American Political Science Review (2020) 1–20
doi:10.1017/S0003055420001045

Deliberation, Single-Peakedness, and Coherent Aggregation

SOROUSH RAFIEE RAD *Bayreuth University*
OLIVIER ROY *Bayreuth University*

*Rational deliberation helps to avoid cyclic or intransitive group preferences by fostering meta-agreements, which in turn ensures single-peaked profiles. This is the received view, but this paper argues that it should be qualified. On one hand we provide evidence from computational simulations that rational deliberation tends to increase proximity to so-called single-*plateaued *preferences. This evidence is important to the extent that, as we argue, the idea that rational deliberation fosters the creation of meta-agreement and, in turn, single-*peaked *profiles does not carry over to single-plateaued ones, and the latter but not the former makes coherent aggregation possible when the participants are allowed to express indifference between options. On the other hand, however, our computational results show, against the received view, that when the participants are strongly biased towards their own opinions, rational deliberation tends to create irrational group preferences, instead of eliminating them. These results are independent of whether the participants reach meta-agreements in the process, and as such they highlight the importance of rational preference change and biases towards one's own opinion in understanding the effects of rational deliberation.*

This paper concerns what we will call the Received View of the role of deliberation in avoiding irrational group preferences (Dryzek and List 2003; List 2002; Miller 1992):

> ***Received View***. *Deliberation helps us to avoid irrational group preferences because it fosters the formation of meta-agreement and, in turn, single-peaked preferences.*[1]

The Received View consists of a combination of two claims.[2] First, it points to the fact that one of the results of deliberation is that it helps us to avoid irrational group preferences—that is, cyclic or intransitive rankings resulting from pairwise majority voting. This is a claim about the result or *effect* of deliberation. Second, the Received View makes an explicit claim regarding the *mechanism* that brings about that effect—namely, the formation of meta-agreement and, in turn, single-peaked preferences. This mechanistic claim is what we might call, following List (2002) and Dryzek and List (2003), the *meta-agreement hypothesis*.

We argue in this paper that the Received View should be qualified. We show, on one hand, that rational deliberation does not necessarily foster coherent aggregation. In certain cases it tends to *create* irrational group preferences. So the Received View's claim regarding the effect of rational deliberation should be qualified. In cases where rational deliberation steers the participants away from intransitive or cyclic group rankings, however, it does so through the creation of a stronger notion than that of single-peaked preferences, viz. the so-called single-plateauedness condition (see first section below). This, we also argue, is an important observation. The meta-agreement hypothesis indeed falls short of supporting the claim that deliberation also leads to an increase in proximity to single-plateauedness. When participants in deliberation are also allowed to express indifference among options, however, single-peakedness is not sufficient to ensure coherent aggregation, whereas single-plateauedness is. So there is an important gap in the Received View regarding the mechanism through which deliberation fosters coherent aggregation.

We argue for this using a computational, minimalistic model of social influence. In this model, the participants go through several rounds in which they exchange their opinions. Each round starts with one of the participants announcing her current preferences and the others updating their opinions accordingly. The next participant then announces her updated preferences, at which point the others once again update their preferences. This goes on until all participants have announced their preferences, and then a new round starts, following the same procedure. The order of the participants is randomly reassigned in each round. The preference updates follow a distance-minimization rule, weighed by possible biases that the participants may have towards their own opinions.

Soroush Rafiee Rad, Postdoctoral Researcher, Department of Philosophy, Bayreuth University, Soroush.R.Rad@gmail.com.
Olivier Roy, Professor, Department of Philosophy, Bayreuth University, Olivier.Roy@uni-bayreuth.de.

We would like to thank the editors and reviewers of the *APSR* for their careful consideration of the paper, especially in the midst of the 2020 pandemic lockdown, and their valuable comments that improved the paper significantly. The paper has also benefited immensely from comments by Alexandru Baltag, Christian List, Stefan Napel, Clemens Puppe, Jan-Willem Romeijn, and the participants in Philosophy Breakfast meeting at Bayreuth University and the ColAForm meetings in London, Paris, and Copenhagen.
Both authors' work is in part supported by the Deutsche Forschungsgemeinschaft (DFG) and Agence Nationale de la Recherche (ANR) as part of the joint project Collective Attitude Formation [RO 4548/8-1].
Received: April 28, 2019; revised: September 03, 2020; accepted: November 10, 2020.

[1] We assume the standard definitions of preference relations, single-peaked preferences, and voting cycles. See Appendix 1 for the technical details.
[2] We thank one of the anonymous reviewers of the *APSR* for urging us to make this distinction precise.

We argue in the Discussion section that this model is broadly compatible with normative views of deliberation, thus grounding our claim that what we call the normative reading of the Received View should be qualified. We argue, furthermore, that the model is general enough to encompass thicker understandings of group deliberation. The model thus provides an alternative "how-possible" explanation of how deliberation can avoid irrational group preferences, one that puts emphasis on rational preference change and openness to changing one's mind upon learning the opinions of others. We again take these findings to bear on the normative reading of the Received View. The paper does not put into question the empirical reading. Nevertheless, in Appendix 2, we do briefly compare our results with empirical findings.

WHY REVISIT THE RECEIVED VIEW

Normative and Descriptive Interpretations

The Received View rests on two observations. First and foremost are the conceptual arguments (Dryzek and List 2003; List 2002; Miller 1992) and the empirical evidence (Farrar et al. 2010; List et al. 2012) that support the meta-agreement hypothesis—that is, the claim that deliberation fosters the creation of meta-agreements and, in turn, single-peaked preferences.

Meta-agreements are agreements regarding the relevant dimensions along which the problem at hand should be conceptualized, as opposed to a full consensus on how to rank the alternatives. To take a concrete example, consider the 1996 British deliberative poll on the future of the monarchy (List et al. 2012). The participants were asked to rank three alternatives according to their preferences: to have a monarchy with a more ordinary royal family, to adopt a republic with a head of state with the same duties as the queen, or to adopt a republic with a head of state with the combined duties of the queen and the prime minister. In that case, the participants might, for instance, agree that the main dimension to consider when deciding on this issue is the trade-off between a more democratic system of checks and balances for the state and the social function of the monarchy, its institutions, and the duties associated with them. This does not entail that they will agree, even after deliberation, on the best way to make that trade-off.

The meta-agreement hypothesis is then coupled with the mathematical fact that pairwise majority voting always delivers a Condorcet winner when the input preference profile is single-peaked (Arrow 1963; Black 1948). With its domain restricted to single-peaked profiles, pairwise majority voting satisfies *rationality*, along with the other Arrowian conditions, when the number of voters is odd. In particular, it generates neither intransitive social preferences nor voting cycles. This means that, to the extent that deliberation fosters the creation of single-peaked preferences, it will tend to have the effect claimed by the Received View—namely, that of making coherent aggregation possible.

It is important to distinguish between two readings of the Received View. The first is the *normative* reading, according to which, under certain favorable conditions frequently associated with idealized or normative theories of deliberation, deliberative processes tend to generate meta-agreement and single-peaked preferences. This is the view, for instance, that is supported by the conceptual arguments presented in Dryzek and List (2003). On the other hand, one could offer an *empirical* reading of the Received View. Here the claim is rather that in real, concrete cases of deliberation, one should expect or even observe increases in meta-agreement that correlate with increases in proximity to single-peakedness. This claim receives some support in Dryzek and List (2003) but has been investigated more thoroughly in Farrar et al. (2010) and List et al. (2012).

Single-Peakedness and Single-Plateauedness

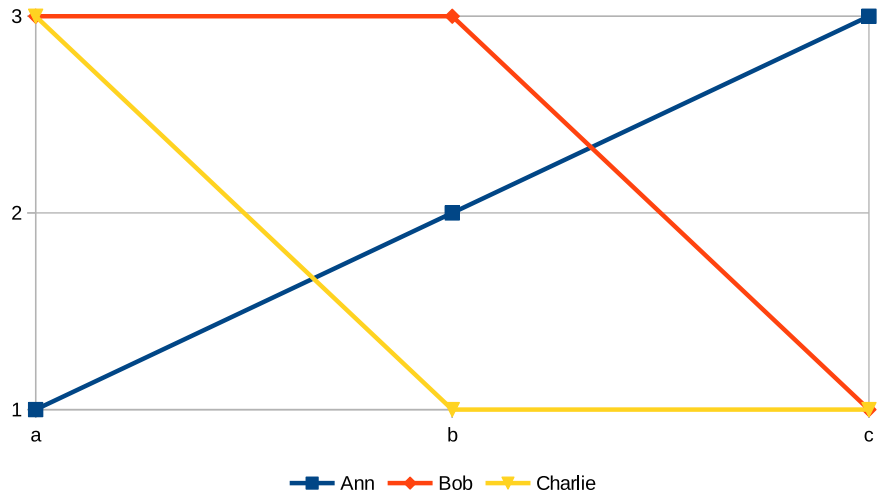
The Received View, as formulated above, is ambiguous regarding the very notion of single-peaked preferences. Black (1948) and Arrow's (1963) classical observation bears on strict rankings and the associated notion of strict single-peaked preferences. If we allow for weak preference rankings, however—viz difference between alternatives—there are at least two notions of single-peaked preferences to consider. This is illustrated in Figure 1.

Recall that single-peakedness requires the existence of a so-called "structuring dimension." In our example, this is the left-to-right ordering of the three alternatives, with a to the left, b in the middle, and c to the right. A given participant's preferences are *strictly* single-peaked with respect to a given structuring dimension whenever there is a unique most preferred option, the "peak," and as one moves away to the left or to the right of the peak, along the given dimension, one always moves to strictly less preferred alternatives.[3] In our example only Ann has strictly single-peaked preferences with respect to the given, left-to-right structuring dimension. We say that a group has strictly single-peaked preferences along a given structuring dimension whenever all participants have strictly single-peaked preferences with respect to that dimension, and we say that a group has single-peaked preferences whenever there exists a structuring dimension along which the group has single-peaked preferences.

When the participants are allowed to express indifference, as in the example where Bob is indifferent between a and b but Charlie is indifferent between b and c, the most general notion of single-peakedness is the so-called *weak single-peak* condition. It generalizes strict single-peakedness by allowing for indifference anywhere along the structuring dimension, only requiring that as one moves away from the most preferred options along the structuring dimension, one never goes "up" again—that is, one always moves to either strictly less-preferred alternatives or indifferent ones. Going back to our example, all three participants have weakly single-peaked preferences.

[3] See again Appendix 1 for the mathematical details.

FIGURE 1. Three Preference Orderings over the Alternatives *a*, *b*, and *c*



Note: Ann's ordering is strictly single-peaked with respect to the left–right geometric ordering. Bob's is single-plateaued but not strictly single-peaked, and Charlie's is weakly single-peaked but not single-plateaued.

Weak single-peakedness is the notion that Miller (1992) refers to, for instance.[4]

Weak single-peakedness is sufficient to avoid cyclic group preferences but can still allow for intransitive ones (Gaertner 2001; Puppe 2018). The profile in Figure 1 provides an example, with the group being indifferent between *a* and *b*, as well as between *b* and *c*, but strictly preferring *a* to *c*. The rationality of group preferences is thus *not* guaranteed by the creation of weakly single-peaked profiles.

To ensure the full rationality of the group preferences with weak rankings, some authors have resorted to a stronger understanding of single-peakedness, called "single-plateauedness" (Moulin 1984). Even though they use "single-peakedness" to refer to it, single-plateauedness is in fact the notion that is used in Dryzek and List (2003) and in the empirical work of Farrar et al. (2010) and List et al. (2012).

Single-plateauedness imposes two additional, structural constraints on the participants' preferences, beyond the possibility of aligning them on a single dimension. First, it only allows for indifference at the top of the ordering. See again Figure 1, where Charlie's ordering is *not* single-plateaued with respect to the given structuring dimension, even though it is weakly single-peaked, because he is indifferent between *b* and *c*, each of which he strictly prefers less than *a*. Second, single-plateauedness rules out complete indifference among the alternatives. Participants must hold some strict preferences. It is indeed easy to construct an example of a single-plateaued profile with full indifference where the resulting group preference is intransitive.[5]

[4] Niemi (1969) uses this notion as well but allows only for strict rankings, so it boils down to strict single-peakedness.

[5] The definition in Dryzek and List (2003) in fact rules out plateaus of more than two alternatives, even at the top. It can easily be generalized, though.

Can Rational Deliberation Create Single-Plateaued Preferences?

The additional structural constraints just mentioned raise doubts about whether the meta-agreement hypothesis could be strengthened so that it also explains the formation of single-plateaued preferences. Recall that the hypothesis, as formulated by List (2002) and Dryzek and List (2003), bears on the formation of single-*peaked* preferences. The argument in support of that hypothesis rests on a three-step mechanism, the last step of which is encapsulated in the following:

> We suggest that *if*, through deliberation, (i) a particular generalizable interest becomes focal and (ii) this generalizable interest can be associated with a single dimension, *then* (a high level of) single-peakedness is a likely consequence. (Dryzek and List, 2003, 16, emphasis in the original)

Now recall that both List (2002) and Dryzek and List (2003) use the expression "single-peakedness" to refer to the condition that, in both papers, formally corresponds to single-plateauedness. In other words, the formal definition they use is the notion of single-plateauedness, not strict or weak single-peakedness. So the question arises whether the argument they present applies equally well to both notions.

The argument has two parts. First, the "reflexive" aspect of deliberation (see the Discussion section below) should help to resolve factual disagreement on how to order the alternatives on the given (unique) dimension that expresses the relevant generalizable interest. We do not take issue with that part of the argument. Dryzek and List claim, however, that in a second step "rationality may finally lead individuals to have single-peaked preferences on the shared dimension" (Dryzek and List 2003, 16). List (2002) makes a similar claim.

We can grant for the sake of argument that rationality requires that there not be multiple "dips" in

preferences on what one explicitly recognizes to be the unique, agreed-upon dimension that expresses the relevant form of a generalizable interest. This will only lead us so far as to ensure the creation of weakly single-peaked profiles, however. As we have seen, this does not guarantee rational group preferences. If the claim is instead that deliberation can create single-plateaued preferences, then it is much less clear, at least at the outset, why rationality should require only one plateau at the top or, for that matter, rule out complete indifference. It seems plausible to assume that such cases of complete indifference, or partial indifference but not on top of the ordering, may naturally arise out of what List and Dryzek call the "social aspect" of deliberation (see again the Discussion section), through which the participants seek compromises when taking each other's opinions into account.

The argument in List (2002) and Dryzek and List (2003) thus does not appear to be strong enough to support a strengthened version of the meta-agreement hypothesis to the effect that, in ideal cases, rational deliberation tends to create single-plateaued preferences. Rationality alone does not seem strong enough to enforce the structural constraints that this notion entails.[6] Something more is needed, which rational deliberation may or may not provide. In other words, when we consider rational deliberation with weak preference rankings, the conceptual arguments developed in List (2002) and Dryzek and List (2003) fall short of providing a mechanism through which rational deliberation can have the effect it is claimed to have by the Received View. It is not clear how the process of rational deliberation alone can foster the creation of single-plateaued preferences, as opposed to "mere" weakly single-peaked ones.

Note that the situation is different for the empirical interpretation of the Received View. There, the evidence provided in Farrar et al. (2010) and List et al. (2012) suggests that deliberation indeed increases proximity to single-plateaued preferences[7] and that this increase inversely correlates with the relative salience of structuring dimensions, which has been taken as a proxy for meta-agreement. In other words, increases in proximity to single-plateauedness are greater in cases where the natural structuring dimension is less salient prior to deliberation, suggesting that deliberation indeed increases salience, and thus meta-agreement. This is evidence for the empirical reading of the Received View, of course, but it leaves open whether deliberation, in idealized circumstances, can be expected to have such a strong structuring effect and help us to go beyond the constraints of pure individual rationality towards the creation of single-plateaued preferences.

[6] The argument here is different from that presented in Ottonelli and Porello (2013), which argues that the first two steps of Dryzek and List's (2003) mechanism require a different, stronger form of substantive agreement.

[7] Recall, again, that these two papers use the expression "single-peaked preferences" to refer to a condition that formally corresponds to single-plateauedness.

Rational Deliberation in Impartial Cultures

A second reason to revisit the Received View is that irrational group preferences, and in particular voting cycles, are only likely in very specific circumstances. This is reflected both in empirical evidence (Feld and Grofman 1992; Radcliff 1994; Regenwetter et al. 2006) and in analytical results (Gehrlein 2004; Niemi 1969). For three alternatives and an odd number of voters, considering only strict rankings, the worst-case scenario for the probability of cycles is so-called *impartial cultures.* These are domains where a voter picked at random is equally likely to have any of the possible strict preference rankings on the alternatives. In impartial cultures, the probability of cycles rises monotonically with the number of voters.

As List (2017) puts it, however, this worst-case scenario is a "knife edge." Even small deviations from the impartial culture make cycles substantially less likely (List and Goodin 2001). The same holds true when moving to weak rankings. An impartial culture over strict rankings maximizes the probability of voting cycles, even when otherwise allowing for indifference (Fishburn and Gehrlein 1980; Jones et al. 1995; Tsetlin, Regenwetter, and Grofman 2003). Since normative views of deliberation impose mostly structural constraints on the process, they do not exclude impartial cultures at the outset. The empirical evidence cited above is of no help, since it is unlikely that participants in deliberative polls will be drawn from an impartial culture. The question of whether deliberative processes are conducive to the formation of rational group preferences in the worst-case scenario is thus open, and computational simulations are a natural tool to address it.

****

The Received View, in its normative interpretation, thus faces two challenges. First is the challenge of showing that rational deliberation can have the effect claimed by the Received View even in impartial cultures—that is, the respective worst-case scenarios for the probability of irrational group preferences with strict and weak rankings. Second is the challenge of showing that this effect is brought about by the creation of single-plateaued preferences, as opposed to weak single-peaked ones.

This paper addresses both challenges, and in the course of doing so provides a more nuanced reading of the Received View. We first show that in a large number of cases, deliberation helps to prevent both intransitive and cyclic social preferences, even in impartial cultures. It does not always do so, however. We point to a number of cases where deliberation actually *creates* cyclic or intransitive group preferences—that is, where it stands in the way of coherent aggregation. We show, however, that single-plateaued preferences track both cases. If the participants are minimally open-minded, deliberation *does* increase proximity to single-plateaued preferences. We argue in the Discussion section that this can be interpreted as supporting a strengthened version of the meta-agreement hypothesis, but that other interpretations that avoid meta-agreements altogether are also possible.

THE MODEL

We model the participants as entering deliberation by holding certain preferences, either weak or strict, over a given set of alternatives. Each participant then publicly announces, in turn, her full preference ranking. The participants do so sincerely and in a random order. After each announcement, they update their ranking, using a distance-minimization rule. That is, their new preference ranking is one that minimizes a given distance measure between their old ranking and the one just announced, possibly with a bias towards their own preferences. Each announcement is thus followed by an update, as opposed to having the participants wait for all preferences to be announced prior to updating. The domain of preference rankings in which they move during deliberation is either the set of all possible strict rankings or the set of all possible weak ones. To avoid having the results hinge on the particulars of a specific distance measure, we use and compare the effects of three well-known such measures between orderings: the Kemeny–Snell (KS; 1962), Cook–Seiford (CS; 1978), and Duddy–Piggins (DP; 2012) distances. Deliberation continues for a fixed number of rounds, after which we check whether the resulting preference profile, if not already consensual, is single-peaked or single-plateaued and whether the group preference generated by pairwise majority voting would be cyclic or intransitive.

The model belongs to the category of what List (2017) calls models of "deliberation as preference transformation."[8] It focuses exclusively on how the participants' preferences change upon learning the preferences of the others. This is thus a model of social influence. It is stripped of many of the features of thicker understandings of deliberation. There is, for instance, no explicit exchange of reasons or no explicit restriction to preferences expressing generalizable interests (Bohman 1997). On the other hand, the model also brackets other factors that might otherwise stand in the way of the positive effects of deliberation, for instance strategic considerations (cf. Landa and Meirowitz [2009] and the references therein). In our model, the participants do not try to convince others to maintain their status or reputation or even to get the right answer. We assess the pros and cons of these modeling choices in further detail in the Discussion section.

We start with a group N of n individuals entering deliberation with a preference ordering over a given set of alternatives, $a_1, \ldots, a_j$. For computational reasons, the results reported here only cover the case of three alternatives.[9] Let $\mathcal{R}$ be the set of all possible rankings over the set of three alternatives. With strict preference rankings over three alternatives, there are six different rankings in $\mathcal{R}$, and 13 if we allow for weak rankings as well. The individuals in N are said to constitute an *impartial culture* whenever, for each $i \in N$ and each ranking $r \in \mathcal{R}$, the probability that i enters the deliberation with ranking r is $1/|\mathcal{R}|$, so $1/6$ in the strict case and $1/13$ otherwise. In order to maximize the probability of cycles, however, for weak rankings we will consider cases in which the participants enter deliberation with strict rankings only but are then allowed to compromise, so to speak, and move to weak rankings in the course of deliberation. As noted earlier, this case maximizes the ex ante probability—that is prior to deliberation—of the cycles.

A *round* of sharing consists of n steps, one for each participant. At step i $(1 \leq i \leq n)$, participant i announces her ranking, and the rest of the group update their opinion accordingly. Each participant announces her preferences once (and only once) every round. The order of speakers is randomly reassigned at each round. This process continues for a fixed number, w, of rounds.

Preference updating is done through distance minimization. Let $R = \langle r_1, \ldots, r_n \rangle$ be a given preference profile. After the i-th member speaks, the rankings are updated to $\langle r'_1, \ldots, r'_n \rangle$, where r'_j is the ranking for which

$$\sqrt{\mathrm{rel}_i d\left(r_i, r'_j\right)^2 + \mathrm{rel}_j d\left(r_j, r'_j\right)^2}$$

is minimal for a given distance measure d. In the strict case, the r'_j are picked from among the domain of strict rankings, and analogously in the weak case.

The parameters rel_j and rel_i represent, respectively, the bias each agent has towards her own opinion. This allows us to capture different degrees to which the participants are open to changing their minds upon learning the opinions of others. We take these two parameters to be real values in the [0,1] interval, with $rel_j = (1 - rel_i)$. Little hinges on this choice of scale; what matters is their relative weight. If $rel_j = rel_i = 0.5$, then j views i as an equal peer: she assigns as much weight to his opinion as to her own. If $rel_j > rel_i$, then j is biased towards her own opinion. Indeed, by putting more "weight" on the parameter $d(r_j, r'_j)^2$, we require that its value be proportionally smaller. The larger the difference between rel_j and rel_i, the stronger the bias. If, at the extreme, $rel_j = 1$ and $rel_i = 0$, then j does not take i's opinion into account at all. The profile r_j will then always trivially minimize the distance to itself. For simplicity, we assume that all agents are biased in the same way towards themselves and all others. We thus keep the values rel_j and rel_i constant for all i and j in the above formula. We return to the interpretation of these bias parameters in the Discussion section.

We use what are arguably the three main approaches to comparing preference rankings: the Kemeny–Snell distance (Kemeny 1959; Kemeny and Snell 1962), the Cook–Seiford distance (Cook 2006), and the Duddy–Piggins distance (Duddy and Piggins 2012). All three

[8] List (2011) proves an impossibility theorem for deliberation as preference transformation. The distance-based approaches that we use below avoid this impossibility by violating the assumption of independence of irrelevant alternatives, arguably the most debatable assumption used in the theorem. See also Kemeny (1959) for a concrete, arguably unproblematic example of such a violation.

[9] Partial results have been obtained for the CS measure for sets of five alternatives. No significant changes from the results presented here have been observed.

have been used in models of preference aggregation (Cook 2006; Kemeny and Snell 1962), belief merge, and democratic deliberation (Duddy and Piggins 2012; Perote-Peña and Piggins 2015). We give their precise definitions in Appendix 1. For now, a few general comments suffice. Kemeny–Snell and CS are very close to one another. They differ mostly in the underlying notion of "betweenness" that they use (see again the Discussion section and Appendix 1). The DP measure is comparatively more recent. It was developed to handle the possible double counting that can occur with KS when comparing vectors with (logically) related components.[10]

Deliberation continues for a fixed number w of rounds, after which we measure closeness to single-peakedness or single-plateauedness in the same way as in Niemi (1969) and List et al. (2012): proximity to single-peakedness/plateauedness is calculated as the relative size of the largest subgroup that is single-peaked or single-plateaued with respect to one ranking. We also calculate how far the group has moved towards consensus through deliberation. This will be important for understanding how deliberation avoids irrational group preferences on this model. We do so by comparing the distances of the initial and the final profiles with their respective closest consensus profiles. We finally check whether the resulting group preference is cyclic or intransitive.

Example of the Deliberative Process

Consider three agents, i, j, and k, deliberating on how to rank three alternatives a, b, and c. Let us assume that deliberation proceeds using the DP distance measure. Suppose further that i, j, and k consider each other epistemic peers—a bias of 0.5—and that their initial preference rankings are as follows:

$$r_i^0 = (a,b,c), \quad r_j^0 = (b,a,c) \quad r_k^0 = (c,b,a).$$

Let us assume that for the first round the agents speak in alphabetical order. So i announces her preference first, and j and k update accordingly.

Since c is the least preferred option for both i and j, it is natural that j would still consider it the least preferred option after updating her judgment by considering i's. On the other hand, j strictly prefers b to a, while i strictly prefers a to b. Since j takes i's judgment to be as reliable as hers, it is natural for her to try to reduce the discrepancy between her judgment and i's. If she gives equal weight to herself and i, this will lead her to rank a and b equally. Her updated ranking should then be $(\{a,b\},c)$. However, k's judgments are the complete opposite of i's on every two alternatives. If she then considers i's judgment to be as reliable as her own, the only natural choice is to take the middle ground and update her ranking to $(\{a,b,c\})$. These are precisely the rankings r_j^1 and r_k^1 that minimize the DP distances

$$\sqrt{d_{DP}\left(r_j^1, r_i\right)^2 + d_{DP}\left(r_j^1, r_j\right)^2}, \text{ and}$$

$$\sqrt{d_{DP}(r_k^1, r_i)^2 + d_{DP}(r_k^1, r_k)^2},$$

respectively—that is, $r_j^1 = (\{a,b\},c)$ and $r_k^1 = (\{a,b,c\})$. The new profile will then be

$$r_i^1 = (a,b,c) \quad r_j^1 = (\{a,b\},c) \quad r_k^1 = (\{a,b,c\}).$$

In the second step, j announces her newly updated ranking $r_j^1 = (\{a,b\},c)$, and i and k update their rankings in the same manner as above.

Since c is ranked last by both i and j, i would still consider it the least preferred option after updating with j's opinion. For a and b, however, the situation is different: i strictly prefers a to b, while j ranks them equally. There is no middle ground for i that would allow her to bring her judgment closer to j's, however. The only options available to her are to either ignore her own preference and adopt j's or to ignore j's and keep her current preference. The situation is similar for k. Both she and j rank a and b equally, but one ranks c strictly below a and b while the other is indifferent between them. Again, for the judgment between a and c (or b and c), there is no middle ground for k. Her options are to either keep her own preferences or adopt j's.

These are again the rankings r_i^2 and r_k^2 that minimize the respective DP distances

$$\sqrt{d_{DP}\left(r_i^2, r_j^1\right)^2 + d_{DP}\left(r_i^2, r_i^1\right)^2}, \text{ and}$$

$$\sqrt{d_{DP}\left(r_k^2, r_j^1\right)^2 + d_{DP}\left(r_k^2, r_k^1\right)^2}.$$

At this stage, i has two options, both of which minimize the above distance. She can keep her current preferences or she can adopt j's. Similarly, for k the options of keeping her current preferences and adopting j's minimize the given distance. So, given the equal weight assumption, the options given by the DP distance minimization above do indeed seem to be the most natural choices available to i and k. In cases like this, our model draws one of the eligible rankings at random.

Let us assume for now that both j and k choose to keep their current preferences, resulting in no change in the profile:

$$r_i^2 = (a,b,c) \quad r_j^2 = (\{a,b\},c) \quad r_k^2 = (\{a,b,c\}).$$

This first round of deliberation then ends with k announcing her preferences and i and k updating theirs accordingly, moving to updated ranking r_i^3 and r_j^3.

Here, we see that in each of the binary judgments between a and b, a and c, or b and c, there is no middle

[10] This possibility is relevant here. The measures will in effect be comparing judgment sets defined from weak or strict rankings, which are all transitive and complete. These properties impose constraints on the resulting judgment sets that are analogous to those stemming from the logical relationships between propositions.

ground on which i can bring her judgment closer to k's. For each of these, she can either keep her own judgment or move to the one announced by k. However, she can still get closer to k's opinion without fully giving up her own ranking by adopting k's ranking for some of these binary choices and keeping her own on others in a consistent way. This will give her two options: $(a,\{b,c\})$, where she maintains her own judgment on pairs a,b and a,c and adopts k's for the pair b,c and $(\{a,b\},c)$. Note that it would be inconsistent for her to adopt k's ranking for two of the pairs and her own for the other. These are precisely the options that minimize the DP distance between i and k:

$$\sqrt{d_{DP}(r_i^3,r_k^2)^2 + d_{DP}(r_i^2,r_i^3)^2}.$$

On the other hand, j will be in the same situation as before: she has no intermediate ranking to move to in order to get closer to k's opinion. Her only options are to either fully adopt the ranking given by k or to completely ignore it and keep her own. This is because they both agree on ranking a and b the same, and thus they only differ in terms of how they rank option c with respect to a and b, and this leaves no room for an intermediate ranking. Again, these are precisely the choices given by the DP distance minimization

$$\sqrt{d_{DP}\left(r_k^3,r_j^2\right)^2 + d_{DP}\left(r_k^2,r_j^2\right)^2}.$$

Depending on the choices of i and k, the new profile will, for example, be

$$r_i^3 = (\{a,b\},c) \quad r_j^3 = (\{a,b\},c) \quad r_k^3 = (\{a,b,c\}).$$

Here, the first round of deliberation ends and the second round starts. Let us assume that j speaks first. After she announces her ranking, i and k will update theirs. Participant i holds the same ranking as the one announced by j, so she will naturally stick to her original choice. And k is in exactly the same situation that j was in the previous step, so her only option is to either stick to her own choice or adopt j's—that is, $(\{a,b\},c)$ or $(\{a,b,c\})$. These, again, are precisely the choices that minimize the corresponding DP distance minimization.

Depending on this choice, the new profile will either be

$$r_i^4 = (\{a,b\},c) \quad r_j^4 = (\{a,b\},c) \quad r_k^3 = (\{a,b\},c),$$

where the group has reached consensus, or

$$r_i^4 = (\{a,b\},c) \quad r_j^4 = (\{a,b\},c) \quad r_k^3 = (\{a,b,c\}).$$

At this point, the deliberation can continue as before; from now on, however, either the agent's choices will remain unchanged (when the announced ranking is the same as that held by the agent) or she will be given a choice between keeping her own preferences and adopting the opinion of the speaker. Depending on how these choices are made, the group can end with a consensus on $(\{a,b\},c)$ or $(\{a,b,c\})$ or stabilize with two agents holding one of these preferences and one holding the other. These choices are similarly supported by the DP distance minimization, and with this strategy the agents are free to choose from among their available options at each stage based on possibly different considerations.

It is worth pointing out that in the analysis of what would be a natural move for our agents in updating their rankings, we interpreted their desire to genuinely consider the opinions of others and accommodate them in their preferences by focusing on how they differ from the speaker in their binary judgments about the alternatives. The agents then updated their rankings by consistently changing some or all of these binary comparisons. It is thus no surprise that the results given by DP distance minimization (or for the same reason with KS) are natural choices for our agents: the number of required binary changes is indeed what DP takes into account in deciding the distance between rankings. The CS distance, which assigns numerical values to the positions in rankings and calculates the distances on that basis, will then correspond to natural choices for agents who use a different strategy for moving closer to the opinions of their group members: for example by averaging their numerical value for each alternative with the numerical value given by the speaker.

RESULTS

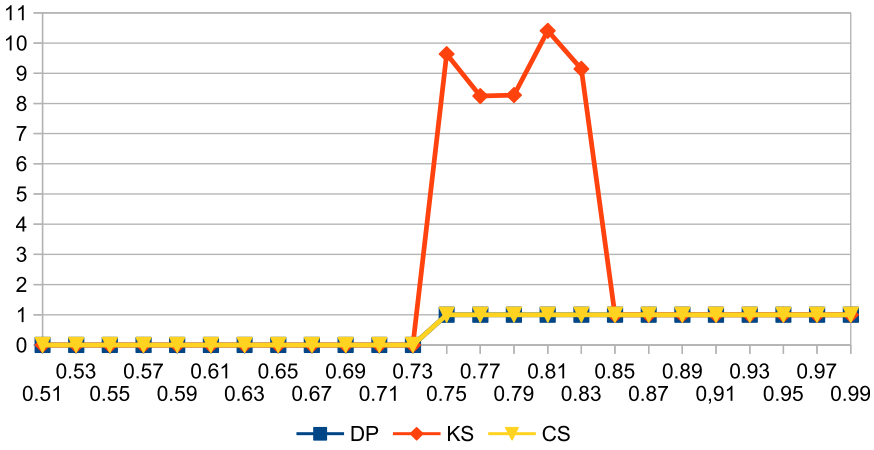
We shall look at the behavior of the model under two different domain assumptions: domains of weak and of strict rankings. For strict rankings, this restriction is for both the input and the output of the deliberation procedure. So, when we talk about deliberation with strict rankings, we mean deliberation in which both the input and the output profiles are linear orders. For the case of weak rankings, however, the initial profiles are strict, but afterwards the participants are allowed to move to pre-orders. Again, we do so in order to maximize the probability of starting cycles.

Deliberation with Strict Rankings Only

For strict rankings, the starting preference profiles are drawn at random under the impartial culture assumption. With this assumption, the probability of cycles is still low for small groups but increases monotonically with the number of agents (Niemi 1969). We observe this as well. Since this increase is relatively slow, however, most of our simulations will be done for groups of 51 agents.

As shown in Figure 2, with strict rankings deliberation completely eliminates voting cycles up to a bias of 0.7 for all three distance measures. For higher biases, all three measures completely stop eliminating cycles. As we shall see presently, for DP and CS this is simply the result of the participants' not changing their minds at all for biases above 0.7. Thus, all and only the initial cycles remain after deliberation. For KS, the situation is different. Deliberation in fact creates a large number

FIGURE 2. Proportion of Initial Cyclic Profiles Still Present after Deliberation (*y*-axis) on Strict Rankings as Bias Increases (*x*-axis; *n* = 51)



of cycles in the 0.75–0.85 bias bracket. For larger groups, this can be observed for all three measures, all in the same bias bracket. Above 0.85, participants who update their preferences using KS, like CS and DP, stop changing their minds altogether.

Deliberation only eliminates cycles through consensus formation when the participants consider all others as peers—that is, with bias 0.5 (Figure 3). For higher bias values, up to 0.7, the participants cluster around two rankings, and relatively fast (2 or 3 rounds).[11] This explains the elimination of cycles up to that value. Any pair of strict rankings on three alternatives is single-peaked. In the range where it creates cycles, KS moves the participants around three clusters. For higher biases, as mentioned above, for all three measures the participants stop changing their minds altogether.

The fact that the participants quickly cluster around a relatively small number of preferences does not necessarily mean that they are closer to consensus than they were at the beginning of the deliberation. This is already suggested by the fact that the clusters are too far away from each other for the participants to move further. Figure 4 makes this suggestion precise. It shows, for each distance measure, how much closer to consensus the final preference profile is in comparison with the starting profile. Since these numbers represent values using three different distance measures, their exact value is not crucial here. What matters is the slope, the relative difference between points on the same line, and the fact that from 0.85 onward, that value is 0.

At the outset of deliberation, we get an average proximity to single-peakedness of 0.74 for groups of 20 to 95 participants, with a slight declining slope as the group size increases. Recall that proximity to single-peakedness is calculated in the same way as in Niemi (1969) and List et al. (2012). As we saw, under all three measures, up to a 0.7 bias, deliberation creates completely single-peaked profiles, as the participants cluster around only two rankings. Above 0.7 for DP and CS, and above 0.85 for KS, the participants do not change their minds at all, so proximity to single-peakedness stays the same. Interestingly, we observe an increase in proximity to single-peakedness of 6% on average for KS in the 0.75–0.85 bracket, where, recall, this measure creates a large number of voting cycles.

Deliberation with Weak Rankings

When allowing for indifference, the impartial culture *on the strict rankings* maximizes the probability of cycles, so this will be our starting point for all simulations. In other words, all participants enter deliberation with strict rankings, but they are able to compromise and settle for indifference in the course of the deliberation.

Deliberation eliminates *all* intransitive profiles and, by the same token, all cyclic ones, up to a bias of 0.75 (cf. Figure 5). Beyond that, we again observe a phase where deliberation with each of the three measures creates irrational group preferences, this time intransitive ones. Note that these are not necessarily cyclic. Figure 6 illustrates this for deliberation with DP distance. These findings are robust as group size increases. The average move to consensus and its correlated clustering follows a pattern that is similar to deliberation with only strict rankings, except that both processes are slower.

Up to bias values of around 0.75, deliberation can foster the creation of single-plateaued preferences (see Figure 7 for KS), but not necessarily (see Figure 7 for DP and CS). The initial proximity to single-plateauedness is stable throughout our simulations, for instance around 0.73 for groups of 51 agents. For KS, deliberation outputs almost universal single-plateauedness up to values of 0.80. This means, for this bias bracket, an average increase of 0.26 in proximity to

[11] Recall that a round consists of updates, one for each participant announcing her preferences. So for large groups, two rounds constitute a substantial number of updates.

FIGURE 3. Average Number of Clusters after Deliberation (*y*-axis) on Strict Rankings, as Bias Increases (*y*-axis; *n* = 51)

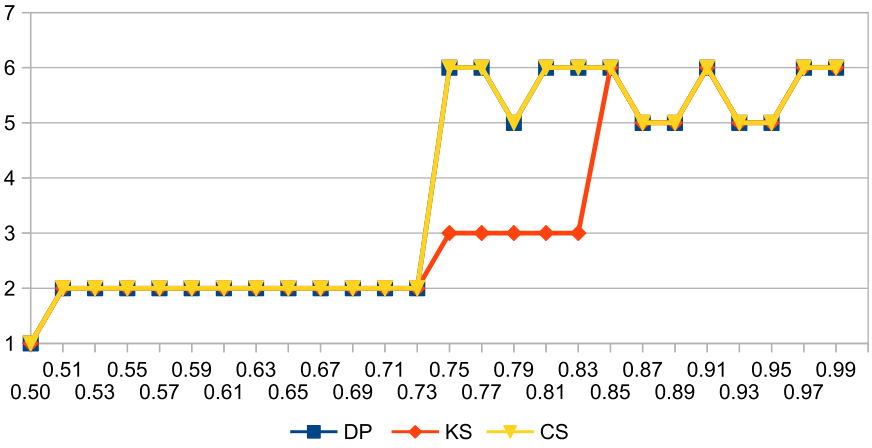
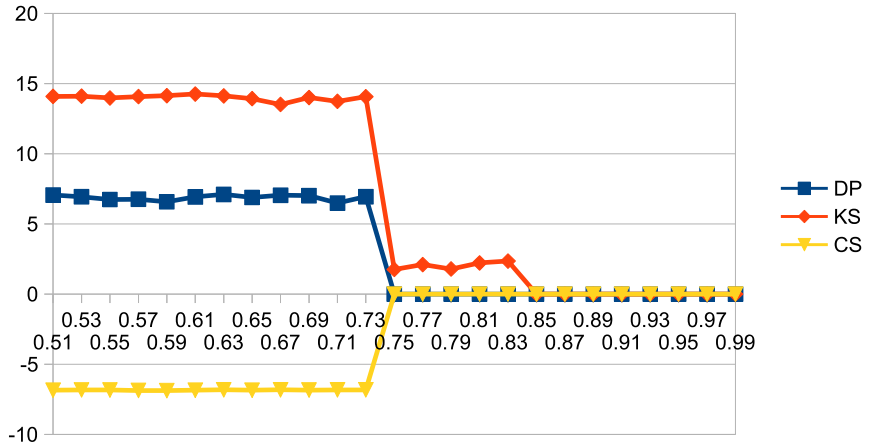


FIGURE 4. Average Move towards Consensus (*y*-axis) as Bias Increases (*x*-axis) for Strict Rankings (*n* = 51)



weak single-plateauedness through deliberation for KS, while for the same bracket DP shows a similar decrease.

We observe, by contrast, a rather sharp decrease in strict proximity to strict single-peakedness in deliberation (Figure 8). Recall that in these simulations the participants enter deliberation with strict rankings only. All the rankings that are identified as being single-plateaued are thus also strictly single-peaked. Here, the participants are allowed to compromise by moving to indifference through deliberation, however. The fact that the number of strictly single-peaked rankings decreases so much suggests that deliberation with all three distance measures tends to create more weak than strict rankings in cases of strong disagreement. This effect appears to be stronger for the DP measure.

The effect of the creation of single-plateaued profiles on irrational—that is, intransitive or cyclic, group preferences is more subtle here than for strict rankings. Recall that up to biases of 0.75, KS deliberation creates high proximity to single-plateauedness. Since this is sufficient to avoid irrational group preferences, up to that bias value all group preferences are rational. However, even for DP, which shows a decrease in the proximity to single-plateauedness in this bracket, deliberation completely eliminates intransitive and cyclic profiles (see Figure 6).

In fact, as Figure 9 (left) shows, KS turns a large number of starting intransitive group preferences into single-plateaued profiles up to even very high bias values.[12] This is so despite the fact that, as we have seen, between 0.75 and 0.9 deliberation with these measures tends to *create* intransitive group preferences. Figure 9 (right) illustrates this for deliberation with the

[12] The data plotted in Figure 9 include weak rankings at the start of deliberation in order to make room for intransitive but acyclic rankings.

FIGURE 5. Percentage of Intransitive Profiles Still Present after Deliberation on Weak Rankings (*y*-axis), as Bias Increases (*x*-axis; *n* = 51)

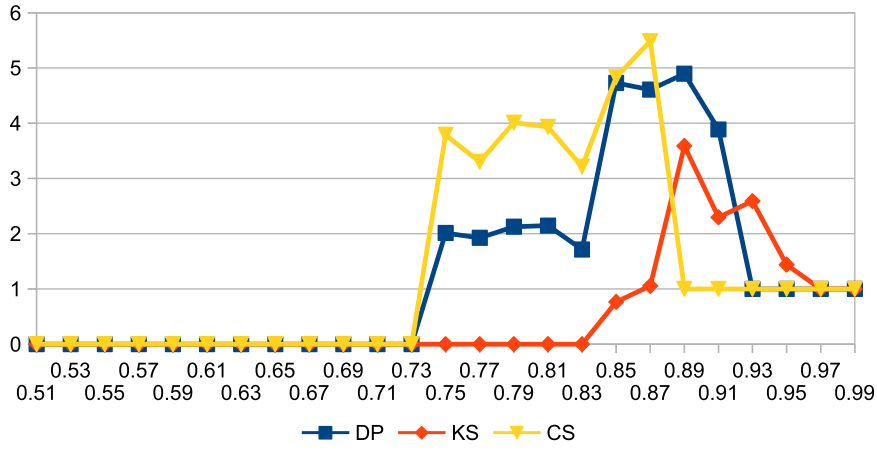


FIGURE 6. Percentage (*y*-axis) of Intransitive (Red Line) and Cyclic Profiles (Blue Line) Still Present after Deliberation on Weak Rankings Using the Duddy–Piggins Distance as Bias Increases (*x*-axis)

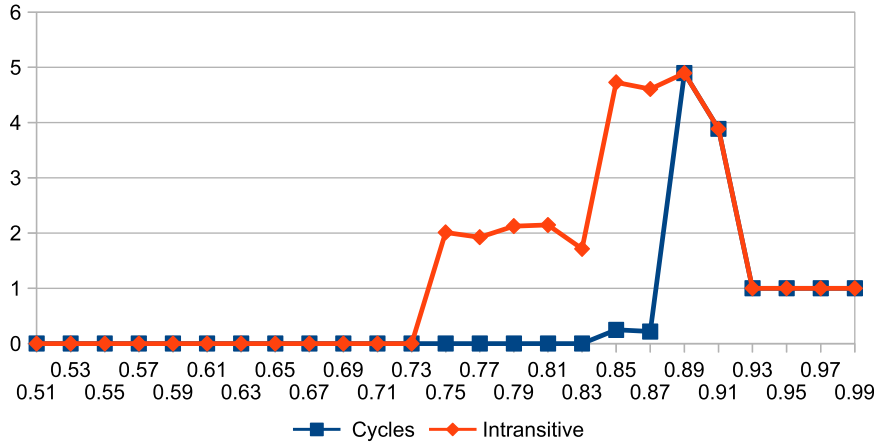


FIGURE 7. Average Proximity to Single-Plateauedness (*y*-axis) as Bias Increases (*x*-axis; *n* = 51)

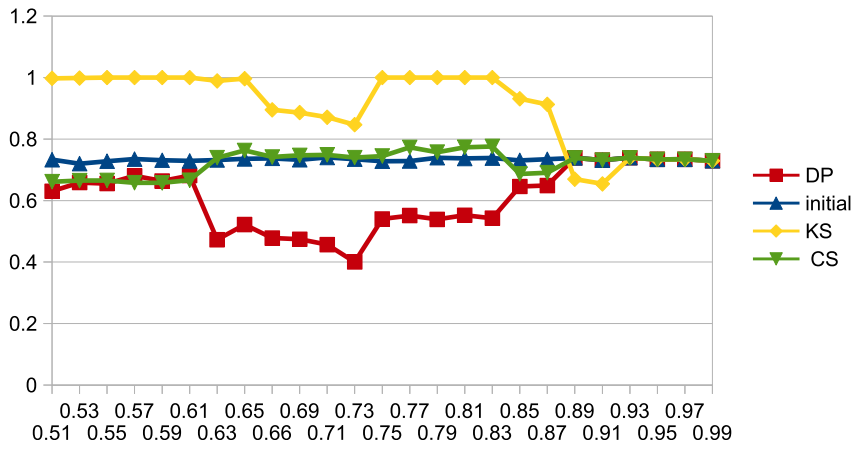


FIGURE 8. Average Proximity to Strict Single-Peakedness (*y*-axis) as Bias Increases (*x*-axis; *n* = 51)

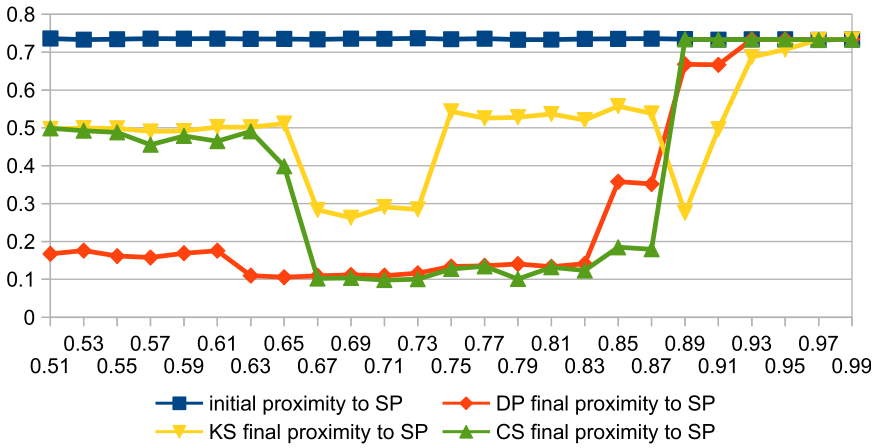
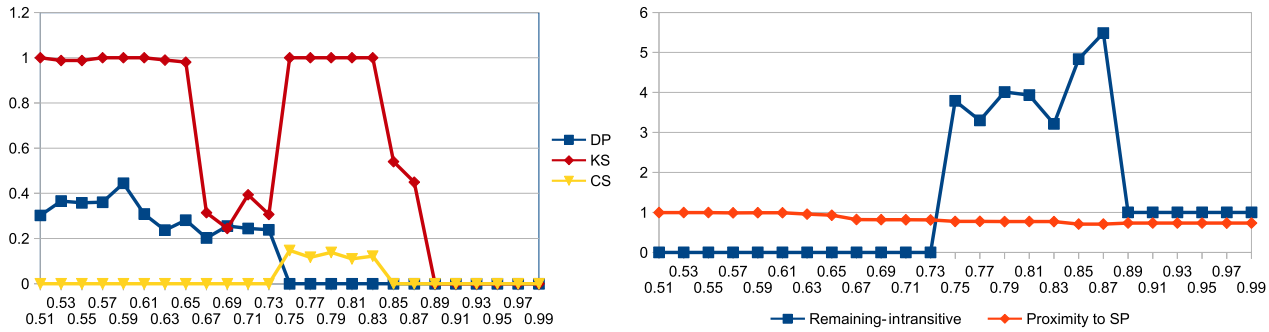


FIGURE 9. Proportion of Rankings Starting as Nontransitive and Ending as Single-Plateaued (*y*-axis, Left) and Proportion of Remaining Intransitive Rankings Together with Resulting Proximity to Single-Plateauedness for CS Distance (*y*-axis, Right) as Bias Increases (*x*-axis, Both; *n* = 51)



CS measures, where we see that between 0.75 and 0.83 we get around 0.7 proximity to single-plateauedness even though deliberation increases the number of intransitive profiles on average more than 500%.

DISCUSSION

Deliberation and Group Preferences: A Qualified Positive Outlook

Political philosophers and social choice theorists have emphasized the importance of deliberation and of single-peaked preferences as a reaction to a rather pessimistic reading of Arrow's impossibility theorem (see e.g., Dryzek and List 2003; Miller 1992). A paradigm of the latter, Riker (1982) claims that this result puts into question the very meaningfulness of voting as a means of capturing social preferences. After Arrow, what is left for democratic voting, Riker claims, is the rather minimal function of periodically removing politicians from office, if necessary. Deliberation seems to offer a natural way to resist this conclusion. Because it can create single-peaked preferences, deliberation allows us to circumvent Arrowian impossibilities in general, and irrational group preferences in particular, and by the same token it reopens the door to a more substantial function of democratic voting.

Our results provide qualified support for this rejoinder. As claimed by the Received View, deliberation has the effect of completely eliminating irrational group preferences, in the case of both weak and strict rankings, *as long as the participants are minimally willing to take into account the opinions of others.* The required threshold of openness appears minimal. In the strict case, the point after which deliberation fails to ensure the rationality of group preferences coincides with the point at which the participants almost completely stop taking others' opinions into account during deliberation. In the weak case, the situation is more subtle. For middle-range biases, the participants still change their minds through deliberation. They tend, however, to move towards profiles that are irrational at the group level.

This positive outlook must thus be qualified by the fact that, against the Received View, for midrange biases towards oneself deliberation tends to create irrational group preferences. As we have seen, this is

true for both weak and strict rankings, and we observe this effect robustly across different group sizes. This happens despite the fact that, as we argue below, we model an idealized case where the participants are using individually rational preference change policies that broadly correspond to two features commonly associated with ideal deliberation. The model thus unveils a novel "paradox" of rationality, viz. another case where individual and collective rationality clash. Note, however, that for weak rankings, many irrational group preferences that are created through deliberation are intransitive but not cyclic. From a normative point of view, this is less damaging than the creation of cyclic group preferences. In those cases, there are still most preferred elements in the group ranking, which in turn allows for the possibility of meaningful social choice, although of course *not* for the construction of a full social welfare ordering.

Returning to low biases, the positive outlook is bolstered by the fact that these results were obtained in the worst-case scenario—that is, under the impartial culture assumption. Recall that this means that each participant is equally likely to enter the deliberation with any of the possible rankings of the alternatives. This probability is, furthermore, independent between agents.[13] Deliberation breaks this symmetry. The participants, as we have seen, end up clustering around a small number of rankings. This turns out to be sufficient to push them off the "knife edge" (List 2017) and ensure the transitivity of the social preference. In other words, even starting with the worst-case scenario, the threat of Arrowian impossibilities and irrational social preferences is alleviated by deliberation when agents are minimally open to the opinions of others.

This positive effect also sheds light on the respective roles of single-peakedness and single-plateauedness in avoiding irrational group preferences. For deliberation on strict rankings, avoiding cycles goes hand in hand with the creation of single-peaked preferences, as claimed by the Received View. For weak rankings, however, the situation is more subtle. As we have observed, for low biases deliberation completely eliminates irrational preferences, and this goes hand in hand with the creation of single-plateaued profiles, at least for the KS distance measure. For DP, however, deliberation slightly decreases the proximity to single-plateauedness, but not sufficiently to threaten the rationality of the resulting group preferences. Deliberation with that measure, at low biases, still completely eliminates intransitive rankings.

This is an important observation since, as we emphasized in the first section, the single-plateau condition imposes two additional constraints on preferences, beyond alignment on a common structuring dimension: the exclusion of complete indifference and of indifference that is not at the top of the ordering. We argued that individual rationality alone does not seem strong enough to enforce these constraints and thus that something more is needed. Our results show that deliberation does help us to meet these constraints.

Since the notion of meta-agreement is absent in our model, the results can be seen as either a complement or an alternative to the meta-agreement hypothesis. The model does not rule out the possibility that the increase in proximity to single-plateauedness might be accompanied by an increase in meta-agreements. From that point of view, the results suggest that rational preference change and openness to changing one's mind upon learning the opinions of others might be the missing parameter that allows us to strengthen the meta-agreement hypothesis so that it also covers single-plateaued preferences. To paraphrase Dryzek and List (2003, 16), the model suggests that

> if (i) the participants are minimally open to changing their minds upon learning the opinions of others and, throughout rational deliberation, (ii) they repeatedly do so using a rational preference change policy, and if, furthermore, (iii) a particular generalizable interest becomes focal and (iv) this generalizable interest can be associated with a single dimension, then (a high level of) single-*plateauedness* is a likely consequence.

This interpretation of our results as being complementary to the Received View could, of course, be replaced by a more radical one that leaves out meta-agreements altogether and instead puts emphasis on rational policies for preference changes and openness to changing one's mind. The crucial hypothesis would then become

> if (i) the participants are minimally open to changing their minds upon learning the opinions of others and, throughout rational deliberation, (ii) they repeatedly do so using a rational preference change policy, then (a high level of) single-*plateauedness* is a likely consequence.

On that reading, instead of aligning their preferences on a salient structuring dimension, increases in proximity to single-plateauedness might instead result from the participants increasingly understanding each other's point of view and, in particular, recognizing the reasonableness of others' preferences.[14] Adjudicating between both hypotheses would require extending the model with a measure of meta-agreement, however, which we leave for future work. Nonetheless, we take it that highlighting the importance of preference change policies and willingness to take others' opinions into account is an important contribution of our model.

It should be furthermore emphasized that, returning to the Received View, our results show that for mid-to high-range biases, relative proximity to single-plateauedness is not even strong enough to prevent the emergence of *new* irrational group preferences. This is

[13] To be precise, what we have used here is the impartial culture assumption, to be distinguished from the impartial anonymous culture assumption (cf. Gehrlein 2004), which allows for correlations between voters. We leave open the question of how our results would generalize to that case.

[14] We again thank one of the reviewers of the *APSR* for urging us to make this possible interpretation explicit.

particularly salient for deliberation with the KS measure, which, in the 0.87–0.95 bracket, increases the number of new irrational group preferences by up to 300%, despite *also* maintaining above 0.8 proximity to single-plateauedness. This shows that if the main reason to value single-plateaued profiles is that they are instrumental to ensuring rational social preferences, this reason is valid only in a restricted class of cases.

Normative Interpretation of the Model

We take our model to be broadly compatible with normative theories of deliberation and thus to have a bearing on the normative reading of the Received View. We argue for this in this subsection and the next. We first explain how the two key parameters in our model are compatible with a normative interpretation. We then move on to situating our modeling choices more generally within the landscape of richer normative theories of deliberation.

Openness to changing one's mind, and the resulting preference dynamics, are the main driving forces in our model. Openness is captured through a simple bias parameter, which induces a possibly unequal weighing of one's opinion in comparison with the opinions of others. The preference dynamic is induced by the minimization of any of the three distance measures.

The bias parameter can be interpreted in at least two ways. On one hand, it can be seen as embodying a form of "cognitive inertia" (Allport and Wylie 2000) through which the participants are resistant to opinion change. To use the standard formulation, they do not necessarily "yield to the force of the better argument" (Steiner 2012). This would yield a non-ideal reading of our model. This is not the only possible interpretation, however. One can instead interpret the bias parameter as a form either of comparative expertise, much in the spirit of (Estlund 1997), or of "mutual respect" (Martini, Sprenger, and Colyvan 2013). In the first case, the parameter represents how much more of an "expert" on the question at hand each participant sees herself as being in comparison to the others. Since for simplicity we have assumed that the bias is the same for all agents, this would mean that we model participants who consider themselves to be at least as much of an expert, on average, as any other. Interpreting the biases in terms of mutual respect, on the other hand, does not presuppose the correctness, in any form, of the attitudes at hand. This interpretation is well suited in the present case because the participants exchange preferences, which have a natural reading in terms of comparative value judgments.

These two interpretations of the bias parameter, in terms of expertise and mutual respect, are broadly in line with what Dryzek and List (2003) call the "reflexive" aspect of rational deliberation:

> **(ref)** Deliberation "induces people to reflect on their preferences, in the knowledge that these preferences have to be justified to others." (9)

Indeed, participants in our model weigh their own preferences against those of others. The different values of the bias parameter can then be seen as the extent to which each participant sees her own view as justifiable to others. The higher the bias towards oneself, the more one sees oneself as an "expert", or the opinion of others as deserving less respect, in which case the need to justify one's view to others will decrease accordingly.

The minimization of any of the three distance measures that we have used, on the other hand, can be seen as embodying a rational preference change policy. At each step of the deliberation, each participant faces an aggregation problem: she must decide how to aggregate the ranking that has just been announced with her own —that is, how to take into account the opinion that has just been shared. It is well known that doing so by minimizing the distance between her profile and the one just announced satisfies all classical Arrowian postulates, except for the independence of irrelevant alternatives, irrespective of whether one uses KS, CS, or DP. The independence of irrelevant alternatives is somewhat controversial, however (see, e.g., Kemeny 1959), and the other Arrowian postulates appear plausible in the present context.

Comparing the three measures we have used, none appears to fare substantially better than the others with regard to internal coherence. Indeed, at the individual level, there is no substantial difference between the axiomatizations of each of the measures. Cook–Seiford differs from KS and DP in terms of the notion of "betweenness" that they use, and the latter differ from each other in terms of how they treat logical redundancies. This does lead to differences in some update scenarios, but not very significant ones.

This means that, to the extent that minimizing distance according to these measures embodies rational preference change policies, they are compatible with a further aspect of rational deliberation identified by Dryzek and List (2003)—namely, the "social" aspect of deliberation:

> **(soc)** deliberation "creates a situation of social interaction where people talk and listen to each other, enabling each person to recognize their interrelation in the group." (9)

The crucial part for us is of course the fact that participants in rational deliberation should *listen* to one another, which we take to imply that they should not only acknowledge the preferences of the other participants but also change their own to the extent that this is rationally required of them. The minimization of our three distance measures substantiates this rationality requirement by taking into account the bias parameter discussed above.

The two main parameters in our model can thus be given a normative interpretation. To put this again in Dryzek and List's (2003) terminology, the model goes beyond the "logical" fact that *if* our two main parameters are compatible with a normative interpretation, *then* our results can be seen as those that could be expected in a normatively ideal case. Rather, the argument we have just made shows that there are reasons that speak in favor of the truth of the antecedent of this

conditional and thus that the model can be interpreted normatively. Of course this normatively ideal interpretation is also compatible with less-than-ideal cases, for instance by interpreting the bias parameter in terms of cognitive inertia.

Modeling Choices and External Validity

Our model is thin in that it leaves out important aspects typically associated with rational deliberation, for instance the process of exchanging reasons for and against holding certain preferences and only expressing generalizable interests or opinions for the common good. For many deliberative theorists, starting in Habermas (1984), these are at the core of genuine group deliberation. From that point of view, one could argue that our model is one of the processes of social influence that may, or may not, accompany deliberation.

We do not see this as a shortcoming or limitation of our model. First, the preferences that the participants hold throughout could be interpreted as what they view as generalizable interests or the common good. Nothing in the model forces an egoistic or self-interested interpretation of the preference rankings. Second, the model is not inconsistent with thicker concepts of deliberation, such as those advocated by deliberative theorists. Indeed, we have just argued that our bias parameter and preference change policy can be seen as substantiating the reflexive and social aspects of rational deliberation. Given this, a plausible conjecture is that structured argumentative processes involving the exchange of reasons, as envisioned by deliberative theorists, would generate rational opinion changes in the sense studied here. In other words, one interpretation of the model is that the rational preference changes that it encodes are brought about by an underlying process of exchanging reasons, which is left implicit in the model. In any case, the model shows that, to the extent that (thin) deliberation can have positive effects on group preferences, it does so even without explicitly assuming the strong constraints that are the hallmarks of "true" deliberation. So the fact that our model is minimal strengthens the traditional arguments in favor of deliberative democratic procedures.

There is of course an important caveat to this: for midrange biases, we observe a new face of the "independence thesis" (Mayo-Wilson, Zollman, and Danks 2011)—that is, cases where individual and collective rationality diverge. Individually rational participants are led through deliberation into collectively irrational preferences. To our knowledge, this possibility has *not* been envisioned in the theory of deliberative democracy and is an interesting observation in itself. Going beyond this observation, however, it raises the question of whether thicker forms of deliberation could avoid this potential pitfall, or whether the normative constraints on thick deliberation could also rule out some bias values as less rational than others. We do not investigate these questions here, but we take it as a valuable contribution of the model that it brings them to the fore.

It may be argued, however, that the positive effects that we observe could be jeopardized by introducing the possibility of strategic interventions. Even if this turns out to be correct, this would not directly affect our claims, since it is unclear whether strategizing should be ruled out at the outset for fully rational deliberation, or whether the latter might at least deter it, as argued for instance in Dryzek and List (2003). But even bracketing this, one should keep in mind that strategizing would only affect *which* preferences are shared, not necessarily, and not even obviously, *how* these preferences are taken into account by others.[15] As long as the resulting opinion dynamic can be viewed as coming from distance minimization, our results show that irrational group preferences would still be eliminated by deliberation with low biases. In the presence of strategizing, the results of deliberation might be problematic for other reasons. It might, for instance, give an unfair advantage to some of the participants. But the resulting social preferences would still be rational under low biases.

All in all, our model can be viewed as making what philosophers of science call "Aristotelian" and "Galilean" idealizations (Frigg and Hartmann 2020) with respect to its target phenomenon, viz. rational deliberation. Aristotelian idealizations isolate the model from irrelevant features of the target phenomena, for instance the participants' hair color or their type of clothing. These are typically idealizations that would *not* improve the model if they were lifted. Both our thin model of deliberation and thicker accounts involve this kind of idealization. Galilean idealizations, on the other hand, explicitly distort the target phenomena, by simplification, typically for computational reasons. Lifting these types of idealizations would actually improve the accuracy of the model with respect to its target phenomena.

The model makes a number of Galilean idealizations. The restriction to deliberation with three alternatives and the fact that biases towards others are stable throughout deliberation and the same for all participants are clear examples of such idealization. More fundamentally, the absence of the explicit exchange of reasons also falls under that category. Note, however, that as far as the positive results of our simulations are concerned, this absence actually strengthens the point, as we have argued above. For the negative results—that is, the creation of irrational group preferences for midrange biases—however, the model could actually profit by making reasons explicit. The absence of strategic considerations is, on the other hand, less clear cut. For the empirical interpretation of the Received View, it would clearly count as a Galilean idealization. To the extent that the model is normative, however, it might also be seen as an Aristotelian one, depending on how normative theories treat strategic interventions.

[15] Of course, if some agents realize that others are systematically misrepresenting their views in order to achieve outcomes that they otherwise prefer, they might adjust their biases towards them. Again, we leave open the question of how such a dynamic in the biases would affect our results.

In general, neither type of idealization is viewed as a principled obstacle to learning about the target—see for instance Grune-Yanoff (2009) for models in economics—and this is the case here as well. Our model in particular provides a how-possible explanation for the claim that the reflexive and social aspects of deliberation can help to create single-plateaued profiles and by the same token can help to prevent irrational group preferences. At the same time, it points to the fact that the very same process that yields rational group preferences at low biases can actually backfire and move the group towards intransitive or even cyclic ones.

Since we are mainly addressing the normative reading of the Received View, the external validity of the model is a less pressing issue. In Appendix 2, we nonetheless present the results of running our model on some of the data that was presented in List et al. (2012). For specific bias values, the results approximate the empirical data and provide an additional interpretation of our bias parameter. Our results are thus complementary to the empirical ones: in both cases, we now have strong evidence for the claim that deliberation creates single-plateaued preferences, despite the strong structural constraints that come with this condition.

CONCLUSION

Both of the claims that constitute the Received View should be qualified. Recall that the Received View consists of (i) the meta-agreement hypothesis, which explains (ii) the fact that deliberation avoids irrational group preferences. Our results suggest either a strengthening of the meta-agreement hypothesis or an independent alternative to it. Indeed, for low biases, even under the unfavorable impartial culture assumption, the simulation results show that deliberation steers the group away from irrational group preferences by increasing proximity to single-*plateauedness*, as opposed to mere weakly single-peaked preferences. This, we argue, could be seen as providing support for a strengthened version of the meta-agreement hypothesis. This also suggests an alternative mechanism for the creation of single-plateau preferences, however, where the key elements are rational preference change and openness to changing one's mind upon hearing the opinions of others, leaving out meta-agreements. On the other hand, for higher biases, our results put into question the second part of the Received View—namely, the claim that rational deliberation helps us to avoid irrational group preferences. As we have seen, for sufficiently high bias values, deliberation actually tends to *create* irrational group preferences.

Avoiding irrational group preferences is, however, just one among many effects that deliberation can have on the participants' and the group's judgments. Our brief discussion of strategizing suggests that whether deliberation will help to promote democratic ideals depends on an intricate trade-off between certain positive effects like avoiding collective irrationality and tracking some procedure-independent standard (Estlund 1997; Perote-Peña and Piggins 2015), on the one hand, and known negative effects like strategizing, groupthink (Janis 1982), pluralistic ignorance (Prentice and Miller 1993), anchoring (Hartmann and Rad 2020), and polarization (Bramson et al. 2017; Hegselmann and Krause 2002) on the other. The latter is particularly interesting here because, as we have seen, all simulations where the agents change their minds in the course of the deliberation process result in clustering around a limited number of preference rankings. If this clustering constitutes a form of polarization, which remains to be seen, this would mean that the latter goes hand in hand with avoiding irrational group preferences.

Even if we restrict ourselves to the relationship between deliberation and rational group preferences, many questions remain. For one thing, for computational reasons, we have limited most of our analysis to sets of three alternatives and groups of at most 343 participants. Cases of more than three alternatives require further investigation. This would require general analytical results, which would also shed light on many of the observations we have made here. It would be interesting, for instance, to determine the robustness of the position of the "tipping point" in biases, after which deliberation ceases to eliminate irrational group preferences. The results provided here are nonetheless important because they help to disentangle the role of deliberation in forming meta-agreement and single-peaked preferences and to prevent cyclic or intransitive group rankings. They provide, we think, a first step towards a comprehensive analysis of the relationship between these three notions and opinion dynamics in group deliberation.

REFERENCES

Allport, Alan, and Glenn Wylie. 2000. "Task Switching, Stimulus Response Bindings, and Negative Priming." *Control of Cognitive Processes: Attention and Performance* 18: 35–70.

Arrow, Kenneth J. 1963. *Social Choice and Individual Values*. New Haven, CT: Yale University Press.

Black, Duncan. 1948. "On the Rationale of Group Decision-Making." *Journal of Political Economy* 56 (1): 23–34.

Bohman, James. 1997. *Deliberative Democracy: Essays on Reason and Politics*. Cambridge, MA: MIT press.

Bramson, Aaron, Patrick Grim, Daniel J. Singer, William J. Berger, Graham Sack, Steven Fisher, Carissa Flocken, and Bennett Holman. 2017. "Understanding Polarization: Meanings, Measures, and Model Evaluation." *Philosophy of Science* 84 (1): 115—59.

Cook, Wade D. 2006. "Distance-Based and Ad Hoc Consensus Models in Ordinal Preference Ranking." *European Journal of Operational Research* 172 (2): 369–85.

Cook, Wade D., and Lawrence M. Seiford. 1978. "Priority Ranking and Consensus Formation." *Management Science* 24 (16): 1721–32.

Dryzek, John S., and Christian List. 2003. "Social Choice Theory and Deliberative Democracy: A Reconciliation." *British Journal of Political Science* 33 (1): 1–28.

Duddy, Conal, and Ashley Piggins. 2012. "A Measure of Distance between Judgment Sets." *Social Choice and Welfare* 39 (4): 855–67.

Estlund, David. 1997. *Beyond Fairness and Deliberation: The Epistemic Dimension of Democratic Authority*. Cambridge, MA: MIT Press.

Farrar, Cynthia, James S. Fishkin, Donald P. Green, Christian List, Robert C. Luskin, and Elizabeth L. Paluck. 2010. "Disaggregating Deliberation's Effects: An Experiment within a Deliberative Poll." *British Journal of Political Science* 40 (2): 333–47.

Feld, Scott L., and Bernard Grofman. 1992. "Who's Afraid of the Big Bad Cycle? Evidence from 36 Elections." *Journal of Theoretical Politics* 4 (2): 231–7.

Fishburn, Peter C., and William V. Gehrlein. 1980. "The Paradox of Voting: Effects of Individual Indifference and Intransitivity." *Journal of Public Economics* 14 (1): 83–94.
Frigg, Roman, and Stephan Hartmann. 2020. "Models in Science." In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zelta. Stanford. https://plato.stanford.edu/archives/spr2020/entries/models-science/
Gaertner, Wulf. 2001. *Domain Conditions in Social Choice Theory*. Cambridge: Cambridge University Press.
Gehrlein, William V. 2004. "Consistency in Measures of Social Homogeneity: A Connection with Proximity to Single Peaked Preferences." *Quality and Quantity* 38 (2): 147–71.
Grüne-Yanoff, Till. 2009. "Learning from Minimal Economic Models." *Erkenntnis* 70 (1): 81–99.
Habermas, Jurgen. 1984. *The Theory of Communicative Action*. Boston: Beacon Press.
Hartmann, Stephan, and Soroush Rafiee Rad. 2020. "Anchoring in Deliberation." *Erkenntnis* 85: 1041–69.
Hegselmann, Rainer, and Ulrich Krause. 2002. "Opinion Dynamics and Bounded Confidence Models, Analysis, and Simulation." *Journal of Artificial Societies and Social Simulation* 5 (3).
Janis, Irving L. 1982. *Groupthink: Psychological Studies of Policy Decisions and Fiascos*. Boston: Houghton Mifflin Press.
Jones, Bradford, Benjamin Radcliff, Charles Taber, and Richard Timpone. 1995. "Condorcet Winners and the Paradox of Voting: Probability Calculations for Weak Preference Orders." *American Political Science Review* 89 (1): 137–44.
Kemeny, John. 1959. "Mathematics without Numbers." *Daedalus* 88 (4): 577–91.
Kemeny, John, and James L. Snell. 1962. "Preference Ranking: An Axiomatic Approach." In *Mathematical Models in the Social Sciences*, eds. John G. Kemeny, and Laurie Snell, 9–23. Cambridge, MA: MIT Press.
Landa, Dimitri, and Adam Meirowitz. 2009. "Game Theory, Information, and Deliberative Democracy." *American Journal of Political Science* 53 (2): 427–44.
List, C., and Robert E. Goodin. 2001. "Epistemic Democracy: Generalizing the Condorcet Jury Theorem." *Journal of Political Philosophy* 9 (3): 277–306.
List, Christian. 2002. "Two Concepts of Agreement." *The Good Society* 11 (1): 72–9.
List, Christian. 2011. "Group Communication and the Transformation of Judgments: An Impossibility Result." *Journal of Political Philosophy* 19 (1): 1–27.
List, Christian, Robert C. Luskin, James S. Fishkin, and Iain McLean. 2012. "Deliberation, Single-peakedness, and the Possibility of Meaningful Democracy: Evidence from Deliberative Polls." *The Journal of Politics* 75 (1): 80–95.
List, Christian. 2017. "Democratic Deliberation and Social Choice: A Review." In *The Oxford Handbook of Deliberative Democracy*, eds. Andre Bächtiger, John S. Dryzek, Jane Mansbridge, and Mark Warren, 463–89. Oxford: Oxford University Press.
Martini, Carlo, Jan Sprenger, and Mark Colyvan. 2013. "Resolving Disagreement through Mutual Respect." *Erkenntnis* 78 (4): 881–98.
Mayo-Wilson, Conor, Kevin. J. Zollman, and David Danks. 2011. "The Independence Thesis: When Individual and Social Epistemology Diverge." *Philosophy of Science* 78 (4): 653–77.
Miller, David. 1992. "Deliberative Democracy and Social Choice." *Political Studies* 40: 54–67.
Moulin, Hervé. 1984. "Generalized Condorcet-Winners for Single Peaked and Single-Plateau Preferences." *Social Choice and Welfare* 1 (2): 127–47.
Niemi, Richard G. 1969. "Majority Decision-Making with Partial Unidimensionality." *American Political Science Review* 63 (2): 488–97.
Ottonelli, Valeria, and Daniele Porello. 2013. "On the Elusive Notion of Meta-Agreement." *Politics, Philosophy & Economics* 12 (1): 68–92.
Perote-Peña, Juan, and Ashley Piggins. 2015. "A Model of Deliberative and Aggregative Democracy." *Economics & Philosophy* 31 (1): 93—121.
Prentice, Deborah A., and Dale T. Miller. 1993. "Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm." *Journal of Personality and Social Psychology* 64 (2): 243–56.
Puppe, Clemens. 2018. "The Single-Peaked Domain Revisited: A Simple Global Characterization." *Journal of Economic Theory* 176: 55–80.
Radcliff, Benjamin. 1994. "Collective Preferences in Presidential Elections." *Electoral Studies* 13 (1): 50—7.
Regenwetter, Michel, Bernard Grofman, Ilia Tsetlin, and A. A. J. Marley. 2006. *Behavioral Social Choice: Probabilistic Models, Statistical Inference, and Applications*. Cambridge: Cambridge University Press.
Riker, William H. 1982. *Liberalism against Populism*. San Francisco: W. H. Freeman.
Steiner, Jürg. 2012. *The Foundations of Deliberative Democracy: Empirical Research and Normative Implications*. Cambridge: Cambridge University Press.
Tsetlin, Ilia, Michel Regenwetter, and Bernard Grofman. 2003. "The Impartial Culture Maximizes the Probability of Majority Cycles." *Social Choice and Welfare* 21 (3): 387–98.

APPENDIX 1: TECHNICAL DEFINITIONS

Weak and Strict Preference Rankings: Let R_i be a preference pre-order or ranking for an agent i over a set of alternatives A. A preference profile R is a set of preference rankings, one for each agent i. When aR_ia', we say that i weakly prefers a over a'. When it is not the case that $a'R_ia$, then we say that i strictly prefers a over a', and we write $a>_ia'$. Indifference is defined as usual. We call a ranking R_i strict whenever for all a, a', either $a>_ia'$ or $a>_ia'$. Otherwise we call it weak.

Voting Cycles: Given a profile R of rankings, an alternative a is a Condorcet winner whenever it wins a majority on any pairwise comparison: for each alternative $a' \neq a$, we have it that a majority of agents strictly prefer a over a'. Pairwise majority yields cyclic social preferences, that is, a voting cycle, whenever there are three alternatives a_1, a_2, and a_3 such that there is a majority of agents who strictly prefer a_1 over a_2, a (possibly different) majority who strictly prefer a_2 over a_3, and a majority who strictly prefer a_3 over a_1.

Transitive Social Preference: A profile R of rankings is said to produce a transitive social preference, if the pairwise-majority-comparison relation, M_R, associated with it is transitive; that is, for all alternatives a, b, and c, if $a\,M_R\,b$ and $b\,M_R\,c$, then $a\,M_R\,c$.

Single-Peaked and Single-Plateaued Profiles: Let $\succ$ be a strict ordering of the alternatives in A. We say that a profile R is strictly single-peaked relative to $\succ$ whenever, for each agent i, there is an alternative a such that for all $a', a'', a''\succ a'\succcurlyeq a$ implies $a''<_ia'$ and $a\succcurlyeq a'\succ a''$ implies $a''<_ia'$. A profile R is single-plateaued relative to $\succ$ 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 aR_ib and cR_ia. A profile R is weakly single-peaked relative to $\succ$ whenever, for each agent i, there is an alternative a such that for all a', a'', $a''\succ a'\succcurlyeq a$ implies $a''R_ia'$ and $a\succcurlyeq a'\succ a''$ implies $a''R_ia'$. Strict single-peakedness implies single-plateauedness, which in turns implies weak single-peakedness, but none of these implications goes the other way around.

When R is single-peaked, weakly or strictly, or single-plateaued, relative to $\succ$ we say that the latter is a structuring dimension for R.

Arrowian Postulates: Let f be the aggregation function and R, R' profiles of individual preferences. Then f satisfies *rationality* if for all R, $f(R)$ is a complete pre-order. It satisfies *weak pareto* if $xf(R)y$ but not $yf(R)x$ whenever $x<_i y$ for every agent i. *Independence* means that for any two profiles R and R' and alternatives x and y, if for all individual i, xR_iy if and only if xR'_iy, then $xf(R)y$ if and only if $xf(R')y$. *Non-dictatorship* means that there is no i such that for all profiles $R, f(R) = R_i$.

Distance measures

The KS distance between rankings are defined as follows. First construct an *agenda* containing, for each pair of alternatives $a_i, a_j\, i \neq j$, the propositions (a_i, a_j) and $\neg(a_i, a_j)$. For a ranking r, define the *judgement set* J_r as follows: if a_i is weakly preferred to a_j according to r, then put $(a_i, a_j) \in J_r$; otherwise, put $\neg(a_i, a_j) \in J_r$. To illustrate this, consider the ranking $r_1 = (a_1, a_2, a_3)$ over three alternatives, meaning that a_1 is strictly preferred to a_2, which is in turn strictly preferred to a_3. The corresponding judgement set J_{r_1} is

$$J_{r_1} = \{(a_1, a_2), \neg(a_2, a_1), (a_1, a_3), \neg(a_3, a_1), (a_2, a_3), \neg(a_3, a_2)\}.$$

If we instead take $r_2 = (a_3, a_2, a_1)$, then J_{r_2} is:

$$J_{r_2} = \{\neg(a_1, a_2), (a_2, a_1), \neg(a_1, a_3), (a_3, a_1), \neg(a_2, a_3), (a_3, a_2)\}.$$

The KS distance between any two rankings r_1 and r_2 is defined as the Hamming distance between J_{r_1} and J_{r_2} — that is, the number of binary changes that one has to make to transform r_1 into r_2. In our example, the distance $d(r_1, r_2)$ between r_1 and r_2 is 6.

Kemeny and Snell (1962) characterize their measure uniquely by a set of intuitive axioms. The first axiom ensures that the measure is mathematically a distance measure. That is, for all rankings r_1, r_2, and r_3,

A1.1. $d(r_1, r_2) \geq 0$,

A1.2. $d(r_1, r_2) = d(r_2, r_1)$,

A1.3. $d(r_1, r_3) \geq d(r_1, r_2) + d(r_2, r_3)$, with equality holding if and only if r_2 is between r_1 and r_3.

The next axiom requires the distance to be invariant under the relabeling of the alternatives. This ensures that the distance depends not on the specific alternatives that we are ranking but only on the way they are ranked.

A2. If r'_1 and r'_2 result from applying the same permutation of objects to r_1 and r_2, then

$$d(r'_1, r'_2) = d(r_1, r_2).$$

Next, it is required that if two rankings agree at the beginning and/or at the end, then the distance should only depend on the middle segment where they differ.

A3. If two rankings r_1 and r_2 agree except for a set S of k elements, which is a segment in both r_1 and r_2, then $d(r_1, r_2)$ may be computed as if these k objects were the only objects being ranked.

Their last axiom sets a unit of measurement for their distance.

A4. The minimum positive distance is 1.

Crucial to this axiomatization is the notion of *betweenness.* The betweenness relationship for rankings in the KS axiomatization is defined in terms of the betweenness of the corresponding judgement sets. We say that the judgement set J is *between* judgement sets J_1 and J_2 if J_1, J_2, and J are distinct and, on each proposition, J agrees with J_1 or with J_2 (or both).

The DP distance measure uses the same representations for agendas and rankings. Now construct the graphs of all possible judgment sets, with an edge between J_1 and J_2 if and only if there are no judgement sets between them. Betweenness is defined in the same way as for the KS distance. The DP distance between r_1 and r_2 is the length of the shortest path between J_{r_1} and J_{r_2} on this graph. To look at an example again, start, as before, with $r_1 = (a_1, a_2, a_3)$ and $r_2 = (a_3, a_2, a_1)$ with the corresponding J_{r_1} and J_{r_2}. The shortest path between r_1 and r_2 is through the judgment sets

$$J_{r_1} \text{----} J_1 \text{----} J_2 \text{----} J_3 \text{----} J_{r_2},$$

defined as follows:

$$J_1 = \{(a_1, a_2), \neg(a_2, a_1), (a_1, a_3), \neg(a_3, a_1), (a_2, a_3), (a_3, a_2)\},$$

$$J_2 = \{(a_1, a_2), (a_2, a_1), (a_1, a_3), (a_3, a_1), (a_2, a_3), (a_3, a_2)\},$$

$$J_3 = \{(a_1, a_2), (a_2, a_1), \neg(a_1, a_3), (a_3, a_1), \neg(a_2, a_3), (a_3, a_2)\},$$

which gives a DS distance of 4 between r_1 and r_2.[16] Thus the DP distance between two rankings r_1 and r_2 also reflects the number of steps required to move from ranking r_1 to r_2. The difference between the KS and DP measures can be seen in what changes are permitted at each step. For the KS measure, only a single binary change is allowed at each step. Notice that this can result in an inconsistent judgment set, which then requires another binary change to become consistent. For example, consider the ranking $r = (a, \{b, c\})$, where alternatives b and c are ranked equally and strictly below the alternative a. The corresponding judgment set will then be

$$J_r = \{(a,b), \neg(b,a), (a,c), \neg(c,a), (b,c), (c,b)\}.$$

A binary change represented by replacing $\neg(b,a)$ with (b,a) will result in an inconsistent judgment set in which a is ranked equal to b and b is ranked equal to c, while a is ranked strictly higher than c. A second binary change represented by replacing $\neg(c,a)$ with (c,a) in the next step will make the judgment set consistent again. The DS measure will, under certain conditions, allow for multiple binary changes in a single step to avoid these inconsistent intermediate judgment sets. To be more specific, in DP we are allowed to change the relative ranking of an alternative with respect to a *set* of alternatives (as opposed to only one in KS), but only if the alternatives in the set are all ranked equally.

The CS distance is calculated by first assigning numbers (call them the *CS-numbers*) to the alternatives in a ranking, starting with 1 for the top alternative, 2 for the next-best alternative, etc. In case of a tie, we assign the average number to the tied alternatives. For instance, if there is one alternative at the top and two alternatives are tied just below, each of the latter gets the average of 2 and 3—that is, 2.5. Let x_i^r be the CS-number of alternative a_i in ranking r. The CS-distance between rankings r_1 and r_2 is the sum of the absolute differences between the CS-numbers of the alternatives:

$$d(r_1, r_2) = |x_1^{r_1} - x_1^{r_2}| + \cdots + |x_n^{r_1} - x_n^{r_2}|.$$

Starting with our two rankings $r_1 = (a_1, a_2, a_3)$ and $r_2 = (a_3, a_2, a_1)$ again, we get $x_1^{r_1} = 1$, $x_2^{r_1} = 2$, $x_3^{r_1} = 3$, and $x_1^{r_2} = 3$, $x_2^{r_2} = 2$, $x_3^{r_2} = 1$:

$$d(r_1, r_2) = |1-3| + |2-2| + |3-1| = 4.$$

The CS measure is characterized uniquely by the same set of axioms as the KS measure. The difference between them hinges on the different underlying notions of "betweenness" that they use.

The notion of betweenness for CS is not defined with respect to the corresponding judgment sets. Instead, it is given in terms of the CS-numbers assigned to the alternatives. In this case, we say that the ranking r is between rankings r_1 and r_2 if for each alternative, its CS-number in r is between its CS-numbers in r_1 and r_2, respectively. That is, for each alternative a_i,

$$x_i^{r_1} \leq x_i^r \leq x_i^{r_2} \quad \text{or} \quad x_i^{r_2} \leq x_i^r \leq x_i^{r_1}.$$

To see how this makes a difference, consider, for example, the profile of three rankings corresponding to the Condorcet Paradox:

$$r_1 = (a, b, c) \quad r_2 = (b, c, a) \quad r_3 = (c, a, b).$$

Here the intuitive result of aggregating this profile is for the group to settle on ranking the three alternatives equally. This option is excluded, however, by the KS measure. It is easy to check that the distance between the equal ranking and the given profile is higher than the distance between each of the rankings r_1, r_2, or r_3 to the profile. This is not the case for the CS measure, however. With this measure, the equal ranking would indeed be the one to minimize the distance. The reason is that for the KS measure, the equal ranking does not fall between any two of the rankings in the profile, while with the CS definition of betweenness it does.

APPENDIX 2: COMPARISON WITH EMPIRICAL RESULTS

We have run our model by taking as our input the data collected in two of the deliberative polls studied in List et al. (2012):[17] the British poll on the status of the monarchy and the Australian poll on the head of state. The data was first translated from SPSS into Java, in which our model is coded. We then used the resulting profile as an input of

[16] This example indirectly shows the kind of "double-counting," mentioned in Duddy and Piggins (2012), that results from taking the Hamming distance between J_{r_1} and J_{r_2}. From the fact that (a_1, a_2) and (a_2, a_3) and their negations are in J_{r_1} and J_{r_1}, respectively, we know by transitivity and completeness that (a_1, a_3) and its negation must be in J_{r_1} and J_{r_2}, respectively. The DP measure ignores this third step, to arrive at a distance of 4, while KS includes it, giving 6.

[17] We thank the authors of this paper for giving us access to their raw data.

1,000 simulations for each bias value between 0.51 and 0.99, by 0.2 increments. The results presented here average over those 1,000 simulations.

The results for the British deliberative poll are presented in Table 1. Our code calculated an initial proximity to single-plateauedness of 0.640 instead of 0.651, as reported in List et al. (2012). This is still within the estimated standard error (0.042). The paper reports a very slight decrease (–0.004) in proximity to single-plateauedness for this poll. This decrease, however, also falls within the estimated standard error. The authors explained this observation by the fact that the issue (the status of the monarchy) was already highly salient prior to the deliberative poll, leading to little to no change of opinion during the event itself.

Our simulations can approximate these results. For all three measures, we observe decreases in proximity to single-plateauedness. For DP, this happens for all bias values except very high ones, and for low and high biases these decreases approximate the empirical data. For KS we only observe decreases at very high values, but again they all approximate the empirical observations. The same holds for CS, but for a much larger bracket of bias values.

The situation is similar for the Australian deliberative poll. The empirical data reported a slight decrease in proximity to single-plateauedness, but again within the estimated standard error. Our model output, as before, decreases for DP that, however, mostly fall outside of that error range, except for biases of 0.85 and 0.87. The results are similar for KS and CS, except that the bias values that approximate the empirical data are higher.

This analysis suggests an alternative interpretation of the bias parameter in terms of the previous salience of the issue at hand, or a correlation between the two notions. To the extent that higher predeliberation salience translates to little to no change of opinion, this can be captured by higher bias values.

Note, furthermore, that the bias ranges that allow the model to approximate the empirical data are also those in which we observed a large number of irrational group preferences being created by deliberation. This observation is repeated here: even though the input profiles—the empirical data for the British and the Australian polls, did not yield irrational group preferences, a large number of output profiles did. Not all of them did, however, reflecting the fact that the resulting empirically observed profiles also did not induce irrational group preferences.

TABLE 1. Simulation Results for the British Deliberative Poll Data.

Bias	DP Proximity to SP	Diff DP	KS Proximity to SP	Diff KS	CS Proximity to SP	Diff CS
0.51	0.629	–0.010	0.985	0.346	0.972	0.333
0.53	0.614	–0.026	0.981	0.341	0.979	0.340
0.55	0.641	0.002	0.982	0.342	0.972	0.333
0.57	0.630	–0.009	0.987	0.347	0.942	0.302
0.59	0.640	0.000	0.983	0.344	0.952	0.312
0.61	0.657	0.018	0.982	0.343	0.941	0.301
0.63	0.591	–0.049	0.999	0.360	0.934	0.295
0.65	0.586	–0.054	0.996	0.357	0.914	0.274
0.67	0.587	–0.052	0.710	0.070	0.619	–0.021
0.69	0.577	–0.063	0.711	0.072	0.626	–0.014
0.71	0.558	–0.081	0.710	0.070	0.620	–0.019
0.73	0.575	–0.065	0.719	0.080	0.623	–0.017
0.75	0.531	–0.109	0.775	0.136	0.617	–0.023
0.77	0.533	–0.106	0.775	0.136	0.615	–0.024
0.79	0.542	–0.098	0.775	0.136	0.618	–0.022
0.81	0.537	–0.102	0.775	0.136	0.615	–0.025
0.83	0.535	–0.105	0.775	0.136	0.614	–0.026
0.85	0.609	–0.031	0.764	0.124	0.625	–0.015
0.87	0.610	–0.030	0.761	0.122	0.626	–0.014
0.89	0.640	0.000	0.638	–0.002	0.640	0.000
0.91	0.640	0.000	0.645	0.006	0.640	0.000
0.93	0.640	0.000	0.679	0.040	0.640	0.000
0.95	0.640	0.000	0.639	–0.001	0.640	0.000
0.97	0.640	0.000	0.640	0.000	0.640	0.000
0.99	0.640	0.000	0.640	0.000	0.640	0.000

Note Initial proximity to single-plateaued preferences is 0.64. The cells in green are those that fall within the estimated error range of the empirical data.

TABLE 2. Simulation Results for the Australian Deliberative Poll Data.

Bias	DP Proximity to SP	Diff DP	KS Proximity to SP	Diff DP	CS Proximity to SP	Diff CS
0.51	0.685	−0.143	0.998	0.170	0.994	0.166
0.53	0.703	−0.125	0.999	0.171	0.995	0.167
0.55	0.714	−0.114	0.999	0.171	0.992	0.164
0.57	0.711	−0.117	0.999	0.171	0.987	0.159
0.59	0.693	−0.135	0.998	0.170	0.984	0.156
0.61	0.691	−0.137	0.999	0.171	0.990	0.162
0.63	0.568	−0.260	0.998	0.170	0.966	0.138
0.65	0.581	−0.247	0.998	0.170	0.937	0.109
0.67	0.592	−0.236	0.905	0.077	0.829	0.001
0.69	0.577	−0.251	0.904	0.076	0.820	−0.008
0.71	0.584	−0.244	0.911	0.083	0.815	−0.013
0.73	0.552	−0.276	0.917	0.089	0.827	−0.001
0.75	0.616	−0.212	0.965	0.137	0.819	−0.009
0.77	0.627	−0.201	0.965	0.137	0.829	0.001
0.79	0.622	−0.206	0.965	0.137	0.818	−0.010
0.81	0.622	−0.206	0.965	0.137	0.824	−0.004
0.83	0.623	−0.205	0.965	0.137	0.823	−0.005
0.85	0.745	−0.083	0.948	0.120	0.800	−0.028
0.87	0.748	−0.080	0.951	0.123	0.801	−0.027
0.89	0.828	0.000	0.796	−0.032	0.828	0.000
0.91	0.828	0.000	0.776	−0.052	0.828	0.000
0.93	0.828	0.000	0.816	−0.012	0.828	0.000
0.95	0.828	0.000	0.779	−0.049	0.828	0.000
0.97	0.828	0.000	0.828	0.000	0.828	0.000
0.99	0.828	0.000	0.828	0.000	0.828	0.000

Note: Initial proximity to single-plateaued preferences is 0.828. The cells in green are those that fall within the estimated error range of the empirical data.