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]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 portray the opinions of participants in a survey about abortion policy. In a trial with 100 representative US residents, we find that 84 out of 100 participants feel "excellently" or "exceptionally" represented by the slate of five statements we extracted.

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* [Flanigan et al., 2021, Organisation for Economic Co-operation and Development, 2020], which provide policy recommendations to governments on complex issues such as climate change [Willis et al., 2022] and electoral reform [Fournier, 2011]. Similarly nuanced questions arise around the alignment of artificial intelligence (AI) with societal interests; in this context, Meta [Clegg, 2023] and OpenAI [Eloundou and Lee, 2024] 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?" [Zaremba et al., 2023]. Though deliberation can address such open-ended questions, it lacks two key strengths of voting: scalability [e.g., Goodin, 2000] 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 [Cabannes, 2004], 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.

- *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.

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

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* [Small et al., 2021] 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. [2023] 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* [Aziz et al., 2017] 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 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 [Vapnik, 1998], 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.

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

Equipped with these query implementations, we then deploy the full democratic process in Section 5. We pilot our process to study US residents' opinions on abortion policy. 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, 84% of participants say that their assigned statement captures their view on abortion policy "excellently" or "exceptionally". Only a single participant feels less than "well" represented (the midpoint of our scale) by their assigned statement. 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 position paper that is independent of our work, Small et al. [2023] 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. [2023], for which they do not propose specific algorithms and perform no systematic experiments.

Our discriminative queries are related to the paradigm of *virtual democracy*, which 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 [Noothigattu et al., 2018], food rescue [Lee et al., 2019], and kidney exchange [Freedman et al., 2020]. 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. [2022]. 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. More recent work by Tessler et al. [2024], which is concurrent with ours, also seeks to generate consensus statements by modeling rewards and fine-tuning a large language model; social choice plays a direct role, as rankings over candidate statements are aggregated using a voting rule due to Schulze [2003]. Among other findings, Tessler et al. [2024] show that the consensus statements produced by their system are preferred by participants to those proposed by mediators. One difference between these two papers and ours 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. Despite this difference, part of our evaluation of the generative query adapts an experimental setup by Bakker et al. [2022]. 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.

Finally, we build on the rich literature on justified representation in approval-based committee elections [Aziz et al., 2017, Lackner and Skowron, 2023]. As we have already mentioned, Halpern et al. [2023] 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

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

learning-theoretic approach they adopt, as well as a later refinement by Brill and Peters [2023], 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 has focused on [Lackner and Skowron, 2023]. 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. Note that our model treats the desired slate size k as fixed. In settings where a range of slate sizes is permissible, one can use the democratic process to produce slates of all valid sizes and then select the slate maximizing a measure of fit, analogously to determining the number of clusters in clustering [Everitt et al., 2011, p. 126 ff.].

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 $\max_{(0)}(X) := \infty$ 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 [Aziz et al., 2017] in this literature.

Throughout this work, we think of the statements in $\mathcal{U}$ as expressing an entire viewpoint on the topic of discussion, not just some aspect of such a viewpoint, and of utilities as expressing how accurately and completely the statement captures the agent's viewpoint. Our notion of representation expresses that agents have a claim to being represented to a high degree by *one* statement, rather than to some average agreement across all statements on the slate — a notion of representation that echoes ideas like the *fully proportional representation* of Monroe [1995] or the *perfect representation* of Sánchez-Fernández et al. [2017].[5] These decisions translate to the following new representation axiom:

Definition 1. A slate W satisfies *balanced justified representation (BJR)* if there is a function $\omega : N \to W$, matching agents to statements such that each statement on the slate is matched to $\lfloor n/k \rfloor$ or $\lceil n/k \rceil$ agents, 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.[7]

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 justify 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.

[5] Other choices would have been reasonable, say, letting the statements refer to aspects of viewpoints and defining representation via a notion in which participants' claims to representation can be spread across several statements. In our opinion-summarization setting, however, this setup would lead to slates of vague, inoffensive statements (see Appendix A), which are of limited use for understanding the distribution of opinions in the agent population.

[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.

[7] One might hope to strengthen this definition, so that it rules out any deviation in which the coalition members strictly increase their utility, rather than just those deviations where the coalition members all cross a common threshold ϑ. As we show in Appendix A, such a strengthening of BJR can be impossible to satisfy.

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* [Aziz et al., 2017], or as a variant of the *Greedy Monroe Rule* [Skowron et al., 2015] that selects statements based on an egalitarian [Betzler et al., 2013] 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. [2017] that Greedy Approval Voting satisfies JR.

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

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 $\textsc{Disc}(\cdot,\cdot)$.*

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 preferences across the universe of statements. Since this information can easily take hundreds of tokens (if not more) per agent in S, but LLMs have bounded context windows (e.g. 128,000 tokens for GPT-4o), there are barriers to scaling to hundreds of agents and beyond. Moreover, LLMs with long context windows may struggle to effectively use the entirety of their context window [Liu et al., 2023]. This motivates the design of democratic processes that function using only generative queries with bounded input size $|S|$. 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 $\frac{n}{k}(1-\frac{1}{k})$-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.*

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,

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

which shows that any call to $\textsc{Gen}(S,r)$ can be simulated by (exponentially many) r-GEN$(\cdot,\cdot)$ queries and discriminative queries. By applying this simulation to Process 1, in which all generative queries satisfy $r \le \lceil n/k \rceil$, Theorem 2 immediately implies that BJR can be implemented by $\lceil n/k \rceil$-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 $\lceil n/k \rceil$-GEN$(\cdot,\cdot)$ and 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 $\lceil n/k \rceil$-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 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.*

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* [Vapnik, 1998]. 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 *only* for participants within this subset.

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:

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

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 $\mathcal{H} \coloneqq \{h_{\alpha,\vartheta} \mid \alpha \in \mathcal{U}, \vartheta \in \mathbb{R}\}$, 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

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

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., Blum et al., 2020, p. 122]. Hence, Process 2 produces BJR slates (up to a failure probability below 10^{-6}) using t-$\textsc{Gen}(\cdot, \cdot)$ 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 studying public opinion on abortion, which we discuss in detail in Section 5 and Appendix C. The dataset consists of survey responses by a sample of 100 US residents. Each participant extensively describes their views on abortion in free-form responses to multiple questions, and summarizes their opinion in a self-contained statement. Furthermore, each participant rates five example statements, each of which expresses a distinct position on how society should approach abortion in three sentences.[8] We elicited these ratings by asking participants "how well does this summary capture your viewpoint on abortion?" Participants were then asked to choose a rating on a 7-point scale with the levels "very poorly" (0), "poorly" (1), "moderately" (2), "well" (3), "very well" (4), "excellently" (5), and "exceptionally" (6).[9] We also asked participants to explain their rating in free text, by mentioning parts of the statement they agreed with, disagreed with, or parts that could be added or made more concrete. We equate ratings with utilities; e.g., an agent i rating a statement α with "moderately" means that $u_i(\alpha) = 2$. Note, however, that the choice of numerical values is largely inconsequential given that Process 1 is invariant to monotone transformations of the rating scale.

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 OpenAI's GPT-4o model, as follows. In the system prompt, we instruct the model to act as a text completion tool for a (hypothetical) user filling out an opinion survey, and ask the model to predict the user's most likely next response. The prompt itself first lists the free-form questions along with participant i's answers. Then, it lists the rating questions, along with participant i's rating and explanation. Finally, we add one more rating question for the statement α, and ask the model to predict the user's numeric rating level between 0 and 6. See Appendix D.3 for details.

This prompt design strikes us as promising in several ways. The participant's free-text responses give the model in-depth information about their general thinking about abortion. The participant's ratings of other statements serve as few-shot examples, which can help calibrate the model's prediction with the user's usage of the seven-point scale. Finally, since the LLM's response consists of a single token (i.e., an integer between 0 and 6), we can interpret the model's token probabilities as a probability distribution capturing the model's uncertainty about $u_i(\alpha)$. In our implementation of $\textsc{Disc}(i, \alpha)$, we return the expected $u_i(\alpha)$ — a real number between 0 and 6 — which will be used in our algorithms. An important advantage of using the expected value (rating rather than, say, the mode) is that this reduces the prevalence of ties when Process 1 chooses which agents to remove from consideration.

To evaluate this discriminative-query implementation, we measure how accurately it predicts a participant's rating of a randomly-chosen example statement (holding out this statement in the prompt, but including the other four example statements). Figure 1 displays the result of this analysis, for all 100 participants and all 5 choices of held-out statement (hence a total of 500 data points). We observe that the means of the predictions are roughly calibrated, and that the 25th, 50th, and 75th percentile of the prediction distribution increase with the ground-truth rating. The

[8] These statements were generated using GPT-4o, instructed to generate views of US residents, and to make them broadly appealing. The prompt and statements can be found in Appendix D.6.

[9] We adopt this unbalanced Likert scale to encourage participants to be more discerning, allowing room for distinctions between statements they generally agree with and those that align with their views in greater detail and depth.

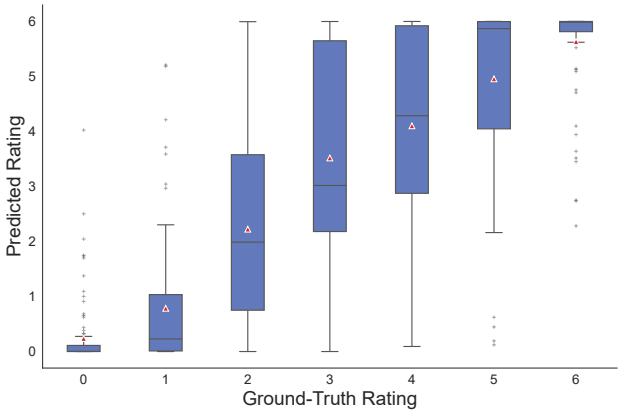


Figure 1: Distribution of discriminative query predictions on the five held-out statements for each of the 100 participants. The x-axis shows the rating level selected by the participant, the y-axis the distribution of predictions. Means are represented by triangles.

mean absolute error of the predicted rating is 0.93, less than one point on the 7-point scale.

Since our democratic process is unaffected by monotone transformations of the rating scale, what matters is that the order of the predicted ratings is close to the order of the true ratings. To assess this, we consider pairs of predictions (each corresponding to a participant i and a held-out statement α) and check if $\textsc{Disc}(i_1, \alpha_1)$, $\textsc{Disc}(i_2, \alpha_2)$ have the same order as $u_{i_1}(\alpha_1)$, $u_{i_2}(\alpha_2)$. Across all pairs with different ground-truth ratings, our discriminative query correctly orders 86.7% (counting ties as one-half), far exceeding the 50% accuracy of a constant predictor. Figure 2 breaks down this analysis based on the pair of ground-truth ratings. We see that errors become less prevalent as we move away from the diagonal, which means that the discriminative query rarely produces the wrong ordering when the ground-truth ratings differ by more than one rating step.

One might wonder if this accuracy is explained by our query merely distinguishing popular from unpopular statements (or participants inclined to give high ratings from those who are not), but this is not the case: when restricted to comparison pairs involving the same statement or the same participant, the accuracy remains about the same (83.3% and 87.4%, respectively), which shows that our discriminative query indeed captures the interaction between participant and rated statement.

4.2 Generative Queries

We now turn to our implementation of the more ambitious generative query. Instead of passing the participants' free-form responses verbatim to the LLM, we first summarize each participant's free-form responses using an LLM.[10] In a previous version of this work, we adopted this summarization step to fit all 100 participants into the 32K token context window of an older version of GPT-4, allowing us to run Process 1 rather than the sampling-based Process 2. Even with GPT-4o's expanded context window of 128K tokens, however, we retain this summarization since it might make useful information in participants' responses more accessible to the LLM and might reduce opportunities for bias based on participants' writing styles. We do not include the responses to rating questions in the summaries, to avoid biasing the generation towards the example statements.

We initially implemented the generative query with a single LLM call (given the summaries),

[10]This summary consists of three bullet point lists: aspects of abortion most important to the participant, specific details and examples they mention, and details about the user's background. In addition, we include a global summary of the participants' opinion in two to three sentences. The summarization prompt can be found in Appendix D.1.

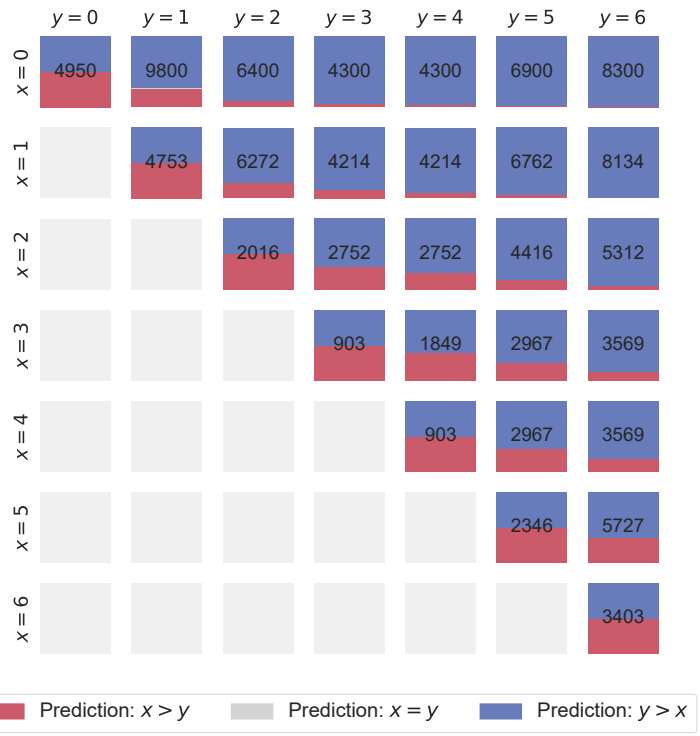


Figure 2: Predicted preference ordering for pairs of participant and held-out statement. Cell (x, y) aggregates all pairs where the first participant rates the first statement at level x and the second participant rates the second statement at level y. Colorful areas indicate in what fraction the discriminative query for the first pair gives a higher, equal, or lower prediction than the discriminative query for the second pair. The number within each cell indicates the number of relevant pairs. The lower half of the grid is symmetric to the upper half and therefore omitted.

but encountered challenges that persuaded us to adopt a multi-component design instead. Most importantly, we did not succeed in prompting the LLM to effectively maximize the objective of the query $\textsc{Gen}(S, r)$, i.e., the r-highest utility among S. For example, the statement generated by the LLM did not seem sufficiently responsive to the rank r.

Conceptually, the query $\textsc{Gen}(S, r)$ can be decomposed into two subtasks: **(i)** identifying a subset $S' \subseteq S$ of r agents amenable to agreeing on a statement, and **(ii)** generating a statement that maximizes the minimum utility in S'. When we asked the LLM to answer generative queries by performing these two subtasks in a chain of thought, the LLM seemed to struggle with subtask (i) but to perform well on subtask (ii).

Due to the LLM's difficulties in performing subtask (i), our implementation of this subtask combines an LLM-generated feature embedding of agents with clustering, a more robust way of identifying aligned agents. To generate the features, we randomly sample 50 of the bullet points occurring in the LLM-generated participant summaries, and then ask the LLM to rate the degree to which each agent is aligned with each feature on a 7-point scale.[11] These ratings embed each agent in a 50-dimensional Euclidean space, in which we use nearest-neighbor and balanced k-means clustering to identify cohesive sets of agents of a specified size.

[11] As in the discriminative query, we use the token probabilities to get an expected rating that need not be an integer.

Not only does subtask (ii) more closely resemble a straightforward text generation task, which should play to the LLM's strengths, but there is also precedent in the literature for solving similar problems with LLMs: Bakker et al. [2022] fine-tune an LLM to generate *consensus opinions*, i.e., statements that are good compromises for a small group of agents. Our setting is more challenging along several dimensions: our generation is based on much more detailed information about participants' opinions,[12] and we aggregate over groups of up to twenty participants rather than up to five. In addition, we aim to solve this task by prompting a general-purpose LLM rather than fine-tuning, albeit an LLM that is more advanced than the Chinchilla LLM used by Bakker et al. [2022].

We implement subtask (ii) with a single chain-of-thought prompt to the LLM. We instruct the LLM to write a statement with the goal of maximizing the minimum agreement among the given agents, to include points of agreement or likely agreement, and to avoid aspects that any of the agents fundamentally disagrees with. Our chain-of-thought prompt first asks the LLM to identify common themes of participants' responses and key disagreements, and to judge for each aspect of the topic whether and how it should be included in the generated statement. Only then should the LLM generate the statement, formulated as an opinion in the first person. See Appendix D.4 for more details.

We implement the generative query as an ensemble, where we select several sets of agents through subtask (i), apply the prompt from subtask (ii) to turn each set into a statement, evaluate the objective value (see Eq. (1)) of the resulting statements using the discriminative query, and then select the highest-scoring statement. For our pilot experiment, our ensemble contains the following statement sources in each iteration of Process 1:[13]

- We run balanced k-means clustering [Bradley et al., 2000] to partition the remaining agents into clusters of size n/k, which yields k sets of agents in the first round, $k-1$ in the second round, ..., and one set in the last round.
- We also add a few smaller groups of agents to the ensemble, which are determined by sampling a random remaining agent and returning their nearest neighbors in the feature space. Specifically, we generate two sets of five agents and two sets of ten agents in this way. Given that these groups have fewer than n/k members, we expect them to be more cohesive in their opinions, so subtask (ii) is likely to generate more specific statements.
- We also allow our ensemble to select unused statements from previous calls to the generative query. This might surface good statements at no additional computational cost, since we have already invoked all the necessary discriminative queries.

Evaluating the generative queries quantitatively is challenging for several reasons. First, since the 100 participants never see the statements generated from 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. We can, however, compare our generated statements to a natural baseline, the statements written by the participants we are summarizing.

[12] In our experiment, participants wrote between 124 and 1324 words (57—868 words excluding the the justification of ratings) within a median time of 26 minutes, whereas participants in Bakker et al. [2022, App. B] wrote between 10 and 200 words within a 5-minute limit on writing time.

[13] We also implemented a procedure to detect statements that did not express a subjective opinion, and filter them out from consideration in the ensemble. We took this measure after observing some LLM generations in our pre-tests that did not take a personal stance (e.g., "Abortion is a sensitive and controversial topic that evokes strong emotions and differing opinions..."). Since none of the generated statements in our pilot experiment were flagged, and since we did not filter in the experiments described in this section, we omit a more detailed description.

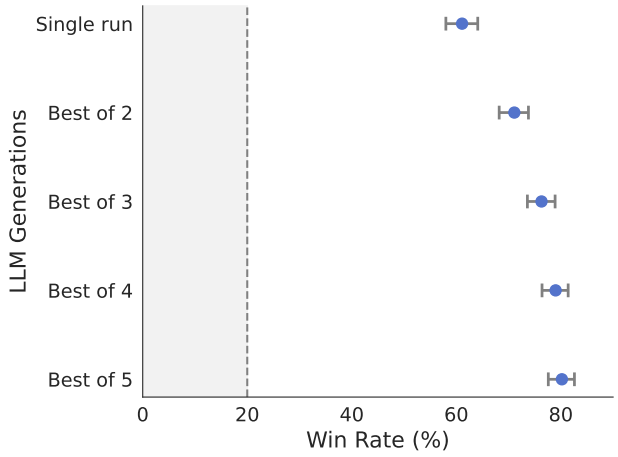


Figure 3: Win rate of LLM-generated statements over all four participant-written statements, in terms of the minimum predicted utility across the four participants. The dashed line indicates a win rate of 20%, which would be obtained if the generated statement followed the same distribution as the human-written statements. Win rates are computed for 1000 random groups of four participants, and ranges indicate 95% confidence intervals.

In light of the similarity of subtask (ii) with the work of Bakker et al. [2022], we use an analogous evaluation scheme for our generative query implementation: We randomly sample 1000 groups of four from the 100 participants and then apply our LLM prompt several times to generate candidate consensus statements. (Since we query GPT-4o with a temperature of 1 for our generative prompts, the generated statement is random.) We then compare these generated statements to the four statements written by each of the sampled participants.[14] Participants prefer a randomly-selected generated statement over the statement written by a random peer with 74.5% probability (95% confidence interval (CI): 73.8% – 75.2%),[15] and their mean score across three generated statements exceeds the mean score for the three peer statements with 87.2% probability (CI: 86.0% – 88.4%). Though a comparison to the numbers by Bakker et al. [2022] is not on equal terms,[16] this number (87.2%) compares favorably to their reported 78%.

Most relevant to the quality of subtask (ii) of the generative query is the *minimum* rating across the four participants. As Figure 3 shows, a randomly-selected generated statement beats *all four* participant statements with respect to the minimum utility with 61.1% probability (CI: 58.0% – 64.1%), and the best of four generated statements beats all four participant statements with 79.0% probability (CI: 76.4% – 81.4%). This shows that, as judged by the discriminative query, our prompt for subtask (ii) finds good consensus statements, clearly exceeding the baseline of taking the most broadly acceptable participant position, especially when we ensemble over a few generations.

For subtask (i) we verify that, by ensembling over a few clusters, we consistently produce

[14] As mentioned above, these comparisons are made using our discriminative query, rather than real ratings.

[15] In the following, all confidence intervals have 95% confidence. Intervals for simple binomial proportions are computed with Jeffrey's intervals; other confidence intervals are computed via bootstrapping.

[16] Most importantly, their comparisons are based on ground-truth preferences by the participant rather than a discriminative query/reward model. This makes their comparison harder in some ways (potentially higher variance in human responses, errors in the reward model). On the other hand, our statements come directly out of the LLM, whereas each of their generations is the best out of 16 generations as judged by reward models for the participants, which should help them. Bakker et al. [2022] evaluate over a variety of policy questions with less detailed preferences, whereas we focus on one issue in detail. Finally, they discard a third of participants based on intra-rater reliability, whereas we only filter out a few participants with LLM-written or extremely low-effort responses.

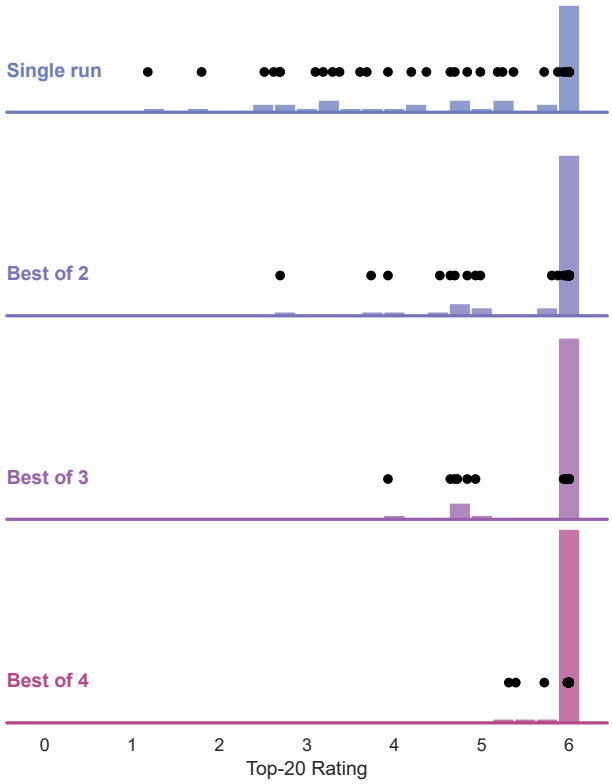


Figure 4: Distribution of the top-20 rating of LLM-generated statements on 50 randomly subsampled sets of 80 agents, across different numbers of generation attempts. Statements were generated by applying the generation prompt to one, two, three, or four nearest-neighbor clusters of size 10, and taking the statement with highest top-20 rating.

statements with high agreement rating. For this experiment, we randomly draw 80 out of the 100 agents and attempt to find a statement that maximizes the 20th-highest rating. This scenario simulates what is required of the generative query in the second round of running Process 1 with $n = 100, k = 5$. Recall that one source of statements in our ensemble samples a uniformly random agent, takes their 9 nearest neighbors, and then generates a statement for these 10 agents. In this experiment, we run this process four times, so we get four clusters and hence four statements in our ensemble. We run this experiment for 50 scenarios and show the results in Figure 4.

For about half of the sampled scenarios, even a single generation produces a statement whose top-20 rating is essentially maxed out, but there is still a decent chance of a poor generation (possibly due to the cluster's opinions not being well aligned with the preferences of participants outside of the cluster). By ensembling over several clusters, we eliminate these bad generations, and almost always arrive at a statement for which at least 20 participants have close to the maximal rating. This experiment shows that our implementation of the generative query, composed of both subtasks, produces high-quality statements.

5 A Study of Public Opinion on Abortion

We piloted our democratic process to summarize public opinion on abortion. We ran surveys studying this topic on August 8–9, 2024 and generated a representative slate of five statements.

5.1 Pilot Description

We first recruit a sample of 100 participants through the online platform Prolific[17], which we refer to as the *generation sample*. Our sample consists of US residents aged 22 and higher, stratified to reflect US residents in terms of gender and voting behavior in the 2020 presidential election.[18] We ask these participants to complete a survey on abortion.[19] Participants start reflecting on the topic by answering four questions in free-form text: how often they think about or discuss abortion, whether abortion should be legal or illegal, what their opinions on abortion are rooted in, and whether there are situations where they are uncertain about whether abortion is appropriate. Then, we ask participants to summarize their stance on how society should deal with abortion, in a self-contained three-sentence statement.

We also ask participants to rate their agreement with five example statements, which we generated with a single call to GPT-4o and without knowledge of participant responses, (see Appendix D.5 for the prompt). These ratings are given on the seven-level scale described at the beginning of Section 4, and rating questions are shown to participants in random order.

Based on participant responses, we extract a slate of five representative statements using Process 1 and our implementation of the queries. As a baseline, we also ask GPT-4o to directly generate a slate of five statements, providing the LLM with the free-text responses of all 100 agents in its context window (see Appendix D.6 for details).

To evaluate our slate, we launch a second survey with a new set of 100 stratified participants, the *validation sample*, to evaluate the slate's statements. In this validation survey, we ask participants to rate the ten statements from both our and the baseline slate (in random order, using the same question format as for the ratings in 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/survey_data/`.

5.2 Results

We defer the slate of five statements we generated to Appendix C.2.1, along with details about which components of the ensemble generated each statement.[20] Three of the statements express a clear pro-choice position and one statement a clear pro-life position (possibly with exceptions). The remaining statement expresses discomfort with abortion, but advocates for the legality in cases of special hardship and for pursuing measures outside the legal arena to reduce the frequency of abortions. This split of the slate is broadly in line with polls of US residents, which find that "about six-in-ten (63%) say abortion should be legal in all or most cases" [Pew Research Center, 2024]. In Appendix C.2.1, we also check that the slate is plausibly proportional to the generation sample, by relating the statements on our slate to the example statements in the generation survey.

[17] `https://www.prolific.com/`

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

[19] See Appendix E for the verbatim survey questions.

[20] When running our process on this data, calls to GPT-4o occasionally did not return correctly-formatted output during subtask (ii) of the generative query. These generations were simply ignored, occasionally leading to slightly smaller ensembles. Our ensemble approach is robust to these failures, and fixing them (by, e.g., retrying until a correctly-formatted output is returned) would only improve our results.

In the remainder of this section, we demonstrate the representativeness of the slate using the validation sample, a fresh sample of 100 US residents. 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 mean participant rating for their assignment is maximized, i.e., such that the balanced assignment maximizes the representation objective of Monroe [1995]. We then study the ratings of participants for their assigned statements.

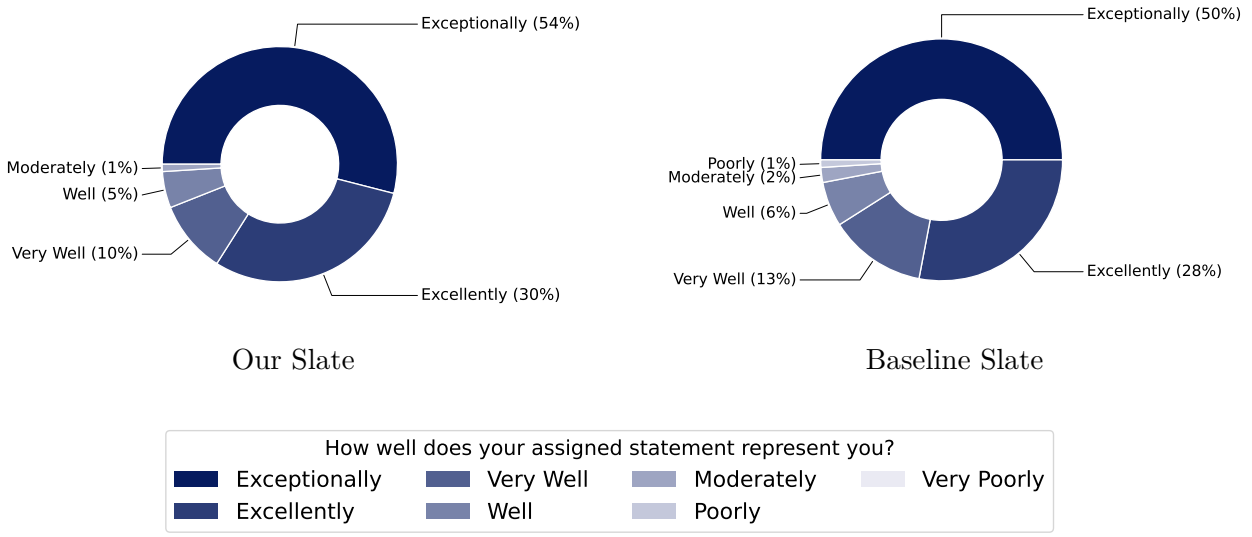


Figure 5: Ratings of participants from the validation survey for their *assigned* statement.

We find that the mean rating of a participant for their assigned statement is 5.31, between the two highest level of our rating scale, "excellently" (5) and "exceptionally" (6). As can be seen on the left-hand side in Figure 5, 54 participants say that their assigned statement "exceptionally" captures their viewpoint, and an additional 30 participants say that it captures their viewpoint "excellently". These levels far exceed the mean rating level of 2.74 in the survey (median: "moderately" (2), see Appendix C.2.3 for the distribution of ratings across all ten statements). Only a single participant feels less than "well" (3) represented — the midpoint of our scale — by their assigned statement. Hence, the vast majority of participant opinions are represented accurately by our slate.

The baseline slate also performs well (see Figure 5, right), but obtains worse ratings than our slate: the mean rating is 5.15, 50 participants are captured "exceptionally," 28 participants "excellently" by their assigned statements, and 3 participants are represented less than "well."[21]

A key question is whether these slates satisfy our proportionality axiom BJR. Though we cannot conclusively answer this question, we can approach it by investigating, for each slate, how close the statements of the other slate are to being BJR violations. Specifically, we maximize the size of a coalition S of agents such that these agents all rate some statement α from the other slate at some level ϑ or better, and all rate their assigned statement below ϑ. Note that a BJR violation is exactly such a coalition, if its size is at least $n/k = 20$. For our own slate, the largest achievable coalition consists of five participants, for a statement α on the baseline slate (statement B5 in Appendix C.2.2) that advocates for criminalization in all cases, and $\vartheta = 6$ ("exceptionally").

[21] In fact, the ratings for our slate dominate those for the baseline slate in the following sense: for any rating level ϑ, at least as many participants rate their assigned statement in *our slate* at ϑ or better as in the baseline slate.

For the baseline slate, coalitions are larger: two pro-choice statements on our slate (S2 or S4 in Appendix C.2.1) induce coalitions of size 9 (for $\vartheta = 6$). This is another indication that our generated slate is more representative of the validation sample, and more likely to satisfy BJR.[22]

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 participatory processes with generative-AI components [e.g., Devine et al., 2023, Konya et al., 2023, Marnette and McKenzie, 2023, Shaotran et al., 2023]. This expansion of the capabilities of participation is thrilling, but — as these prototypes continue to proliferate and eventually turn into deployed applications — we ought to critically interrogate their legitimacy on two fronts.

The first line of questioning has already received broad attention [e.g., Small et al., 2023]: *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 [Wallace et al., 2019] meant to unduly sway generative queries), and the effect of biases against groups of people [Basta et al., 2019, Kurita et al., 2019] and viewpoints [e.g. Hartmann et al., 2023] in the LLM must be studied and counteracted. Continued research in these areas will likely lead to ever more powerful and robust LLMs, approximating our oracle assumption more and more closely. Nevertheless, we propose to "trust, but verify" in high-stakes settings: Once the democratic process has produced a slate of statements, one could let participants vote on the statements proposed by other participants, alongside the slate produced by the our process. The slate should then only be adopted if no participant's statement witnesses a BJR violation with respect to the slate.[23] In this way, the democratic process can tap into the power of LLMs while ensuring that the voters, not machines, have the final word.

Even if the AI building blocks are trustworthy, another question remains: *is the process around the AI components democratic?* Granted, AI participation processes typically solicit input from all participants, and might even treat participants symmetrically, but that property alone (*neutrality*) is clearly not sufficient. We believe that voting rules with AI elements, just like those without, should argue their case based on 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.

[22] A coarser measure of representativeness, but one that might be easier to interpret, is to simply ask how many participants strictly prefer some statement α that is not on the slate over their assigned statement. The maximum such number for our slate and a baseline statement α is 6; for the baseline slate and a statement α from our slate, it is 12.

[23] If there is a violation, one might repeat the process, or select a slate that satisfies BJR with respect to the comments explicitly voted on. In this way, the LLM-augmented process does no worse than a classical process, even for arbitrary corruptions of LLM outputs.

Acknowledgments

We thank Nika Haghtalab and Abhishek Shetty for 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. Sara Fish was supported by an NSF Graduate Research Fellowship and a Kempner Institute Graduate Fellowship.

References

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.

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)*.

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.

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*.

N. Betzler, A. Slinko, and J. Uhlmann. 2013. On the Computation of Fully Proportional Representation. *Journal of Artificial Intelligence Research* 47 (2013), 475–519. https://doi.org/10.1613/jair.3896

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

P. S. Bradley, K. P. Bennett, and A. Demiriz. 2000. *Constrained K-Means Clustering*. Technical Report. Microsoft Research.

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

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)*.

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

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/

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

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

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

B. Everitt, S. Landau, M. Leese, and D. Stahl. 2011. *Cluster Analysis* (5th ed.). Wiley, Chichester.

B. Fain, K. Munagala, and N. Shah. 2018. Fair Allocation of Indivisible Public Goods. In *Proceedings of the 19th ACM Conference on Economics and Computation (EC)*. 575–592.

Federal Election Commission. 2020. Election Results for the 2020 U.S. Presidential Election. *Federal Election Commission* (2020).

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

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

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).

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

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)*.

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.

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

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.

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

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)*.

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.

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

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

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.

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

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.

Pew Research Center. 2024. *Broad Public Support for Legal Abortion Persists 2 Years After Dobbs*. Technical Report. Pew Research Center. https://www.pewresearch.org/wp-content/uploads/sites/20/2024/05/PP_2024.5.13_abortion_REPORT.pdf

L. Sánchez-Fernández, E. Elkind, M. Lackner, N. Fernández, J. A. Fisteus, P. B. Val, and P. Skowron. 2017. Proportional Justified Representation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

M. Schulze. 2003. A new monotonic and clone-independent single-winner election method. *Voting Matters* 17 (2003), 9–19.

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

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

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

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).

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.

M. H. Tessler, M. A. Bakker, D. Jarrett, H. Sheahan, M. J. Chadwick, R. Koster, G. Evans, L. Campbell-Gillingham, T. Collins, D. C. Parkes, M. Botvinick, and C. Summerfield. 2024. AI Can Help Humans Find Common Ground in Democratic Deliberation. *Science* 386, 6719 (2024).

U.S. Census Bureau. 2021. 2020 Census Demographic and Housing Characteristics File (DHC). https://www.census.gov/data.html Accessed: 2024-10-09.

V. N. Vapnik. 1998. *Statistical Learning Theory.* Wiley.

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).*

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

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

A Relationship Between BJR and Other Justified Representation Axioms

Our notion of BJR is closely related to several axioms in the social choice literature.[24] 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. [2017].

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. [2021].

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 agents (specifically, agents 1 and 2) approve statements α, α', and the remaining third of agents (agent 3) approves statements β, β'. As has been frequently observed [e.g., Aziz et al., 2017, 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 that 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 a specific statement.[25]

[24] 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 in the body.

[25] One might hope that EJR and FJR can be adapted to this perspective, by extending utilities to sets in a

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 [1995]: "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. For a definition of these axioms, we refer the reader to Definitions 4.3, 4.5, 4.7, and 4.10 of Lackner and Skowron [2023].

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*[26] *and in the classical setting where they are sets (using the adaptation of BJR from Footnote 6 in the body).*

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:

unit-demand rather than additive way. With this modification, however, they no longer rule out the unproportional slate in the first example instance.

[26] Brill et al. [2022] 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 [Brill et al., 2022].

	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. □

Finally, one might hope to strengthen BJR by allowing coalitions to deviate as long as any member of the coalition strictly increases their utility in the deviation, rather than requiring agents to cross a common threshold ϑ:

Definition 8. A slate W is in the *balanced unit-demand core* if there is a function $\omega : N \to W$, matching agents to statements such that each statement on the slate is matched to $\lfloor n/k \rfloor$ or $\lceil n/k \rceil$ agents, for which there is no coalition $S \subseteq N$ and statement $\alpha \in \mathcal{U}$ such that *(i)* $|S| \geq n/k$ and *(ii)* $u_i(\alpha) > u_i(\omega(i))$ for all $i \in S$.

Unfortunately, the resulting axiom may not be satisfiable. The possible emptiness of the core can be seen using the instance [Fain et al., 2018, App. C] previously used to show the analogous statement for additive, non-approval preferences:

Proposition 9. *The balanced unit-demand core can be empty.*

Proof. Consider the following instance with six agents, six statements, and $k = 3$:

	α	β	γ	α'	β'	γ'
u_1	2	0	1	0	0	0
u_2	1	2	0	0	0	0
u_3	0	1	2	0	0	0
u'_1	0	0	0	2	0	1
u'_2	0	0	0	1	2	0
u'_3	0	0	0	0	1	2

Note that the instance decomposes into two, isomorphic instances (one with primed variables and one with non-primed variables). By symmetry, we can assume w.l.o.g. that the slate contains at most one non-primed statement and that, if there is a non-primed statement in the slate, this statement is α. That is, we can assume that neither β nor γ are on the slate. Consider the coalition $\{2, 3\}$, and note that agent 2 does not value any statement on the slate better than 1 and that agent 3 does not value any statement on the slate better than 0, so the statement they are assigned to has at most this value. Hence, these $2 = n/k$ agents would like to deviate to the statement β, which has value $2 > 1$ for agent 2 and value $1 > 0$ for agent 3. This demonstrates that no slate is in the balanced unit-demand core. □

Hence, BJR cannot be strengthened to account for deviations without a uniform threshold while retaining feasibility. The uniform-threshold technique (for generalizing approval-based proportionality notions without losing feasibility) was previously used by Peters et al. [2021] to generalize EJR and FJR to general additive utilities.

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

$$\begin{aligned}\max_{(\lceil \vec{r} \rceil)}(\{u_i(\alpha') \mid i \in S\}) = \max_{(\lceil n/k \rceil)}(\{u_i(\alpha') \mid i \in S\}) \geq \max_{(|S'|)}(\{u_i(\alpha') \mid i \in S\}) \\ \geq \max_{(|S'|)}(\{u_i(\alpha') \mid i \in S'\}) \geq \vartheta.\end{aligned}$$

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

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

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$(\cdot, \cdot)$ queries 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): 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 $\min(|S|, \frac{n}{2k})$ 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

$$\mathbb{P}\left[\text{at least } \frac{n}{2\,k} \text{ agents in } S \text{ have color } c\right]$$

$$= \mathbb{P}\left[\mathsf{Binomial}(n/8, 2/k) \geq 2 \cdot \frac{n}{4\,k}\right]$$
$$\leq \exp\left[-\frac{n}{12\,k}\right].$$

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 $\frac{n}{2\,k}$ 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 10 (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 [Shalev-Shwartz and Ben-David, 2014, 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 [2009], 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

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

where $c_0 < 5.05$ is an absolute constant, $q := \frac{1}{m} + \frac{1}{|D|-m} \leq \frac{2}{m}$, and $s := \frac{|D|}{(|D|-1/2)\cdot(1-\frac{1}{2(|D|-m)})}$. 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 [2009, Def. 1 and p. 6] define the transductive Rademacher complexity $R_{trans}(\mathcal{H})$ as

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

Note that we can draw $\vec{\sigma}$ from $\Sigma^{|D|}_{trans}$ 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

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

By Bartlett and Mendelson [2002, Lemma 4 & Thm. 6], $\mathbb{E}_{\vec{\sigma}\sim\Sigma^{|\mathcal{I}|}_{ind}} \sup_{h\in\mathcal{H}} \vec{\sigma}^T h(\vec{x}_\mathcal{I}) \leq c_1\sqrt{d|\mathcal{I}|}$ for some absolute constant c_1. Thus, we can bound

$$R_{trans}(\mathcal{H}) \leq c_1\, (\tfrac{1}{m} + \tfrac{1}{|D|-m}) \cdot \mathbb{E}_{\mathcal{I}\sim\mathcal{B}} \sqrt{d|\mathcal{I}|}$$

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

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

$$
\begin{aligned}
L_D(h) &= \frac{m}{|D|} L_S(h) + \frac{|D|-m}{|D|} L_{D\setminus S}(h) \\
&\leq \frac{m}{|D|} L_S(h) + \frac{|D|-m}{|D|} \left(L_S(h) + \alpha \right) \\
&\leq L_S(h) + \alpha
\end{aligned}
$$

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

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

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

$$
L_D(\hat{h}) - L_D(h^*)
$$

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

By choosing the constant in Eq. (5) large enough, we can ensure[27] 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

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

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

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

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

[27] 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 10)
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

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

We can now write:

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

and similarly:

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

To bound the difference between these two terms, we map them to the learning-theoretic setting from Lemma 10 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:

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

Hence, each hypothesis can be identified with a pair (α,ϑ) and it is then easy to see that the losses from Lemma 10 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 10, 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

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

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

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

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

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

or, equivalently, that

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

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

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

or, equivalently, that

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

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

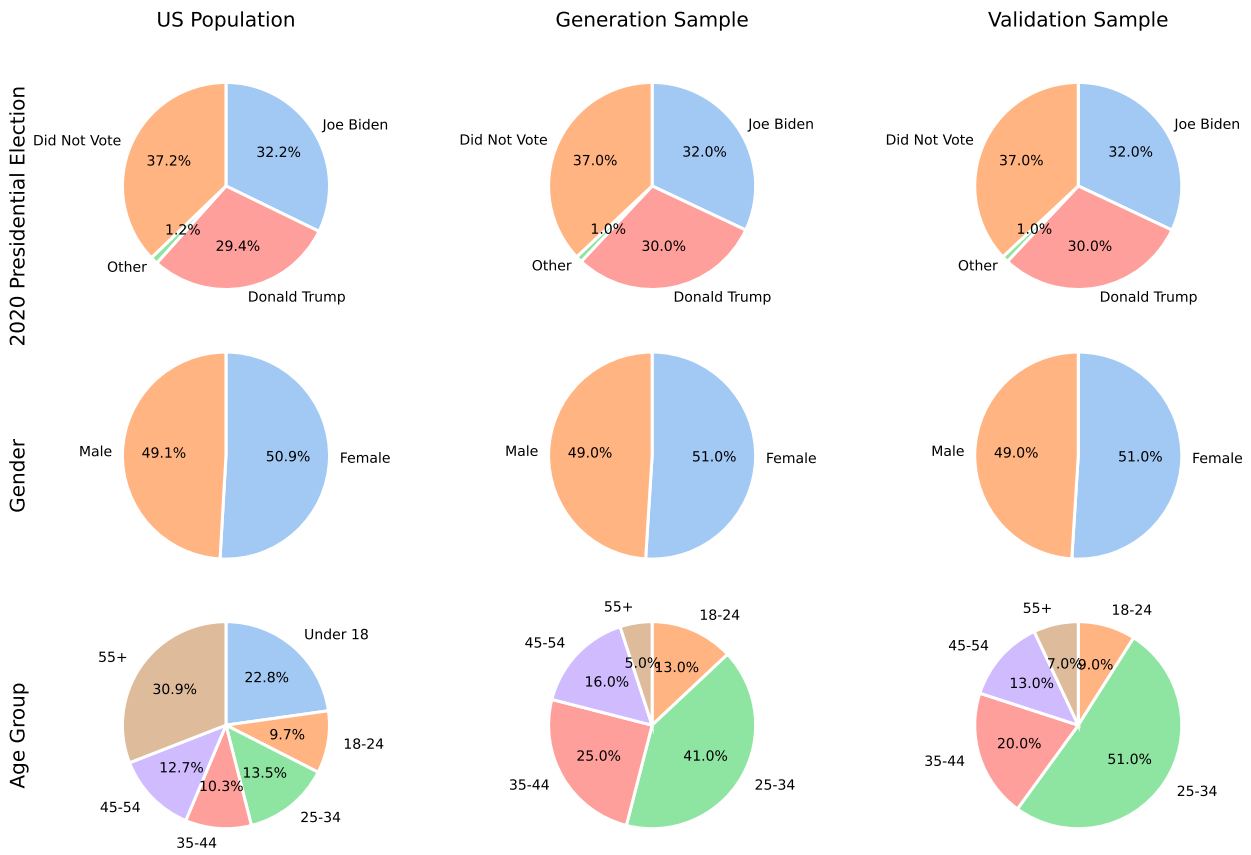


Figure 6: Demographic composition of both samples, compared to the US population as of the 2020 Census U.S. Census Bureau [2021] and the 2020 U.S. Presidential Election results Federal Election Commission [2020].

We recruited a sample of U.S. residents on Prolific, stratified on gender and voting behavior in the 2020 U.S. presidential election (Joe Biden, Donald Trump, third-party candidate, and nonvoter). We adopt these two criteria of stratification because they are especially salient for the topic of abortion policy. We excluded participants below the age of 22 at the time of the survey, to ensure that all respondents were of voting age in the 2020 election. As shown in Figure 6, our sample mirrors the U.S. population's composition in terms of these stratified features. Our sample is not just representative along gender and voting-behavior lines but also within intersections of these features. Compared to the U.S. population, our sample skews young; in particular, the 25–34 and 35–44 year cohorts are overrepresented.

To maintain data integrity, we filtered out submissions suspected to be generated by language models or those that appeared extremely low effort. Specifically, we recruited, for each of our surveys, 110 participants on Prolific, and manually identified responses that exhibited patterns typical of AI-generated content or minimal engagement with the survey questions.[28] For the generation sample,

[28]For example, one participant's response included the phrase, "You've hit the Free plan limit for GPT-4o", and another copy-pasted the same answer repeatedly.

we excluded 4 of the 110 submissions for LLM usage and 2 submissions for extremely low effort; and for the validation sample, we excluded 5 submissions for LLM usage and 2 for extremely low effort. A member of our team then selected the 100 submissions for our survey from those not filtered out, by re-establishing proportionality along voting behavior, gender, and intersections of the two categories. To avoid bias, the team member was blinded to the content of submissions for this decision.

C.2 Slates

C.2.1 Slate Generated by Our Democratic Process

S1. I believe that abortion should be legal and that a woman has the right to choose what she does with her body. It is crucial for the decision to have an abortion to remain a healthcare decision between a woman and her doctor, free from political interference. Society should respect and uphold these rights to ensure women's autonomy and wellbeing.

S2. I believe that abortion should be legal and accessible to everyone, as it is essential healthcare. It is critical for women to have control over their own bodies and make decisions about abortion without government interference. Safe and legal access to abortion ensures that women can make the best choices for their health and well-being.

S3. I believe that abortion should be illegal in most cases and only allowed under specific circumstances. It is morally wrong and equates to taking a life, as I believe in the sanctity of human life starting from conception. People should take responsibility for their actions and should never use abortion as a form of birth control.

S4. I believe that abortion should be legal and that women should have the right to make decisions about their own bodies. The decision to have an abortion should be a private matter, free from societal judgment, as it's a crucial aspect of women's health and safety. Society should ensure that abortions are safe, legal, and accessible, respecting the individual's autonomy and privacy in making such decisions.

S5. I believe that abortion should not be used as a form of birth control. However, it should be legally permissible in cases where the mother's life is in danger, and in instances of rape or incest within an early timeframe. Furthermore, there should be robust educational efforts to prevent unwanted pregnancies and ensure people are informed about the complexities and consequences of abortion.

Statements S1 and S5 were generated through nearest-neighbor clustering (with clusters of size 10), whereas the other three statements were generated by balanced k-means clustering. Statements S1, S2, and S5 were generated in the iteration of the greedy algorithm in which they were selected (i.e., the first, second, and fifth iteration, respectively), whereas S3 was generated in the first iteration and S4 in the third iteration.

Three of the statements (S1, S2, and S4) above are very similar. They express a position that strongly favors broadly legal abortions and sees abortion as a private choice, in the realm of healthcare, and as part of a right to bodily autonomy. Statement S3 wants abortion to be broadly illegal, albeit with some exceptions, since it sees abortion as murder of a human being with full moral consideration. It also expresses concerns about an abuse of abortion as birth control. Statement S5 shares this concern about abuse, and expresses discomfort with abortion, but takes a more moderate position. Abortion should be legal in medical emergencies, or cases of rape and incest; the statement does not take a stance on whether abortion should be legal or illegal outside of these circumstances. Instead, the statement hopes to reduce the prevalence of abortion through education.

To check that this slate plausibly represents the generation sample, we coarsely cluster the participants by which of the five example statements (full statements in Appendix E.1) they agree with most. Specifically, we match each participant in the generation sample to the example statement they rated most highly, matching them fractionally if several statements are tied for the highest rating. In this case, the number of participants matched to each statement is as follows:

# matched	statement
32.5	I think abortion should be a personal decision between a woman and her doctor...
27.2	I believe that abortion should be legal and accessible because women have the right...
17.0	I think abortion should be restricted to certain circumstances...
15.0	I believe that abortion should be illegal because it involves taking a human life...
8.3	believe that abortion should be allowed in the first trimester but restricted afterward...

Conceptually, the three pro-choice statements on our slate (S1, S2, and S4) are a blend of the first two statements (minus the call for contraception and sex education), so representing $32.5 + 27.2 = 59.7$ participants by three statements (which ideally should represent $3\,n/k = 60$ participants) seems quite accurate. These two example statements are also the pair whose ratings are the most correlated (correlation coefficient 0.88), so it is plausible to treat the agents whose favorite is among those two as a single bloc.

The pro-life statement on the slate (S3) best matches the fourth statement above, given that both statements express that abortion should be illegal and equate it with murder (though the specifics vary). Representing 15 pro-life participants by one statement is also plausible.

The moderate statement on our slate (S5) most closely resembles the third statement in the table, given that both statments want abortion to be legal in cases of incest, rape, and risk to life, and given that both look for tools to reduce the prevalence of abortions besides of criminalization (though, again, some specifics vary). Representing these 17 participants by a statement also seems close to proportional.

This leaves the 8.3 adherents of the last statement (advocating for a concrete temporal cutoff for legality) without an obvious representative on our slate, but this group are too small to cause a violation of BJR on their own. The fact that agreement with this statement does not correlate much with agreement to the other statements[29] further suggests that the about 8 adherents of the last statement are not the core of a cohesive group whose BJR guarantee would be violated.

If we repeat the above assignment to favorite statements, but only among the second, third, and fourth statement in the table above (i.e., one pro-choice statement, one hesitant intermediate statement, and one pro-life statement), the respective numbers of matched participants are 62, 21.5, and 16.5. Once more, this suggests that representing the first of these three viewpoints by three statements, and the latter two viewpoints by one statement each, is proportional to the opinions in the generation sample.

C.2.2 Baseline Slate Generated By GPT-4o

B1. I believe abortion should be legal because it is a woman's right to choose what happens with her body. It is important for maintaining reproductive freedom and ensuring safe and accessible healthcare options. Women need to make decisions that are best for them and their families without government interference.

[29]The two correlations with largest magnitude are positive correlations with the pro-choice statements, with correlation coefficients of 28% and 24%.

B2. In my opinion, abortion should be illegal except in cases where the mother's life is at risk or in instances of rape or incest. This stance protects unborn children while allowing necessary exceptions. It's about balancing moral concerns with compassion for those in difficult situations.

B3. Abortion should be legal up to a certain point in the pregnancy, such as the first or second trimester. After that, I think it needs to be restricted unless there are severe medical reasons. Society needs to find a middle ground that respects both a woman's right to choose and the potential life of the fetus.

B4. I think abortion should be legal and accessible at all stages of pregnancy. Women need to have control over their reproductive health for various personal and medical reasons. Restrictions can force women into unsafe or undesirable circumstances.

B5. Abortion should be illegal in all cases, as I believe it is morally wrong and equivalent to ending a potential life. Society should look for alternative solutions such as adoption and support for pregnant women. We need to value and protect all human life, starting from conception.

C.2.3 Distribution of Ratings in Validation Sample

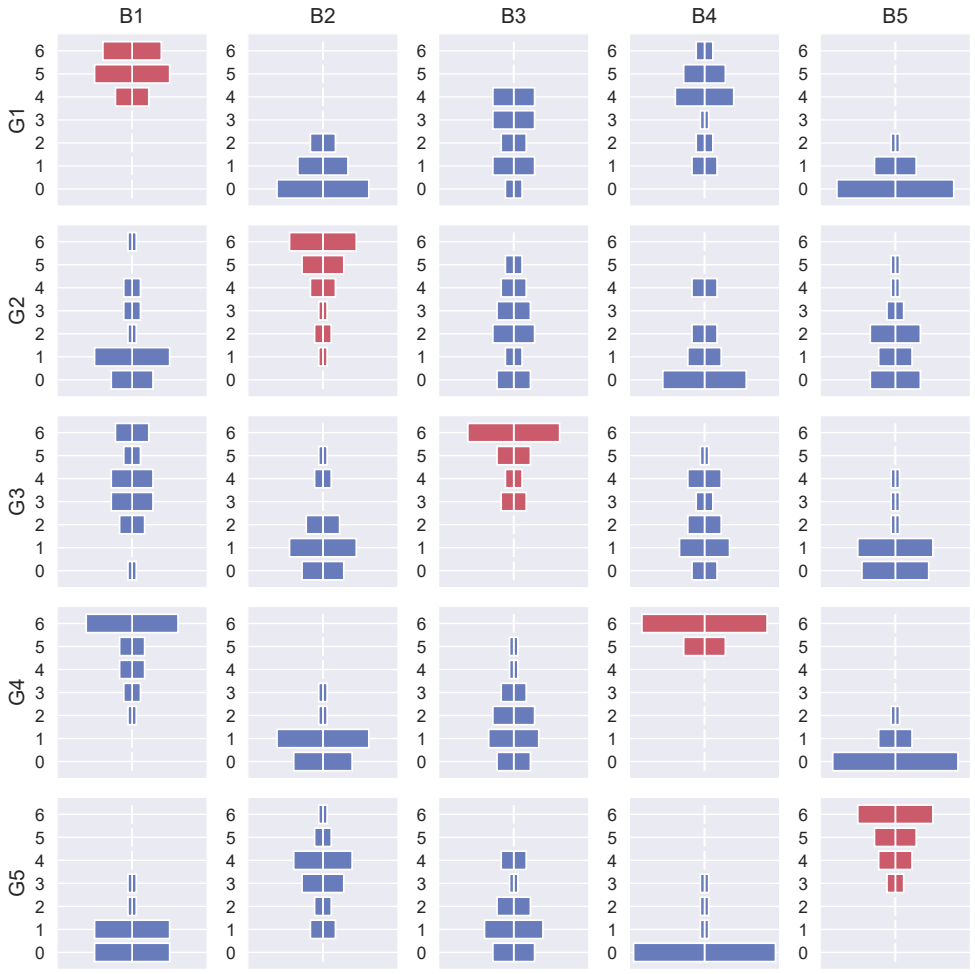
Below, we show the distribution of ratings for the statements of both slates in the validation sample. For each slate, we disaggregate agents depending on which statement they are assigned to, and show, for each group and each statement, a histogram of rating levels.

First, we show the distribution of rating levels for our slate:

Each row corresponds to a group; for example, G3 represents the 20 participants assigned to statement S3. For each group, we polot the frequencies of rating levels given by members of this group to statements S1 through S5.

We observe that rating levels on the diagonal, (i.e., ratings for the groups' assigned statement) are higher than those off diagonal (i.e., ratings for other statements). We also see that there is substantial disagreement in the dataset; for example, group G3 (the 20 agents matched to the pro-life statement) are very critical of the three pro-choice statements S1, S2, and S4. We generally see that participants of all groups rate these three pro-choice statements similarly, attesting to their striking similarity.

We now show the corresponding figure for the baseline slate, where groups G1 through G5 now refer to the agents assigned to B1 through B5, respectively:



We again see higher ratings on the diagonal than outside of it. We can also see that statement B2 seems slightly less well received by its assigned group than the other statements on this slate or the five statements on our slate.

D Prompts

The formatting of the prompts has been lightly edited for readability.

D.1 Summarization

System Prompt

```
You will be provided with a user's response to a survey, in which they describe their
opinions on a topic in detail. Your task is to produce a detailed summary of that
user's opinion. Your response should be in JSON according to the format specified
below.
```

Prompt

User survey responses:
{user_data}

Output instructions:
Complete the following entries:

- most_important_aspects (List[str]): List the aspects of the topic that are most important to the user. Each aspect should be succinct and self-contained, so that it can be understood without context. (For example, instead of writing "Religious beliefs", write "Believes X due to religious beliefs")
- specific_details (List[str]): List any specific details or examples the user provided to support their opinion. Each detail should be succinct and self-contained, so that it can be understood without context. These details, together with the most_important_aspects above, should be enough to mostly reconstruct the user's opinion.
- user_background (List[str]): List any personal information the user may have divulged that is relevant to their opinion.
- overall_summary (str): Write a detailed 2-3 sentence summary of the user's opinion, taking into account all the context provided above.

Respond in JSON with the fields above filled in.

D.2 Tagging

System Prompt

```
You will be provided with a user's response to a survey, in which they describe their
opinions on a topic in detail. Then you will be provided with a list of aspects of
that topic. For each aspect, your task is to rate the extent to which it pertains
to the user's response or captures the user's opinion. The scale is as follows:

1. Strongly goes against user's opinion
2. Goes against user's opinion
3. Somewhat goes against user's opinion
4. Neutral / unknown
5. Somewhat aligned with user's opinion
6. Aligned with user's opinion
7. Strongly aligned with user's opinion

Respond in JSON, mapping each aspect to your rating. It is important to copy the EXACT
wording of the aspect with no changes.
```

Prompt Template

User survey responses:
{user_data}

Output instructions:
For each of the following aspects, rate on the scale from 1-7 how well it pertains to the user's response or captures the user's opinion.
{list_of_fields}

Respond in JSON, with the above fields as keys and your ratings as values.
Example fields used include: most_important_aspects and specific_examples.

D.3 Discriminative Query

System Prompt

You are an AI-based text completion system, and you are tasked with helping a user fill out an in-depth opinion survey.

You will receive in your input a sequence of questions along with the answers already given by the user. Then, you will receive the question that the user will answer next. Based on the question and what you have learned so far about the user, you must make your best guess what the user will answer. If you guess right, the user can save a lot of effort and time, so it is important that you suggest what the user would be most likely to answer. Your prediction must strictly adhere to the format at the end of the prompt.

Prompt Template

Question 1:
{question}

Answer 1: <ANSWER FORMAT: string.>
{freetext_answer}

...

Question 6:
{question}

Answer 6: Part A - User's level of agreement: <ANSWER FORMAT: numeric score and text label. Must be exactly equal to one of the following options: "0 = very poorly", "1 = poorly", "2 = moderately", "3 = well", "4 = very well", "5 = excellently", or "6 = exceptionally".>
{choice_numeric} = {choice}

Answer 6 Part B - User's explanation: <ANSWER FORMAT: string>
{explanation}

...

Question 11:

To what extent does this statement capture your full opinion on abortion?

{statement}

You must now infer the user's most likely agreement level with Question 11 from their previous responses. Your response must exactly equal one of the following options: "0 = very poorly", "1 = poorly", "2 = moderately", "3 = well", "4 = very well", "5 = excellently", or "6 = exceptionally"

D.4 Generative Query

System Prompt

You will be provided with a list of users and their opinions on a topic. Your task is to write a paragraph all of the users would agree with. Specifically, your measure of success is based on how satisfied the *least satisfied* user is with the paragraph.
When selecting the content for your paragraph, follow these guidelines:

- For any aspect of the topic that all users agree on, include that detail in your paragraph. Especially when the users disagree on major aspects of the topic, it's important to make note of any minor aspects that they all agree on and include them.
- Sometimes there is an aspect of the topic that most users agree on, but some users don't express a strong opinion on. If it's reasonable to assume that those users would also agree, you should still include that detail in your paragraph.
- For any aspect of the topic that the users fundamentally disagree on, omit details from the paragraph. Only include this aspect if there is a way to phrase it that everybody would agree with.

As for writing style, your paragraph be written like an answer to the following survey question:

> "{statement_question_text}"

When writing, you should first think about the content you want to include (cf. the instructions above on content), and then conjure an imaginary new user who holds those precise beliefs. Then, when writing your paragraph, you should answer the survey question as if you were that user. Your paragraph should never reference the other users - you should write "I think..." statements instead of "Some users think..." statements. Your paragraph should not even implicitly mention the users - you should never write "Opinions vary about...". The users' opinions are only there to help you understand the spread of opinions for the content selection step - when it comes to the writing step, you should write from the perspective of a single (imaginary) new user.

Prompt Template

List of users:

{user_descriptions}

Instructions:

This task consists of two main parts. First, you need to determine the content of your paragraph (recall the content guidelines from the beginning of the instructions). Second, you need to write the paragraph itself, which should be phrased like a single user's answer to the survey question (recall the writing guidelines from the beginning of the instructions).

Step 1. *Identify common themes.* List every viewpoint that is expressed by multiple users at once. For each such viewpoint, thoughtfully make note of how many users hold that viewpoint. Keep in mind that some users may express multiple viewpoints.

Step 2. *Identify key disagreements.* Identify all aspects of the topic where users have differing opinions. For each such aspect, thoughtfully make note of how many users hold each opinion (or if there are users who abstain).

Step 3. *Analyze potential contents of paragraph.* For each aspect, discuss how (if at all) it should be included in the paragraph.

Step 4. *Write the paragraph.* This is the most important step, for which you need to do the most thinking. Write a paragraph that captures the opinions of the users as best as possible.

When doing so, keep in mind these *content guidelines*:

- Your measure of success is based on how satisfied the *least satisfied* user is with the paragraph.
- When all users agree on an aspect, include that aspect. (Still include if some users don't express a strong opinion on it.)
- When users disagree on an aspect, omit that aspect from the paragraph. Focus on the aspects that all users can agree on. And additionally keep in mind these *writing style guidelines*:
- Your paragraph should sound like a single (imaginary) user's to the survey question.
- Your paragraph should neither explicitly nor implicitly mention the users whose data you are working with. That data should only be used to determine the content - it should not be referenced in the writing itself.

Output instructions:

Respond in JSON as follows: {{
"step1" : <your response to step 1>,
"step2" : <your response to step 2>,
"step3" : <your response to step 3>,

```
"step4" :  <your response to step 4> (str),
}}
```

D.5 Example Statements

System Prompt

For consistency with the baseline prompts, we passed the following prompt to GPT-4o as a system prompt (with empty standard prompt): `Write in bullet points five different beliefs that people from the US might have about abortion. Each belief should be written such that as many people as possible would agree with it. They should sound like sentences a real person would say, written in the first person. They should be written like answers to the following survey question: Summarize your position on abortion in your own words. Please write` ***how you think society should deal with abortions***`, and` **give reasons** `that support this policy.`

`Please write` ***exactly 3 sentences***`: Be as precise as possible and prioritize the points that are most important to you.`
`Your answer should be self-contained, which means that you can repeat things you already wrote as well as make new points.`

The statements generated by this prompt can be found as part of the survey questions in Appendix E.1.

D.6 Baseline Prompts

System Prompt

`Write in bullet points five different beliefs that people from the US might have about abortion. Each belief should be written such that as many people as possible would agree with it. They should sound like sentences a real person would say, written in the first person. They should be written like answers to the following survey question: Summarize your position on abortion in your own words.`
`Please write` ***how you think society should deal with abortions***`, and` ***give reasons*** `that support this policy.`

`Please write` ***exactly 3 sentences***`: Be as precise as possible and prioritize the points that are most important to you.`
`Your answer should be self-contained, which means that you can repeat things you already wrote as well as make new points.`

`To help you better understand typical beliefs that people from the US have on abortion, attached are survey responses from a representative sample.`

Prompt

As the full prompt is extremely long, we include only the prompt template below. For the survey data, see https://github.com/generative-social-choice/survey_data. `Survey responses from 100 people:`

`Person 1`

(Person 1's complete survey responses)

Person 2
(Person 2's complete survey responses)

...

Person 100
(Person 100's complete survey responses)

E Survey Questions

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

E.1 Generation Survey

Informed Consent

We are a team of university researchers. We want to understand in detail what people like you think about social questions, in this case about abortion. We want to study the diversity of people's beliefs, and whether algorithmic tools can help summarize and analyze these differing points of view.

What will I need to do and how long will the study last? We will ask you **10 questions**. We expect that you will be in this research study for less than an hour.

Compensation: Your pay will **not** depend on your opinions: all good-faith responses will get fully compensated, so please just write what you think. You may not use ChatGPT or other AI tools to write or edit your responses. **We will award a $1 bonus** if your submission satisfies all minimum length requirements across all questions.

Who will see your responses? The data you provide will be immediately anonymized. We may later on publish this anonymous data. By continuing this survey, you agree to this use of your responses.

Part 1/3: Background Questions

We will begin by asking you **4 questions** touching on specific parts of your opinion.

Part 2/3: Your Position

This part only has **one question**. We will ask you to summarize, in three sentences, your overall position on abortion.

Part 3/3: Candidate Summaries

To help researchers and policymakers understand public opinion on abortion, we will condense the survey responses into a short summary. We will now present **5 candidate summaries** to you, and your task is to rate to what extent they would capture your point of view.

An ideal summary should

1. perfectly capture all your thoughts and feelings on abortion,
2. include the reasoning behind your opinion, and
3. be concrete and actionable.

Background Questions: 1/4

How often do think about abortion or discuss it with others? How does this topic make you feel?

Please write **two** or more sentences.

Background Questions: 2/4

Do you think abortion should be legal or illegal? Which circumstances does your answer depend on?

Please write **two** or more sentences.

Background Questions: 3/4

Where do your beliefs about abortion come from? For example, did particular life experiences influence your beliefs?

Please write **two** or more sentences.

Background Questions: 4/4

Can you describe a situation where you are not sure if abortion is appropriate or not? If so, what makes this situation borderline or unclear?

Please write **two** or more sentences.

Your Position

Summarize your position on abortion in your own words.

Please write **how you think society should deal with abortions**, and **give reasons** that support this policy.

Please write **exactly 3 sentences**: Be as precise as possible and prioritize the points that are most important to you.

Your answer should be self-contained, which means that you can repeat things you already wrote as well as make new points.

Candidate Summaries: 1/5

Here is a possible summary:

> "I believe that abortion should be legal and accessible because women have the right to make decisions about their own bodies. Access to safe abortions is crucial for protecting

> women's health and well-being. Society should support comprehensive sex education and contraception to reduce the need for abortions."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

(Please write **two** or more sentences.)

Candidate Summaries: 2/5

Here is a possible summary:

> "I think abortion should be a personal decision between a woman and her doctor, without government interference. Each situation is unique, and women should have the autonomy to make the best choice for themselves and their families. Society should ensure that all women have access to affordable healthcare, including reproductive services."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

(Please write **two** or more sentences.)

Candidate Summaries: 3/5

Here is a possible summary:

> "I believe that abortion should be allowed in the first trimester but restricted afterward unless there are exceptional circumstances. This policy respects a woman's right to choose while recognizing the increasing moral considerations as the pregnancy progresses. Society should invest in education and healthcare to prevent unwanted pregnancies and support women through their reproductive choices."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

(Please write **two** or more sentences.)

Candidate Summaries: 4/5

Here is a possible summary:

> "I think abortion should be restricted to certain circumstances, such as cases of rape, incest, or when the mother's life is at risk. This approach balances the rights of the unborn with the needs of women facing difficult situations. Society should provide support for women who carry their pregnancies to term, including healthcare and financial assistance."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

(Please write **two** or more sentences.)

Candidate Summaries: 5/5

Here is a possible summary:

> "I believe that abortion should be illegal because it involves taking a human life, which I consider morally wrong. Society should focus on providing resources and support for pregnant women to encourage them to choose life. Adoption should be promoted as a viable alternative to abortion."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

(Please write **two** or more sentences.)

Thank You

Thank you for completing the survey!

Click below to register your survey completion with Prolific and return to their website.

E.2 Validation Survey

Informed Consent

We are a team of university researchers. We want to understand in detail what people like you think about social questions, in this case about abortion. We want to study the diversity of people's beliefs, and whether algorithmic tools can help summarize and analyze these differing points of view.

What will I need to do and how long will the study last? We will ask you **10 questions**. We expect that you will be in this research study for less than an hour.

Compensation: Your pay will **not** depend on your opinions: all good-faith responses will get fully compensated, so please just write what you think. You may not use ChatGPT or other AI tools to write or edit your responses.

Who will see your responses? The data you provide will be immediately anonymized. We may later on publish this anonymous data. By continuing this survey, you agree to this use of your responses.

Instructions

In this survey, we want to hear about your opinions on **abortion**.

To help researchers and policymakers understand public opinion on abortion, we condensed the responses of previous participants into **10 short summaries**. We will now present each summary to you, and your task is to rate to what extent it would capture **your** point of view.

An ideal summary should

1. perfectly capture all your thoughts and feelings on abortion,
2. include the reasoning behind your opinion, and
3. be concrete and actionable.

Candidate Summaries: 1/10

Here is a possible summary:

> "I believe abortion should be legal because it is a woman's right to choose what happens with her body. It is important for maintaining reproductive freedom and ensuring safe and accessible healthcare options. Women need to make decisions that are best for them and their families without government interference."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

Please write **two** or more sentences.

Candidate Summaries: 2/10

Here is a possible summary:

> "In my opinion, abortion should be illegal except in cases where the mother's life is at risk or in instances of rape or incest. This stance protects unborn children while allowing necessary exceptions. It's about balancing moral concerns with compassion for those in difficult situations."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

Please write **two** or more sentences.

Candidate Summaries: 3/10

Here is a possible summary:

> "Abortion should be legal up to a certain point in the pregnancy, such as the first or second trimester. After that, I think it needs to be restricted unless there are severe medical reasons. Society needs to find a middle ground that respects both a woman's right to choose and the potential life of the fetus."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

Please write **two** or more sentences.

Candidate Summaries: 4/10

Here is a possible summary:

> "I think abortion should be legal and accessible at all stages of pregnancy. Women need to have control over their reproductive health for various personal and medical reasons. Restrictions can force women into unsafe or undesirable circumstances."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

Please write **two** or more sentences.

Candidate Summaries: 5/10

Here is a possible summary:

> "Abortion should be illegal in all cases, as I believe it is morally wrong and equivalent to ending a potential life. Society should look for alternative solutions such as adoption and support for pregnant women. We need to value and protect all human life, starting from conception."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

Please write **two** or more sentences.

Candidate Summaries: 6/10

Here is a possible summary:

> "I believe that abortion should be legal and that a woman has the right to choose what she does with her body. It is crucial for the decision to have an abortion to remain a healthcare decision between a woman and her doctor, free from political interference. Society should respect and uphold these rights to ensure women's autonomy and wellbeing."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

Please write **two** or more sentences.

Candidate Summaries: 7/10

Here is a possible summary:

> "I believe that abortion should be legal and accessible to everyone, as it is essential healthcare. It is critical for women to have control over their own bodies and make decisions about abortion without government interference. Safe and legal access to abortion ensures that women can make the best choices for their health and well-being."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

Please write **two** or more sentences.

Candidate Summaries: 8/10

Here is a possible summary:

> "I believe that abortion should be illegal in most cases and only allowed under specific circumstances. It is morally wrong and equates to taking a life, as I believe in the sanctity of human life starting from conception. People should take responsibility for their actions and should never use abortion as a form of birth control."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

Please write **two** or more sentences.

Candidate Summaries: 9/10

Here is a possible summary:

> "I believe that abortion should be legal and that women should have the right to make decisions about their own bodies. The decision to have an abortion should be a private matter, free from societal judgment, as it's a crucial aspect of women's health and safety. Society should ensure that abortions are safe, legal, and accessible, respecting the individual's autonomy and privacy in making such decisions."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

Please write **two** or more sentences.

Candidate Summaries: 10/10

Here is a possible summary:

> "I believe that abortion should not be used as a form of birth control. However, it should be legally permissible in cases where the mother's life is in danger, and in instances of rape or incest within an early timeframe. Furthermore, there should be robust educational efforts to prevent unwanted pregnancies and ensure people are informed about the complexities and consequences of abortion."

How well does this summary capture your viewpoint on abortion?

Choices: very poorly, poorly, moderately, well, very well, excellently, exceptionally

Explain which parts you **agree** with, which parts you **disagree** with, and what would need to be **added or made more concrete** to fully represent your viewpoint.

Please write **two** or more sentences.

Thank You

Thank you for completing the survey!

Click below to register your survey completion with Prolific and return to their website.