Language Agents as Digital Representatives in Collective Decision-Making

Daniel Jarrett[*]
Google DeepMind

Miruna Pîslar[*]
Google DeepMind

Michiel A. Bakker
Google DeepMind

Michael Henry Tessler
Google DeepMind

Raphael Köster
Google DeepMind

Jan Balaguer
Google DeepMind

Romuald Elie
Google DeepMind

Christopher Summerfield
Google DeepMind

Andrea Tacchetti
Google DeepMind

Abstract

Consider the process of collective decision-making, in which a group of individuals interactively select a preferred outcome from among a universe of alternatives. In this context, “representation” is the activity of making an individual’s preferences present in the process via participation by a proxy agent—i.e. their “representative”. To this end, learned models of human behavior have the potential to fill this role, with practical implications for multi-agent scenario studies and mechanism design. In this work, we investigate the possibility of training *language agents* to behave in the capacity of representatives of human agents, appropriately expressing the preferences of those individuals whom they stand for. First, we formalize the setting of *collective decision-making*—as the episodic process of interaction between a group of agents and a decision mechanism. On this basis, we then formalize the problem of *digital representation*—as the simulation of an agent’s behavior to yield equivalent outcomes from the mechanism. Finally, we conduct an empirical case study in the setting of *consensus-finding* among diverse humans, and demonstrate the feasibility of fine-tuning large language models to act as digital representatives.

1 Introduction

Collective decision-making is a hallmark of intelligent behavior [1, 2], and is ubiquitous in economic, social, and political spheres of society [3, 4]. Consider any process in which a group of individuals interactively select a preferred outcome from a universe of alternatives: For instance, consider a diverse group of humans communicating in a mediated environment, seeking to arrive at a consensus opinion on a topic of debate [5, 6]. In this context, “representation” is the activity of making an (otherwise absent) individual’s preferences present in the process through participation by a proxy agent—i.e. their “representative” [7]. To this end, learned models of human behavior have the potential to fill this role, with practical implications for scenario studies and mechanism design. For instance, it can facilitate personalized and detailed simulations of collective interactions, allowing iterative refinement of mechanisms before real-world deployment. The key question is: *What makes a good representative*?

Simulation for Representation In this work, we explore the possibility of training *language agents* to behave in the capacity of representatives of human agents, appropriately expressing the preferences of those individuals whom they stand for. On the one hand, this has natural connections to existing work in simulating human behavior, such as learning from demonstrations [8–11], generating synthetic data, [12–15], and creating plausible simulacra of social agents [16–21]. On the other hand, unlike such prior work, our objective in simulating human behavior is specifically for “representation”. This entails unique requirements: (a) grounding in the context of *collective interaction*, (b) attending to the granularity of *individual fidelity*, and (c) operating in the high-dimensional domain of *language space*. Specifically, in this work, we are exploring what representation means, how to measure representativity, and whether language agents can be trained to act as representatives on behalf of humans.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

[*] Equal contribution.

Contributions We make three key contributions in this investigation. First, we formalize the setting of *collective decision-making* as the episodic process of interaction between a group of agents and a decision mechanism (Section 2). Building upon this, our second contribution is to characterize the problem of *digital representation* as the simulation of an agent's behavior so as to yield equivalent outcomes when interacting with the mechanism (Section 3). Finally, we conduct an empirical case study in the setting of *consensus-finding* in natural language among diverse humans, and demonstrate the feasibility of fine-tuning large language models to act as such digital representatives (Section 4).

2 Collective Decision-Making

Consider a general setting for collective decision-making, where interactions between a group of individuals are mediated by a *decision mechanism*, yielding a process that produces a final outcome. By way of preliminaries, we first formalize the notion of a *social choice function*, which is an abstract mapping from preferences to outcomes that is implemented by a mechanism:

Definition 1 (Social Choice Function) Let $N := \{1, \dots, n\}$ denote a set of *participants*, who are engaged in the process of making a collective decision. Denote with Ω the space of *outcomes*, from which the participants are required to make a final selection $\omega \in \Omega$. Denote with $\theta_i \in \Theta_i$ the *type* characterizing each participant $i \in N$, which captures their preferences over different outcomes. Moreover, let $\theta := (\theta_i)_{i \in N} \in \Theta := \Theta_1 \times \cdots \times \Theta_n$ indicate the *type profile* of all participants. Then a *social choice function* h is a mapping from the space of type profiles into the space of outcomes,

$$h : \Theta \to \Omega \tag{1}$$

In line with related fields, "participants" may also be referred to as "agents" and "players", and an "outcome" may be synonymous with an "alternative" and a "collective decision". ◇

We operate in the standard setting for discrete-time Markov decision processes. Let $x \in \mathcal{X}$ denote the *state* variable: While we keep notation simple, x may play the role of "contexts" and "histories" that capture all observations prior to the current time step. Let $u_i \in \mathcal{U}_i$ denote the *action* variable for $i \in N$, and let $u := (u_i)_{i \in N} \in \mathcal{U} := \mathcal{U}_1 \times \cdots \times \mathcal{U}_n$ indicate the *action profile* for all participants. In our setting, u will largely play the role of "utterances" in language space. Denote with T the finite length of each decision episode. The outcome of each episode is precisely the terminal state $\omega := x^T$.

Definition 2 (Decision Mechanism) Each participant $i \in N$ is associated with a *behavior* denoted with $\pi_i \in \Pi_i \subseteq \Delta(\mathcal{U}_i)^{\mathcal{X}}$, and let $\pi := (\pi_i)_{i \in N} \in \Pi := \Pi_1 \times \cdots \times \Pi_n$ give the *behavior profile* for all participants. A *decision mechanism* τ is a mapping from states and action profiles to next states,

$$\tau \in \mathcal{T} \subseteq \Delta(\mathcal{X})^{\mathcal{X} \times \mathcal{U}} \tag{2}$$

and an *outcome function* f maps behavior profiles and mechanisms into distributions over outcomes. In our Markov decision process setting, the outcome function is simply the result of "rolling out" an episode of the interactive process between π and τ to arrive at (a distribution over) terminal states,

$$f : \Pi \times \mathcal{T} \to \Delta(\Omega) \tag{3}$$

A "behavior" may synonymously be referred to as a "policy" or "strategy", and a "mechanism" may be referred to as a "game" or "environment". Our setting is in general non-stationary: We may use superscripts $t \in \{1, \dots, T\}$ to indicate time steps, but omit them unless explicitly required. ◇

Finally, a mechanism τ is said to *implement* a social choice function h when $h(\theta) = f(\pi, \tau)$ for all $\pi \in \Pi$. Both π, τ may depend on θ, though not required. Often, mechanisms are designed with an *optimization objective* in mind. For instance, let each participant be associated with a *payoff function*,

$$g_i : \Omega \times \Theta_i \to \mathbb{R} \tag{4}$$

such that $g_i(\omega, \theta_i)$ gives the payoff that participant $i \in N$ enjoys from the collective decision ω. As before, for convenience, let $g := (g_i)_{i \in N}$ and (with some abuse of notation) write $g : \Omega \times \Theta \to \mathbb{R}^n$. Then a "utilitarian" social choice function is implemented as $h(\theta) = f(\pi, \tau^*)$, by the mechanism:

$$\tau^* \in \arg\max_{\tau \in \mathcal{T}} \mathbb{E}_{\omega \sim f(\pi, \tau)} G(\omega, \theta) \qquad \text{where} \qquad G(\omega, \theta) := \frac{1}{n} \sum_{i=1}^{n} g_i(\omega, \theta_i) \tag{5}$$

Note that a "payoff function" is often interchangeable with a "reward function" or "utility function". In this work, we use the scenario of *consensus-finding* among individuals as our illustrative example.

Case Study (Consensus-Finding) In this setting, a group of human participants shares their thoughts on a debated topic using natural language. Through interaction with a mediation mechanism, they aim to reach a consensus on the matter [5,6]. Specifically, an episode begins with a set of N participants observing a *question* of interest, sampled from a corpus of questions. For instance, this may be: "Should we adopt a universal basic income (UBI) policy?" (see Figure 1). Each participant expresses their *opinion* in written text. Next, the mediator mechanism processes these opinions and outputs a *draft consensus* statement. Each participant then provides their own *critique* of the draft consensus in written text. Note that participants do not observe each other's opinions or critiques. Finally, the mediator mechanism processes these critiques to produce a *revised consensus* statement. The episode ends with participants viewing the revised consensus, followed by an individual demographics *questionnaire*.

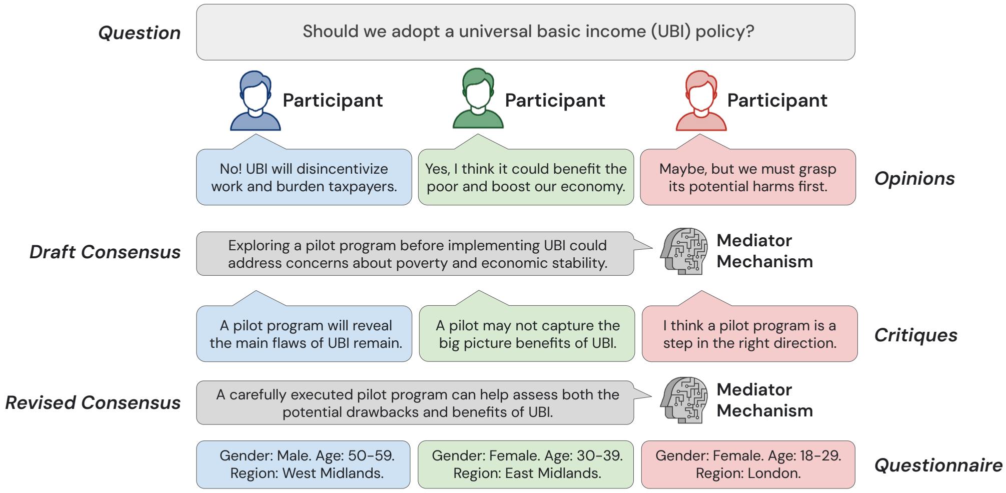
Question
Should we adopt a universal basic income (UBI) policy?
Participant
Participant
Participant
No! UBI will disincentivize work and burden taxpayers.
Yes, I think it could benefit the poor and boost our economy.
Maybe, but we must grasp its potential harms first.
Opinions
Draft Consensus
Exploring a pilot program before implementing UBI could address concerns about poverty and economic stability.
Mediator Mechanism
A pilot program will reveal the main flaws of UBI remain.
A pilot may not capture the big picture benefits of UBI.
I think a pilot program is a step in the right direction.
Critiques
Revised Consensus
A carefully executed pilot program can help assess both the potential drawbacks and benefits of UBI.
Mediator Mechanism
Gender: Male. Age: 50–59. Region: West Midlands.
Gender: Female. Age: 30–39. Region: East Midlands.
Gender: Female. Age: 18–29. Region: London.
Questionnaire

Figure 1: *Consensus-Finding*. Observe that this is an instance of a collective decision-making setting:

- N is the set of participants, typically 3–5 for each episode;
- $\Omega = \mathcal{X}$ is the space of revised consensus statements;
- $x^1 \in \mathcal{X}$ is the *question* of interest;
- $u_i^1 \in \mathcal{U}$ is the *opinion* of participant $i \in N$ on the question;
- $x^2 \in \mathcal{X}$ is the *draft consensus* statement taking those into account;
- $u_i^2 \in \mathcal{U}$ is the *critique* of participant $i \in N$ on the draft consensus;
- $x^3 \in \mathcal{X}$ is the *revised consensus* statement taking those into account;
- $u_i^3 \in \mathcal{U}$ is the dummy action that terminates the episode with a demographics questionnaire;
- $\theta \in \Theta$ is the participants' preference profile over different consensus opinions;
- $\pi_i(\cdot|x)$ is the utterance behavior of participant $i \in N$ in response to $x \in \mathcal{X}$; and
- $\tau(\cdot|x, u)$ is the mediation mechanism's transition function at $x \in \mathcal{X}$, $u \in \mathcal{U}$.

Remark (Mediator Mechanism) Two comments deserve brief mention. Firstly, for added context, the mediator mechanism τ is a 70B-parameter Chinchilla [22] model, fine-tuned as described in [5,6] to best assist humans in collectively producing written consensus statements with maximal approval (viz. Equation 5). In particular, these statements are preferred more strongly than several high-quality baselines, including individual human opinions and consensus statements from human mediators.[1] Secondly, we want to emphasize that for our purposes, we treat the mechanism as a black-box transition function. Our goal is to investigate how *digital representatives* can be trained to operate on behalf of humans in collective decision-making settings, with the consensus-finding scenario as our primary example. In this sense, the inner details of the mechanism are not of importance to us, save for recognizing that $\mathcal{T}$ (and Π, as we shall see) are families of pre-trained and fine-tuned large language models.

[1] The mediator mechanism was not trained to endorse a specific perspective or persuade others of any viewpoint; rather, it was trained purely to produce consensus statements based on the utterances provided by participants.

3 Digital Representatives

Suppose the true policy profile $\pi^* \in \Pi$, and we are interested in a model policy profile $\tilde{\pi}$ that approximates some or all of π^* with digital representatives, in the context of decision mechanisms $\tau \in \mathcal{T}$. (For example, this might involve substituting a part of a single participant's policy, or multiple participants' policies, within π^*). *What defines the set of profiles $\tilde{\pi}$ that we can consider as "equivalent" to π^*?*

3.1 Digital Clones

One obvious choice is to require that $\pi^*(u|x) = \tilde{\pi}(u|x)$ for all $x \in \mathcal{X}$ and $u \in \mathcal{U}$, essentially seeking an identical clone of π^*. This is fine as an objective for *training* language agents to serve as digital representatives—in fact, likelihood-based training is exactly what we perform in Section 4. However, we argue that matching conditionals alone is incomplete for *evaluating* digital representatives, as it may be deemed both "too strong" and "too weak". On one hand, it is "too strong" as $\tilde{\pi}$ may not need to clone the entirety of π^*—*in the context of* τ. Intuitively, suppose there were some decomposition,

$$\mathcal{U} = \mathcal{U}_{||} \times \mathcal{U}_{\perp} \tag{6}$$

such that $\mathcal{U}_{||}$ somehow encapsulated the "relevant" dimensions of utterances, whereas $\mathcal{U}_{\perp}$ encapsulated the "irrelevant" dimensions. It may be helpful to draw an analogy with the "content" vs. "style" distinction in the context of image classification [23], or the "meaning" vs. "form" distinction in the context of semantic clustering [24]. If this is true, we would only care about matching the conditional,

$$\pi(u_{||}|x) = \textstyle\int_{\mathcal{U}_{\perp}} \pi(u|x) du_{\perp} \tag{7}$$

That is, we would simply require that $\pi^*(u_{||}|x) = \tilde{\pi}(u_{||}|x)$, for $x \in \mathcal{X}$ and $u_{||} \in \mathcal{U}_{||}$. If we had access to such a decomposition, the matter would be settled. But such a factoring is seldom apparent.

On the other hand, defining representativity purely on the basis of conditionals is also "too weak", as the policies in $\tilde{\pi}$ are ultimately unrolled in one or more transitions by interacting through the mechanism. Any discrepancy in conditionals may lead to compounding errors through such interactions.

3.2 Digital Representatives

A complete measure of representativity should also consider the dynamics of interaction. Here, we draw connections with value-aware learning [25, 26] and value equivalence [27–30] in model-based reinforcement learning. This allows us to define representation in a way that addresses both concerns at once. First, some more notation: Given any $\pi \in \Pi$, $\tau \in \mathcal{T}$, the vector of *expected payoffs* to participants at state x taking action u at time t is given by the "value function" $Q^t_{\pi,\tau} : \mathcal{X} \times \mathcal{U} \to \mathbb{R}^n$:

$$Q^t_{\pi,\tau}(x, u) = \mathbb{E}_{x^T \sim f(\pi,\tau)}\big[g(x^T, \theta)|x^t = x, u^t = u\big] \tag{8}$$

Moreover, define the Bellman operator $\mathbb{B}_{\pi,\tau}$ over the space of functions $Q : \mathcal{X} \times \mathcal{U} \to \mathbb{R}^n$ such that

$$(\mathbb{B}_{\pi,\tau}Q)(x, u) := \mathbb{E}_{x' \sim \tau(\cdot|x,u)}\mathbb{E}_{u' \sim \pi(\cdot|x')}Q(x', u') \tag{9}$$

Then the sequence of functions $Q^{1:T}_{\pi,\tau}$ is the unique solution to backward recursions $Q^t = \mathbb{B}^t_{\pi,\tau}Q^{t+1}$ for $t \in \{1, \ldots, T-1\}$, and $Q^T(x, u) := g(x, \theta)$. Now we examine three notions of equivalence:

Definition 3 (Representational Equivalence) Fix Π and $\mathcal{T}$. Define first the set of policy profiles $\pi \in \Pi$ for which actions profiles $u \sim \pi(\cdot|x)$ are equal in distribution to $u \sim \pi^*(\cdot|x)$ given any state:

$$\Pi(\pi^*) := \{\pi \in \Pi : \pi^*(\cdot|x) = \pi(\cdot|x) \;\; \forall\, x \in \mathcal{X}\} \tag{10}$$

Instead of matching *conditionals* in isolation, we may focus on *transitions*: Define the set of policy profiles $\pi \in \Pi$ whose operators $\mathbb{B}_{\pi,\tau}$ have equal effect to $\mathbb{B}_{\pi^*,\tau}$ on any function $Q \in \mathcal{Q} \subseteq (\mathbb{R}^n)^{\mathcal{X} \times \mathcal{U}}$:

$$\Pi(\pi^*, \mathcal{T}, \mathcal{Q}) := \{\pi \in \Pi : \mathbb{B}_{\pi^*,\tau}Q = \mathbb{B}_{\pi,\tau}Q \quad \forall\, \tau \in \mathcal{T} \text{ and } Q \in \mathcal{Q}\} \tag{11}$$

Finally, we may assess equivalence based on payoffs of *trajectories*: Define the set of policy profiles whose operators $\mathbb{B}^{1:T}_{\pi,\tau}$ have identical effect to $\mathbb{B}^{1:T}_{\pi^*,\tau}$ when all applied onto any $Q^T \in \mathcal{Q} \subseteq (\mathbb{R}^n)^{\mathcal{X} \times \mathcal{U}}$:

$$\Pi^T(\pi^*, \mathcal{T}, \mathcal{Q}) := \{\pi \in \Pi : \mathbb{B}^1_{\pi^*,\tau} \circ \cdots \circ \mathbb{B}^T_{\pi^*,\tau}Q^T = \mathbb{B}^1_{\pi,\tau} \circ \cdots \circ \mathbb{B}^T_{\pi,\tau}Q^T \; \forall\, \tau \in \mathcal{T}, Q^T \in \mathcal{Q}\} \tag{12}$$

Since repeated application of a Bellman operator T times turns any function Q^T into a value function for the beginning of an episode, this condition effectively asks for equality in expected payoffs. $\diamond$

Expression 10 is simply the singleton class of digital clones. However, Expressions 11 and 12 both capture a notion of *invariance* in Π in the sense that they define potentially larger classes of policy profiles. Moreover, they are both defined on the basis of interaction with the mechanism. It turns out that Expression 11 still demands too much, and 12 provides the better definition for "representation":

Proposition 1 (Representational Equivalence) Fix Π and $\mathcal{T}$, and let $\mathcal{Q}$ be closed under Bellman updates. Consider the equivalence classes of policy profiles induced by π^* from Definition 3. We have

$$\Pi(\pi^*) \subseteq \Pi(\pi^*, \mathcal{T}, \mathcal{Q}) \subseteq \Pi^T(\pi^*, \mathcal{T}, \mathcal{Q}) \tag{13}$$

In particular, consider the maximal space of functions $\mathcal{Q} = (\mathbb{R}^n)^{\mathcal{X}\times\mathcal{U}}$. If the space of mechanisms is also maximal, that is $\mathcal{T} = \Delta(\mathcal{X})^{\mathcal{X}\times\mathcal{U}}$, then we have that the first subset relation is an equality, that is

$$\Pi(\pi^*) = \Pi(\pi^*, \mathcal{T}, \mathcal{Q}) \subseteq \Pi^T(\pi^*, \mathcal{T}, \mathcal{Q}) \tag{14}$$

Let action profiles be decomposable as $\mathcal{U} = \mathcal{U}_{||} \times \mathcal{U}_\perp$ with $\text{card}(\mathcal{U}_\perp) > 1$, and the interaction between the mechanism and policies be such that values $Q^t_{\pi,\tau}(x,u) = Q^t_{\pi,\tau}(x,(u_{||},u_\perp)) = Q^t_{\pi,\tau}(x,u_{||})$ for all $t < T$, $x \in \mathcal{X}$, $u \in \mathcal{U}$, $\pi \in \Pi$, and $\tau \in \mathcal{T} \subseteq \Delta(\mathcal{X})^{\mathcal{X}\times\mathcal{U}}$. Then the second subset relation is proper,

$$\Pi(\pi^*) \subseteq \Pi(\pi^*, \mathcal{T}, \mathcal{Q}) \subset \Pi^T(\pi^*, \mathcal{T}, \mathcal{Q}) \tag{15}$$

Proof. Appendix A. □

If the class of mechanisms is maximal, then the transition-based equivalence class $\Pi(\pi^*, \mathcal{T}, \mathcal{Q})$ is reduced to a singleton, whereas it is still possible for the trajectory-based $\Pi^T(\pi^*, \mathcal{T}, \mathcal{Q})$ to be larger (viz. Expression 14). More importantly, if the class of mechanisms is such that expected payoffs are invariant to some $\mathcal{U}_\perp \subseteq \mathcal{U}$, then the equivalence class $\Pi^T(\pi^*, \mathcal{T}, \mathcal{Q})$ is *strictly* the largest. The intuition is that $\tilde{\pi} \in \Pi^T(\pi^*, \mathcal{T}, \mathcal{Q})$ are free to define arbitrary behavior within $\mathcal{U}_\perp$ without affecting expected payoffs of final outcomes. As an example, suppose the class of mechanisms $\mathcal{T} \subset \Delta(\mathcal{X})^{\mathcal{X}\times\mathcal{U}}$ is itself invariant to some $\mathcal{U}_\perp \subseteq \mathcal{U}$ (that is, $\tau(\cdot|x,u) = \tau(\cdot|x,(u_{||},u_\perp)) = \tau(\cdot|x,u_{||})$ for all $x \in \mathcal{X}$, $u \in \mathcal{U}$, $\tau \in \mathcal{T}$). It is easy to see—by induction from $Q^T_{\pi,\tau}(x,u) := g(x,\theta)$—that the invariance assumption on value functions required for Expression 15 holds (but not necessarily for Expression 14).

Value Equivalence In summary, this motivates a simple estimate of "representativity" that captures how much a given model profile $\tilde{\pi}$ deviates from the true profile π^* through their expected payoffs:

$$\textit{Representativity}(\pi^*, \tilde{\pi}) := \max_{\tau \in \mathcal{T}, Q^T \in \mathcal{Q}} \mathcal{L}\big(\mathbb{E}_{\omega \sim f(\pi^*,\tau)} Q^T(\omega), \mathbb{E}_{\omega \sim f(\tilde{\pi},\tau)} Q^T(\omega)\big) \tag{16}$$

where $\mathcal{L}$ is some measure of discrepancy between its vector arguments. In other words, $\tilde{\pi}$ is a "good" representative of π^* if behaviors elicited from each profile are such that, when aggregated through mechanisms $\tau \in \mathcal{T}$, they yield outcomes that (in expectation) are equivalent as measured by $Q^T \in \mathcal{Q}$.[2]

4 Case Study: Consensus-Finding

In the context of consensus-finding (viz. Figure 1) as collective decision-making, we now turn to the empirical question: *Is it feasible to train language agents to act as digital representatives of humans?* Indeed, large language models are often pre-trained on datasets rich in diverse preferences, and recent work has shown that language agents can generate plausible behavior in interactive settings [21], and to simulate subpopulations of humans in studies in economics [20], psychology [17], and social science [19]. However, while socio-demographically prompted models can capture the sentiment of general subpopulations of humans [31, 32], they are not as reliable on more granular levels [18, 33]. Is it possible to fine-tune language models on a personalized level to represent *individuals* in a group?

4.1 Experiment Setup

In this work, we train and evaluate digital representatives using a consensus-finding dataset from [5, 6]. For completeness, we briefly recall the input questions and data collection process [5, 6]. Then we give an overview of the final dataset and the training process for digital representatives in our experiment.

[2] Note that this is reminiscent of how environment models are judged based on value equivalence [25–30]. However, instead of defining equivalent environments based on a reference class of policies for reinforcement learning, here we are defining equivalence classes of multi-agent policy profiles based on a reference class of mechanisms.

Questions The corpus of input questions concern reasonably divisive political issues relevant in the United Kingdom. Questions were generated using a pre-trained 70B-parameter Chinchilla [22] language model. First, 152 "seed" questions on contemporary topics of debate were written by hand. Random samples of 10 seed questions at a time were used to prompt the model to generate final questions, for a total of 3,500 unique questions. Questions likely to prompt extremist views or discriminatory language were removed by hand, leaving a corpus of 2,922 questions. See [5,6] for more detail.

Data Collection In each episode of data collection, a group of 3–5 participants based in the United Kingdom participated in an implementation of the protocol described in Figure 1. To begin, each participant π_i^* viewed a question x^1 sampled from the corpus. Next, they each wrote an opinion u_i^1 by typing free text into an input box in a custom web application. Then, the opinions of the group were included in a prompt provided to the mediator mechanism τ to generate a draft consensus statement x^2. Each participant then wrote a critique u_i^2 by typing free text into another input box. Finally, these were all included in another prompt provided to the mediator mechanism to generate a revised consensus statement x^3, after which participants answered a short demographic questionnaire u_i^3. Separately, before writing their initial opinion, participants were asked to provide a "position" score indicating their agreement on the declarative form of the question (so a question of the form "Should we [...]" will have declarative form "We should [...]"), to allow measuring baseline disagreement among the group. An average session took 45–60 minutes, and each set of participants were engaged in three different episodes (i.e., corresponding to three different questions).[3] See [5, 6] for more detail.

Experiment Dataset The final dataset we use for experiments is the result of 1,662 episodes of data collection, corresponding to a total of 2,290 participants and 6,872 written opinions and critiques. For each episode, the dataset includes the text of each question, written opinions, draft consensus statement, written critiques, and revised consensus statement. For each participant in a episode, the metadata recorded include the other questions that the participant had engaged in from a past debate, as well as the written opinions and critiques of that participant on those other questions. The questionnaire each participant completed covers a range of demographic information, including gender identity, age (in decades), ethnicity, region of domicile, education, political and religious affiliation, immigration status, and approximate household income. For each question, the initial "position" score of each participant was also recorded. We split this final dataset into training and validation sets on the basis of *both* data collection episodes and participant identity. This ensured that each participant and each question only appears in either the training set or the validation set, but not both. This produced a training set with 1,332 episodes of data collection, corresponding to 1,828 participants and 5,484 instances of written opinions and written critiques, and a validation set with 330 episodes of data collection, corresponding to 462 participants and 1,386 instances of written opinions and critiques.

Digital Representatives In our experiments, we focus on learning digital representatives of human participants for the *critique* step (u^2). Let $\hat{\pi}_i$ be a model trained to produce a written critique on the basis of a question, participant i's opinion, a draft consensus, and (optionally) any additional information about that participant. Then the digital representative $\tilde{\pi}_i$ for that participant is defined as:

$$\tilde{\pi}_i(\cdot|x^t) := \begin{cases} \hat{\pi}_i\,(\cdot|x^t) & \text{if } t = 2 \\ \pi_i^*(\cdot|x^t) & \text{otherwise} \end{cases} \tag{17}$$

Our training pipeline follows a standard supervised learning procedure for fine-tuning large language models to learn from demonstrations. We train 1B- and 30B-parameter adaptations of Chinchilla [22] models as digital representatives $\hat{\pi}_i$, using the dataset to estimate and maximize the training objective:

$$\mathcal{J}_{\text{training}}(\phi) := \mathbb{E}_{\substack{\pi_i^*, \tilde{\pi}_i \sim \mathcal{D}_{\text{training}} \\ \vdots \\ u_i^2 \sim \pi_i^*(\cdot|x^2)}} \big[\log \tilde{\pi}_i(u_i^2|x^2) \big] \tag{18}$$

where ϕ denotes the trainable parameters of the model, and the vertical ellipsis indicates the sampling of intermediate states and actions omitted for brevity. We fine-tune all model parameters using a constant learning rate of $2 \cdot 10^{-6}$, a batch size of 128, and the AdamW optimizer [34]. During training, we evaluate the model's performance by calculating the perplexity on the validation set. This metric is employed for early stopping to prevent overfitting. Following this procedure, we select the best checkpoints after 150 training steps for the 1B model and after 10 steps for the 30B model.

[3] Full details of the data collection procedure were first reviewed by an independent ethics review committee. No personally identifiable data was collected, and no offensive content was shown to participants. All participants had provided their informed consent before engaging, and were compensated £15 per hour for their involvement.

Table 1: *Example Critique*. Critiques sampled from a vanilla/fine-tuned digital representative ("DR").

Ground Truth	I agree that some sort of tax should be mandatory but as I said earlier if its too severe these wealthy people wont [**sic**] stay on these islands and take their wealth elsewhere, its [**sic**] one thing having wealth but it has to attractive for these people.
Vanilla DR 1B	It is really quite unclear why the government wants to make this change. It does not make sense to make such a drastic change to the wealth tax. The logic of the government is that by introducing such a tax they will be able to raise £45 billion in revenues. The same money will be used to build infrastruct-[**...**]
Fine-Tuned DR 1B	I agree with this opinion and think it captures the nuances of the issue well.
Vanilla DR 30B	It is a little unclear to me why we should have this tax at all. I understand the problem of inequality, but I'm not sure that a flat tax would help at all. We could be doing more good by helping people on lower incomes to find jobs.
Fine-Tuned DR 30B	I agree with this. However the tax should not be so severe as to make people leave the country and therefore the gap would get wider.

4.2 Results: Critique Evaluation

First, we evaluate digital representatives at the level of individual critiques. Our first performance measure is the log-likelihood of ground-truth critiques in the validation set, under the learned model:

$$\mathcal{J}^{(\text{critique})}_{\text{likelihood}}(\phi) := \mathbb{E}_{\substack{\pi_i^*,\tilde{\pi}_i\sim\mathcal{D}_{\text{validation}} \\ \vdots \\ u_i^2\sim\pi_i^*(\cdot|x^2)}}\big[\log\tilde{\pi}_i(u_i^2|x^2)\big] \tag{19}$$

Another measure involves prompting a large version of PaLM 2 [35] (i.e. over 1–2 orders of magnitude larger than digital representatives) as an autorater, reporting the win-rate against a baseline critique:

$$\mathcal{J}^{(\text{critique})}_{\text{autorater}}(\phi) := \mathbb{E}_{\substack{\pi_i^*,\tilde{\pi}_i\sim\mathcal{D}_{\text{validation}} \\ \vdots \\ u_i^2\sim\tilde{\pi}_i(\cdot|x^2) \\ v_i^2\sim\pi_{\text{baseline}}}}\big[\textit{Autorater}(u_i^2 \succ v_i^2|x^2)\big] \tag{20}$$

where the golden baseline is the ground-truth critique (i.e. $\pi_{\text{baseline}} := \pi^*$). See Table 1 for some critique samples. See Figure 2 for results on ablations and baseline (additional results in Appendix B).

4.3 Results: Consensus Evaluation

Second, we evaluate digital representatives at the level of consensus outcomes. Our first performance measure is the discrepancy in expected payoff (cf. Equation 16) relative to some baseline consensus, where payoffs are estimated using a payoff model accompanying the work in [5, 6], which is a model based on a 1B-parameter Chinchilla [22] that outputs a scalar "agreement" score[4] for each participant:

$$\mathcal{J}^{(\text{outcome})}_{\text{payoff}}(\phi) := \mathbb{E}_{\substack{\pi^*,\tilde{\pi}\sim\mathcal{D}_{\text{validation}} \\ \vdots \\ u^2\sim\tilde{\pi}(\cdot|x^2) \\ x^3\sim\tau(\cdot|x^2,u^2) \\ y^3\sim\tau_{\text{baseline}}}}\big[\mathcal{L}\big(Q^3(x^3),Q^3(y^3)\big)\big] \tag{21}$$

Here, $\mathcal{L}$ computes the average element-wise difference between its two inputs, for participants who are substituted by their digital representatives. We test two regimes of model policy profiles $\tilde{\pi}$: One in which a single participant is substituted, that is $\tilde{\pi}_{-i} := (\pi_1^*,\ldots,\tilde{\pi}_i,\ldots,\pi_n^*)$, and one in which all of them are substituted, that is $\tilde{\pi}_N := (\tilde{\pi}_1,\ldots,\tilde{\pi}_n)$. As before, another measure involves an autorater:

$$\mathcal{J}^{(\text{outcome})}_{\text{autorater}}(\phi) := \mathbb{E}_{\substack{\pi^*,\tilde{\pi}\sim\mathcal{D}_{\text{validation}} \\ \vdots \\ u^2\sim\tilde{\pi}(\cdot|x^2) \\ v^2\sim\pi^*(\cdot|x^2) \\ x^3\sim\tau(\cdot|x^2,u^2) \\ y^3\sim\tau_{\text{baseline}}}}\big[\textit{Autorater}(x^3 \succ y^3|x^2,v^2)\big] \tag{22}$$

where we report the win-rate against a baseline consensus (and the golden baseline is the ground-truth consensus). See Figure 3 for results on ablations and baselines (and further results in Appendix B).

[4] The use of a payoff model serves as a proxy for human endorsement without replicating the entire study of [5, 6].

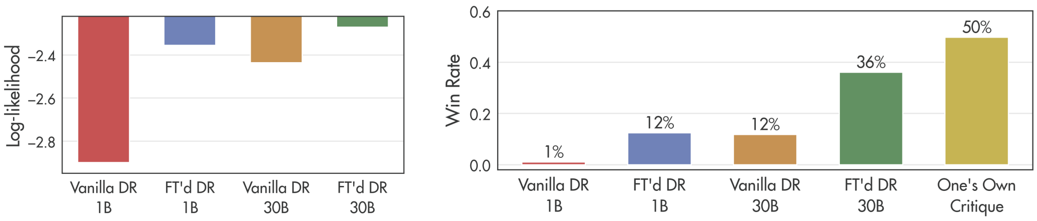

Figure 2: *Critique Evaluation*. **Left**: Mean log-likelihood of ground-truth critiques from human participants (from the validation set), evaluated under models $\hat{\pi}_i$ with *fine-tuned* ("FT'd") or *vanilla* digital representatives ("DRs") with 1B or 30B parameters. The fine-tuned models consistently exhibit higher log-likelihoods compared to their vanilla counterparts, indicating a superior representation of participants' ground-truth critiques. **Right**: The autorater's win-rate of a sampled critique (for different models $\hat{\pi}_i$ across the x-axis) against one's own ground-truth critique (i.e. the baseline). The golden bar represents π^* against itself, serving as a reference for ceiling performance. **Conclusion**: *Fine-tuning and scale both improve the representativeness of DRs for held-out participants' critiques.*

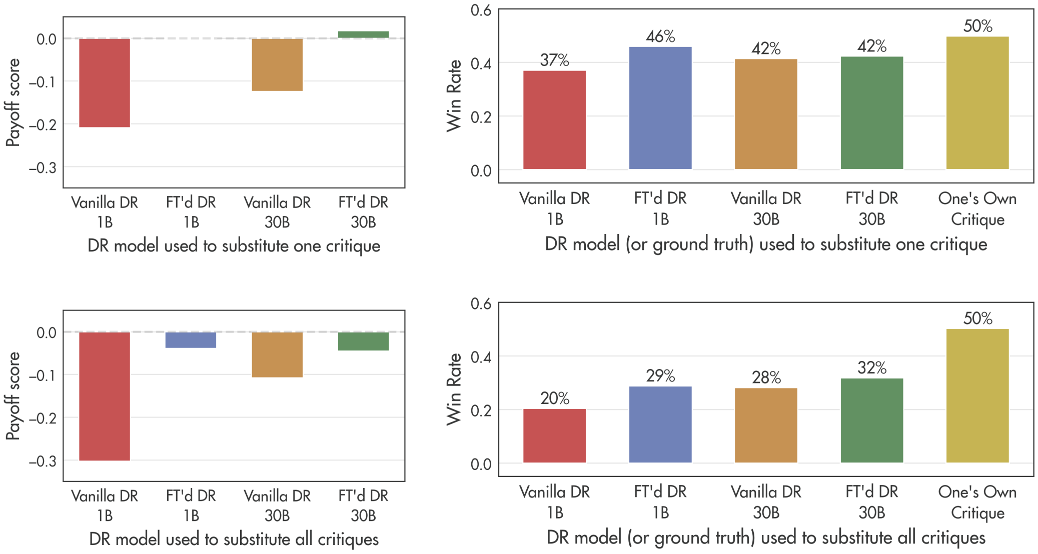

Figure 3: *Consensus Evaluation*. Either one participant's critique (**top**), or all participants' critiques (**bottom**) are substituted with critiques sampled from their respective digital representatives $\hat{\pi}_i$. **Left**: Mean discrepancy in payoffs between the revised consensus generated by the mediator mechanism using participants' ground-truth critiques, versus using critiques sampled from DRs. Replacing either one or all ground truth critiques with vanilla DR samples results in significantly degraded payoffs, whereas fine-tuned DR samples yield payoffs with little or no degradation, indicating that those consensuses are more or less equivalent. **Right**: The autorater's win-rate of a generated revised consensus (using critiques sampled from different models $\hat{\pi}_i$) against the revised consensus generated using ground-truth critiques (i.e. the baseline). The golden bar represents the baseline consensus against itself, serving as a reference for ceiling performance. Substituting a single critique (chosen at random) results in consensuses perceived as roughly equally similar across all DRs by the autorater (13% difference between ceiling and Vanilla 1B DRs). This is likely a property of the mediation mechanism, which we empirically observe disregards outlier critiques. However, substituting the entire group's critiques significantly influences the revised consensus output (30% difference between ceiling and Vanilla 1B DRs), with the 30B fine-tuned DRs having a notably higher win-rate here. **Conclusion**: *Using the notion of representativity motivated earlier, in both top and bottom panels we also observe that fine-tuning and scale both improve the representativity of DRs for held-out episodes.*

5 Discussion

We fine-tune language agents to act as digital representatives in the context of collective decision-making, offering practical implications for scenario studies and mechanism design. In the sequel, we provide an overview of related work and conclude with a discussion on limitations and future work.

Related Work Our work in "simulation for representation" straddles two domains of research: in (1) how human behavior is simulated with models, and (2) how human behavior is represented by models.

Simulation: First, *imitation learning* deals with training an artificial agent to mimic a demonstrator [8–11], to match some notion of performance [36–39], or to recover some underlying motive for their behavior [40–43]. Multiple artificial agents have also been cloned [44,45], such as for simple control tasks [46], bandit environments [47], and driving simulation [48]. This contrasts with our high-dimensional language domain, and with our focus on "representativity" in collective decision-making. Second, *synthetic data generation* deals with sampling from learned distributions to approximate real phenomena, such as in sequential data [49–51], medical environments [14,52], driving simulation [15], as well as in language space for social science [12], privacy preservation, [53], and to augment text mining [13]. However, this contrasts with our emphasis on learning models to act in an interactive context. Third, recent work has explored creating plausible simulacra of *social agents* using language models, such as in creating populated prototypes for social computing [16], in prompting models to simulate human sub-populations [19] and replicate human subject studies [17, 18], in simulating economic agents [20], and in creating believable social behavior in sandbox environments [21]. In contrast, we fine-tune models on a personalized level to represent specific individuals within a group.

Representation: Since large language models are often pre-trained on human data with diverse perspectives, a recent line of work has seeks to measure the preferences that pre-trained models *inherently reflect* with prompting, such as the representation of broad-based global demographic groups [54], the alignment of their views with demographic categories in the United States [31], as well as to uncover the intrinsic ideological biases that pre-trained models exhibit [55]. Secondly, a related strand of research systematically probes pre-trained models to *actively simulate* biased perspectives, such as by prompting with socio-demographic profiles [33], as well as by prompting with liberal or conservative profiles to sample text with corresponding moral biases [32]. Some models have also been *explicitly fine-tuned* to query the opinions and worldviews of demographic categories of people [56,57]. In contrast, however, in this work we seek to simulate behavior at the granularity of individual representativity, specifically in the context of collective interaction. Finally, while we draw an analogy with value-aware [25, 26] and value equivalent learning [27–30], our work has close connections with broader notions of representation [7] in collective interaction [3], consensus-based decision-making [4], and has potential implications for real and simulated deliberative democracy [58–61].

Future Work We began with the question: "*What makes a good representative?*". Our argument contends that a representative should not only capture the conditional dynamics of behavior but also ensure that its interaction through a mechanism preserves the trajectory-wise dynamics of outcomes. Several caveats are in order: First, there is no escaping the fact that the fidelity of any notion of "representative" is most solidly grounded in human endorsement. While we use a black-box payoff model as proxy, future work would greatly benefit from human validation of digital representatives. Second, in this work we focused on learning representatives for the critique-writing step only. Future work would benefit from naturally extending this to the opinion-writing step as well, yielding fully-simulated representatives. Thirdly, although we defined and evaluated representatives based on an equivalence objective, in our experiments the models were trained on a standard likelihood-based fine-tuning objective. Future work could explore directly training for the equivalence objective.

Broader Impact It is important to emphasize that our training of digital representatives is geared toward *simulation*, not advocating for their deployment as substitutes for human accountability in decision-making. For instance, suppose we were interested in using digital representatives for the purpose of simulating outcomes to improve some downstream tasks. Then evaluation of *those* tasks must be performed with real humans. This is crucial, since we do not train representatives on any notion of factuality or transparency. In fact, representatives are specifically trained to mimic their human inputs per se, thereby carrying over any biases from those corresponding humans. That said, this research was conducted with a focus on societal benefit. The ultimate goal is to facilitate granular simulations of collective interactions for policy design, allowing decision mechanisms to benefit from scalable and cost-effective iterative refinement before real-world deployment.

References

[1] Thomas Bose, Andreagiovanni Reina, and James AR Marshall. Collective decision-making. *Current opinion in behavioral sciences*, 16:30–34, 2017.

[2] Richard P Mann. Collective decision making by rational individuals. *Proceedings of the National Academy of Sciences*, 115(44):E10387–E10396, 2018.

[3] Yadati Narahari. *Game theory and mechanism design*, volume 4. World Scientific, 2014.

[4] Darcy K Leach. When freedom is not an endless meeting: A new look at efficiency in consensus-based decision making. *The Sociological Quarterly*, 57(1):36–70, 2016.

[5] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

[6] Anonymous Authors. Using ai to reduce the divisivness of human opinions. *Submission Under Review by Venue*, 2023.

[7] Suzanne Dovi. Political representation. *The Stanford Encyclopedia of Philosophy*, 2006.

[8] Hoang M Le, Andrew Kang, Yisong Yue, and Peter Carr. Smooth imitation learning for online sequence prediction. *International Conference on Machine Learning (ICML)*, 2016.

[9] Yisong Yue and Hoang M Le. Imitation learning (presentation). *International Conference on Machine Learning (ICML)*, 2018.

[10] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 2018.

[11] Alihan Hüyük, Daniel Jarrett, Cem Tekin, and Mihaela van der Schaar. Explaining by imitating: Understanding decisions by interpretable policy learning. *International Conference on Learning Representations (ICLR)*, 2021.

[12] Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*, 2023.

[13] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*, 2023.

[14] Alex J Chan, Ioana Bica, Alihan Huyuk, Daniel Jarrett, and Mihaela van der Schaar. The medkit-learn (ing) environment: Medical decision modelling through simulation. *arXiv preprint arXiv:2106.04240*, 2021.

[15] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[16] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, 2022.

[17] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

[18] Jacqueline Harding, William D'Alessandro, NG Laskowski, and Robert Long. Ai language models cannot replace human research participants. *AI & SOCIETY*, pages 1–3, 2023.

[19] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[20] John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

[21] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

[22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[23] Jason Wang, Luis Perez, et al. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11(2017):1–8, 2017.

[24] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

[25] Amir-massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Artificial Intelligence and Statistics*, pages 1486–1494. PMLR, 2017.

[26] Amir-massoud Farahmand. Iterative value-aware model learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[27] Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 33:5541–5552, 2020.

[28] Christopher Grimm, André Barreto, Greg Farquhar, David Silver, and Satinder Singh. Proper value equivalence. *Advances in Neural Information Processing Systems*, 34:7773–7786, 2021.

[29] Christopher Grimm, Andre Barreto, and Satinder Singh. Approximate value equivalence. *Advances in Neural Information Processing Systems*, 35:33029–33040, 2022.

[30] Dilip Arumugam and Benjamin Van Roy. Deciding what to model: Value-equivalent sampling for reinforcement learning. *Advances in Neural Information Processing Systems*, 35:9024–9044, 2022.

[31] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.

[32] Gabriel Simmons. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv preprint arXiv:2209.12106*, 2022.

[33] Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. How (not) to use sociodemographic information for subjective nlp tasks. *arXiv preprint arXiv:2309.07034*, 2023.

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019.

[35] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[36] Umar Syed and Robert E Schapire. A reduction from apprenticeship learning to classification. *Advances in neural information processing systems (NeurIPS)*, 2010.

[37] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. *International conference on Machine learning (ICML)*, 2004.

[38] Gergely Neu and Csaba Szepesvári. Apprenticeship learning using irl and gradient methods. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

[39] Monica Babes, Vukosi Marivate, and Michael L Littman. Apprenticeship learning about multiple intentions. *International conference on Machine learning (ICML)*, 2011.

[40] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. *AAAI Conference on Artificial Intelligence (AAAI)*, 2008.

[41] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. *International conference on machine learning (ICML)*, 2016.

[42] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems (NeurIPS)*, 2016.

[43] Daniel Jarrett, Alihan Hüyük, and Mihaela Van Der Schaar. Inverse decision modeling: Learning interpretable representations of behavior. In *International Conference on Machine Learning*, pages 4755–4771. PMLR, 2021.

[44] Eric Zhan, Stephan Zheng, Yisong Yue, Long Sha, and Patrick Lucey. Generative multi-agent behavioral cloning. *arXiv*, 2018.

[45] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9329–9338, 2019.

[46] Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. Multi-agent generative adversarial imitation learning. *Advances in neural information processing systems*, 31, 2018.

[47] Andy Shih, Stefano Ermon, and Dorsa Sadigh. Conditional imitation learning for multi-agent games. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 166–175. IEEE, 2022.

[48] Raunak P Bhattacharyya, Derek J Phillips, Blake Wulfe, Jeremy Morton, Alex Kuefler, and Mykel J Kochenderfer. Multi-agent imitation learning for driving simulation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1534–1539. IEEE, 2018.

[49] Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. Generating high-fidelity, synthetic time series datasets with doppelganger. *ACM Internet Measurement Conference (IMC)*, 2019.

[50] Tianlin Xu, Li K Wenliang, Michael Munn, and Beatrice Acciaio. Cot-gan: Generating sequential data via causal optimal transport. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[51] Daniel Jarrett, Ioana Bica, and Mihaela van der Schaar. Time-series generation by contrastive imitation. *Advances in Neural Information Processing Systems*, 34:28968–28982, 2021.

[52] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

[53] Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*, 2023.

[54] Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

[55] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*, 2023.

[56] Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. Communitylm: Probing partisan worldviews from language models. *arXiv preprint arXiv:2209.07065*, 2022.

[57] Patrick Haller, Ansar Aynetdinov, and Alan Akbik. Opiniongpt: Modelling explicit biases in instruction-tuned llms. *arXiv preprint arXiv:2309.03876*, 2023.

[58] John S Dryzek, André Bächtiger, and Karolina Milewicz. Toward a deliberative global citizens' assembly. *Global Policy*, 2(1):33–42, 2011.

[59] Fredrik Engelstad. The assignment of political office by lot. *Social Science Information*, 28(1):23–50, 1989.

[60] Joseph M Bessette. *The mild voice of reason: Deliberative democracy and American national government*. University of Chicago Press, 1994.

[61] Jan Leike. A proposal for importing society's values. *Musings on the Alignment Problem*, 2023.

A Proof of Proposition 1

Proposition 1 (Representational Equivalence) Fix Π and $\mathcal{T}$, and let $\mathcal{Q}$ be closed under Bellman updates. Consider the equivalence classes of policy profiles induced by π^* from Definition 3. We have

$$\Pi(\pi^*) \subseteq \Pi(\pi^*, \mathcal{T}, \mathcal{Q}) \subseteq \Pi^T(\pi^*, \mathcal{T}, \mathcal{Q}) \tag{13}$$

In particular, consider the maximal space of functions $\mathcal{Q} = (\mathbb{R}^n)^{\mathcal{X}\times\mathcal{U}}$. If the space of mechanisms is also maximal, that is $\mathcal{T} = \Delta(\mathcal{X})^{\mathcal{X}\times\mathcal{U}}$, then we have that the first subset relation is an equality, that is

$$\Pi(\pi^*) = \Pi(\pi^*, \mathcal{T}, \mathcal{Q}) \subseteq \Pi^T(\pi^*, \mathcal{T}, \mathcal{Q}) \tag{14}$$

Let action profiles be decomposable as $\mathcal{U} = \mathcal{U}_{||} \times \mathcal{U}_{\perp}$ with $\text{card}(\mathcal{U}_{\perp}) > 1$, and the interaction between the mechanism and policies be such that values $Q^t_{\pi,\tau}(x,u) = Q^t_{\pi,\tau}(x,(u_{||},u_{\perp})) = Q^t_{\pi,\tau}(x,u_{||})$ for all $t < T$, $x \in \mathcal{X}$, $u \in \mathcal{U}$, $\pi \in \Pi$, and $\tau \in \mathcal{T} \subseteq \Delta(\mathcal{X})^{\mathcal{X}\times\mathcal{U}}$. Then the second subset relation is proper,

$$\Pi(\pi^*) \subseteq \Pi(\pi^*, \mathcal{T}, \mathcal{Q}) \subset \Pi^T(\pi^*, \mathcal{T}, \mathcal{Q}) \tag{15}$$

Proof. The first subset in Expression 13 is immediate from Definitions 9 and 11. The second subset can be seen as follows: Fix $\tilde{\pi} \in \Pi(\pi^*, \mathcal{T}, \mathcal{Q})$, so for any $\tau \in \mathcal{T}$ and $Q \in \mathcal{Q}$ we have the following,

$$\mathbb{B}_{\pi^*,\tau}Q = \mathbb{B}_{\tilde{\pi},\tau}Q \tag{23}$$

But $\mathbb{B}_{\pi^*,\tau}Q$ and $\mathbb{B}_{\tilde{\pi},\tau}Q$ are also in $\mathcal{Q}$ from the closure assumption, so we can repeatedly apply the Bellman operators $\mathbb{B}_{\pi^*,\tau}$ and $\mathbb{B}_{\tilde{\pi},\tau}$ on the left and right hand sides for a total of T times, therefore

$$\mathbb{B}^1_{\pi^*,\tau}\circ\cdots\circ\mathbb{B}^T_{\pi^*,\tau}Q = \mathbb{B}^1_{\tilde{\pi},\tau}\circ\cdots\circ\mathbb{B}^T_{\tilde{\pi},\tau}Q \tag{24}$$

so $\tilde{\pi} \in \Pi^T(\pi^*, \mathcal{T}, \mathcal{Q})$. Next, the equality in Expression 14 is obvious. Finally, the proper subset in Expression 15 can be shown by picking the following model policy $\tilde{\pi}$ as proof (note that $\tilde{\pi} \notin \Pi(\pi^*)$):

$$\tilde{\pi}((u_{||},u_{\perp})|x) := \mathbb{1}_{\{u_{\perp}=\tilde{u}_{\perp}\}}\pi^*(u_{||}|x) \tag{25}$$

First, note that the value function $Q^{1:T}_{\pi^*,\tau}$ for the true policy π^* profile satisfies the following recursion:

$$\begin{aligned}
&Q^t_{\pi^*,\tau}(x,(u_{||},u_{\perp})) && (26)\\
&\quad= (\mathbb{B}^t_{\pi^*,\tau}Q^{t+1}_{\pi^*,\tau})(x,(u_{||},u_{\perp})) && (27)\\
&\quad= \mathbb{E}_{x'\sim\tau(\cdot|x,(u_{||},u_{\perp}))}\mathbb{E}_{(u'_{||},u'_{\perp})\sim\pi^*(\cdot|x')}Q^{t+1}_{\pi^*,\tau}(x',(u'_{||},u'_{\perp})) && (28)\\
&\quad= \mathbb{E}_{x'\sim\tau(\cdot|x,(u_{||},u_{\perp}))}\textstyle\int_{\mathcal{U}_{||}}\int_{\mathcal{U}_{\perp}}\pi^*((u'_{||},u'_{\perp})|x')Q^{t+1}_{\pi^*,\tau}(x',(u'_{||},u'_{\perp}))du'_{||}du'_{\perp} && (29)\\
&\quad= \mathbb{E}_{x'\sim\tau(\cdot|x,(u_{||},u_{\perp}))}\textstyle\int_{\mathcal{U}_{||}}\int_{\mathcal{U}_{\perp}}\pi^*(u'_{||}|x')\pi^*(u'_{\perp}|x')Q^{t+1}_{\pi^*,\tau}(x',(u'_{||},u'_{\perp}))du'_{||}du'_{\perp} && (30)\\
&\quad= \mathbb{E}_{x'\sim\tau(\cdot|x,(u_{||},u_{\perp}))}\textstyle\int_{\mathcal{U}_{||}}\pi^*(u'_{||}|x')Q^{t+1}_{\pi^*,\tau}(x',u'_{||})du'_{||} && (31)\\
&\quad= \mathbb{E}_{x'\sim\tau(\cdot|x,(u_{||},u_{\perp}))}\mathbb{E}_{u'_{||}\sim\pi^*(\cdot|x')}Q^{t+1}_{\pi^*,\tau}(x',u'_{||}) && (32)
\end{aligned}$$

for $t \in \{1,\ldots,T-1\}$, and $Q^T_{\pi^*,\tau}(x,u) = g(x,\theta)$. Likewise, the value function $Q^{1:T}_{\pi^*,\tau}$ for the model policy profile $\tilde{\pi}$ satisfies the following recursion:

$$\begin{aligned}
&Q^t(x,(u_{||},u_{\perp})) && (33)\\
&\quad= (\mathbb{B}^t_{\tilde{\pi},\tau}Q^{t+1})(x,(u_{||},u_{\perp})) && (34)\\
&\quad= \mathbb{E}_{x'\sim\tau(\cdot|x,(u_{||},u_{\perp}))}\mathbb{E}_{(u'_{||},u'_{\perp})\sim\tilde{\pi}(\cdot|x')}Q^{t+1}(x',(u'_{||},u'_{\perp})) && (35)\\
&\quad= \mathbb{E}_{x'\sim\tau(\cdot|x,(u_{||},u_{\perp}))}\textstyle\int_{\mathcal{U}_{||}}\int_{\mathcal{U}_{\perp}}\tilde{\pi}((u'_{||},u'_{\perp})|x')Q^{t+1}(x',(u'_{||},u'_{\perp}))du'_{||}du'_{\perp} && (36)\\
&\quad= \mathbb{E}_{x'\sim\tau(\cdot|x,(u_{||},u_{\perp}))}\textstyle\int_{\mathcal{U}_{||}}\int_{\mathcal{U}_{\perp}}\mathbb{1}_{\{u'_{\perp}=\tilde{u}_{\perp}\}}\pi^*(u'_{||}|x')Q^{t+1}(x',(u'_{||},u'_{\perp}))du'_{||}du'_{\perp} && (37)\\
&\quad= \mathbb{E}_{x'\sim\tau(\cdot|x,(u_{||},u_{\perp}))}\textstyle\int_{\mathcal{U}_{||}}\pi^*(u'_{||}|x')Q^{t+1}(x',(u'_{||},\tilde{u}_{\perp}))du'_{||} && (38)\\
&\quad= \mathbb{E}_{x'\sim\tau(\cdot|x,(u_{||},u_{\perp}))}\mathbb{E}_{u'_{||}\sim\pi^*(\cdot|x')}Q^{t+1}(x',(u'_{||},\tilde{u}_{\perp})) && (39)
\end{aligned}$$

for $t \in \{1,\ldots,T-1\}$, and $Q^T(x,u) = g(x,\theta)$. But $Q^{1:T}_{\pi^*,\tau}$ is a solution to this recursion, and as solutions to backward recursions are unique, we have that $Q^{1:T}_{\tilde{\pi},\tau} = Q^{1:T}_{\pi^*,\tau}$, hence $\tilde{\pi} \in \Pi^T(\pi^*, \mathcal{T}, \mathcal{Q})$.

Next, we show $\tilde{\pi} \notin \Pi(\pi^*, \mathcal{T})$: Pick $Q(x, (u_{||}, u_\perp)) := \mathbb{1}_{\{u_\perp \neq \tilde{u}_\perp\}}$. Then we have that

$$
\begin{aligned}
(\mathbb{B}_{\pi^*,\tau}Q)(x, (u_{||}, u_\perp)) &= \mathbb{E}_{x' \sim \tau(\cdot|x,(u_{||},u_\perp))} \mathbb{E}_{(u'_{||},u'_\perp) \sim \pi^*(\cdot|x')} Q(x', (u'_{||}, u'_\perp)) && (40) \\
&= \mathbb{E}_{x' \sim \tau(\cdot|x,(u_{||},u_\perp))} \mathbb{P}(u'_\perp \neq \tilde{u}_\perp | x', \pi^*) && (41) \\
&= \mathbb{P}(u'_\perp \neq \tilde{u}_\perp | x, (u_{||}, u_\perp), \pi^*, \tau) && (42)
\end{aligned}
$$

and likewise,

$$
\begin{aligned}
(\mathbb{B}_{\tilde{\pi},\tau}Q)(x, (u_{||}, u_\perp)) &= \mathbb{E}_{x' \sim \tau(\cdot|x,(u_{||},u_\perp))} \mathbb{E}_{u'_{||} \sim \pi^*(\cdot|x')} Q(x', (u'_{||}, \tilde{u}_\perp)) && (43) \\
&= \mathbb{E}_{x' \sim \tau(\cdot|x,(u_{||},u_\perp))} \mathbb{P}(\tilde{u}_\perp \neq \tilde{u}_\perp) && (44) \\
&= 0 && (45)
\end{aligned}
$$

Suppose it were true that $\tilde{\pi} \in \Pi(\pi^*, \mathcal{T}, \mathcal{Q})$, so that for all Q we have that $(\mathbb{B}_{\pi^*,\tau}Q)(x, (u_{||}, u_\perp)) = (\mathbb{B}_{\tilde{\pi},\tau}Q)(x, (u_{||}, u_\perp))$, which means $\mathbb{P}(u'_\perp = \tilde{u}_\perp | x, (u_{||}, u_\perp), \pi^*, \tau) = 1$. But this is only possible if $\mathcal{U}_\perp = \{\tilde{u}_\perp\}$ hence $\text{card}(\mathcal{U}_\perp) = 1$, which is a contradiction with the premise that $\text{card}(\mathcal{U}_\perp) > 1$. □

B Further Experiment Detail

Variants of Digital Representatives As a reminder, a digital representative $\hat{\pi}_i$ is a language model trained to produce a critique based on a question, participant i's opinion, a draft consensus (all of which form the *base* information), and optionally any supplemental information about participant i. The main text discusses various information sources within the task pipeline, which we outline in Table 2 along with text examples from the dataset. We experimented with different combinations of conditioning information along various axes of variation to create diverse datasets for fine-tuning. In Figure 4 we show a handful of variants of digital representatives when conditioning on various data in addition to the base information. Guided by the log-likelihood of the ground truth critiques (in the validation set), we empirically determined the most effective additional information, which included the participant i's opinions and critiques on other debate questions discussed within the same data collection pipeline. An input prompt example from the dataset used to fine-tune the digital representatives as presented in the main text is illustrated in text box 1.

Table 2: *Possible prompt information.* A list of sections (with examples) that may be included in the prompt for fine-tuning digital representatives.

Alias	Description	Example section as part of the prompt
Base	Current question under debate	Current question under debate: Should the government spend more on science and technology research?
Base	Opinion of participant i about the question under debate	Opinion of the participant about the question under debate: I think its really important that the government spends more on Science and Technology. If you think about the problems we have as a society and as a planet - the answers lie with Science and Technology.
Base	Draft consensus statement	Group draft consensus statement (to be critiqued): The government should spend more on science and technology research as this would lead to better technology and medical discoveries to help the world.
D	Demographic profile of participant i (based on questionnaire)	You represent a participant with the following demographic profile. Gender Identity: Male. Age (in decades): 50 to 59. Region: West Midlands. Ethnicity: White. Party voted for in the most recent General Election: Labour. Highest level of education completed: Higher or secondary or further education (A-levels, BTEC, etc.). Religious affiliation: No religion. Approximate household income: £20k to £40k per year. Immigration status in the UK: British Passport holder.
O	Participant i's opinions about other questions (1 to 3)	Question 1 from a past debate: Should the government spend more on science and technology research? Opinion of the participant about question 1: I believe that the government should be allocating more money for science and technology research.
C	Participant i's critiques of other questions (1 to 3)	Question 1 from a past debate: Should the government spend more on science and technology research? Critique of the participant about question 1: I totally agree. Science and Technology can help us out of our current problems and make life better for everyone.
P	Participant i's position score (as text) about other questions (1 to 3)	Question 1 from a past debate: Should the government spend more on science and technology research? Position of the participant about question 1: agree. Question 2 from a past debate: Should students be required to pass a literacy and numeracy test before graduating from primary school? Position of the participant about question 2: strongly disagree.

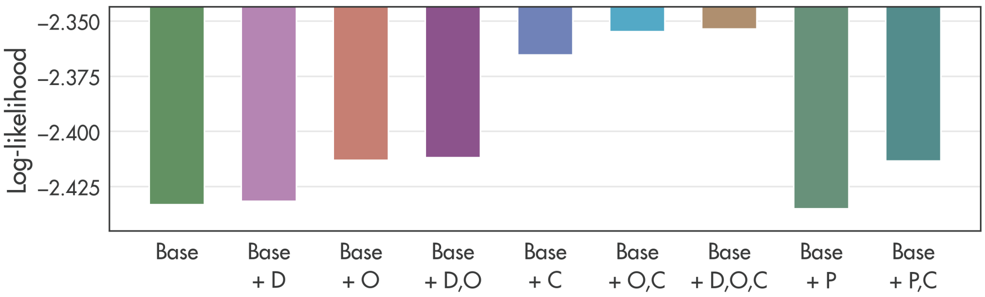

Figure 4: *Variants of Digital Representatives*. Mean log-likelihood of ground-truth critiques from human participants (from the validation set), evaluated under various digital representatives. All these DRs have 1B parameters and were fine-tuned on datasets conditioned on diverse additional information, as indicated on the x-axis (refer to Table 2 for details and examples). The optimal variant (*Base*+*O*+*C*) incorporates participant i's opinions and critiques to other questions. This variant performs very similarly to its counterpart that additionally includes demographic information (*Base*+*O*+*C*). However, we opted for the former for simplicity. Note that variants based solely on demographics (*Base*+*D* or position scoring (*Base*+*P*) perform much worse. This suggests that integrating participant-specific few-shot information enhances both task- and self-consistency.

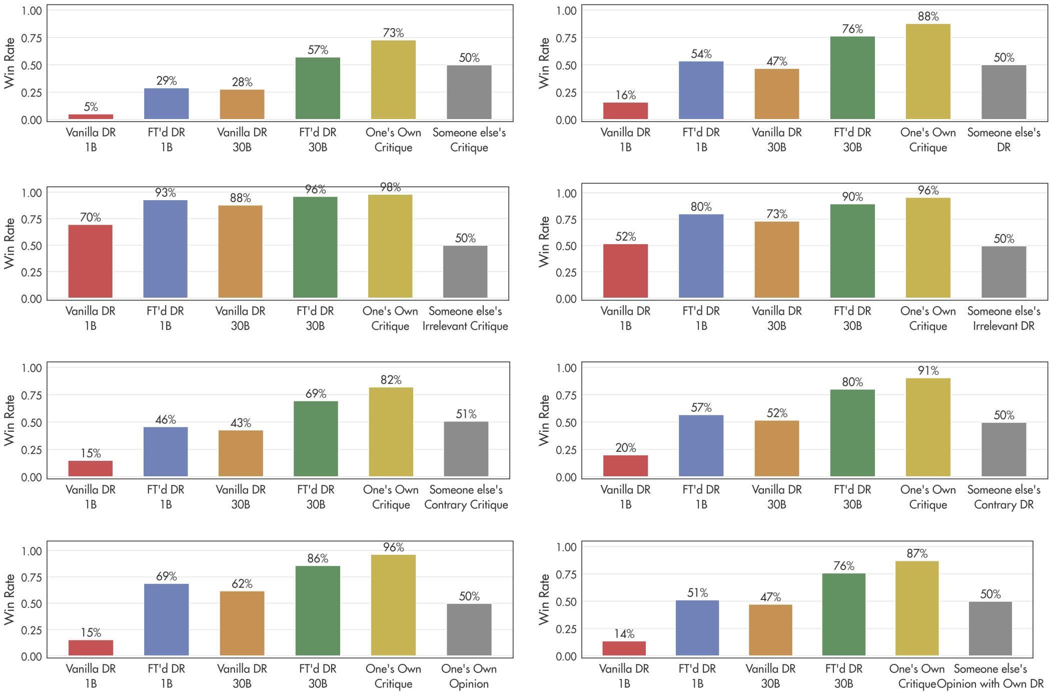

Figure 5: *Autorater Critique Evaluations: Ablations*. The autorater's win-rate of sampled critiques against a custom baseline, as specified by the grey bar within each plot (included as a sanity check, expected around 50%). The ground truth (*One's Own Critique*) is also included to set the performance ceiling. The main take-away is that the 30B fine-tuned DR (green) closely matches the performance of ground truth critiques (golden), suggesting its ability to learn both task-specific aspects (e.g., critique-ness, topic) and participant-specific nuances (e.g., individual preferences, style). A similar trend is observed for the 1B fine-tuned DR, albeit to a lesser extent.

A group of participants is engaged in debates regarding important questions. They have been tasked with coming up with a consensus for a new debate question.
This participant has shared their opinions on other open debate questions, as outlined below.
Question 1 from a past debate: Should students be required to pass a literacy and numeracy test before graduating from primary school?
Opinion of the participant about question 1: I'm against this, children should not be under pressure at such a young age and should concentrate more on being children. Obviously monitor literacy and numeracy but to make them pass a test is unnecessary. Some people develop academically as they get older, so a test at this young age is meaningless. Children will have enough pressure as adults, let them enjoy being children.
Critique of the participant about question 1: Totally agree one hundred per cent with this, pupils should not have to pass a test before graduating but they should be monitored and helped if necessary. Let children be children. Question 2 from a past debate: Should we require people to pay a fixed tax on any amount of wealth they own above £10 million?
Opinion of the participant about question 2: There is a growing wealth gap in the UK and the world and a tax like this would help narrow that gap, help the country pay its way and help to invest in important infrastructure. However we do not want to set this fixed amount of tax too high that it puts wealthy individuals off from living here, investing here and spending money. I do think though £10 million is the right figure to apply this tax.
Critique of the participant about question 2: Like the group opinion, it aligns well with what my response was. Totally agree with the money helping to fund the public sector and poorest in society. My concern was that the tax should not be set to high and 1% seems a fair figure. That 1% would still generate a substanial sum for the country.
Please offer a critique that accurately represents the participant's opinion on a new debate question in relation to the group draft consensus statement.
Current question under debate: Should the government spend more on science and technology research?
Opinion of the participant about the question under debate: I think its really important that the government spends more on Science and Technology. If you think about the problems we have as a society and as a planet - the answers lie with Science and Technology. We have just had a serious pandemic and we were helped by scientists coming up quickly with a vaccination. One of the ways out of the Climate Change crisis is through science and technology coming up with solutions as we as individuals do not want to change pur habits. More money going into Science and Technology will help us counter these crises and future problems. If we don't put the money in, the next pandemic maybe more severe and costly.
Group draft consensus statement (to be critiqued): The government should spend more on science and technology research as this would lead to better technology and medical discoveries to help the world.
Critique of the participant:

Box 1: Prompt example for the digital representative used in the main text (*Base+O+C*).

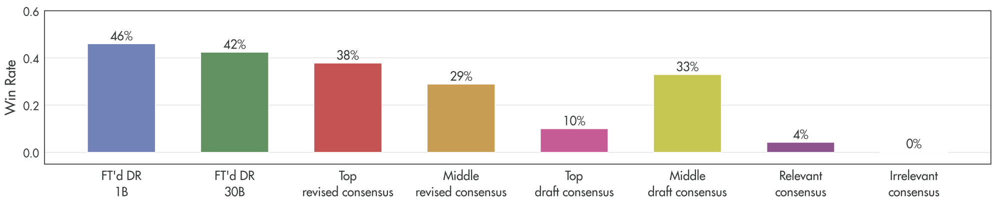

Figure 6: *Autorater Consensus Evaluations: Ablations*. The autorater's win-rate of generated consensuses (based on single participant substitutions with DRs) against the baseline (based solely on ground truth critiques). For comparison, various baseline consensus statements (as specified on the x-axis) are included from the original dataset. This ablation demonstrates that DRs contribute to generating consensuses that are competitively similar to human-selected consensuses (i.e. red bar).