Erkenntnis (2025) 90:2879–2907
https://doi.org/10.1007/s10670-024-00814-7

ORIGINAL RESEARCH

Anchoring as a Structural Bias of Deliberation

Sebastian Till Braun[1] · **Soroush Rafiee Rad**[2] · **Olivier Roy**[3]

Received: 22 February 2023 / Accepted: 29 February 2024 / Published online: 30 July 2024

Abstract

We study the anchoring effect in a computational model of group deliberation on preference rankings. Anchoring is a form of path-dependence through which the opinions of those who speak early have a stronger influence on the outcome of deliberation than the opinions of those who speak later. We show that anchoring can occur even among fully rational agents. We then compare the respective effects of anchoring and three other determinants of the deliberative outcome: the relative weight or social influence of the speakers, the popularity of a given speaker's opinion, and the homogeneity of the group. We find that, on average, anchoring has the strongest effect among these. We finally show that anchoring is often correlated with increases in proximity to single-plateauedness. We conclude that anchoring can constitute a structural bias that might hinder some of the otherwise positive effects of group deliberation.

Anchoring in group deliberation occurs when the opinions expressed early in the process have more influence on the deliberative outcome than those expressed later. Thus, when there is anchoring the order of speech matters. Anchoring can be seen as a special case of a more general tendency of decision processes, whether individual or collective, to be biased toward information that is presented early (Tversky & Kahneman, 1974). This phenomenon is well documented in social psychology and collective decision making (Tversky & Kahneman, 1974; Mussweiler & Strack, 1997; Chapman & Johnson, 1999), and appears to be both resilient and pervasive,

✉ Soroush Rafiee Rad
soroush.r.rad@gmail.com; s.raeerad@uva.nl

Sebastian Till Braun
sebastian.braun@uni-bayreuth.de

Olivier Roy
olivier.roy@uni-bayreuth.de

[1] Faculty of Law, Business and Economics, University of Bayreuth, Bayreuth, Germany

[2] Dutch Institute for Emergent Phenomena & Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands

[3] Department of Philosophy, University of Bayreuth, Bayreuth, Germany

with numerous illustrations in a variety of domains (Chapman & Johnson, 1999; Epley & Gilovich, 2001; McElroy & Dowd, 2007; Mussweiler et al., 2004; Mussweiler & Strack, 1989, 1999, 2001, 2005).

Anchoring can be problematic. To the extent that deliberation is central to *procedural* explanations of democratic legitimacy (Habermas, 1984; Bohman & Rehg, 1997; Peter, 2020), anchoring may provide an unfair and arbitrary advantage to those who speak first. It can furthermore exacerbate other known structural biases and negative outcomes of group deliberation, for instance polarization (Bramson et al., 2017; Dorst, 2023), the formation of spurious unanimity (Prentice & Miller, 1993), or hidden profiles, that is, when participants refrain from sharing relevant private information (Stasser and Titus, 2003). Finally, it can be seen as giving an unfair advantage to those with good argumentative skills or strategic sophistication because they might tend to intervene early in the discussion.

Several mechanisms have been proposed to explain anchoring, the majority of which emphasize various forms of cognitive biases of the participants (Furnham & Boo, 1997). Tversky and Kahneman (1974), for instance, explain it as the result of decision makers' failure to correctly adjust for the initial information that they receive. Others explain it as resulting from confirmation bias (Chapman & Johnson, 1999; Mussweiler & Strack, 1999).

These explanations leave open the question of whether anchoring could be avoided if the decision makers were less prone to different forms of biases and, in the limit, whether it could occur among fully rational individuals. Given the pervasiveness of cognitive biases, answering these questions requires moving away from the lab, toward studying theoretical models of rational deliberation. Hartmann and Rafiee Rad (2020) have taken the first step in that direction, by studying a computational model of deliberation where the participants repeatedly exchange and update probabilistic opinions. They observe that anchoring can also emerge in such cases. Even fully rational agents with no shortcomings in processing the evidence at hand still give more weight to the opinions expressed early. Hartmann and Rafiee Rad conclude that anchoring should be seen as a structural bias of deliberation, not necessarily as the result of some failure of the individual participants.

This first evidence that anchoring can occur even among fully rational agents also leaves open important questions. First is the question whether the effect identified by Hartmann and Rafiee Rad is specific to deliberation on *probabilistic* judgments, or also occurs when the participants exchange, update, and ultimately aggregate *preferences*. The latter is important, because public deliberation is often aimed at supporting social choice (Miller, 1992; List, 2002; Dryzek & List, 2003). In that case deliberation does not primarily bears on individual and collective beliefs, but rather on preferences or value judgments. This is what we study in this paper.

Second, although the findings reported by Hartmann and Rafiee Rad strongly suggest that order of speech is an important determinant of the deliberative outcome, it leaves open the question of how strong that effect is in comparison with other determinants like the relative social influence of the participants, the popularity of certain opinions, or how homogenous the group is in the first place. On average, does being an opinion leader matter more than speaking early? Does anchoring still occur when

the opinions expressed first are only held by a small minority in the group? These questions are not addressed at all by Hartmann and Rafiee Rad.

Finally, the observation that anchoring can occur among rational agents raises the question of how this affects the otherwise very positive outlook on deliberation that deliberative democrats often hold (Cohen, 1989a, b; Estlund, 1993, 1997; Manin, 1987). In particular, a number of contributions have pointed out that deliberation on preference rankings might help increase proximity to single-plateaued preferences, and thus avoid incoherent group preferences or Arrowian impossibilities (List, 2002; Dryzek & List, 2003; List et al., 2012; Farrar et al., 2010; Rafiee Rad & Roy, 2021). Is anchoring correlated with increases in proximity to single-plateauedness? If yes, does this affect existing arguments for deliberative forms of democracy?

This paper addresses these questions by building on the computational model of deliberation on preference rankings presented by Rafiee Rad and Roy (2021). We show that anchoring has, on average, a stronger effect on the deliberative outcome than social influence, popularity of opinion, and, to a lesser extent, homogeneity of the group. This further supports the idea that anchoring is a structural bias of deliberation. We show furthermore that the increases in proximity to single-plateauedness reported by Rafiee Rad and Roy are positively correlated with anchoring, and assess the relevance of this finding for existing arguments for deliberative democracy.

1 The Model

Our results are based on the model introduced in Rafiee Rad and Roy (2021). This model consists of groups of agents who sequentially exchange opinions and are, to different degrees, consensus-seeking in the sense that they are prone to move toward the opinions expressed by others. We review the main points of this model here, and mark the parameters we treat differently. We refer again to Rafiee Rad and Roy for a more detailed presentation, including concrete examples of deliberative processes.[1]

We consider groups between 3 and 99 participants that enter deliberation holding certain preferences, represented by complete rankings and allowing for indifference, over a set of three alternatives.[2] There are 13 such possible rankings. Each participant is assigned one at the beginning of deliberation under the impartial culture assumption (Tsetlin et al., 2003), viz., by drawing from a uniform distribution.

Deliberation proceeds in rounds, each round being a sequence of steps. Each step consists of one announcement and, possibly, preference updates. Only one participant announces her preference at each step, and each participant announces her preference exactly once in a round. The order of announcements is fixed throughout deliberation.

[1] The simulation code and replication data are available at: https://doi.org/10.7910/DVN/AQRYIL

[2] We consider only the case of three alternatives for computational reasons. By doing so we follow Rafiee Rad and Roy (2021). Abou Zeid (2021) reports on results regarding increases in proximity to single-peakedness for five and six alternatives. For reasons explained in Sect. 2.3, we conjecture that the findings reported here would replicate with more alternatives.

At each step, each participant updates her preferences to move closer to the one that has just been announced. The updates are modeled using distance minimization. Upon hearing someone else's opinion, an individual updates her preference by moving to a ranking[3] that minimizes a weighted version of the squared distance between her ranking and the one just announced. Updates can thus be seen as weighted two-person distance-based preference aggregations (Eckert & Klamler, 2011). For robustness we compare the result obtained using the Kemeny–Snell (KS; Kemeny and Snell (1962)), the Cook–Seiford (CS; Cook and Seiford (1978)), and the Duddy–Piggins (DP; Duddy and Piggins (2012)) distances.[4]

The participants are thus taken to be rational in the sense that, to the extent that they are prone to move towards the opinions of others, these moves are in some sense minimal. Distance-based aggregation procedures have been characterized axiomatically already by Kemeny (1959), who also argues that they are most naturally seen as expressing willingness to reach consensus. Minimality (or conservativeness) is a standard requirement of rational attitude changes, both in the case of beliefs (Makinson, 1993; Dietrich et al., 2016) and of preferences (Grüne-Yanoff & Hansson, 2009; Alechina et al., 2013).

While minimality constraints guide preference changes, the model is noncommittal regarding the rationality or even the propensity of being consensus-seeking. This is captured by the fact that the participants minimize a *weighed* version of the squared distance between rankings. Slightly more formally, when participant i announces her preferences, k's preferences r_k are updated to the ranking r'_k that minimizes the following:

$$\sqrt{w_{ki}d(r_i, r'_k)^2 + \hat{w}_{ki}d(r_k, r'_k)^2},$$

where $w_{ki} \in [0, 1]$ is the weight that k assigns to i, and $\hat{w}_{ki} = 1 - w_{ki}$ is the weight that k assigns to herself, relative to i.

Different participants can be weighted differently by the others: the weight that k assigns to i need not to be the same as the weight that k assigns to another agent i'. If $w_{ki} = \hat{w}_{ki}$, then k sees i as a peer and will be willing to meet her "in the middle," so to speak. The more weight k assigns to herself compared to i, i.e., the lower

[3] If there is more than one ranking that minimizes the squared weighted distance, then the participant picks one at random, with equal probabilities for each. See Abou Zeid (2021) for a more detailed discussion of that point.

[4] See the Appendix of Rafiee Rad and Roy (2021) for more details about each of these measures, including a discussion of their respective axiomatizations. For now, it is sufficient to point out that the KS measure is essentially a Hamming distance, counting the number of pairs of alternatives on which any two ranking differs. Its minimum value is 0, and its maximum is 6, in one-unit discrete increments. The DP measure is structurally very similar but has been designed to avoid some double-counting inherent in the KS measure when rankings are assumed to be complete and transitive. It ranges from 0 to 4, again in one-unit increments. CS, on the other hand, assigns numbers to alternatives according to their rank, i.e., in a similar fashion as the Borda voting rule, c.f. Pacuit (2019), and the distance is calculated by adding the absolute values of the rank differences for each alternative. Like DP, CS's minimum value is 0, and maximum value is 4.

the value of w_{ki}, the less she will move towards i. In the limiting case where i gets assigned zero weight, k will not update her preferences at all when i announces hers.

The participants can be assigned arbitrary weights between 0 and 1 by the others.[5] Here we depart from Rafiee Rad and Roy (2021) who made the simplifying assumption that each agent is "immodest" in the sense that they assign at least as much weight to themselves as to any other, i.e., that $\hat{w}_{ki} \geq 0.5$ for all i and k. We consider the case of immodest agents again in Sect. 2.2. The weight that others assign to each agent is drawn at random from an uniform distribution at the beginning of the deliberation, and stays constant throughout the process.

The extent to which participants are consensus-seeking is thus an exogenous parameter that we vary across simulations. We do not assume that certain weights, e.g. only strictly positive ones, are more rational than others. The only case that this modeling of weights excludes is that of consensus-*averse* participants, who would move *away* from the opinion of, say, someone they deeply distrust. In Sect. 2.3 we discuss this assumption in more detail. For now it is sufficient to stress that, in this paper, the propensity to update, and the extent to which the participants update their preferences, are taken as an "arational" parameters. *If* updates take place, however, they are rationally constrained, in terms of minimal changes.

While weights can be non-uniform—different participants can have different weights—we make the simplifying assumption that the weights are common: that all participants agree on everyone else's weight. Upon hearing i's opinion, any other participant updates by giving the same weight to i's opinion. Technically this means that for all agents i, k, and k', $w_{ki} = w_{k'i}$. For that reason, in what follows we simply write w_i for the weight that i gets assigned by (all) others. Under this assumption, each w_i can be seen as a measure of i's relative social influence in the group.

We take weights to be non-uniform but common because we are interested in comparing the effect of order of speech with other determinants of the deliberative outcome, one of them being the respective weights of the agents' opinions. Having all agree on everyone else's weight allows us to identify more clearly those agents who carry more weight in the group, i.e., those that have the strongest relative social influence, and then to compare the impact of social influence on the outcome of deliberation with the impact of speaking early. This simplifying assumption furthermore allows to maximize the impact of the agents with high weights on the deliberative process. If, as it turns out, order of speech still trumps the impact of individual weights in that case, this speaks more strongly for viewing anchoring as a structural bias of deliberation.

Deliberation proceeds for a fixed number of rounds. Rafiee Rad and Roy already observe that, in this model, deliberation stabilizes rather quickly. After three rounds, on average, the participants either reach consensus or stabilize on rankings that are too close to one another to make any further move possible. For that reason,

[5] This assumption rules out possible correlations between weight assignments, e.g. capturing different levels of trust or distrusts for participants sharing certain preferences or other characteristics. We thank the anonymous reviewers of this paper for pointing this out, but leave the study of this case for future work.

deliberation is bounded to a relatively small number of rounds, typically up to five, after which virtually all simulations have already stabilized. Observe, however, that depending on the size of the group, five rounds can consist of a large number of steps, since each participant speaks exactly once in each round.

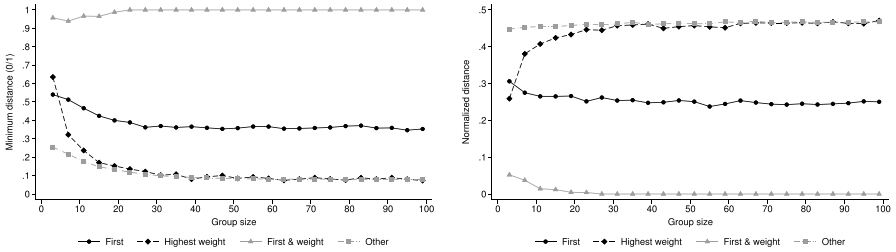
At the end of deliberation, we measure two dependent variables: anchoring and increase in proximity to single-plateauedness. Anchoring is measured in two ways: a Boolean and a continuous one. The Boolean measure $Minimum_{ij}$ takes the value of 1 for agent i in a particular simulation j if there is no other agent i' for whom the collective ranking after deliberation, determined by pairwise majority voting, is strictly closer to i''s initial ranking. The continuous variable $Distance_{ij}$ measures the distance between speaker i's initial ranking and the collective ranking after deliberation. We compute both variables under the three distance measures DP, KS, and CS. $Distance_{ij}$ is normalized by the maximum of each measure (DP: 4; KS: 6; CS: 4), so that it ranges from zero to one.[6]

Single-plateauedness is a property of preference profiles that generalizes the classical notion of single-peaked preferences (Moulin, 1984). Formally, it is defined as follows: a profile of preferences R is single-plateaued relative to a given "ordering dimensions" $\succ$ of the alternatives whenever, for each agent i and triple of alternatives a, b, c, such that $a \succ b \succ c$ or $c \succ b \succ a$, it is not the case that i both strictly prefers a to b and c to a. Single-plateaued preferences are sufficient to avoid so-called Condorcet cycles, and intransitive group preferences more generally. See Gaertner (2001) for an overview, Dryzek and List (2003) for a prominent argument to the effect that deliberation can foster the formation of single-peaked preferences, and Rafiee Rad and Roy (2021) for a discussion of the importance of studying single-pleateauedness in that context. In the present paper we measure *proximity* to single-plateauedness, which is calculated as the relative size of the largest subgroup that is single-plateaued with respect to some ranking (Niemi, 1969; List et al., 2012). Increase in proximity to single-plateauedness is then calculated as the difference in proximity to single-plateauedness before and after deliberation.

2 Results

We first report summary statistics indicating that anchoring indeed occurs in the model. The first speakers have a considerably greater impact on the deliberative outcome than the participants that get assigned highest weights by others. We then turn to multivariate regression analyses. This allow us to study how the strength of the anchoring effect depends on, and compares to other determinants of deliberative outcomes, such as group size or the proportion of participants that share the preferences of the first speaker.

[6] For technical reasons, in Sect. 2, when calculating $Distance_{ij}$, we drop observations from model runs that yield incoherent group preferences. This applies to less than 0.001 % of the observations.



(a) Minimum (0/1) **(b)** Distance

Fig. 1 Average distance to original ranking by agent type and group size (DP measure). *Notes*: The figure plots the average distance between speakers' original ranking and the collective post-deliberation ranking by agent type and group size. Panel (**a**) plots the average probability of having the lowest distance among all agents (ties allowed), i.e., the mean of $Minimum_{ij}$. Panel (**b**) plots the average normalized distance between the collective preference ranking after deliberation and an agent's initial profile, i.e., the mean of $Distance_{ij}$. Both measures are calculated using the DP measure. We distinguish between agents that speak first (black, solid), have the highest weight (black, dashed), speak first *and* have the highest weight (,gray, solid), and neither speak first nor have the highest weight (gray, dashed-totted)

2.1 Summary Statistics

Figure 1 shows the average values of $Minimum_{ij}$ (Panel (a)) and $Distance_{ij}$ (Panel (b)) by agent type and group size under the DP measure. We distinguish between agents that speak first but do not have the highest weight (black solid line), who have the highest weight but do not speak first (black dashed), who speak first and have the highest weight (gray solid), and those who neither speak first nor have the highest weight (gray dashed-dotted). Figure 4 in Appendix shows similar results for the KS and CS measures.

The first key result of Fig. 1 is that speaking first carries a large advantage that is relatively stable across groups of all sizes. Panel (a) shows that the probability of the first speaker having the smallest distance of all agents is about 54 % in small groups of three agents. In comparison, the probability for agents who are neither the first to speak nor have the highest weight is 25.4 % (remember that $Minimum_{ij}$ allows for ties). Since first speakers are randomly selected among all speakers, the post-deliberation advantage results from their speaking position. As group size increases, the mean of $Minimum_{ij}$ for first speakers decreases somewhat and stabilizes around 35–37% for group sizes of 26 or more agents. This compares to a large-group mean of about 8% for those who are neither first speakers nor speakers with the highest weight. Panel (b) shows similar patterns for the continuous distance measure. In large groups, the mean distance between the initial ranking of first speakers and the collective ranking after deliberation is about 24–25% (relative to the maximum), much less than the 46–47% observed for those who neither speak first nor have the highest weight.

The second important finding is that being assigned the highest weight is a major advantage in small groups, but this quickly disappears as group size increases. Panel (a) shows that in groups of three, the agent with the highest weight has the

smallest distance among all agents in 63.6% of all model runs. This is even higher than the 54% observed for first speakers. Already in groups of seven, the advantage of having the highest weight halves to 32.2%—and is thus significantly smaller than that of the first speaker. Agents with the highest weight then quickly become indistinguishable—in terms of their mean $Minimum_{ij}$—from agents who neither have the highest weight nor are first speakers. Panel (b), using the continuous distance measure, confirms that anchoring trumps individual weight in all but very small groups. In fact, the relative advantage of the first speaker over agents with the highest weight is the same as over others in groups of 31 or more agents.

The third key result from Fig. 1 is that agents who happen to be the first to speak and have the highest weight have greater advantages than the separate effects of anchoring and weight would suggest. In other words, order of speech and weight reinforce each other. This is most evident in large groups where weight alone does not confer an advantage. Panel (a) shows that first speakers who also happen to have the highest weight always have the lowest distance in groups of 23 or more agents. Moreover, in those groups, the distance between the initial and collective rankings is zero for first speakers who also have the highest weight, as Panel (b) shows.[7]

2.2 Multivariate Regression Analyses

Regression equation Our regression analysis focuses on $Distance_{ij}$, calculated under the DP measure, as the key outcome variable. In additional robustness checks, we compute $Distance_{ij}$ under the KS and CS measures and consider $Minimum_{ij}$ as an alternative outcome variable (see below).

Our baseline regression model is as follows:

$$Distance_{ij}^{k} = \alpha + \beta First_{ij} + \gamma Weight_{ij} + \mathbf{X}_j\delta + GS_j + u_{ij}. \quad (1)$$

$First_{ij}$ is a Boolean variable set to 1 when i speaks first in run j. $Weight_{ij}$ is also a Boolean variable set to 1 when i has the highest weight in run j. The vector $\mathbf{X}_j$ contains determinants of deliberative outcomes that vary at the model-run level, such as preference similarity. GS_j is a full set of dummy variables for group size,[8] and u_{ij} is an error term. We cluster standard errors at the level of model runs to allow for arbitrary correlation within model runs. Our key parameter of interest is β, the effect of speaking first on $Distance_{ij}$.

Baseline results Table 1 shows the baseline regression results for the DP measure. Column (1) regresses distance on the Boolean variables for speaking first and having the highest weight. Conditioning on a full set of controls for group size ensures that the parameters are only identified from comparisons within groups of the same

[7] It becomes very unlikely in large groups that the first speaker also has the highest weight. Due to the large number of model runs, we still observe 30 such cases in groups of 99 agents.

[8] Results are identical if we add model-run fixed effects to eliminate any unobserved characteristics at the model-run level. This is to be expected as order of speech and weight are randomly assigned in the simulations.

Table 1 Baseline results (DP measure)

	(1)	(2)	(3)	(4)
First speaker (0/1)	−0.217***	−0.217***	−0.217***	−0.217***
	(0.001)	(0.001)	(0.001)	(0.001)
Highest weight (0/1)	−0.024***		−0.024***	−0.024***
	(0.001)		(0.001)	(0.001)
Weight (0–1)		−0.019***		
		(0.000)		
Share w/ first speaker's preferences			−0.212***	
			(0.009)	
Preference similarity (standardized)				−0.009***
				(0.001)
Observations	3,824,976	3,824,976	3,824,976	3,824,976

The dependent variable is $Distance_{ij}$, calculated under distance measure DP. The variable, which ranges from zero to one, measures the normalized distance between the collective preference ranking after deliberation and an agent's initial profile. All regressions include a full set of group size indicator variables. Standard errors clustered at the level of a model run are in parentheses. *** denotes statistical significance at the 1% level

size. The estimates confirm our result from Sect. 2.1 that order of speech trumps weight. Speaking first reduces the distance measure by, on average, 21.7 percentage points (pp, relative to the maximum value). In comparison, having the highest weight reduces the distance by only 2.4 pp, and is thus much less important than speaking first, at least on average.

To assess the effect of weight more generally, we add a continuous variable indicating each agent's weight on a zero-to-one scale. The estimates in Column (2) suggest that the distance between the initial and the collective ranking decreases by only 1.9 pp as weight increases from zero to one. Therefore, when using this more general measure, we again find that the benefits of weight are relatively small compared to speaking first.

Columns (3) and (4) consider the impact of group characteristics. Column (3) shows that a higher proportion of agents who share the first speaker's preferences reduces distance. In other words, all agents benefit if the first speaker's preferences are common in the group. Moving from a share of 0–100% decreases distance by 21.2 pp on average. This effect looks large, but very high proportions of agents sharing the first speaker's preferences are unlikely in large groups.[9] Column (4) illustrates that general preference similarity between agents reduces the distance between initial and collective preferences after deliberation. We measure preference similarity using the Hirschman–Herfindahl index (HHI).[10] We standardize the index so that

[9] When averaged over all agents in our simulations, the average proportion of agents sharing the first speaker's preferences is 7.7% with a standard deviation of 3.8.

[10] The HHI is defined as $\mathrm{HHI}_j = \sum_{l=1}^{12} (s_j^l)^2$, where s_j^l is the share of agents in model run j that have preference profile l. There are 13 possible preference profiles. For sufficiently large groups, the measure is bounded between 1/13 (if agents' profiles are equally distributed across profiles) and 1 (if all agents have

Table 2 Interactions between anchoring and other characteristics (DP measure)

	(1)	(2)	(3)	(4)
First speaker (0/1)	−0.209***	0.042***	−0.215***	−0.220***
	(0.001)	(0.002)	(0.002)	(0.001)
× Highest weight (0/1)	−0.202***			
	(0.002)			
× Weight (0–1)		−0.516***		
		(0.002)		
× Share w/ first speaker's preferences			−0.028*	
			(0.016)	
× Preference similarity (standardized)				0.004***
				(0.000)
Highest weight (0/1)	−0.016***		−0.024***	−0.024***
	(0.001)		(0.001)	(0.001)
Weight (0–1)		−0.009***		
		(0.000)		
Share w/ first speaker's preferences			−0.210***	
			(0.009)	
Preference similarity (standardized)				−0.010***
				(0.001)
Observations	3,824,976	3,824,976	3,824,976	3,824,976

The dependent variable is $Distance_{ij}$, calculated under distance measure DP. The variable, which ranges from zero to one, measures the normalized distance between the collective preference ranking after deliberation and an agent's initial profile. All regressions include a full set of group size indicator variables. Regressions add interaction terms between speaking first and (a) having the highest weight (Column (1)), (b) the 0–1 weight continuum (Column (2)), (c) the proportion of agents sharing the first speaker's preferences (Column (3)), and (d) our index of preference similarity (Column (4)). Standard errors clustered at the level of a model run are in parentheses. *** and * denote statistical significance at the 1% and 10% level, respectively

it has a mean of zero and a standard deviation of one. Column (4) indicates that a one-standard-deviation increase in preference similarity reduces distance for all agents by 0.9 pp on average.

Overall, then, the effect of speaking first stands out among the determinants of deliberative outcomes. Appendix Tables 4 and 5 replicate the baseline results using the KS and CS measures, respectively. The results are similar across all three measures. If anything, the anchoring effect becomes stronger under the alternative distance measures.

Mediators of anchoring Next, we investigate whether other agents or group characteristics mediate the effect of speaking first. To this end, we add interaction

Footnote 10 (continued)
the same profile). If the group size is less than 13, the lower bound is one over the group size. When averaged over all agents, the mean value in our data is 0.095 with a standard deviation of 0.022.

terms between speaking first and these characteristics to the baseline regression (1). Table 2 presents the interaction effects for the DP measure.

The first column adds an interaction term between speaking first and weight. The regression confirms our results from Sect. 2.1 that the anchoring effect is much larger when the first speaker also has the highest weight. The parameter estimate on the interaction term implies that the reduction in distance for the first speaker is 20.2 pp greater if she also has the highest weight in the group. Thus, the anchoring effect doubles from −20.9 pp to −41.1 pp when the first speaker and the speaker with the highest weight are the same.[11]

Column (2) shows that the anchoring effect depends crucially on the weight of the first speaker. According to the estimates, the impact of the first speaker ranges from +4.2 pp for speakers with zero weight to −47.4 pp ($4.2 + 51.6 \times 1$) for speakers with maximum weight. For agents with very low weight, being the first to speak is thus not at all advantageous or even somewhat disadvantageous. A natural explanation for this observation is that we allow for arbitrary weights. Agents with weights close to zero have essentially no influence on the deliberative outcome. For minimally large groups, it is virtually certain that someone speaking later will have large enough weight to steer the group's opinion her way, canceling the relative advantage that the first speaker would otherwise have had.

In contrast, the anchoring effect hardly varies with other agents' preferences. Column (3) shows that the anchoring effect is stronger, i.e., more negative, if the proportion of other agents who share the first speaker's preferences increases. However, the interaction effect is small. Moving from a proportion of 0–10% slightly decreases the impact of speaking first from −21.5 pp to −21.8 pp. In contrast, speaking first has less effect if agents have similar preferences anyway. A one-standard-deviation increase in preference similarity increases the impact of speaking first from −22.0 pp to −21.6 pp. The anchoring effect is thus slightly less pronounced in more homogeneous groups.

Tables 6 and 7 in the Appendix replicate the results for the CS and KS measures, respectively. The results are similar across all three measures. The only significant difference concerns the interaction between the first speaker and the proportion of agents who share the first speaker's preferences. This interaction is significantly larger for the KS than for the DP measure, but not statistically significant for the CS measure.

Agent position So far, we have focused on the effect of speaking first, comparing it to all other positions. Figure 2 shows that it is not beneficial for agents to speak earlier, unless they speak at the very beginning of deliberation. The figure shows the effect of speaking at different positions on the distance between the initial individual rankings and the final collective ranking. All effects are measured relative to the last speaker. We focus on groups of 51 agents, but the results for other group sizes are

[11] The overall effect of −41.1 refers to the difference in the distance measure between agents who are first speakers and have highest weight, and those who have the highest weight but are not first speakers. Compared to agents who neither are the first speaker nor have the highest weight, the average difference is as high as 42.7 pp (−20.9 − 20.2 − 1.6).

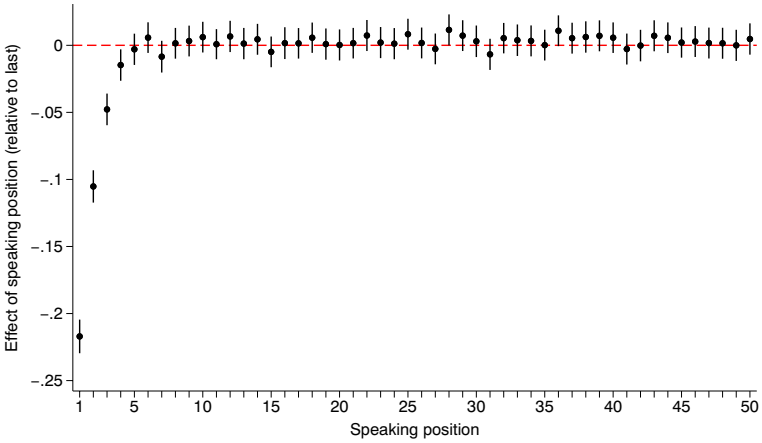


Fig. 2 Effect of speaking position on distance relative to last speaker (DP measure). *Notes*: The figure plots the effect of speaking at different positions on the distance between original and collective ranking after deliberation under the DP measure. All effects are measured relative to the last speaker and are estimated for groups of 51 agents. Point estimates are marked by a dot. The vertical bands indicate the 95% confidence interval of each estimate

similar. Speaking first rather than last reduces the distance by 21.7 pp. The second speaker still sees her distance measure reduced by 10.5 pp on average. Speaking in third and fourth place is associated with a small distance reduction of 4.8 pp and 1.5 pp, respectively. However, after that, the order of speech does not affect the relative distance.

Robustness checks We report on two robustness checks. First, we consider $Minimum_{ij}$ as an alternative outcome variable. Appendix Tables 8 and 9 replicate our main regression tables and show that all our main results also hold for the Boolean outcome measure. In particular, speaking first has a much larger (positive) effect on $Minimum_{ij}$ than weight. Speaking first increases the probability that there is no other agent for whom the collective ranking is closer to her original one by 29.3 pp. Agents with the highest weight have, on average, only a 3.9 pp higher probability. When an agent is both the first speaker and the one with the highest weight, she has a 77.9 pp higher probability of having the smallest distance in the group. So, just as for the continuous measure, the status of the first speaker and their weight strongly reinforce each other.

Second, we restrict the weight that agents assign to others to at most 0.5, as in Rafiee Rad and Roy (2021). Thus, we assume that agents are "immodest" in that they give at least as much weight to themselves as to others.[12] Appendix Table 10 shows, for all three distance measures DP, CS, and KS, regressions of the

[12] In Rafiee Rad and Roy (2021) this modeling choice was motivated by the fact that the participants reach consensus almost universally when at least one of them gets assigned more than 0.5 weight by the others. In those circumstances, the meta-agreement hypothesis, the main object of study in Rafiee Rad and Roy (2021), is moot.

continuous distance measure $Distance_{ij}$ on Boolean variables for the first speaker, highest weight, and the interaction between the two (as in Column (1) of Table 2). We again find that anchoring trumps weight and that the effects of speaking first and weight strongly reinforce each other. However, the impact of speaking first is less pronounced for immodest agents than for agents with arbitrary weights. This is especially true for the DP measure, where speaking first decreases the distance between collective and initial preference ranking by "only" 4.7 pp. Nevertheless, the effect is five times as large as that of having the highest weight. A plausible explanation for this effect is that immodest agents are by definition less consensus-seeking and as such move less towards the opinions of others, decreasing the overall impact of each individual announcement.

2.3 Discussion

The results just reported show that anchoring is both significant and robust in the present model. Speaking first or, to a lesser extent, second provides a strong advantage in that the result of deliberation will be, on average, closer to the opinion of these participants than to the opinion of any other. The multivariate regression analyses have allowed fine-graining this observation by comparing the strength of the effect of the order of speech with relative weights, group size, popularity, and homogeneity of opinions. As we have seen, anchoring is the strongest among those possible determinants of the deliberative outcome. The anchoring effect only slightly diminishes in groups that are already very homogeneous prior to deliberation, i.e., where the participants already hold very similar preferences.

This observation provides strong additional support for the idea, already formulated by Hartmann and Rafiee Rad (2020), that anchoring is a structural bias of deliberation. The effect stems from the structure of the deliberation process itself, not necessarily from participants' mistakes or cognitive biases. The agents in this model indeed have no cognitive or computational limitations. Furthermore, they are rational in that, to the extent that they are consensus-seeking, they update their preferences by minimizing the distance between their ranking and the one just announced. There is, finally, no random "noise" or shocks in the model, which would capture mistakes or unexpected changes in the participants' preferences.

The fact that, in this paper, the order of speech is fixed throughout deliberation might, at first sight, appear as a natural explanation of anchoring. There are good reasons, however, to believe that this is not the main one. Even if, like in Rafiee Rad and Roy (2021), the order of speech is reshuffled in each round, in all but very small groups there will be a high number of announcements made, one for each participant, before the next round starts and a new first speaker can announce her preferences again. By then, the participants have already—and cumulatively (see next paragraph)—moved towards the first and second speakers. Reshuffling the order of speech would only result in someone who had already moved closer to the opinion of the first and the second speaker announcing her preferences first in the next round. This new first speaker might still affect the deliberative outcome, but this effect will still be much smaller than the effect of the original two announcements.

The observation that order of speech is trumped by weight only in groups of three participants can be seen as indirect evidence for this.

A more plausible explanation of anchoring is that preference updates are sequential or cumulative in this model. Indeed, the participants update their preferences at each step in a round, i.e., after each separate announcement. This means that the only "initial" ranking that gets announced is the very first one. Each subsequent announcement is one of an updated ranking. The second speaker will already have moved closer to the ranking of the first speaker before announcing her (updated) preferences. What she announces is thus, in most cases, *not* her initial preference, but a ranking that lies somewhere between this initial ranking and one announced by the first speaker. This effects of the first announcement compounds, so to speak, as the round continues. What the third speaker announces has already been updated twice, and at that time, the different rankings are close to each other, with a strong bias towards the first ranking announced. The effect of the third and the subsequent announcements is thus comparatively smaller.[13]

Eliminating the order of speech altogether is an obvious, but in our view too idealized, solution to anchoring. Instead of updating sequentially, the agents could of course have been modeled as receiving all the rankings of the others simultaneously, and as updating by moving to the ranking that minimizes the average weighted distance from their own ranking. For agents with limited attention and memory, however, to simultaneously consider all rankings appears rather unrealistic, especially in large groups. In the idealized case where the agents do not have such cognitive limitations, they would then be conceived as waiting until all others have announced their opinions. We do not view this as a plausible representation either. Given the pervasiveness and often unconscious effect of social influence (Nolan et al., 2008), having the agents stoically refraining from getting influenced by what they hear until everyone has spoken seems an idealization that goes beyond assuming full rationality. In other words, although patience is certainly a virtue that would be instrumental in preventing anchoring, it is an open question whether it is also a feature of fully rational agents.

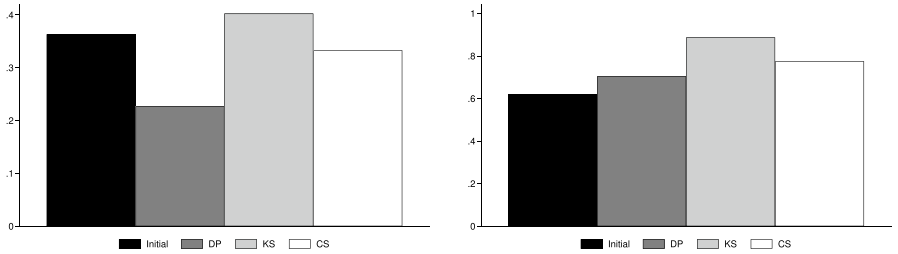
More plausible solutions might be to restrict announcements to subgroups, and possibly to reshuffle these groups after each round, or to introduce stochasticity into the update process. The first idea here is to divide the participants into smaller groups in each round, and have them deliberate sequentially as above. The outcome of each round of deliberation within the subgroups would of course remain anchored to the respective first speakers, but their influence on the overall deliberative outcome would, we conjecture, become smaller as the number of subgroups increases. This procedure would furthermore match more closely some deliberative designs implemented, for instance, in deliberative polls (Fishkin & Luskin, 2005). The second idea is to model participants as updating their preferences only with some probability after each announcement, such that this probability increases the longer they wait to update, and such that it reaches 1 at the

[13] Note that if this explanation is correct, then anchoring should also occur in cases where the agents deliberate over larger sets of alternatives.

end of each round. This would generalize the current model by allowing different degrees of resistance to social influence, ranging from the case modeled here, where participants potentially update after every announcement, to the other extreme case, where they wait until everyone has announced their preferences. We conjecture that in this generalized model, anchoring would decrease in direct proportion to the degree to which the participants are resistant to social influence. However, we leave the systematic study of these two solutions for future work.

Two modeling choices made in this paper deserve to be highlighted before closing the section. First, the model we studied leaves out many aspects usually associated with rational deliberation, most importantly the process of exchanging reasons for and against certain judgments. For many deliberative theorists, starting already with Habermas (1984), this aspect is central for an exchange of opinions to count as a genuine deliberation. Of course, the model we study here is consistent with such richer understandings of deliberation. However, the question remains whether one could avoid anchoring by incorporating an exchange of reasons in a richer model. We leave this for future work, but conjecture that the effect would not disappear: if the preference update process remains the same, the reasons presented first in deliberation might carry more weight than those presented later, and this in turn would be reflected by the hypothetical reason-based preference changes.

Second, to the extent that the model might also be used to represent more realistic deliberative scenarios, a number of idealizing assumptions about the weights could be lifted. First, in the present model, the weights remain constant throughout the deliberation. This rules out participants who might be able to dynamically adjust the weight they assign to others. Various methods have been proposed in related models to capture such changes, either endogenously (Jackson, 2010, chap. 7) or exogenously (Hartmann & Rafiee Rad, 2020). Second, weights are assigned to each participant independently, ruling out possible correlations. Allowing for correlations in weight assignment could serve as a proxy for group identity or partisanship. Participants could, for instance, associate certain preferences with particular interest groups or political orientation, and base their weight assignments on this rather than on the personal identity of the speaker. Finally, by assuming that the weights take values between 0 and 1, we model participants as being consensus-seeking, or perhaps more strongly put, as *not* being consensus-*averse*. Complete distrust can only be modeled here by having a participant assign a weight of 0 to another, in which case the first simply disregards the opinion of the second. Thus, the model cannot capture a stronger form of polarizing dynamics where some participants want to distance themselves from others and thus move away from them. Lifting these three assumptions about the weight would certainly make the current model more realistic, which in turn might provide possible explanations for concrete cases of anchoring in deliberation. However, since the goal of the present paper was to investigate whether anchoring can be seen as a structural bias of deliberation, i.e., whether it can occur in idealized contexts, we leave open here the concrete implementation of these more realistic assumptions.



(a) Single-peakedness (b) Single-plateauedness

Fig. 3 Average distance to single-peakedness and single-plateauedness by measure. *Notes*: The figure plots the average distance to single-peakedness (Panel (**a**)) and single-plateauedness (Panel (**b**)) before and after deliberation. The after-deliberation averages are calculated separately for distance measures DP, KS, and CS

3 Anchoring and Increase in Proximity to Single-Plateauedness

We now investigate whether anchoring is correlated with increases in proximity to single-plateauedness. Rafiee Rad and Roy (2021) have already reported that, in the present model, deliberation increases proximity to single-plateauedness and, to the extent that the agents are minimally consensus-seeking, that this increase goes together with the elimination of incoherent group preferences. These results have been replicated and generalized to the case of 5 and 6 alternatives by Abou Zeid (2021), who observes an even more widespread elimination of incoherent group preferences. The question we address in this section is whether, and if so to what extent, this positive effect can be traced back to anchoring, which we have just seen is both a significant and pervasive in the model. We then discuss the consequences of our findings for the overall assessment of the value of group deliberation.

3.1 Results

Recall that proximity to single-plateauedness is defined as the relative size of the largest subgroup that is single-plateaued with respect to some ranking (Niemi, 1969; List et al., 2012). Increase in proximity to single-plateauedness is then calculated as the difference in proximity to single-plateauedness before and after deliberation. In the following, the unit of analysis is now a model run (rather than an agent).

Figure 3 shows the average distance to strict single-peakedness (Panel (a)) and single-plateauedness (Panel (b)) before and after deliberation, across both group size and weights. As already observed in Rafiee Rad and Roy (2021), distance to (strict) single-peakedness is lower after deliberation for the CS and DP measures than before, but slightly higher for the KS measure. The intuitive explanation is that both CS and DP favor indifference when two opposed pairs of strict preferences are being compared. This makes it significantly less likely that the final profile is strictly single-peaked. Distance to single-plateauedness is higher after deliberation than before for all three measures. The lower increase observed for DP and CS has

Table 3 Anchoring and distance to single-peakedness & single-plateauedness (DP measure)

	Single-peakedness		Single-plateauedness	
	(1)	(2)	(3)	(4)
Anchoring measure:				
Continuous (−1–0)	0.351***		0.514***	
	(0.004)		(0.004)	
Boolean (0/1)		0.225***		0.234***
		(0.003)		(0.003)
Observations	74,971	74,992	74,971	74,992

The dependent variable is the distance to single-peakedness (Columns (1) and (2)) and single-plateauedness (Columns (3) and (4)) under distance measure DP. All regressions include a full set of group-size indicator variables. Robust standard errors are in parentheses. *** denotes statistical significance at the 1% level

a similar explanation as in the strict case: updating with either measure instead of KS results more frequently in complete indifference, which is, by definition, not single-plateaued. All differences are statistically significant in two-sided t-tests (not shown).

We study the relationship between anchoring and distance to single-peakedness and single-plateauedness using the following regression equation:

$$Peak_j = \alpha + \beta Anchoring_j + GS_j + u_j. \tag{2}$$

Here, $Peak_j$ is the distance to single-peakedness at the end of the model run j, $Anchoring_j$ is a measure of the strength of the anchoring effect in that run, GS_j is a full set of group-size Boolean variables, and u_j is an error term. We use a similar regression framework to examine the effect of anchoring on proximity to single-plateauedness. $Anchoring_j$ is defined as follows:

$$Anchoring_j = -\frac{Distance_{1j}}{\max_i(Distance_{ij})}.$$

Here $Distance_{1j}$ is the distance between the preference ranking and the final profile for the first speaking agent. Thus, the anchoring measure ranges from −1 (when the distance of the first speaker is equal to the maximum distance in this model run) to zero (when the distance of the first speaker is zero). We multiply the ratio by −1 for ease of interpretation. As an alternative measure, we use a Boolean variable that takes the value of 1 iff the first speaker's distance equals the minimum distance in a given model run.

Table 3 shows the results of estimating Eq. (2) under the DP measure. Columns (1) and (2) show that anchoring is positively associated with distance to single-peakedness. The coefficient estimate in Column (1) implies that distance to single-peakedness increases by 0.351 points as our continuous anchoring measures move from −1 to zero. Column (2) shows that proximity to single-peakedness

is 0.225 points higher when the first speaker has the lowest distance (compared to when she has not). These effect sizes are large relative to the sample mean of single-peakedness of 0.227 (standard deviation of 0.418). Similarly, Columns (3) and (4) report strong positive associations between anchoring and single-plateauedness.

Appendix Tables 11 and 12 show that anchoring and single-peakedness/-plateauedness are also positively associated when using the KS and CS measures. Effect sizes are of comparable magnitude for the continuous anchoring measure, but smaller for the Boolean one, especially for the KS measure. These results hold even if we add additional control variables for preference similarity, the proportion of agents who share the preferences of the first speaker, and single-peakedness/-plateauedness before deliberation (not shown).

3.2 Discussion: Anchoring and Coherent Aggregation

The analysis in the previous section shows that anchoring is an important determinant of increase in proximity to single-plateauedness, and ultimately to the efficacy of deliberation to steer away from incoherent group preferences. How much is this a concern? For one thing, our result can be seen as showing that anchoring in fact *supports* coherent aggregation. It indeed helps increase proximity to single-plateauedness. So, as far as one focuses on avoiding incoherent group preferences, anchoring does not appear to be a *direct* concern.

Anchoring, however, can indirectly affect our overall positive assessment of deliberation and, in the end, outweigh the sheer avoidance of incoherent group preferences. Here it is important to distinguish two different contexts of collective decision making: those in which there is what Estlund (1997) calls a "procedure-independent" standard for evaluating the outcome of deliberation and voting, and those in which there is no such procedure-independent standard, or in simpler terms, cases where there is a correct or right answer to the question under deliberation, and those in which there is not.[14]

If there is no procedure-independent standard, the value of deliberation is procedural: it aims at supporting fairness and equal participation. In these cases anchoring, as a structural bias, is problematic. Even if deliberation leads to coherent group preferences, the results show that this might be partly due to anchoring, which in turn introduces a form of arbitrary advantage to the first and the second speaker. Anchoring furthermore opens the door to the possibility of strategizing, by allowing one to increase or decrease the impact of some opinions in the final verdict by changing the speakers' positions within the group. To the extent that the value of deliberation is procedural, it introduces its own structural bias and thus makes its ability to avert incoherent group preferences less important.

The situation is less clear-cut in cases where there *are* procedure-independent standards, because weights might or might not track individual expertise. Assuming that this expertise is difficult to recognize, i.e., that weights only poorly track

[14] Note that the first case need not imply that the participants are deliberating on empirical questions or matters of fact. They could also be deliberating on value judgments (Rabinowicz, 2016).

actual expertise, anchoring can diminish the epistemic value of the deliberative outcome. As we have seen, anchoring trumps weights on average. Recall, however, that we have also observed that the effect of the order of speech and weight mutually amplify each other. Speaking first and having a higher weight ensures, on average, a stronger influence on the outcome than the expected separate effect of each of these determinants. If those weights reliably track expertise, one can use this fact to bolster the epistemic value of the deliberation outcome. Assessing this in detail would, however, require to study thoroughly the truthtracking and verisimilitude properties (Rabinowicz, 2016) of deliberation as modeled here. We leave this for future work.

4 Conclusion

Anchoring appears to be a strong and pervasive aspect of deliberation, serious enough to counterbalance some of its otherwise positive features. Both the summary statistics and the multivariate regression analyses show that speaking first or second provides a comparatively strong influence on the deliberative outcome. The regression analyses have also revealed that order of speech trumps relative weight and popularity of opinion, two of the most obvious other determinants of deliberation in the model. Furthermore, we observed that simultaneously being the first speaker and having the strongest relative influence gives an advantage beyond the separate effects of either aspects. In the last section, we then turned to the relation between anchoring and increases in proximity to single-plateaued preferences. We showed that the latter often goes hand in hand with the former.

These findings have broader consequences for the understanding of collective decision-making processes. On the one hand, the possibility of coherent aggregation through the creation of single-peaked preferences removes a barrier to collective decision making. This might help in redistributing responsibility to its members (List et al., 2011). On the other hand, anchoring shows that equal participation in the deliberation does not translate into equal influence on its outcome. This is so even for rational agents who are not bound by any hierarchical or power relations. This inequality in influence only results from a structural bias that gives more influence to those who speak earlier, even if the order of speakers is not under the control of any group member or an external planner. As we have argued, this can overshadow other benefits of deliberation, for instance, that it fosters coherent aggregation. This observation, in turn, opens the door to a more general study of how one can best balance these advantages and drawbacks of deliberation, and ultimately improve our democratic decision-making practices.

Appendix: Additional Tables

Baseline Results for KS and CS Measures

See Tables 4 and 5.

Table 4 Baseline results (KS measure)

	(1)	(2)	(3)	(4)
First speaker (0/1)	−0.246***	−0.246***	−0.246***	−0.246***
	(0.001)	(0.001)	(0.001)	(0.001)
Weight (0/1)	−0.020***		−0.020***	−0.020***
	(0.001)		(0.001)	(0.001)
Weight (0–1)		−0.021***		
		(0.000)		
Share w/ first speaker's preferences			−0.271***	
			(0.005)	
Preference similarity (standardized)				−0.011***
				(0.000)
Observations	3,824,970	3,824,970	3,824,970	3,824,970

The dependent variable is $Distance_{ij}$, calculated under distance measure KS. The variable indicates the normalized distance between the preference ranking and the final profile of an agent and ranges from zero to one. All regressions include a full set of group-size indicator variables. Standard errors clustered at the level of a model run are in parentheses. *** denotes statistical significance at the 1% level

Table 5 Baseline results (CS measure)

	(1)	(2)	(3)	(4)
First speaker (0/1)	−0.293***	−0.293***	−0.293***	−0.293***
	(0.001)	(0.001)	(0.001)	(0.001)
Highest weight (0/1)	−0.028***		−0.028***	−0.028***
	(0.001)		(0.001)	(0.001)
Weight (0–1)		−0.028***		
		(0.001)		
Share w/ first speaker's preferences			−0.306***	
			(0.007)	
Preference similarity (standardized)				−0.013***
				(0.000)
Observations	3,824,979	3,824,979	3,824,979	3,824,979

The dependent variable is $Distance_{ij}$, calculated under distance measure CS. The variable indicates the normalized distance between the preference ranking and the final profile of an agent and ranges from zero to one. All regressions include a full set of group-size indicator variables. Standard errors clustered at the level of a model run are in parentheses. *** denotes statistical significance at the 1% level

Interactions Between Anchoring and Other Characteristics for KS and CS Measures

See Tables 6 and 7.

Table 6 Interactions between anchoring and other characteristics (KS measure)

	(1)	(2)	(3)	(4)
First speaker (0/1)	−0.239***	0.007***	−0.250***	−0.251***
	(0.001)	(0.002)	(0.001)	(0.001)
× Highest weight (0/1)	−0.169***			
	(0.002)			
× Weight (0–1)		−0.504***		
		(0.002)		
× Share w/ first speaker's preferences			0.051***	
			(0.013)	
× Preference similarity (standardized)				0.007***
				(0.000)
Weight (0/1)	−0.013***		−0.020***	−0.020***
	(0.001)		(0.001)	(0.001)
Weight (0–1)		−0.012***		
		(0.000)		
Share w/ first speaker's preferences			−0.273***	
			(0.005)	
Preference similarity (standardized)				−0.012***
				(0.000)
Observations	3,824,976	3,824,976	3,824,976	3,824,976

The dependent variable is $Distance_{ij}$, calculated under distance measure KS. The variable indicates the normalized distance between the preference ranking and the final profile of an agent and ranges from zero to one. All regressions include a full set of group-size indicator variables. Regressions add interaction terms between speaking first and (a) weight as a binary variable (Column (1)), (b) weight on the full 0–1 continuum (Column (2)), (c) the proportion of agents sharing the first speaker's preferences (Column (3)), and (d) our index of preference similarity (Column (4)). Standard errors clustered at the level of a model run are in parentheses. *** denotes statistical significance at the 1% level

Table 7 Interactions between anchoring and other characteristics (CS measure)

	(1)	(2)	(3)	(4)
First speaker (0/1)	−0.284***	0.052***	−0.294***	−0.299***
	(0.001)	(0.002)	(0.002)	(0.001)
× Highest weight (0/1)	−0.235***			
	(0.002)			
× Weight (0–1)		−0.689***		
		(0.003)		
× Share w/ first speaker's preferences			0.011	
			(0.018)	
× Preference similarity (standardized)				0.007***
				(0.000)
Weight (0/1)	−0.019***		−0.028***	−0.028***
	(0.001)		(0.001)	(0.001)
Weight (0–1)		−0.015***		
		(0.001)		
Share w/ first speaker's preferences			−0.307***	
			(0.008)	
Preference similarity (standardized)				−0.014***
				(0.001)
Observations	3,824,979	3,824,979	3,824,979	3,824,979

The dependent variable is $Distance_{ij}$, calculated under distance measure CS. The variable indicates the normalized distance between the preference ranking and the final profile of an agent and ranges from zero to one. All regressions include a full set of group size indicator variables. Regressions add interaction terms between speaking first and (a) weight as a binary variable (Column (1)), (b) weight on the full 0–1 continuum (Column (2)), (c) the proportion of agents sharing the first speaker's preferences (Column (3)), and (d) our index of preference similarity (Column (4)). Standard errors clustered at the level of a model run are in parentheses. *** denotes statistical significance at the 1% level

Boolean Distance Measure

See Tables 8 and 9.

Table 8 Baseline results for Boolean distance measure (DP measure)

	(1)	(2)	(3)	(4)
First speaker (0/1)	0.293***	0.293***	0.293***	0.293***
	(0.002)	(0.002)	(0.002)	(0.002)
Highest weight (0/1)	0.039***		0.039***	0.039***
	(0.001)		(0.001)	(0.001)
Weight (0–1)		0.030***		
		(0.001)		
Share w/ first speaker's preferences			0.358***	
			(0.006)	
Preference similarity (standardized)				0.023***
				(0.001)
Observations	3,825,000	3,825,000	3,825,000	3,825,000

The dependent variable is $Minimum_{ij}$, calculated under distance measure DP. The Boolean measure $Minimum_{ij}$ takes the value of 1 for agent i in a particular simulation j if there is no other agent for whom the collective ranking after deliberation, determined by pairwise majority voting, is closer to the original ranking. All regressions include a full set of group-size indicator variables. Standard errors clustered at the level of a model run are in parentheses. *** denotes statistical significance at the 1% level

Table 9 Interactions between anchoring and other characteristics for Boolean distance measure (DP measure)

	(1)	(2)	(3)	(4)
First speaker (0/1)	0.275***	−0.204***	0.283***	0.293***
	(0.002)	(0.002)	(0.003)	(0.002)
× Highest weight (0/1)	0.484***			
	(0.004)			
× Weight (0–1)		0.992***		
		(0.004)		
× Share w/ first speaker's preferences			0.141***	
			(0.026)	
× Preference similarity (standardized)				0.001
				(0.001)
Highest weight (0/1)	0.020***		0.039***	0.039***
	(0.001)		(0.001)	(0.001)
Weight (0–1)		0.010***		
		(0.001)		
Share w/ first speaker's preferences			0.351***	
			(0.006)	
Preference similarity (standardized)				0.023***
				(0.001)
Observations	3,825,000	3,825,000	3,825,000	3,825,000

The dependent variable is $Minimum_{ij}$, calculated under distance measure DP. The Boolean measure $Minimum_{ij}$ takes the value of 1 for agent i in a particular simulation j if there is no other agent for whom the collective ranking after deliberation, determined by pairwise majority voting, is closer to the original ranking. All regressions include a full set of group-size indicator variables. Regressions add interaction terms between speaking first and (a) weight as a binary variable (Column (1)), (b) weight on the full 0–1 continuum (Column (2)), (c) the proportion of agents sharing the first speaker's preferences (Column (3)), and (d) our index of preference similarity (Column (4)). Standard errors clustered at the level of a model run are in parentheses. *** denotes statistical significance at the 1% level

Immodest Agents

See Table 10.

Table 10 The effect of anchoring and weight with immodest agents

	DP (1)	KS (2)	CS (3)
First speaker (0/1)	−0.047***	−0.150***	−0.167***
	(0.001)	(0.001)	(0.001)
× Highest weight (0/1)	−0.056***	−0.108***	−0.126***
	(0.003)	(0.003)	(0.004)
Highest weight (0/1)	−0.009***	−0.020***	−0.026***
	(0.001)	(0.001)	(0.001)
Observations	3,824,418	3,824,579	3,823,460

The dependent variable is $Distance_{ij}$, calculated under distance measure DP (Column (1)), KS (Column (2)), and CS (Column (3)). The variable, which ranges from zero to one, indicates the normalized distance between the collective preference ranking after deliberation and an agent's initial profile. All regressions include a full set of group-size indicator variables. "Immodest" agents assign at least as much weight to themselves as to others. Standard errors clustered at the level of a model run are in parentheses. *** denotes statistical significance at the 1% level

Anchoring and Single-Peakedness & Single-Plateauedness (KS and CS Measures)

See Tables 11 and 12.

Table 11 Anchoring and distance to single-peakedness & single-plateauedness (KS measure)

	Single-peakedness		Single-plateauedness	
	(1)	(2)	(3)	(4)
Anchoring measure:				
Continuous (−1–0)	0.228***		0.301***	
	(0.006)		(0.005)	
Boolean (0/1)		0.036***		0.009***
		(0.004)		(0.002)
Observations	74,971	74,992	74,971	74,992

The dependent variable is the distance to single-peakedness (Columns (1) and (2)) and single-plateauedness (Columns (3) and (4)) under distance measure KS. All regressions include a full set of group-size indicator variables. Robust standard errors are in parentheses. *** denotes statistical significance at the 1% level

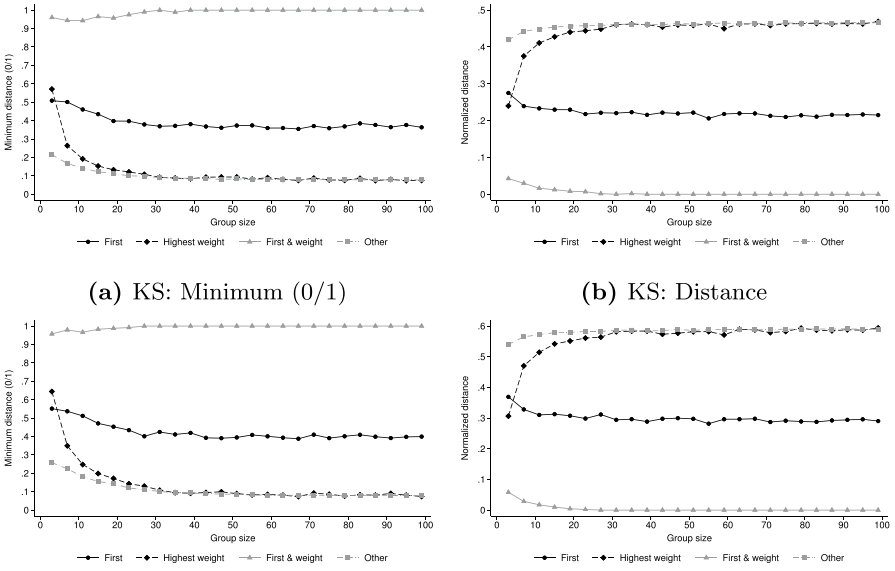
Table 12 Anchoring and distance to single-peakedness & single-plateauedness (CS measure)

	Single-peakedness		Single-plateauedness	
	(1)	(2)	(3)	(4)
Anchoring measure:				
Continuous (−1–0)	0.289***		0.564***	
	(0.004)		(0.004)	
Boolean (0/1)		0.111***		0.137***
		(0.004)		(0.003)
Observations	74,971	74,992	74,971	74,992

The dependent variable is the distance to single-peakedness (Columns (1) and (2)) and single-plateauedness (Columns (3) and (4)) under distance measure CS. All regressions include a full set of group-size indicator variables. Robust standard errors are in parentheses. *** denotes statistical significance at the 1% level

Additional Figures

See Fig. 4.



(a) KS: Minimum (0/1)

(b) KS: Distance

(c) CS: Minimum (0/1)

(d) CS: Distance

Fig. 4 Average distance to original ranking by agent type and group size (KS/CS measures). *Notes*: The figure plots the average distance between speakers' original ranking and the collective post-deliberation ranking by agent type and group size. Panels (**a**) and (**c**) plot the average probability of having the lowest distance among all agents (ties allowed), i.e., the mean of $Minimum_{ij}$. Panels (**b**) and (**d**) plot the average normalized distance between the collective preference ranking after deliberation and an agent's initial profile, i.e., the mean of $Distance_{ij}$. Measures in the upper and lower panels are calculated using the KS and CS measures, respectively. We distinguish between agents that speak first (black, solid), have the highest weight (black, dashed), speak first *and* have the highest weight (gray, solid), and neither speak first nor have the highest weight (gray, dashed-dotted

Acknowledgements The authors would like to thank the participants of the "Philosophy Breakfast" at the University of Bayreuth, the "LIRA Seminar" at the University of Amsterdam, and the "Choice Group" at the London School of Economics for their helpful comments and suggestions. Rafiee Rad and Roy's work on the paper was partly supported by the DFG-ANR project "Collective Attitudes Formation" (ColA-Form, RO-4548/8-1). Rafiee Rad's work was also partly supported by the Dutch Institute for Emergent Phenomena (DIEP) cluster at the University of Amsterdam.

References

Abou Zeid, M. J. (2021). Collective rationality and deliberation over five and more alternatives. Master's thesis, University of Bayreuth.

Alechina, N., Liu, F., & Logan, B. (2013). Minimal preference change. In D. Grossi, O. Roy, and H. Huang (Eds.), *International workshop on logic, rationality and interaction*, (pp. 15–26). Springer.

Bohman, J., & Rehg, W. (1997). *Deliberative democracy: Essays on reason and politics*. Cambridge: MIT Press.

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Flocken, C., & Holman, B. (2017). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science, 84*(1), 115–159.

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes, 79*, 1–39.

Cohen, J. (1989). Deliberation and democratic legitimacy. In A. Hamlin & P. Pettit (Eds.), *The Good Polity: Normative Analysis of the State* (pp. 17–34). New York: Basil Blackwell.

Cohen, J. (1989). The economic basis of deliberative democracy. *Social Philosophy and Policy, 6*(2), 25–50.

Cook, W. D., & Seiford, L. M. (1978). Priority ranking and consensus formation. *Management Science, 24*(16), 1721–1732.

Dietrich, F., List, C., & Bradley, R. (2016). Belief revision generalized: A joint characterization of Bayes' and Jeffrey's rules. *Journal of Economic Theory, 162*, 352–371.

Dorst, K. (2023). Rational polarization. *Philosophical Review, 132*(3), 355–458.

Dryzek, J. S., & List, C. (2003). Social choice theory and deliberative democracy: A reconciliation. *British Journal of Political Science, 33*(1), 1–28.

Duddy, C., & Piggins, A. (2012). A measure of distance between judgment sets. *Social Choice and Welfare, 39*(4), 855–867.

Eckert, D., & C. Klamler (2011). Distance-based aggregation theory. In E. Herrera-Viedma, J. L. Garcí a Lapresta, J. Kacprzyk, M. Fedrizzi, H. Nurmi, and S. Zadrożny (Eds.), *Consensual processes*, Volume 267 of *Studies in Fuzziness and Soft Computing*, pp. 3–22. Berlin: Springer.

Epley, N., & Gilovich, T. (2001). Putting adjustment back into the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12*, 391–396.

Estlund, D. (1993). Who's afraid of deliberative democracy? On the strategic/deliberative dichotomy in recent constitutional jurisprudence. *Texas Law Review, 71*, 1437–1477.

Estlund, D. (1997). Beyond fairness and deliberation: The epistemic dimension of democratic authority. In J. Bohman & W. Rehg (Eds.), *Deliberative Democracy: Essays on Reason and Politics* (pp. 173–204). Cambridge MA, London: MIT Press.

Farrar, C., Fishkin, J. S., Green, D. P., List, C., Luskin, R. C., & Paluck, E. L. (2010). Disaggregating deliberation's effects: An experiment within a deliberative poll. *British Journal of Political Science, 40*(2), 333–347.

Fishkin, J. S., & Luskin, R. C. (2005). Experimenting with a democratic ideal: Deliberative polling and public opinion. *Acta Politica, 40*(3), 284–298.

Furnham, A., & Boo, H. (1997). A literature review of the anchoring effect. *The Journal of Socio-Economics, 40*(1), 35–42.

Gaertner, W. (2001). *Domain conditions in social choice theory*. Cambridge: Cambridge University Press.

Grüne-Yanoff, T., & Hansson, S. O. (2009). From belief revision to preference change. In T. Grüne-Yanoff & S. O. Hansson (Eds.), *Preference change, theory and decision library* (Vol. 42, pp. 159–184). Dordrecht: Springer.

Habermas, J. (1984). *The theory of communicative action* (Vol. 1). New York: Beacon Press.

Hartmann, S., & Rafiee Rad, S. (2020). Anchoring in deliberation. *Erkenntnis, 85*, 1041–1069.

Jackson, M. O. (2010). *Social and economic networks*. Princeton NJ: Princeton University Press.

Kemeny, J. (1959). Mathematics without numbers. *Daedalus, 88*(4), 577–591.

Kemeny, J. G., & Snell, J. L. (1962). Preference ranking: An axiomatic approach. In *Mathematical models in the social sciences*. Hafner.

List, C. (2002). Two concepts of agreement. *The Good Society, 11*(1), 72–79.

List, C., Luskin, R. C., Fishkin, J. S., & McLean, I. (2012). Deliberation, single-peakedness, and the possibility of meaningful democracy: Evidence from deliberative polls. *The Journal of Politics, 75*(1), 80–95.

List, C., Pettit, P., et al. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford: Oxford University Press.

Makinson, D. (1993). Five faces of minimality. *Studia Logica, 52*(3), 339–379.

Manin, B. (1987). On legitimacy and political deliberation. *Political Theory, 15*(3), 338–368.

McElroy, T., & Dowd, K. (2007). Susceptibility to anchoring effects: How openness-to-experience influences responses to anchoring cues. *Judgment and Decision Making, 2*, 48–53.

Miller, D. (1992). Deliberative democracy and social choice. *Political Studies, 40*, 54–67.

Moulin, H. (1984). Generalized condorcet-winners for single peaked and single-plateau preferences. *Social Choice and Welfare, 1*, 127–147.

Mussweiler, T., Englich, B., & Strack, F. (2004). Anchoring effect. In R. F. Pohl (Ed.), *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement, and Memory* (pp. 183–196). Hove: Psychology Press.

Mussweiler, T., & Strack, F. (1989). Thinking the unthinkable: The effects of anchoring on likelihood estimates of nuclear war. *Journal of Applied Social Psychology, 19*, 67–91.

Mussweiler, T., & Strack, F. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology, 73*(3), 437–446.

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology, 35*, 136–164.

Mussweiler, T., & Strack, F. (2001). The semantics of anchoring. *Organizational Behavior and Human Decision Processes, 86*, 234–255.

Mussweiler, T., & Strack, F. (2005). Subliminal anchoring: Judgmental consequences and underlying mechanisms. *Organizational Behavior and Human Decision Processes, 98*, 133–143.

Niemi, R. G. (1969). Majority decision-making with partial unidimensionality. *American Political Science Review, 63*(2), 488–497.

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and Social Psychology Bulletin, 34*(7), 913–923.

Pacuit, E. (2019). Voting methods. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Fall 2019 ed.). Metaphysics Research Lab, Stanford University.

Peter, F. (2020). The grounds of political legitimacy. *Journal of the American Philosophical Association, 6*(3), 372–390.

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology, 64*(2), 243.

Rabinowicz, W. (2016). Aggregation of value judgments differs from aggregation of preferences. In A. Kuzniar & J. Odrowaz-Sypniewska (Eds.), *Uncovering Facts and Values: Studies in Contemporary*

Epistemology and Political Philosophy, Volume 107 of Poznan Studies in the Philosophy of the Sciences and the Humanities (pp. 9–40). Brill Rodopi.

Rafiee Rad, S., & Roy, O. (2021). Deliberation, single-peakedness, and coherent aggregation. *American Political Science Review, 115*(2), 629–648.

Stasser, G., & Titus, W. (2003). Hidden profiles: A brief history. *Psychological Inquiry, 14*(3–4), 304–313.

Tsetlin, I., Regenwetter, M., & Grofman, B. (2003). The impartial culture maximizes the probability of majority cycles. *Social Choice and Welfare, 21*(3), 387–398.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.