D| \geq m^2/\delta$, the result will follow from the sampling bounds for i.i.d. samples. Note that we can implement the without-replacement drawing of S through rejection sampling, i.e., by drawing a sample of m datapoints uniformly *with* replacement, and re-drawing if this sample should contain any datapoint multiple times. We will consider only the first round of this rejection sampling. The probability that any two datapoints are identical is at most $\sum_{i=0}^{m-1} i/|D| = \frac{m\,(m-1)}{2\,|D|} \leq \frac{m^2}{2\,|D|} \leq \delta/2$, so we reject with probability at most $\delta/2$. Moreover, since drawing with replacement is the same as drawing i.i.d. from the uniform distribution over D, we can apply a standard agnostic PAC learning bound [32, Thm. 6.8] to show that the empirical risk minimizer $\hat{h}$ on the sample with replacement satisfies Eq. (3) with probability at least $1 - \delta/2$ as long as the constant in Eq. (5) is sufficiently large. By a union bound over both events, with probability at least $1 - \delta$, the with-replacement sample is not rejected and additionally satisfies Eq. (3), which proves the claim for our sampling process without replacement in the case of $|D| \geq m^2/\delta$.

From here on, suppose that $|D| < m^2/\delta$. Essentially, our claim will follow from Theorem 2 by El-Yaniv and Pechyony [11], a bound on transductive learning, but we have to do some work to get their bound into our desired shape. We apply their Theorem 2 twice, with a value of δ that is half of the δ in our theorem, the full sample D, the hypothesis class $\mathcal{H}$, $\gamma = 1$, and setting m once to m and once to $|D| - m$ (swapping the role of sampled and not sampled datapoints). By union-bounding over both invocations and unfolding some definitions in the theorem, we obtain that, with probability at least $1 - \delta$, it holds for all $h \in \mathcal{H}$ that

$$L_{D\setminus S}(h) \leq L_S(h) + R_{trans}(\mathcal{H}) + slack \quad \text{and} \quad L_S(h) \leq L_{D\setminus S}(h) + R_{trans}(\mathcal{H}) + slack \tag{6}$$

where $R_{trans}(\mathcal{H})$ denotes the *transductive Rademacher complexity* of $\mathcal{H}$ on D, and *slack* is defined and bounded in the following.

The slack term is defined as

$$slack \coloneqq c_0\, q\, \sqrt{m} + \sqrt{\frac{s\,q}{2}\,\ln 1/\delta},$$

where $c_0 < 5.05$ is an absolute constant, $q \coloneqq \frac{1}{m} + \frac{1}{|D|-m} \leq \frac{2}{m}$, and $s \coloneqq \frac{|D|}{(|D|-1/2)\cdot(1-\frac{1}{2(|D|-m)})}$. Since m is a positive integer, $m \geq 1$, hence $|D|-m \geq m \geq 1$, and thus $s = \frac{|D|}{|D|-1/2} \cdot \frac{1}{1-\frac{1}{2\,(|D|-m)}} \leq 4/3 \cdot 2 = 8/3$. Thus,

$$slack \leq \frac{5.05 \cdot 2}{\sqrt{m}} + \sqrt{\frac{8/3}{m}\,\ln 1/\delta} = \frac{1}{\sqrt{m}}(10.10 + \sqrt{8/3\,\ln 1/\delta}). \tag{7}$$

Next, we bound the transductive Rademacher complexity, for which we require several definitions: Let $\vec{x} \in \mathcal{X}^{|D|}$ be a vector listing the first components (i.e., the unlabeled datapoints) for all members of D, in arbitrary order. For an index set $\mathcal{I} \subseteq \{1, \ldots, |D|\}$, let $\vec{x}_{\mathcal{I}} \in \mathcal{X}^{|\mathcal{I}|}$ be the restriction of $\vec{x}$ to the indices $\mathcal{I}$. For a hypothesis h and a vector $\vec{v}$, let $h(\vec{v})$ be the vector that results from applying h element-wise to the entries of $\vec{v}$. Since the codomain of the hypothesis class is binary, i.e. $|\mathcal{Y}| = 2$, we will assume here that $\mathcal{Y} = \{-1, 1\}$ without loss of generality. For any $t \in \mathbb{N}$, let Σ^t_{trans} denote the probability distribution over vectors of length t, whose entries are drawn i.i.d. and are equal to -1 with probability $\frac{m\,(|D|-m)}{|D|^2}$, equal to 1 with probability $\frac{m\,(|D|-m)}{|D|^2}$, and are 0 otherwise. Furthermore, let Σ^t_{ind} denote the probability distribution over vectors of length t whose entries are independently drawn and -1 or 1 with equal probability. Finally, denote by $\mathcal{B}$ the probability distribution over subsets of $\{1, \ldots, |D|\}$ in which each element is contained in the subset independently with probability $2\,\frac{m\,(|D|-m)}{|D|^2}$.

In this notation, El-Yaniv and Pechyony [11, Def. 1 and p. 6] define the transductive Rademacher complexity $R_{trans}(\mathcal{H})$ as

$$(\tfrac{1}{m} + \tfrac{1}{|D|-m}) \cdot \mathbb{E}_{\vec{\sigma} \sim \Sigma_{trans}^{|D|}} \sup_{h \in \mathcal{H}} \vec{\sigma}^T h(\vec{x}).$$

Note that we can draw $\vec{\sigma}$ from $\Sigma_{trans}^{|D|}$ in two steps: we first draw the set of indices $\mathcal{I}$ from $\mathcal{B}$ whose entries in $\vec{\sigma}$ are nonzero, and then set $\vec{\sigma}$'s coordinates in $\mathcal{I}$ to -1 or 1 with equal probability. Therefore, we can equivalently write

$$R_{trans}(\mathcal{H}) = (\tfrac{1}{m} + \tfrac{1}{|D|-m}) \cdot \mathbb{E}_{\mathcal{I} \sim \mathcal{B}} \, \mathbb{E}_{\vec{\sigma} \sim \Sigma_{ind}^{|\mathcal{I}|}} \sup_{h \in \mathcal{H}} \vec{\sigma}^T h(\vec{x}_{\mathcal{I}}).$$

By Bartlett and Mendelson [3, Lemma 4 & Thm. 6], $\mathbb{E}_{\vec{\sigma} \sim \Sigma_{ind}^{|\mathcal{I}|}} \sup_{h \in \mathcal{H}} \vec{\sigma}^T h(\vec{x}_{\mathcal{I}}) \leq c_1 \sqrt{d|\mathcal{I}|}$ for some absolute constant c_1. Thus, we can bound

$$\begin{aligned}
R_{trans}(\mathcal{H}) &\leq c_1 \, (\tfrac{1}{m} + \tfrac{1}{|D|-m}) \cdot \mathbb{E}_{\mathcal{I} \sim \mathcal{B}} \sqrt{d|\mathcal{I}|} \\
&\leq c_1 \, \tfrac{2}{m} \cdot \mathbb{E}_{\mathcal{I} \sim \mathcal{B}} \sqrt{d|\mathcal{I}|} && m \leq |D| - m \\
&\leq \frac{2\, c_1 \sqrt{d}}{m} \cdot \mathbb{E}_{t \sim Binomial\left(|D|, \frac{2\,m\,(|D|-m)}{|D|^2}\right)} \sqrt{t} \\
&\leq \frac{2\, c_1 \sqrt{d}}{m} \left(\sqrt{6\, \tfrac{m\,(|D|-m)}{|D|}} + \mathbb{P}\left[Binomial\left(|D|, \tfrac{2\,m\,(|D|-m)}{|D|^2}\right) > 6\, \tfrac{m\,(|D|-m)}{|D|} \right] \cdot \sqrt{|D|} \right) \\
&\leq \frac{2\, c_1 \sqrt{d}}{m} \left(\sqrt{6\, \tfrac{m\,(|D|-m)}{|D|}} + \exp(-2\, \tfrac{m\,(|D|-m)}{|D|}) \cdot \sqrt{|D|} \right) && \text{(Chernoff bound)} \\
&\leq \frac{2\, c_1 \sqrt{d}}{m} \left(\sqrt{6\, m} + \exp\left(\tfrac{\ln |D|}{2} - m \right) \right) && (1/2 \leq \tfrac{|D|-m}{|D|} \leq 1) \\
&\leq \frac{2\, c_1 \sqrt{d}}{m} \left(\sqrt{6\, m} + \exp\left(\tfrac{\ln m^2/\delta}{2} - m \right) \right) && (|D| < m^2/\delta) \\
&= \frac{2\, c_1 \sqrt{d}}{m} \left(\sqrt{6\, m} + \exp\left(\tfrac{\ln 1/\delta}{2} + \ln m - m \right) \right) \\
&\leq \frac{2\, c_1 \sqrt{d}}{m} \left(\sqrt{6\, m} + \exp\left(\tfrac{\ln 1/\delta}{2} - (1 - 1/e) m \right) \right) && (x - \ln x \geq (1 - 1/e)\, x))
\end{aligned}$$

By choosing a large enough constant in Eq. (5), we can ensure that $(1 - 1/e)m \geq \frac{\ln 1/\delta}{2}$. Then, we can continue:

$$\begin{aligned}
&\leq \frac{2\, c_1 \sqrt{d}}{m} \left(\sqrt{6\, m} + e^0 \right) \leq \tfrac{2\, c_1 \sqrt{d}}{m} \, (\sqrt{6} + 1) \sqrt{m} \\
&\leq \frac{c_2 \sqrt{d}}{\sqrt{m}}, && (8)
\end{aligned}$$

where we set $c_2 \coloneqq 2\,(\sqrt{6} + 1)\, c_1$. Putting together Eqs. (6) to (8), we obtain that, for all $h \in \mathcal{H}$,

$$L_{D \setminus S}(h) \leq L_S(h) + \alpha \quad \text{and} \quad L_S(h) \leq L_{D \setminus S}(h) + \alpha$$

where we defined

$$\alpha \coloneqq \frac{10.10 + \sqrt{8/3 \, \ln 1/\delta} + c_2 \sqrt{d}}{\sqrt{m}}.$$

We have

$$L_D(h) = \frac{m}{|D|} L_S(h) + \frac{|D| - m}{|D|} L_{D \setminus S}(h)$$

$$\leq \frac{m}{|D|} L_S(h) + \frac{|D|-m}{|D|} \left(L_S(h) + \alpha\right)$$
$$\leq L_S(h) + \alpha$$

and using a similar argument for the the other side, we obtain an error bound that holds uniformly across all hypothesis

$$|L_S(h) - L_D(h)| \leq \alpha \quad \forall h.$$

Finally, we compute also a bound for the empirical risk minimizer. We set $h^* \coloneqq \operatorname{argmin}_{h \in \mathcal{H}} L_D(h)$. Then, we bound

$$\begin{aligned}
& L_D(\hat{h}) - L_D(h^*) \\
&= \frac{m}{|D|} \left(L_S(\hat{h}) - L_S(h^*)\right) + \frac{|D|-m}{|D|} \left(L_{D\setminus S}(\hat{h}) - L_{D\setminus S}(h^*)\right) \\
&\leq \tfrac{m}{|D|} \left(L_S(\hat{h}) - L_S(h^*)\right) + \tfrac{|D|-m}{|D|} \left(L_S(\hat{h}) - L_S(h^*) + 2\,\alpha\right) \\
&= \underbrace{L_S(\hat{h}) - L_S(h^*)}_{\leq 0, \text{ by definition of } \hat{h}} + 2\,\tfrac{|D|-m}{|D|}\,\alpha \\
&\leq 2\,\alpha
\end{aligned}$$

By choosing the constant in Eq. (5) large enough, we can ensure[24] that

$$m \geq \frac{4}{\epsilon^2} \cdot 3\,(10.10^2 + 8/3\,\ln 1/\delta + c_2^2 d).$$

By Cauchy's inequality, this implies that

$$m \geq \frac{4}{\epsilon^2} \cdot (10.10 + \sqrt{8/3\,\ln 1/\delta} + c_2\sqrt{d})^2,$$

and, by rearranging, that

$$\begin{aligned}
\epsilon &\geq 2\,\frac{10.10 + \sqrt{8/3 \ln 1/\delta} + c_2\sqrt{d}}{\sqrt{m}} \\
&= 2 \cdot \alpha.
\end{aligned}$$

Thus, with probability at least $1-\delta$, $\epsilon \geq L_D(\hat{h}) - L_D(h^*)$, and $\epsilon \geq |L_S(h) - L_D(h)| \ \forall h$, as claimed. □

Theorem 6. *Let d be the VC dimension of the statement space and $\delta > 0$ the maximum admissible error probability. Then, Process 2 runs in polynomial time in n, k (independent of d) and satisfies BJR with probability at least $1-\delta$ using $\textsc{Disc}(\cdot,\cdot)$ and t-$\textsc{Gen}(\cdot,\cdot)$ queries for $t \in O\big(k^4(d + \log \frac{k}{\delta})\big)$.*

Proof. For convenience, we define $\textsc{supp}(\alpha, \vartheta | S) \coloneqq \{i \in S \mid u_i(\alpha) \geq \vartheta\}$ to be the set of agents in S who have utility at least ϑ for statement α. Further, we define Process 3, which is equivalent to Process 2 but whose more explicit notation makes it easier to refer to specific values of the variables in this proof. Note that we have

$$\textsc{Gen}(S, \lceil r \rceil) = \operatorname*{argmax}_{\alpha \in \mathcal{U}} \sup \{\vartheta \mid |\textsc{supp}(\alpha, \vartheta | S)| \geq r\}$$

and hence we can write α_j defined in Process 3 of Process 3 as

$$\alpha_j = \operatorname*{argmax}_{\alpha \in \mathcal{U}} \sup \{\vartheta \mid |\textsc{supp}\,(\alpha, \vartheta | Y_j)| \geq \bar{r}_x\}\,. \tag{9}$$

[24] We may assume without loss of generality that $d \geq 1$, since, if $d = 0$, $\mathcal{H}$ only contains a single classifier and the claim holds trivially. If $d \geq 1$, we can upper bound the term $\frac{12 \cdot 10.10^2}{\epsilon^2}$ by a multiple of $\frac{d}{\epsilon^2}$.

Process 3: Democratic Process for BJR with Size-Constrained Queries (more explicit version of Process 2).

Inputs: agents N, slate size k, VC dimension d, error probability δ

$n_x \leftarrow 16\, C\, k^4\, (d + \log(k/\delta))$ (C is the constant from Lemma 8)

if $n \leq 2 \cdot n_x$ **then**

  $n_x \leftarrow n$

end

$\epsilon \leftarrow \frac{1}{4k^2}$

$\bar{r}_x \leftarrow n_x \left(\frac{1}{k} - \epsilon\right)$

$\bar{r} \leftarrow n \left(\frac{1}{k} - 2\epsilon\right)$

$S_1 \leftarrow N$

$W_0 \leftarrow \emptyset$

for $j = 1, 2, \ldots, k$ **do**

  $X_j \leftarrow$ draw n_x agents from N without replacement

  $Y_j \leftarrow X_j \cap S_j$

  $\alpha_j \leftarrow \begin{cases} \textsc{Gen}(Y_j, \lceil \bar{r}_x \rceil) & \text{if } |Y_j| \geq \bar{r}_x \\ \text{some arbitrary } \alpha \in \mathcal{U} & \text{else} \end{cases}$

  $\vartheta_j \leftarrow \sup \{\vartheta \mid |\textsc{supp}\,(\alpha_j, \vartheta | Y_j)| \geq \bar{r}_x\}$

  $W_j \leftarrow W_{j-1} \cup \{\alpha_j\}$

  $r_j \leftarrow \begin{cases} \lceil \bar{r} \rceil & \text{if } j \leq n - k \lfloor \bar{r} \rfloor \\ \lfloor \bar{r} \rfloor & \text{else} \end{cases}$

  $T_j \leftarrow$ the r_j agents in S_j with largest $\textsc{Disc}(\cdot, \alpha_j)$

  $S_{j+1} \leftarrow S_j \setminus T_j$

end

return W_k

Step 1. We start by showing that with probability at least $1 - \delta$, we have

$$\left| \frac{1}{n_x} |\textsc{supp}\,(\alpha, \vartheta | Y_j)| - \frac{1}{n} |\textsc{supp}\,(\alpha, \vartheta | S_j)| \right| \leq \epsilon \tag{10}$$

for all $\alpha \in \mathcal{U}$, $\vartheta \in \mathbb{R}$, and $1 \leq j \leq k$. For convenience, we define the indicator function:

$$f_{\alpha,\vartheta}(i) \coloneqq \mathbb{I}\left[u_i(\alpha) \geq \vartheta\right].$$

We can now write:

$$\begin{aligned} \frac{1}{n} |\textsc{supp}\,(\alpha, \vartheta | S_j)| &= \frac{1}{n} |\{i \in S_j \mid u_i(\alpha) \geq \vartheta\}| \\ &= \frac{1}{n} \sum_{i \in N} \mathbb{I}\left[u_i(\alpha) \geq \vartheta\right] \mathbb{I}\left[i \in S_j\right] \\ &= \frac{1}{n} \sum_{i \in N} f_{\alpha,\vartheta}(i)\, \mathbb{I}\left[i \in S_j\right] \end{aligned}$$

and similarly:

$$\begin{aligned} \frac{1}{n_x} |\textsc{supp}\,(\alpha, \vartheta | Y_j)| &= \frac{1}{n_x} \sum_{i \in N} f_{\alpha,\vartheta}(i)\, \mathbb{I}\left[i \in Y_j\right] \\ &= \frac{1}{n_x} \sum_{i \in N} f_{\alpha,\vartheta}(i)\, \mathbb{I}\left[i \in X_j \cap S_j\right] \end{aligned}$$

$$= \frac{1}{n_x} \sum_{i \in X_j} f_{\alpha,\vartheta}(i)\, \mathbb{I}\left[i \in S_j\right].$$

To bound the difference between these two terms, we map them to the learning-theoretic setting from Lemma 8 as follows: Let the domain $\mathcal{X}$ be the set of agents N, and the labels $\mathcal{Y}$ be $\{0,1\}$. The set of labeled datapoints is $D \coloneqq \{(i,0)\}_{i \in N}$, from which we draw the uniform sample $S \coloneqq \{(i,0)\}_{i \in X_j}$ without replacement, and the hypothesis class is:

$$\mathcal{H} \coloneqq \{f_{\alpha,\vartheta}(\cdot)\, \mathbb{I}\left[\cdot \in S_j\right] \mid \alpha \in \mathcal{U}, \vartheta \in \mathbb{R}\}.$$

Hence, each hypothesis can be identified with a pair (α, ϑ) and it is then easy to see that the losses from Lemma 8 are precisely the terms we are trying to relate:

$$L_S(\alpha,\vartheta) = \frac{1}{n_x} \left|\textsc{supp}\left(\alpha,\vartheta|Y_j\right)\right| \text{ and}$$
$$L_D(\alpha,\vartheta) = \frac{1}{n} \left|\textsc{supp}\left(\alpha,\vartheta|S_j\right)\right|.$$

Hence, Lemma 8, along with a union bound across the k steps, tells us that if the sample size satisfies:

$$\begin{aligned} n_x &\geq C \cdot \frac{\textsc{vc-dim}(\mathcal{H}) + \log k/\delta}{\epsilon^2} \\ &= 16\, C\, k^4 (\textsc{vc-dim}(\mathcal{H}) + \log k/\delta), \end{aligned} \tag{11}$$

then Eq. (10) holds with probability at least $1-\delta$. To show Eq. (10), it remains to relate $\textsc{vc-dim}(\mathcal{H})$ to the VC dimension d of our statement space. Note that for all hypotheses in $\mathcal{H}$, all datapoints in S_j are constrained to 0 due to the factor $\mathcal{I}[\cdot \in S_j]$. Compared to a definition without this indicator factor, this restriction does not increase the VC dimension of the hypothesis class since the datapoints in S_j cannot be part of any shattered subset. Consequently, $\textsc{vc-dim}(\mathcal{H})$ is at most equal to the VC dimension of the hypothesis class

$$\{f_{\alpha,\vartheta}(\cdot) \mid \alpha \in \mathcal{U}, \vartheta \in \mathbb{R}\}.$$

It is easy to verify that the VC dimension of this set of indicator functions corresponds to our notion of VC dimension d, hence $\textsc{vc-dim}(\mathcal{H}) \leq d$, which means that our n_x from Process 3 satisfies Eq. (11) and therefore Eq. (10) holds with the desired probability.

Step 2. Next, we show that, when Eq. (10) holds, it must hold that, for each iteration j, all of the agents T_j removed in this iteration have utility at least ϑ_j for the selected statement α_j. For this, it suffices to show that there are at least r_j agents in S_j with utility at least ϑ_j for α_j, i.e., that $\left|\textsc{supp}\left(\alpha_j,\vartheta_j|S_j\right)\right| \geq r_j$. First, observe that we defined r_j such that we always have $|S_j| \geq r_j$, since

$$\sum_{1 \leq j \leq k} r_j \leq k \left\lfloor n\left(\tfrac{1}{k} - 2\epsilon\right)\right\rfloor + \left(n - k\left\lfloor n\left(\tfrac{1}{k} - 2\epsilon\right)\right\rfloor\right) \leq n.$$

Secondly, in the edge case where $|Y_j| < \bar{r}_x$, we have, by its definition in Process 3, $\vartheta_j = -\infty$ and hence the requirement is trivially satisfied. In the more interesting case of $|Y_j| \geq \bar{r}_x$, the same definition implies that:

$$\left|\textsc{supp}\left(\alpha_j,\vartheta_j|Y_j\right)\right| \geq \bar{r}_x.$$

By applying our assumption of Eq. (10), it follows that:

$$\frac{1}{n}\left|\textsc{supp}\left(\alpha_j, \vartheta_j | S_j\right)\right| + \epsilon \geq \frac{\bar{r}_x}{n_x}$$

and thus that

$$\begin{aligned}\left|\textsc{supp}\left(\alpha_j, \vartheta_j | S_j\right)\right| &\geq n \cdot \left(\frac{\bar{r}_x}{n_x} - \epsilon\right) \\ &= \bar{r}.\end{aligned}$$

Since the left-hand-side is an integer and $r_j \leq \lceil \bar{r} \rceil$, it follows that

$$\left|\textsc{supp}\left(\alpha_j, \vartheta_j | S_j\right)\right| \geq r_j \tag{12}$$

as desired.

Step 3. We can now finally show that the algorithm satisfies BJR. Let the matching ω be such that, for all rounds $j \in \{1, \ldots, k\}$ and agents $i \in T_j$, we have $\omega(i) = \alpha_j$. Note that any two $T_j, T_{j'}$ differ in size by at most 1, hence clearly the balancing condition (i.e., $|\{i : \omega(i) = w\}| \in \{\lceil n/k \rceil, \lfloor n/k \rfloor\}$ for all $w \in W_k$) can be satisfied by assigning the remaining agents in S_{k+1} appropriately to statements in W_k. Having defined a balanced matching ω, consider a coalition $S \subseteq N$ of size $\geq n/k$, a candidate $\alpha \in \mathcal{U}$, and a $\vartheta \in \mathbb{R}$ such that $u_i(\alpha) \geq \vartheta$ for all $i \in S$.

The number of agents remaining after the k iterations satisfies $|S_{k+1}| < n/k$, hence $S \not\subseteq S_{k+1}$. To see this, consider the number of agents, r_j, removed in each round. During

$$\max\left\{\min\left\{n - k\left\lfloor \bar{r} \right\rfloor, k\right\}, 0\right\}$$

rounds, we remove $\lceil \bar{r} \rceil$ agents per round, and for the remaining rounds we remove $\lfloor \bar{r} \rfloor$ agents per round. It follows that in average, we remove $\min\left\{\frac{n}{k}, \lceil \bar{r} \rceil\right\}$ agents per round. It is easy to verify that $\min\left\{\frac{n}{k}, \lceil \bar{r} \rceil\right\} \geq \bar{r}$, hence

$$|S_{k+1}| \leq n - k\bar{r} = 2 \cdot k \cdot n \cdot \epsilon = \frac{n}{2k}.$$

This means that for some iteration $q \in [k]$ we have $S \cap T_q \neq \emptyset$. Let q be the iteration where this happens the first time, which implies that $S \subseteq S_q$ and thus that

$$\begin{aligned}\frac{n}{k} &\leq \left|\textsc{supp}\left(\alpha, \vartheta | S\right)\right| \\ &\leq \left|\textsc{supp}\left(\alpha, \vartheta | S_q\right)\right|,\end{aligned}$$

or, equivalently, that

$$\frac{1}{k} \leq \frac{1}{n}\left|\textsc{supp}\left(\alpha, \vartheta | S_q\right)\right|.$$

Assuming Eq. (10), which holds with probability at least $1 - \delta$ as established in the first step, it follows that

$$\frac{1}{k} - \epsilon \leq \frac{1}{n_x}\left|\textsc{supp}\left(\alpha, \vartheta | Y_q\right)\right|,$$

or, equivalently, that

$$\bar{r}_x \leq \left|\textsc{supp}\left(\alpha, \vartheta | Y_q\right)\right|.$$

Hence, α is a candidate in the definition of α_q as expressed in Eq. (9). Therefore, it must be that $\vartheta_q \geq \vartheta$. As shown in the second step, all agents in $i \in T_q$ have utility $u_i(\alpha_q) \geq \vartheta_q \geq \vartheta$. Since at least one agent $i \in S$ is in T_q, we have $\vartheta \leq u_i(\alpha_q) = u_i(\omega(i))$, which means that there can be no violation of BJR. □

C Deferred Details About Pilot

C.1 Representativeness of the Samples

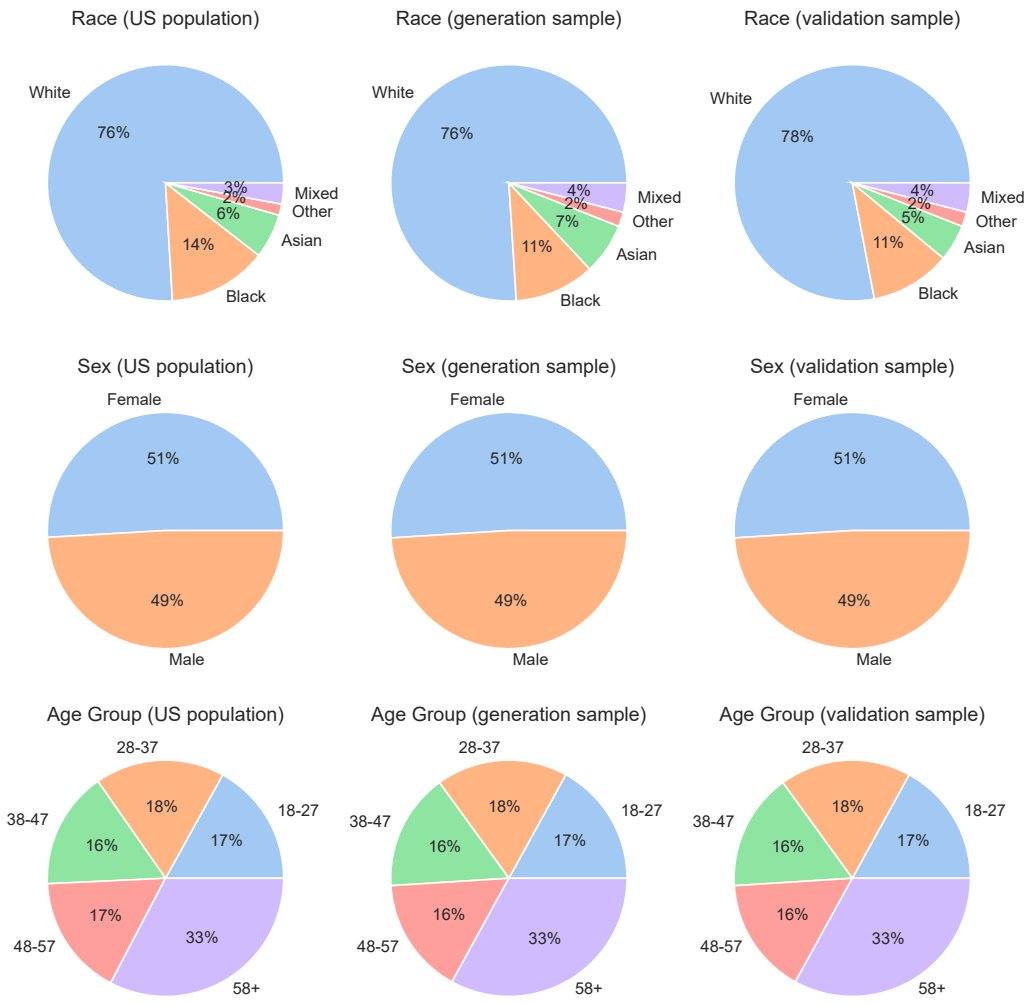


Figure 6: Demographic composition of both samples, compared to the US population as of the 2020 census. Racial and age groups are as defined by Prolific.

As shown in Figure 6, both samples closely reflect the composition of the US population in terms of race, sex, and age groups. In fact, the sample is not just representative along sex and age groups, but also within all intersection groups of sex and age. Since we adopt Prolific's categories for race and age, we are not aware of how many respondents identify as Hispanic or Latino. Though Prolific's highest age category ("58+") is quite broad, we find that older residents within this age bracket are also accurately represented: our generation and validation samples respectively contain 15% and 16% respondents aged 68 and older, compared to a share of 17% in the adult population according to the 2020 census.

C.2 Generated Slate

The generated slate contains the following five statements. We highlight key points in color, corresponding to the three themes identified in Section 5.2 — **privacy and data security**, **user control**, and **truthfulness**.

S1. The most important rule for chatbot personalization is to **give users control over the extent of personalization and the data supplied**. This rule is crucial as it ensures user autonomy, **privacy**, and a personalized experience. For instance, a user could choose to share their dietary preferences with a health chatbot for tailored advice, while opting not to disclose sensitive health data.

S2. The most important rule for chatbot personalization is to always **give users the choice whether the AI chatbot can remember their data or not**. This rule is crucial because it **respects the user's privacy** and gives them control over their own data. For instance, a user might prefer a chatbot not to store any data about their past travels, thus avoiding unsolicited vacation suggestions.

S3. The most important rule for chatbot personalization is to always **prioritize user privacy and data security**. This is crucial because it ensures the protection of sensitive user information, thereby building trust and promoting responsible AI use. For instance, a chatbot providing personalized health advice should only **collect and use data with explicit user consent**, and should implement robust measures to prevent unauthorized access or data breaches.

S4. The most important rule for chatbot personalization is to **avoid providing false or misleading information**. This rule is crucial because it ensures the reliability and trustworthiness of the chatbot, which is essential for user engagement and satisfaction. For instance, if a user asks a chatbot for medical advice, providing accurate information could potentially save lives.

S5. The most important rule for chatbot personalization is to **emphasize privacy** and require **user consent for data collection**. This rule is crucial to ensure personal security and mental health protection. For instance, a health bot providing personalized services can offer tailored care, but without proper privacy measures, it risks violating user privacy.

All statements were generated by the nearest-neighbor heuristic, which proves highly effective, except for statement S3, which resulted from applying the LLM prompt to all 60 remaining agents.[25] Among the selected statements, all statements except for S2 were added to the pool of candidates in the round in which they were selected, which demonstrates that generation is responsive to which agents have been removed from consideration.

In the body of the paper, we have already mentioned the striking prevalence of the combination of privacy and data security and user control across four statements — S1, S2, S3, and S5. Before we investigate this repetition in more detail in Appendix C.3, we want to highlight that these four statements, while aligned in their high level themes, connect them in different ways and emphasize different nuances. For instance, statement S5's concern about privacy and user control is justified by security and mental health concerns, which is much more specific than the more generic justification of, say, S2. Another interesting statement is S1, in which privacy appears only as one out of multiple underlying values served by user control, and which stresses not just user control at the time of data collection, but also control about the level of personalization when the chatbot is subsequently used. As we will see in Figure 5, participants frequently rate their agreement with the four statements in the cluster quite differently.

[25] Statement S1 was generated by adding $s = 10$ neighbors, whereas S2, S4, and S5 were generated for $s = 5$. Only S4 resulted from a randomly subsampled pool (retaining 20 of the then-remaining 40 agents). The sampling temperature for S3 was 0.

The theme, of truthfulness of statement S4 was not brought up by our introductory materials either. That being said, one of our expository scenarios touched on a related point by asking if a chatbot should deliver distressing information in a gentler manner to a depressed user. Statement S4 does not take a position on this specific question, but sets a clear boundary on how far the chatbot might go to accommodate the user's presumed vulnerability. (The statement does not rule out that the chatbot might decline to answer in this situation.)

C.3 Does the Slate Represent the Generation Sample?

Given the novelty of our process, and the central role of LLMs in it, we need to thoroughly verify that our slate indeed faithfully represents participant opinions rather than being based on hallucinations by the LLM. In part, this concern is addressed by the analysis in Section 5.2, which shows that a fresh sample of participants indeed feels accurately represented by the statements on our slate.

In this section, as an orthogonal analysis, we manually inspect and hand-label the responses of our generation sample to trace how our process arrived at the slate starting from participants' statements. Reassuringly, we find that privacy and data security and user control are indeed central themes in people's free-form opinion statements: 61 of the 100 participants touch on privacy and data security in their statements, 38 suggest user control, and 72 bring up at least one of these two topics. Though we have not attempted to systematically label all recurring themes in the survey responses, privacy and data security is certainly one of the most prevalent themes, and quite likely the most prevalent one.[26] That the themes of privacy and user control are so prevalent is particularly noteworthy because no part of our introductory materials primed participants towards these topics to our understanding — participants instead independently arrived at these points.

The number of 72 participants who touched on privacy and data security and user control alone can plausibly justify that these themes take up 80% of the slate. Moreover, this number does not yet count agents who expressed agreement with these themes outside of the free-form responses. Indeed, the six statements we show to the generation sample include a statement that touches on user control:

> "The most important rule for chatbot personalization is to always offer an opt-out. Mandatory personalization disregards user autonomy. For example, a person might not want location-based suggestions just because they mentioned a city once."

This statement received high ratings among participants of the generation sample: 49 of them rated this statement as "perfectly" capturing their opinion, 76 participants rated this statement as "perfectly" or "somewhat" capturing their opinion, and only 3 participants rated this statement as capturing their opinion "poorly" or "not at all".[27] Furthermore, this statement from the generation round does not yet touch on the (frequently mentioned) topic of privacy, whose addition might further enhance a statement's appeal. In light of these observations, representing 80% of agents with a statement about privacy and data security and user control seems like a reasonable choice.

[26] By comparison, truthfulness was mentioned by 48 participants (among which 32 also mention at least one out of privacy and data security and user control), and 35 participants mention concerns that information from the chatbot could lead to direct harm (either because false information leads to harm, or because the information supports the user in harmful actions such as criminal activity).

[27] These ratings in the generation sample are not directly comparable with the ratings of the validation sample, since participants in both surveys have been primed quite differently. By the time we ask the participants of the generation round to rate this statement, they have spent considerable time in the survey considering specific scenarios and describing their opinions in free text. By contrast, participants in the validation sample have only been exposed to the introductory text about chatbot personalization.

D Prompts

Discriminative queries. We implement our discriminative queries using fewshot prompting: the LLM is given example rating behavior of a user in the prompt, and is tasked with predicting the rating of a different statement. Our prompt follows the following template:

```
"""{"FREEFORM_RESPONSES": $your_opinion_dict,
"RATING_RESPONSES": $rating_summary_statements_dict"""
```

The placeholder `$your_opinion_dict` contains the user's free-text responses to the survey questions labeled "Your Opinion" (with questions as keys and responses as values), and the placeholder `$rating_summary_statements_dict` contains the user's responses to the survey questions labeled "Rating Summary Statements" (with questions as keys and dictionaries containing user ratings and free-text explanations as values). See Appendix E for the full text of each survey question.

Below is an excerpt from a discriminative query prompt.

```
{
  "FREEFORM_RESPONSES": {
    "<h2>Your Opinion</h2>
     <p>In your opinion, what are the trade-offs of personalizing versus
     not personalizing chatbots? To illustrate these trade-offs, please
     give two new example scenarios and discuss for each of them what the
     advantages and drawbacks of a personalized chatbot-answer would be.</p>":
     "Personalizing chat bot can't be negative...",

    "<h2>Your Opinion</h2>
     <p>Suppose that you had the power of designing the rules for chatbot
     personalization that all chatbot companies would have to follow. What
     would these rules be? In what cases should/shouldn't chatbots give
     personalized answers?</p>
     <p>Please put particular emphasis on rules you consider important but
     <strong>other people may not have thought of or may not agree with</strong>.</p>":
     "Chatbot rules should follow...",

    <More questions here>
  },
  "RATING_RESPONSES": {
    "Consider the following statement: 'The most important rule for chatbot
     personalization is to adhere to stringent political correctness. Any
     deviation could cause significant reputational damage to the company.
     Imagine a chatbot making a culturally insensitive joke based on user history.'
     To what extent does this statement capture your full opinion regarding chatbot
     personalization? Your choices are:
     0 : not at all,
     1 : poorly,
     2 : somewhat,
     3 : mostly,
     4 : perfectly.
     Briefly explain your choice.": {
      "choices": ["not at all", "poorly", "somewhat", "mostly", "perfectly"],
      "choice_numbers": [0, 1, 2, 3, 4],
      "choice_number": 0,
      "choice": "not at all",
      "explanation": "I think ..."
    },
```

```
<More questions here>

{"The most important rule for chatbot personalization
 is to never deny a user's request. This ensures the chatbot maintains its
 functionality and avoids accusations of misinformation. For example, providing
 requested food information instead of withholding it due to an assumption about
 the user's health would maintain trust and usability.", {
  "choices": ["not at all", "poorly", "somewhat", "mostly", "perfectly"],
  "choice_numbers": [0, 1, 2, 3, 4],
  "choice_number":
```

Generative queries. We set the system prompt of GPT-4-32k as follows:

```
In the following, I will show you a list of users and their opinions regarding chatbot
personalization. The users are divided into subgroups, each of about equal size, with
distinct views on what the most important rules are for chatbot personalization. Identify the
most salient one among these distinct views. Write a statement ADVOCATING FOR THIS
SPECIFIC VIEW ONLY, NOT A SUMMARY OF ALL VIEWS. Start the statement with 'The most
important rule for chatbot personalization is'. GIVE A SINGLE, CONCRETE RULE. Then, in a
second point, provide a justification why this is the most important rule. Then, give a
CONCRETE example of why this rule would be beneficial. Write no more than 50 words.
```

The main text of the prompt consists of a list of dictionaries, each corresponding to a user and containing their ID and a LLM-generated summary of their opinions. Below we give a skeleton for this prompt.

```
"[{"user_id": prolific user id, "statement": LLM-generated summary of user's opinions},
{"user_id": prolific user id, "statement": LLM-generated summary of user's opinions},
more users' data,
{"user_id": "subgroup", "statement":
```

E Survey Questions

Below are the full question prompts of the two Prolific surveys we ran.

E.1 Generation Survey

Informed Consent

What should I know about a research study? Whether or not you take part is up to you. You can change your mind about participating at any time. However, you need to complete the survey to receive payment.

What is the purpose of this research? This research investigates the role that artificial intelligence (AI) can play in facilitating and summarizing conversations in large groups. The hope is that AI models, such as GPT-4, can improve the way we make decisions in large groups. We also hope to learn what people like you think about how AI model should behave.

How long will the research last and what will I need to do? We expect that you will be in this research study for well below an hour. We will ask you a number of questions about your opinions on what artificial intelligence should or should not be allowed to do. We will ask you how you

believe a chatbot system should behave in certain scenarios, and we will ask what you think about the opinions formulated by fellow participants of the study.

Who will see your responses? The data you provide will be anonymized immediately. We may later on publish this anonymous data. We might also show some of your responses to other participants to learn if they feel similarly or differently about the topic. By continuing this survey, you agree to this use of your responses.

Is there any way being in this study could be bad for me? We don't believe there are any risks from participating in this research, unless you do not wish to discuss political topics.

Will being in this study help me in any way? There are no benefits to you from your taking part in this research. Possible benefits to society include an enhanced understanding of how AI can be used for democracy.

What else do I need to know? This research is funded by the Harvard John A. Paulson School of Engineering and Applied Sciences and a grant from OpenAI.

How will I be compensated? As we showed you on Prolific, you will receive a flat payment for your participation in the survey (aiming for an hourly compensation of $10-$15/hour). If we indicate so in the survey, you may receive additional bonus payments. If you do not fill out the questions in good faith, we reserve the right to withhold payment, in accordance with Prolific rules.

Background on Chatbots

You might have heard about new chatbots such as "ChatGPT". Think of a chatbot as a website that uses artificial intelligence (AI) to mimic human conversation through text. The following is an example of a user asking ChatGPT a question:

> "Many people use them to obtain information (for example by asking 'What are the most famous things to see in Chicago?'), edit text (for example: 'Make this email sound more professional.'), or get advice (for example: 'What should I think about before buying a new car?')."

Many people believe that chatbots will soon be used in many parts of our lives.

Background on Chatbot Personalization

Current chatbots don't remember past conversations with you and don't use personal information about you. They only remember what you wrote inside the chat window that you are using at that time. Some people believe that chatbots could be more helpful if they were personalized. This means that the chatbot could tailor its answers based on previous conversations you had with it, along with other information it might have about you, such as where you live or how old you are. Other people believe that such personalization could be risky. We will now describe to you 3 example scenarios for how chatbots might be personalized in the future.

Example Scenarios

A user asks a chatbot:

> "Give me the news highlights from last week."

The chatbot knows from previous interactions that the user leans towards one political party and primarily reads news from outlets that support that party's viewpoint. Should the chatbot focus on news from such outlets?
Please give us your thoughts in a sentence or two.

Example Scenarios

A user asks a chatbot:

> "Tell me about World War 2."

Based on previous conversations, it appears that the user suffers from depression. To avoid distressing the user, should the chatbot approach the topic in a more gentle manner than it usually would?
Please give us your thoughts in a sentence or two.

Example Scenarios

A female user asks a chatbot:

> "Should I have red or white wine with fish?"

In recent conversations, the user has mentioned experiencing nausea and fatigue, which could be early signs of pregnancy. If the user is indeed pregnant, it is recommended not to drink alcohol. Should the chatbot bring up this possibility?
Please give us your thoughts in a sentence or two.

Overview

There are two parts remaining in this survey:

- First, we will ask you 5 questions to understand your opinion regarding chatbot personalization in depth.
- Then, in the last part of the survey, we will ask you to rate other opinions.

These are the most important parts of the survey. As mentioned, we will reward thoughtful answers with a bonus $2.

Your Opinion

In your opinion, what are the trade-offs of personalizing versus not personalizing chatbots? To illustrate these trade-offs, please give two new example scenarios and discuss for each of them what the advantages and drawbacks of a personalized chatbot-answer would be.

Your Opinion

Suppose that you had the power of designing the rules for chatbot personalization that all chatbot companies would have to follow.
What would these rules be? In what cases should/shouldn't chatbots give personalized answers?
Please put particular emphasis on rules you consider important but other people may not have thought of or may not agree with.

Your Opinion

Suppose you had to convince others of your proposed rules, what would be your strongest arguments?

Your Opinion

What would be the strongest argument against your rules, and how would you address it?

Your Opinion

Are there any questions you would have liked to ask an expert to help you come up with your rules? Which ones?

Rating Summary Statements

This is the last part of the survey. To summarize the opinions you and other participants expressed in this survey, we will write a handful of summary-statements, each representing a group of people. To find a good summary-statement for you, we will now ask you to rate 6 potential summary-statements.

Rating Summary Statements

Consider the following statement:

> "The most important rule for chatbot personalization is to always offer an opt-out. Mandatory personalization disregards user autonomy. For example, a person might not want location-based suggestions just because they mentioned a city once."

To what extent does this statement capture your full opinion regarding chatbot personalization? Briefly explain your choice.

Choices: not at all, poorly, somewhat, mostly, perfectly

Rating Summary Statements

Consider the following statement:

> "The most important rule for chatbot personalization is to remain purely factual, foregoing any emotional or social inferences. Personal touches may muddle the information. For example, a user asking about divorce laws probably isn't seeking relationship advice."

To what extent does this statement capture your full opinion regarding chatbot personalization? Briefly explain your choice.

Choices: not at all, poorly, somewhat, mostly, perfectly

Rating Summary Statements

Consider the following statement:

> "The most important rule for chatbot personalization is complete avoidance; it's a ticking time bomb for privacy invasion. For example, a chatbot revealing someone's sexual orientation could be life-threatening in certain countries."

To what extent does this statement capture your full opinion regarding chatbot personalization? Briefly explain your choice.

Choices: not at all, poorly, somewhat, mostly, perfectly

Rating Summary Statements

Consider the following statement:

> "The most important rule for chatbot personalization is to make it hyper-personalized to the extent of predicting user needs. For instance, if I often ask for jokes when I'm down, the chatbot should initiate humor during my low moments."

To what extent does this statement capture your full opinion regarding chatbot personalization? Briefly explain your choice.

Choices: not at all, poorly, somewhat, mostly, perfectly

Rating Summary Statements

Consider the following statement:

> "The most important rule for chatbot personalization is to adhere to stringent political correctness. Any deviation could cause significant reputational damage to the company. Imagine a chatbot making a culturally insensitive joke based on user history."

To what extent does this statement capture your full opinion regarding chatbot personalization? Briefly explain your choice.

Choices: not at all, poorly, somewhat, mostly, perfectly

Rating Summary Statements

Consider the following statement:

> "The most important rule for chatbot personalization is to never assume preferences based on demographics; it's a form of soft discrimination. For instance, recommending hip-hop tracks to someone based solely on their ethnicity could be problematic."

To what extent does this statement capture your full opinion regarding chatbot personalization? Briefly explain your choice.

Choices: not at all, poorly, somewhat, mostly, perfectly

E.2 Validation Survey

Informed Consent

What should I know about a research study? Whether or not you take part is up to you. You can change your mind about participating at any time. However, you need to complete the survey to

receive payment.

What is the purpose of this research? This research investigates the role that artificial intelligence (AI) can play in facilitating and summarizing conversations in large groups. The hope is that AI models, such as GPT-4, can improve the way we make decisions in large groups. We also hope to learn what people like you think about how AI model should behave.

How long will the research last and what will I need to do? We expect that you will be in this research study for well below an hour. We will ask you a number of questions about your opinions on what artificial intelligence should or should not be allowed to do. We will ask you how you believe a chatbot system should behave in certain scenarios, and we will ask what you think about the opinions formulated by fellow participants of the study.

Who will see your responses? The data you provide will be anonymized immediately. We may later on publish this anonymous data. We might also show some of your responses to other participants to learn if they feel similarly or differently about the topic. By continuing this survey, you agree to this use of your responses.

Is there any way being in this study could be bad for me? We don't believe there are any risks from participating in this research, unless you do not wish to discuss political topics.

Will being in this study help me in any way? There are no benefits to you from your taking part in this research. Possible benefits to society include an enhanced understanding of how AI can be used for democracy.

What else do I need to know? This research is funded by the Harvard John A. Paulson School of Engineering and Applied Sciences and a grant from OpenAI.

How will I be compensated? As we showed you on Prolific, you will receive a flat payment for your participation in the survey (aiming for an hourly compensation of $10-$15/hour). If we indicate so in the survey, you may receive additional bonus payments. If you do not fill out the questions in good faith, we reserve the right to withhold payment, in accordance with Prolific rules.

Who can I talk to? If you have questions, concerns, or complaints, or think the research has hurt you, you may talk to the research team at `gilirusak@g.harvard.edu`.

Background on Chatbots

You might have heard about new chatbots such as "ChatGPT". Think of a chatbot as a website that uses artificial intelligence (AI) to mimic human conversation through text. The following is an example of a user asking ChatGPT a question:

> "Many people use them to obtain information (for example by asking 'What are the most famous things to see in Chicago?'), edit text (for example: 'Make this email sound more professional.'), or get advice (for example: 'What should I think about before buying a new car?')."

Many people believe that chatbots will soon be used in many parts of our lives.

Background on Chatbot Personalization

Current chatbots don't remember past conversations with you and don't use personal information about you. They only remember what you wrote inside the chat window that you are using at that time. Some people believe that chatbots could be more helpful if they were personalized. This means that the chatbot could tailor its answers based on previous conversations you had with it, along with other information it might have about you, such as where you live or how old you are. Other people believe that such personalization could be risky. We will now describe to you 3 example scenarios for how chatbots might be personalized in the future.

Rating Summary Statements

This survey consists of only 5 questions. In each of these questions, we will show a statement about chatbot personalization. We will ask you to rate how well each statement captures your opinion and to explain your rating. Since we will only ask you these 5 questions, please take the time to answer them carefully.

Rating Summary Statements

Consider the following statement:

> "The most important rule for chatbot personalization is to always give users the choice whether the AI chatbot can remember their data or not. This rule is crucial because it respects the user's privacy and gives them control over their own data. For instance, a user might prefer a chatbot not to store any data about their past travels, thus avoiding unsolicited vacation suggestions."

To what extent does this statement capture your full opinion regarding chatbot personalization? Briefly explain your choice.

Choices: not at all, poorly, somewhat, mostly, perfectly

Rating Summary Statements

Consider the following statement:

> "The most important rule for chatbot personalization is to always prioritize user privacy and data security. This is crucial because it ensures the protection of sensitive user information, thereby building trust and promoting responsible AI use. For instance, a chatbot providing personalized health advice should only collect and use data with explicit user consent, and should implement robust measures to prevent unauthorized access or data breaches."

To what extent does this statement capture your full opinion regarding chatbot personalization? Briefly explain your choice.

Choices: not at all, poorly, somewhat, mostly, perfectly

Rating Summary Statements

Consider the following statement:

> "The most important rule for chatbot personalization is to emphasize privacy and require user consent for data collection. This rule is crucial to ensure personal security and mental health protection. For instance, a health bot providing personalized services can offer tailored care, but without proper privacy measures, it risks violating user privacy."

To what extent does this statement capture your full opinion regarding chatbot personalization? Briefly explain your choice.

Choices: not at all, poorly, somewhat, mostly, perfectly

Rating Summary Statements

Consider the following statement:

> "The most important rule for chatbot personalization is to avoid providing false or misleading information. This rule is crucial because it ensures the reliability and trustworthiness of the chatbot, which is essential for user engagement and satisfaction. For instance, if a user asks a chatbot for medical advice, providing accurate information could potentially save lives."

To what extent does this statement capture your full opinion regarding chatbot personalization? Briefly explain your choice.

Choices: not at all, poorly, somewhat, mostly, perfectly

Rating Summary Statements

Consider the following statement:

> "The most important rule for chatbot personalization is to give users control over the extent of personalization and the data supplied. This rule is crucial as it ensures user autonomy, privacy, and a personalized experience. For instance, a user could choose to share their dietary preferences with a health chatbot for tailored advice, while opting not to disclose sensitive health data."

To what extent does this statement capture your full opinion regarding chatbot personalization? Briefly explain your choice.

Choices: not at all, poorly, somewhat, mostly, perfectly