Erkenntnis
https://doi.org/10.1007/s10670-024-00913-5

ORIGINAL RESEARCH

# Ill-informed Consensus or Truthful Disagreement? How Argumentation Styles and Preference Perceptions Affect Deliberation Outcomes in Groups with Conflicting Stakes

Jonas Stein[1] · Jan-Willem Romeijn[2] · Michael Mäs[3]

Received: 18 April 2024 / Accepted: 26 November 2024

## Abstract

In groups where members deliberate with limited information, consensus can emerge where, under complete information, fundamental disagreement would prevail. Using an agent-based model, we explore the factors contributing to group consensus by comparing argumentation styles in two types of groups: agents in groups of *advocates* communicate arguments for options perceived as personally beneficial. Agents in groups of *diplomats* do the same but avoid disagreement in that they bring up arguments supporting a second-best option whenever their interaction partner perceives to benefit the least from what the sender finds best. Results show that consensus depends on argumentation style, but also on what members initially perceive as preferred. Diplomats are more likely to form consensus when initial perceptions accurately align with full information preferences, which diverge within the group. Conversely, and perhaps counterintuitively, in the presence of inaccurate initial perceptions, groups of advocates converge while diplomats part in disagreement. Our results imply that the ideal argumentation style must be considered carefully in light of both the desired outcome and the initial information distribution: when conflicting stakes produce a trade-off between consensus and truthful perceptions, polite versus selfish ways of deliberation may produce one or the other outcome, depending on the initial information members are equipped with.

**Keywords** Collective deliberation · Conflicting stakes · Group decision-making · Agent-based modeling

✉ Jonas Stein
j.d.stein@rug.nl

[1] Department of Sociology, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Rozenstraat 31, 9712 TG Groningen, The Netherlands

[2] Faculty of Philosophy, University of Groningen, Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands

[3] Department of Sociology, Karlsruhe Institute of Technology, Douglasstrasse 24, 76133 Karlsruhe, Germany

Published online: 27 January 2025

## 1 Introduction

Since James March's seminal work on 'Exploration and exploitation in organizational learning' (1991), much research has investigated the notion that optimal group performance requires a delicate balance between individual's efforts to seek new solutions and their ability to adopt existing approaches (Bernstein et al., 2018; Lazer & Friedman, 2007; Levinthal, 1997; Mason & Watts, 2012). Across the contexts studied, a common denominator is that individuals do not exhibit conflicting stakes, in that the benefit of a solution is the same to everyone. Disagreement among group members arises from heterogenous access to information, but not from group members evaluating the same information differently.

Yet, in many groups, different solutions have inherently different qualities to individual group members. Here, disagreement about a 'best' solution may prevail even when everyone is faced with the same information: For example, members of a hiring committee may prefer different qualities in candidates, and therefore disagree on the choice of the right applicant. In the same vein, organizational board members can have strategic motives to weigh executive decisions differently, or the assessment of a political decision may depend heavily on what stakeholder is involved in it.

In such situations, decisions are often obtained through voting. However, unless the decision is unanimous, voting requires that everyone accepts the decision irrespective of their own preferences. Alternatively, some or all actors must compromise or bargain, knowingly disregarding what they prefer the most in exchange for an alternative choice. Such compromises can be difficult to obtain as they depend on actors' willingness to sacrifice and may come along with prolonged negotiations, tension, and conflict (Priem et al., 1995).

Here, we point to another possibility: ill-informed consensus. In the process of deliberating arguments pro and con alternative decision options, there may be a phase when all group members agree on the same option, simply because they are not aware of all existing information. If the group makes the decision at this moment, it would be based on a consensus despite diverging preferences. This approach is different from compromising because group members do not settle for an option they deem suboptimal. Instead, every group member individually comes to the—warranted yet faulty—conclusion that a given decision option is best for themselves, given the information they currently possess.

We propose an agent-based model to investigate such an emergence of consensus despite diverging preferences. In the model, agents communicate arguments to form and adapt perceptions of their preferences over three decision options. These agents are categorized into two subgroups, each with distinct stakes that, under full information, lead them to prefer different options. Preferences are modeled as a zero-sum situation: the more one subgroup gains from an option, the less does the other and hence, the greater the divergence of preferences. A third option represents middle ground in the sense that it is a second-best choice with equal value to everyone. A group reaches consensus when deliberation has led everyone to perceive that they prefer the same option. Notably, since truthful (full information) preferences diverge, consensus can only be present when at least some group members have inaccurate perceptions.

Our paper is structured as follows. In the next section we motivate our modeling approach and spell out our specific goals with it. The model itself is presented in Sect. 3, and in Sect. 4 we specify the simulations that we carry out on this model. Section 5 contains the results of the simulations, and these are further discussed and evaluated in Sect. 6. Finally, Sect. 7 summarizes the main conclusions.

## 2 Motivating the Model

The main concern of this study is to investigate how argumentation styles affect decision-making in settings with conflicting stakes and incomplete information. In addressing this question, we do not aim to determine how, in actual fact, argumentation styles impact on decision-making, by making the models maximally descriptively adequate. Instead, we offer possible scenarios and deliberative mechanisms for how such argumentation styles might impact on decision making under the given circumstances, with the aim of drawing tentative normative conclusions about them. In line with simulation studies elsewhere in philosophy, we intend to explore the conditions under which social deliberation is beneficial or detrimental to collective opinion formation, focusing on distinctions between conditions that are both empirically and theoretically salient.

To be sure, our model is certainly not free from empirical constraints. To the contrary, and much in line with the sociological literature that our research is embedded in, our modeling choices are underpinned and motivated by empirical studies. In fact, we believe our model fares relatively well in approximating the empirical facts of social deliberation on a number of relevant aspects. As evidenced by the references below, the argumentation styles that we distinguish are ideal–typical simplifications of how people have been observed to deliberate in sociological experiments and field studies. Moreover, as argued in the next section, the agents in our model advance arguments in a deliberation according to procedures that resemble behaviors observed among actual deliberators. As said, we do not strive for the full descriptive adequacy of our models. But since we want to use the models for exploring possible deliberative mechanisms and, ultimately, for drawing tentative normative conclusions about forms of social deliberation, we need to ensure the approximate descriptive adequacy of our models in particular respects.

In what follows, we will review a number of key modeling choices and provide further motivations for them, mixing theoretical and empirical considerations. Further empirical motivations can be found in the next section, in which the model specifications are reviewed more elaborately. Towards the end, the current section also briefly discusses our models in view of a broader philosophical literature.

When asking how argumentation styles affect decision-making in settings with conflicting stakes and incomplete information, what styles are we taking into consideration? Considering an exhaustive list of argumentation styles hardly seems possible and at any rate exceeds the scope of a single study. Instead, the model is inspired by empirical and theoretical research on human behavior in deliberative settings

(Cialdini & Goldstein, 2004; Deffuant et al., 2000; Mercier & Sperber, 2011; Wittenbaum et al., 2004) and sets out to compare populations of agents using either of two ideal–typical argumentation styles. *Advocates* represent individuals who communicate arguments supporting what they currently prefer, given the information they possess. In this sense, advocates represent individuals who raise information that is consistent with their own beliefs, without taking into account any characteristics of the agent they are talking to. This specific type of agent is inspired by empirical research on discussion settings (Mercier & Sperber, 2011; Stasser & Titus, 2003; Wittenbaum et al., 2004), suggesting that individuals are usually inclined to raise arguments in favor of their own opinion. Theoretical models on social influence (Flache et al., 2017; Hegselmann & Krause, 2002) and collective deliberation (Madsen et al., 2018; Olsson, 2013) usually assume similar behavior according to which individuals freely bring forth arguments supporting what they find best.

Intuitively, one would expect that groups of advocates rarely end up with consensus when conflicting stakes are present. Pushing for what one finds selfishly beneficial when a conversation partner has little to gain seems an unlikely way to convince them. Communicating arguments with others who share the same interests, on the other hand, will amplify latent preferences: new arguments will fall on fertile ground and strengthen their belief in that option. As both types of interactions are repeated many times within the group, patterns should emerge where members with identical stakes become more similar in their convictions but fail to agree on a decision with those opposed to it.

Despite the theoretical (Flache et al., 2017) and empirical (Mercier & Sperber, 2011) basis motivating our choice of advocate-type agents, many situations can be thought of in which individuals deviate from arguing for what they find best. Social conformity, for example, is a strong force in human behavior (Asch, 1956; Cialdini & Goldstein, 2004) and may prompt individuals to steer clear of disagreements with their conversation partners. Discussions in 'Hidden Profile' settings (Stasser & Titus, 1985) show that group members often fail to raise dissenting information because they underestimate its significance (Lu et al., 2012; Wittenbaum et al., 2004). Drawing on Social Judgement Theory (Sherif & Hovland, 1961), literature on 'bounded confidence' (Deffuant et al., 2000; Hegselmann & Krause, 2002) argues that individuals will reject or ignore information that is too different from their own convictions. Prominent works on repulsive influence (Baldassarri & Bearman, 2007; Flache & Macy, 2011; Mark, 2003) assume that trying to influence someone with information perceived as too dissonant may provoke even greater opposition in them. In all of these cases, strategic considerations would lead individuals to express convictions more similar to their conversation partner than they actually are, either out of fear of being sanctioned, or because a more honest expression would get rejected right away.

For these reasons, we introduce a second argumentation style: *Diplomats* aim to convince others of the option they currently prefer as well but are cautious to not offend their conversation partner. Such agents arguably follow an ideal of reasonable discussion that has deep roots in pragmatist philosophy: they are guided by a communicative rather than a strictly instrumental rationality in their social deliberations (Habermas, 1985a, b) and occupy a shared space of reasons (Brandom, 1994). In our simulations, as further discussed below, we attempt to capture these ideas on reasonable debate in a specific deliberative format.

To be clear, philosophers like Habermas, Rawls and Brandom have argued for an ideal of reasonable discussion primarily because of its potential to overcome diverging preferences. But the same reasonableness may also prove its worth when preferences are irreconcilable. In particular, we might expect diplomats to reach a state of ill-informed consensus more often than advocates. This is so because diplomats will raise arguments supporting a second-best option instead of trying to convince their conversation partner of an option they prefer the least. On a group level, this should enhance the circulation of arguments in favor of an option both subgroups find neither worst nor best. In consequence, states may emerge where so many arguments in favor of what is in fact the second-best option have been shared that everyone ends up convinced that this is the most beneficial option.

Further developments in the philosophical understanding of social deliberation, both in social epistemology and in political theory, lead us to question this intuitive connection between an ideal of reasonable debate on the one hand and the feasibility of consensus on the other. In the social epistemology of science, researchers have discovered that a certain degree of fragmentation and dissensus, fueled by constraints on information sharing, may be beneficial to the results of social deliberation (Zollman, 2010). In political theory, the broadly Habermasian ideal of reasonable debate has been challenged by so-called agonistic pluralism (Mouffe, 1999), i.e., the view that a focus on reasonable debate cannot properly accommodate the depth of disagreement between deliberators and in fact hampers political representation. In short, both formal social epistemology and activist political theory have offered arguments towards the overall idea that social deliberation benefits from vocal participants that sustain and act out dissensus.

While these arguments against reasonable diplomats and in favor of vocal advocates pertain in first instance to settings in which there is, at the level of the collective, a shared goal, like finding a correct scientific theory or achieving adequate political representation in a pluralist society. They are not principally geared towards the phenomenon of ill-informed consensus. Nevertheless, it is conceivable that the beneficial effects of dissensus and advocacy once again carry over to settings in which preferences are irreconcilable and in which consensus can only be reached through faulty preference perceptions. It seems clear that if the positions of the deliberators are fundamentally at loggerheads and entirely transparent to themselves, consensus formation is impossible. However, if deliberators have incomplete information about what serves their interest best, then following a communication strategy that embraces rather than eschews conflict may offer advantages.

All in all, we conclude that the debate over social deliberation offers arguments pulling in opposite directions. Speaking generally and acknowledging that there will be other ways to partition the debate into camps, we can place a Habermas-inspired ideal of reasonable debate, represented by diplomatic interlocutors, against a Mouffe-inspired ideal of agonistic debate, represented by vocal advocates. Both camps have arguments going for them and this leads us to ask: which of these strategies is more conducive to achieving an outcome of truthful disagreement or ill-informed consensus?

In what follows we study such questions by means of simulation studies. We describe how argumentation styles influence groups to either arrive at ill-informed consensus or part ways in truthful disagreement, and what factors play a role in the deliberation dynamics leading up to it. Because plausible settings exist where one or the other outcome is more advantageous, we refrain from interpreting simulation results normatively. For example, issues such as impending health crises or natural catastrophes can be pressing enough that taking any kind of decision—even if partially based on inaccurate knowledge—is preferable over further disagreement. Conversely, 'agreeing to disagree' may be preferable in simple transactional situations when disagreeing group members can find more fitting decision partners elsewhere, and when failure to obtain consensus bears little consequences in the first place. For this reason, lessons from our simulations must be assessed in the light of the specific contexts in which decisions are made.

To this day, to our knowledge at least, the model we built is the first to study collective deliberation in a context where group members' individual preferences diverge. There are of course game-theoretic models and simulations in which agents are self-interested. But our focus is not on agents coordinating their actions to each other, but rather on agents deliberating, i.e., exchanging arguments and investigating the possibility of consensus. Most prominent models of normative deliberation (Hegselmann & Krause, 2006; Mason & Watts, 2012; Olsson, 2013; Zollman, 2010) assume that the best option is the same to everyone, and that full information will lead to natural convergence around this option. In our model, consensus only emerges when at least some group members have faulty perceptions about their preference. Full information, on the other hand, prompts agents to discern their individual stakes and preferences, and then disagreement is the natural outcome. In light of the many plausible contexts where insurmountable disagreement is the most truthful conclusion, we deem the approach taken here an important yet understudied setting.

In addition, our model is among the few to study how argumentation styles affect consensus-making in groups (see van Veen et al., 2020 for a related study). Simulation research has investigated how network structure (Bernstein et al., 2018; Lazer & Friedman, 2007; Mason & Watts, 2012; Shore et al., 2015), homophilous interaction preferences (Stein et al., 2024) and cognitive characteristics of the recipient of an information (Madsen et al., 2018; Zollman, 2010) shape collective deliberation. However, less is known about the possible group-level consequences on how people choose information they disclose—although it appears that the latter should greatly affect the former. By comparing groups of agents who advocate for what they think is best with agents who avoid harsh disagreement, our study takes a first step in this direction. The choice of advocates and diplomats can also be seen as a comparison of direct versus indirect communication styles across organizational or cultural contexts (Hall, 1976). Of course, the two argumentation styles we outline do not remotely capture all ways according to which humans reason with each other. Instead, the focus of this study is to compare the status quo in how most deliberation models assume actors to argue (Hegselmann & Krause, 2006; Madsen et al., 2018; Olsson, 2013; Zollman, 2010) with a more nuanced, and perhaps more realistic argumentation style (Hahn & Harris, 2014). As mentioned, competing theoretical expectations about whether diplomats or advocates form consensus more often exist, underlining the importance of simulating discussion outcomes in groups with different argumentation styles.

## 3 Model Description

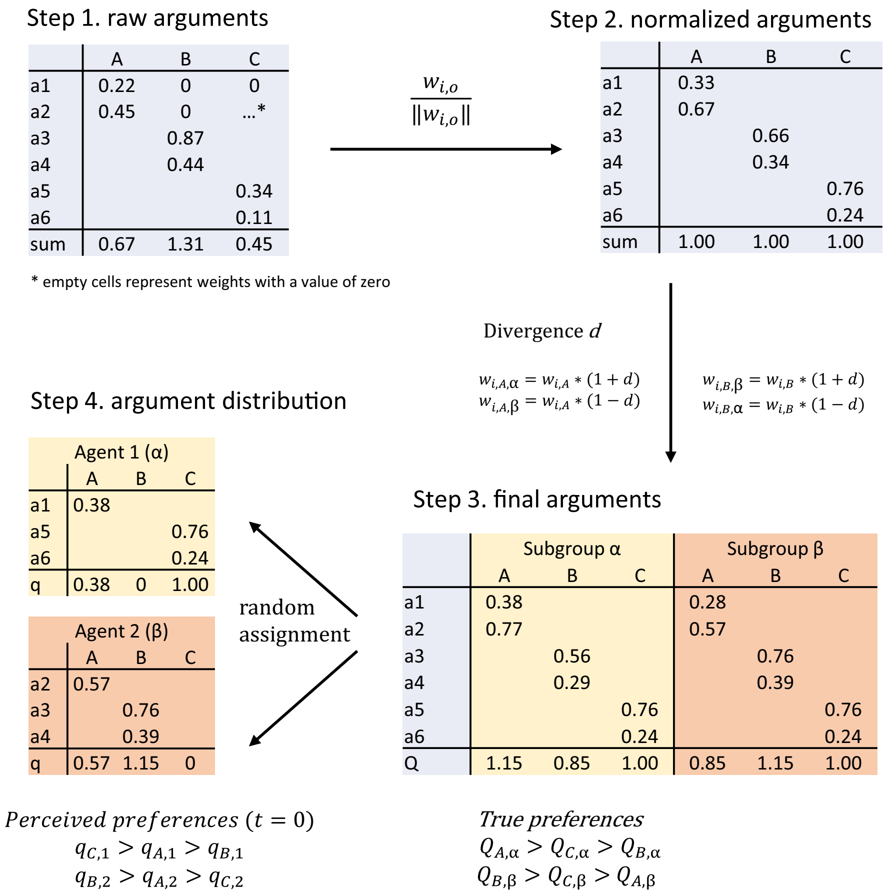
To represent a discussion setting where group members face diverging preferences, we assume a group of $N$ agents consisting of two subgroups $\{\alpha, \beta\}$. Every agent $i$ affiliates with one of the subgroups $g$. Agents are endowed with arguments that support either option $o$ out of three decision options $\{A, B, C\}$. Taken together, all arguments pertaining to a decision option reveal the *true preferences* of a given subgroup of agents for that particular option. Arguments contain subgroup-specific weights so that the true preferences for a given option of one subgroup are different from the true preferences of the other. The creation of options, arguments and argument weights is outlined in the description below. Subgroups are assumed to be of equal size.

During group deliberation, agents operate under limited information and form *preference perceptions* based on the arguments they currently possess. At each round $t$ of the simulation, one agent attempts to influence another agent by sharing an argument. The receiving agent then integrates the argument into their argument set and updates their preference perceptions.

The discussion ends when the group forms consensus, i.e., all agents have identical perceptions in terms of which option they prefer the most. Note that the model is set up such that consensus is only possible before all agents possess all arguments. Under full information, agents' perceptions equal their true preferences, which are different for the two subgroups. Unless consensus is obtained, we stop the simulation when enough arguments have spread so that agents' perceptions approximate their true preferences, and no argument combination they receive would be capable of changing their conviction.

### 3.1 Initialization

Prior to the deliberation process, we assume a fixed set of three decision options $O = \{A, B, C\}$ and $I$ available arguments $A = \{a_1, a_2, \ldots, a_I\}$. Fixed sets exclude the possibility that agents redefine options or create new ones, which would be hard to design, track and explain in a simulation. Each argument $a_i$ contains weights for each decision option, i.e. $a_i = \{w_{i,A}, w_{i,B}, w_{i,C}\}$ representing information about the benefits of the different options. Following related models on consensus-making and argumentation in groups (Stein et al., 2024; van Veen et al., 2020), we assume that each argument cannot support more than one option simultaneously. That is, we assign each argument a positive weight to only one of the options and a weight of zero for the other options. We further assume that each decision option has an equal number of arguments with a positive weight pointing towards them. For a given simulated group, weights are first randomly drawn from a uniform distribution so that



Step 1. raw arguments

| | A | B | C |
|---|---|---|---|
| a1 | 0.22 | 0 | 0 |
| a2 | 0.45 | 0 | …* |
| a3 | | 0.87 | |
| a4 | | 0.44 | |
| a5 | | | 0.34 |
| a6 | | | 0.11 |
| sum | 0.67 | 1.31 | 0.45 |

* empty cells represent weights with a value of zero

$$\xrightarrow{\frac{w_{i,o}}{\|w_{i,o}\|}}$$

Step 2. normalized arguments

| | A | B | C |
|---|---|---|---|
| a1 | 0.33 | | |
| a2 | 0.67 | | |
| a3 | | 0.66 | |
| a4 | | 0.34 | |
| a5 | | | 0.76 |
| a6 | | | 0.24 |
| sum | 1.00 | 1.00 | 1.00 |

Divergence $d$

$w_{i,A,\alpha} = w_{i,A} * (1+d)$
$w_{i,A,\beta} = w_{i,A} * (1-d)$

$w_{i,B,\beta} = w_{i,B} * (1+d)$
$w_{i,B,\alpha} = w_{i,B} * (1-d)$

Step 3. final arguments

| | Subgroup α | | | Subgroup β | | |
|---|---|---|---|---|---|---|
| | A | B | C | A | B | C |
| a1 | 0.38 | | | 0.28 | | |
| a2 | 0.77 | | | 0.57 | | |
| a3 | | 0.56 | | | 0.76 | |
| a4 | | 0.29 | | | 0.39 | |
| a5 | | | 0.76 | | | 0.76 |
| a6 | | | 0.24 | | | 0.24 |
| Q | 1.15 | 0.85 | 1.00 | 0.85 | 1.15 | 1.00 |

random assignment

Step 4. argument distribution

| | Agent 1 (α) | | |
|---|---|---|---|
| | A | B | C |
| a1 | 0.38 | | |
| a5 | | | 0.76 |
| a6 | | | 0.24 |
| q | 0.38 | 0 | 1.00 |

| | Agent 2 (β) | | |
|---|---|---|---|
| | A | B | C |
| a2 | 0.57 | | |
| a3 | | 0.76 | |
| a4 | | 0.39 | |
| q | 0.57 | 1.15 | 0 |

*Perceived preferences* $(t = 0)$
$q_{C,1} > q_{A,1} > q_{B,1}$
$q_{B,2} > q_{A,2} > q_{C,2}$

*True preferences*
$Q_{A,\alpha} > Q_{C,\alpha} > Q_{B,\alpha}$
$Q_{B,\beta} > Q_{C,\beta} > Q_{A,\beta}$

**Fig. 1** Creation and distribution of arguments, argument weights and preferences

$w_{i,o} \sim U\{0,1\}$. Subsequently, weights are normalized ($w_{i,o}$ / $\| w_{i,o} \|$), such that each option's sum of weights equals that of another (Fig. 1, Step 1 and 2).

In Step 3 of the argument creation process, we manipulate all arguments pertaining to option *A* and *B* such that their weights are different for the two subgroups. In doing so, we capture the aspect that options have divergent benefits for members of different subgroups, and that this is reflected in the weights of the arguments pertaining to them. We introduce a divergence parameter $d \in [0,1]$. All weights pertaining to option *A* are multiplied by $1 + d$ for α agents and by $1 - d$ for β agents, and all weights pertaining to option *B* are multiplied by $1 + d$ for β agents and by $1 - d$ for α agents. Weights pertaining to option *C* remain unchanged, i.e. they have the same value for members of either subgroup.

For an agent of a given subgroup, the sum of weights associated with a decision option represent their *true preference*: $Q_{o,g} = \sum_{i=1}^{I} w_{i,o,g}$. Summation reveals that for any value of divergence $0 < d < 1$, preferences of α members will correspond to $A > C > B$, while preferences of β members correspond to $B > C > A$. True preferences thus crucially capture the theoretical scope of the study, namely, a divergence of preferences among group members under full information, with a higher value of $d$ implying higher divergence. Note that although every individual agent has a strict preference ranking, the incomparability of interpersonal utility (Hausman, 1995; Robbins, 1938) makes it problematic to rank options on an aggregate (group) level. We therefore do not make assumptions about the collective benefits (or 'optimality') of either of the decision options.

The last step of the initialization procedure is to assign arguments to group members (Step 4 in Fig. 1). Here, assuming systematic biases in agents' information acquisition prior to discussion would make it possible to assign arguments such that they mainly correspond to agents' true preferences. However, since we do not have a particular theoretical or empirical motivation that would lead us to assume such assignment, we let agents take turns at randomly drawing from the total set of arguments, one at a time without replacement, until all arguments have been assigned. The Appendix includes additional analyses assuming lopsided initial argument distributions. Note that because argument weights for agents of different subgroups diverge, assigning arguments at random still generates initial preference perceptions that correlate with agents' true preferences. How agents form perceptions is outlined below.

### 3.2 Argument Processing and Communication

Similar to how a true preference $Q_{o,g}$ is computed, an agent forms a *preference perception* $q_{o,x,t}$ for each option $o$ by summing over the weights of the arguments they possess at round $t$. This allows agents to rank options from being perceived as most to least preferred. Preference perceptions are based on incomplete information and do not necessarily overlap with the true preferences of a group member. Because preference perceptions of an agent can shift over time, they are denoted with a subscript $t$. Subscript $x$ denotes the individual agent.

Over the course of the simulation, agents communicate arguments, influencing other group members with the arguments they share. Agents who receive arguments integrate them and update their perceptions, using the weights that correspond to their subgroup.[1] Each round $t$ of the simulation consists of the following steps:

1. A *sending agent* and a *receiving agent* are activated.
2. The sending agent selects an argument to share with the receiving agent.

[1] Take the following example as a short illustration of this process: if Agent 1 from Fig. 1 shared argument *a1* with Agent 2, Agent 2 would add a value of 0.28 to their perceived preference for option $A$, even though the same argument has a stronger weight (0.38) to Agent 1. The perceived preference for option $A$ of Agent 2 would thus be $0.57 + 0.28 = 0.85$.

3. The receiving agent integrates the argument and updates their perceptions.

We compare groups in which the sending agent selects an argument according to either of two argumentation styles: Agents who are *advocates* select arguments that tend to strongly support the option they perceive to prefer the most, regardless of the perceptions of the receiving agent. *Diplomats* do the same but avoid selecting an argument that supports the option their receiver perceives to prefer the least. Diplomats thus only differ from advocates when preference perceptions between agents are opposed, and otherwise behave in an identical fashion.

Similar to many empirical settings, both diplomats and advocates are unaware of the exact set of arguments of other agents, making it possible that agents communicate an argument that others are already aware of. However, we do assume that group members are aware of each others' perceived preferences. This reflects the notion that real-world decision-makers often do have an intuition of each other's positions but that underlying arguments remain private information. Awareness of perceptions enables diplomatic agents to know what options to avoid during interaction, and lets agents realize when consensus is present.

Following standard procedure of canonical social influence models (Deffuant et al., 2000; Flache & Macy, 2011; Keijzer et al., 2018), we randomly select a sending and a receiving agent at each round of the simulation. The sending agent selects an argument according to a two-step softmax function (Daw et al., 2006). Softmax functions are commonly used for modeling human decision-making across a range of fields (Guo & Yu, 2019; Harlé et al., 2015; Sutton & Barto, 2018),[2] are capable of predicting observed decision-making in experimental tasks (Daw et al., 2006; Witt et al., 2024; Wu et al., 2024), and have properties that make them plausible approximations of human choices (Reverdy & Leonard, 2015): a softmax function assigns the highest choice probability to the option with the highest reward, choice probabilities are sensitive to distances between options, and choices are influenced by an adjustable degree of random deviation and noise. Similar to related collective deliberation models (Stein et al., 2024; van Veen et al., 2020), we implement our softmax procedure as follows. In the first step of the procedure, the agent chooses an option $o^*$ they want to support. If the agent is an *advocate*, they consider their preference perceptions and choose an option $o$ according to the probability

$$p(o) = exp(\tau * q_{o,x,t}) / \sum_{o \in \{A,B,C\}} exp(\tau * q_{o,x,t}) \tag{1}$$

where the slope parameter $\tau$ controls the degree of adherence (as opposed to randomness) in agents' decisions. The higher $\tau$, the more their choice is determined by selecting the option they perceive to prefer the most.

*Diplomats* select an option in a very similar manner, but with one important difference: they exclude what their receiver perceives to prefer the least from the set of

[2] Although the name varies across fields. 'Softmax' is used within biology, cognitive and neuroscience studies, economics, sociology and consumer studies use mathematically equivalent 'discrete choice' functions (Blume et al., 2011; Greene, 2009).

options considered by Eq. 1. Thus, when a diplomat considers which option to argue for, they act as if the receiver's least preferred option did not exist and choose from the remaining options instead. In consequence, whenever the option a diplomatic sender perceives to prefer the most is simultaneously the option that the receiver perceives to prefer the least, she is most likely to choose the option corresponding to her perceived second preference instead.

The second step of the discrete choice procedure concerns the selection of the argument to be shared. This step is the same for advocates and diplomats alike. Here, an agent regards the set of arguments $A_{x,t}$ they currently hold and considers those argument weights $w_{i,o^*,g}$ that correspond to their chosen option $o^*$. They pick one of their arguments with the probability given by

$$p(a_i) = exp(\tau * w_{i,o^*,g}) / \sum_{i \in A_{x,t}} exp(\tau * w_{i,o^*,g}) \tag{2}$$

Again, the parameter $\tau$ determines agents' adherence to choosing stronger versus weaker arguments pertaining to her chosen option. We assume that the value of $\tau$ in Eqs. 1 and 2 is the same across simulations. By default, we set the adherence parameter to $\tau = 2$ such that agents make choices neither deterministically nor randomly, but according to probabilities that lie somewhere in between these two extremes. This behavioral assumption is consistent with empirically validated models employing similar choice functions (Daw et al., 2006) and coherent with our understanding of human deliberation (Wittenbaum et al., 2004). As shown in the Appendix (Fig. 5 Panel A), main results do not depend on the exact value of the adherence parameter but hold for any $\tau > 0$.

After the sending agent has chosen an argument to be shared, the receiving agent integrates the argument and concludes the round by updating her perceived benefits. Since arguments represent information about the benefits of different decision options, which diverge among subgroups, a receiving agent always integrates an argument using the weights corresponding to their own subgroup. These weights can be different from the sender's weights.

The process of randomly activating an agent, sharing an argument, and updating benefit perceptions of the receiver is repeated until either of two states are reached: (1) all agents align in their perception of what they prefer the most, i.e. they form consensus. (2) Two agents of opposing interest groups perceive to prefer what they actually prefer, and no argument combination they receive can possibly change their perception. In this case, consensus becomes impossible, and the simulation would continue until all agents had received all arguments. Note that because this state becomes more likely the more arguments are exchanged, consensus must emerge early enough for all group members to be able to align on one of the options.

## 4 Setup of Simulation Experiments

We conducted simulation experiments to investigate deliberation outcomes in groups with diverging preferences and different argumentation styles, tracking whether the simulated groups reached one of the following states: (1) everyone in the group perceives their second-best option $C$ to be best. (2) Everyone perceives $A$ or $B$ as best, which is the best option for half of the group but the worst option for the other. (3) The group disagrees about which option is best and the discussion has reached a point where perceptions cannot be altered. Argumentation styles were represented by groups that either consist of advocates or diplomats. We operationalized preference divergences by the parameter $d$, with higher values of $d$ implying higher divergence. Theoretical intuition led us to expect that diplomats more often form consensus than advocates but gives no indication of the strength of preference divergence to assume. For this reason, we started the analyses with a simple comparison of the probability of ill-informed consensus in groups of diplomats and advocates under high ($d = 0.6$) and low ($d = 0.2$) divergence.[3] Because this simple comparison revealed that discussion outcomes crucially depended on the level of divergence, subsequent analyses vary $d$ from very low (0.05) to very high (0.95) levels in steps of 0.05, and compare discussion outcomes in groups of advocates versus diplomats at each level.

For every parameter combination, we simulated 1000 independent discussion processes. We assume groups of $N = 6$ agents, which is not an unrealistic size in decision-making settings. A set of $I = 90$ arguments was used (30 arguments per decision option) to create a set of arguments that is sufficiently large for deliberation outcomes to not be determined by the coincidental spread of single arguments. The adherence parameter $\tau$ is set to 2, meaning that agents choose options and send arguments that correspond well to their argumentation style while still allowing for a small degree of randomness in argument communication. Figure 6 in the Appendix demonstrates robustness of the main results at different values for $\tau$. Main results include additional analyses in which we vary group size between 4 and 12 in steps of two, and the number of arguments between 30 and 300 in steps of 30.

The findings of this paper rest on a setting where arguments are initially distributed at random. Yet, contexts can be thought of where cognitive heuristics (Mercier & Sperber, 2011) or homophilous information networks (McPherson et al., 2001) would lead individuals to acquire information selectively prior to discussion. For this reason, additional analyses reported in the Appendix elucidate that discussion outcomes may differ when initial argument distributions correlate with subgroup membership, but that the underlying mechanisms stay the same. Concluding sensitivity analyses investigate if mixed groups of advocates and diplomats result in more or less consensus compared to groups using either argumentation style exclusively.

[3] For reference, high divergence means that the true preference score of group members' most preferred option is 400% higher than their least preferred (1.6 and 0.4, respectively). Low divergence, on the other hand, implies an increase of 50% (1.2 vs. 0.8).

**Table 1** Percentage of groups reaching consensus on either option, by group's argumentation style and level of preference divergence *d* (1000 replications per treatment)

| | Preference divergence | |
|---|---|---|
| | Low<br>($d=0.2$) | High<br>($d=0.6$) |
| Advocates | 95% | 8% |
| Diplomats | 16% | 41% |

## 5 Results

We start off by comparing how often agent groups of advocates versus diplomats find ill-informed consensus under two levels of preference divergence (Table 1). Under high divergence, groups of diplomats converge much more often on either option (41%) than advocates (8%), which is consistent with the intuition that Diplomats' avoidance of harsh disagreement fosters consensus formation. Counter this intuition, however, only 16% of diplomat groups converged on a consensus under low divergence while more than 90% of advocate groups do. Intuitively, lower divergence should foster consensus because weights are more similar for members of opposing subgroups, making it easier to converge. While this is clearly the case for advocates, why can the same not be said of diplomats?

Figure 3 hints at a possible explanation for this puzzling finding, presenting a more fine-grained examination of discussion outcomes across the divergence parameter range. The dashed lines in Fig. 3A suggest that both advocates and diplomats are less likely to experience consensus on *A* or *B* as divergence levels rise. This is explained by the fact that under higher divergence, weights of arguments in favor of *A* or *B* are increasingly different for members of opposing subgroups, making it harder for the group to converge around either of these options. Advocates are especially likely to find consensus on *A* or *B* under low divergence because their argumentation style has self-reinforcing characteristics: random initial majorities in favor of an option will convince other group members, who will then advocate for this option as well, leading to swift convergence. Diplomats, on the other hand, are limited by what their interaction partner perceives to prefer least, making it hard to find consensus when single individuals find worst what a majority finds best.

Among advocates, consensus on option *C* becomes increasingly rare under higher divergence as well (Fig. 2A, solid red line), despite arguments weights in favor of *C* being unaffected by *d*. An explanation for this finding becomes apparent from the solid black line in Fig. 2B: as divergence gets larger, agents' initial perceptions increasingly align with their true preferences, meaning that members of different subgroups start the discussion with diametrically opposed perceptions already. As discussions evolve, advocates amplify this initial disagreement when raising arguments according to what they perceive to benefit from the most, up to a point where discussions are deadlocked and consensual agreement becomes impossible. Figure 2C supports this explanation, showing that the number of *C* arguments sent in groups of advocates decreases in higher divergence (solid red line).

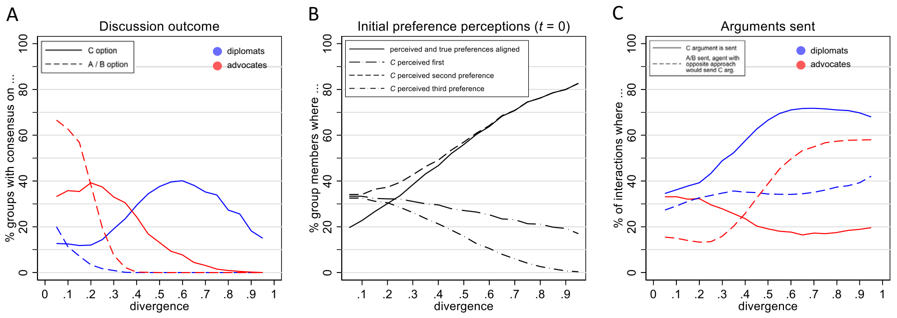


**Fig. 2** Discussion outcome, initial preference perceptions and agent sending behavior, by divergence

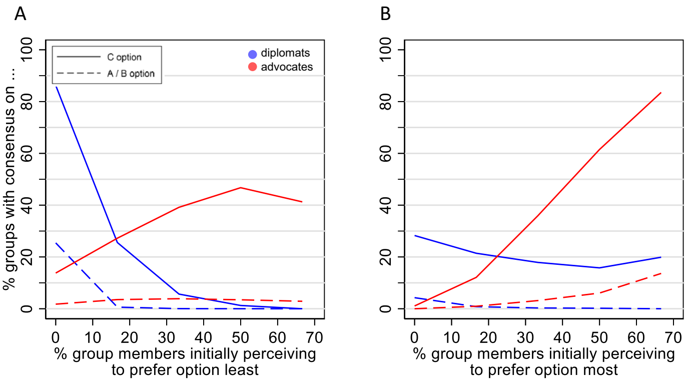


**Fig. 3** Discussion outcome at $d = 0.30$, by initial perceptions of group members

Ill-informed consensus on option $C$ in groups of diplomats, on the other hand, follows a complex pattern (Fig. 2A, solid blue line): higher divergence implies greater chances of consensus until $d = 0.6$. But divergence levels beyond 0.6 negatively impact the proportion of groups with consensus on $C$ again. Contrary to intuition, consensus in groups of diplomats occurs less often than among advocates until $d < 0.4$. A closer look at Fig. 3B offers an intuition why diplomats rarely form consensus: at low divergence levels, the proportion of group members initially preferring $C$ the least is relatively high (short dash-dotted line). Because of their argumentation style, diplomats will avoid sending arguments favoring $C$ to such agents, making consensus on $C$ unlikely. Higher divergence levels, on the other hand, make it easier for diplomats to form consensus: more agents start the discussion with perceptions that correspond to their true preferences, meaning that agents of different groups will have opposing perceptions about options $A$ and $B$, and more agents perceive $C$ as second best. As divergence rises, diplomats raise more arguments in favor of $C$ (Fig. 2C, solid blue line), explaining a greater fraction of groups with consensus on this option until $d = 0.6$.

Divergence does not affect the initial distribution of arguments, but changes the way arguments are perceived by an agent. Higher divergence implies that an *A* argument will have a much greater impact on an α-agent's perception that *A* is best, but a smaller impact on the perception of a β-agent. From this follows that at low divergence, assigning arguments at random creates perceptions that are more random, while random argument assignment at high divergence implies perceptions that correspond closely to the ranking of true preferences among group members.

In sum, the results obtained here reveal a striking finding: at low divergence levels ($<0.2$), advocate groups are more than three times as likely to find an ill-informed consensus on either option, but the opposite is the case for higher divergence levels ($>0.6$). The explanation we propose is that divergence impacts the initial distribution of perceptions, which in turn interacts with agents' argumentation style and their chances of finding consensus: advocates are more likely to find consensus on an option when more agents initially perceive to prefer this option the most. Diplomats, on the other hand, are more likely to establish consensus on an option the fewer agents perceive to prefer this option the least.

We test the proposed explanation by zooming in on groups of diplomats and advocates at a moderate level of $d=0.3$ and analyze how often they find consensus, depending on the proportion of agents in the group initially perceiving an option to be least (Fig. 3, left panels) or most beneficial (right panels). The results overwhelmingly support the proposed explanation: the proportion of diplomat groups finding consensus sharply declines the more group members initially perceive an option as worst, while advocates become more likely to build consensus the higher the proportion of agents initially perceiving an option as best. Note that the two variables are positively correlated: due to the random assignment of arguments during initialization, allocating disproportionately few arguments in favor of an option to some agents by chance implies that others will receive disproportionately many. Because of this, almost 50% of advocate groups find consensus on *C* even when two thirds of the group initially perceive this option as least preferred.

Figure 2 showed that divergence affects the consensus-making capacities of advocates and diplomats through its influence on the initial distribution of perceptions. We now investigate if the same is the case for other substantial features of the discussion setting, namely the number of available arguments and group size. If the general explanation holds, any factor leading to a closer alignment between initial perceptions and true preferences should positively impact consensus-making for diplomats but negatively for advocates. Figure 4 shows that this is indeed the case. As the number of arguments grows, agents hold more arguments at the start of the discussion. Here, the law of great numbers implies that initial perceptions that are based on more arguments will more closely resemble their expected value (i.e. $E(q_{x,t0,o}) = Q_{o,g}/N$) and are hence more often aligned with agents' true preferences (Fig. 4B, left side). In consequence, more groups of diplomats and fewer groups of advocates find consensus as the number of arguments grows (Fig. 4A, left side). Group size, on the other hand, has an opposite effect (Fig. 4, right side): as groups get larger, the same number of arguments is distributed over more agents, leading to

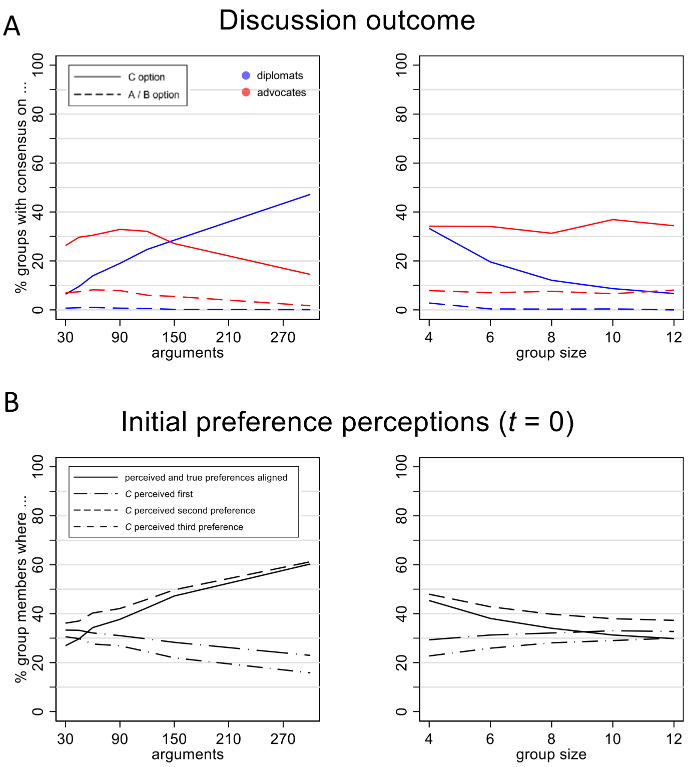


**Fig. 4** Discussion outcome and initial preference perceptions of group members, by argument pool and group size ($d = 0.3$)

less alignment of initial perceptions with the ranking of true benefits, and, in turn, less frequent consensus among diplomats.

In sum, the results presented here reveal striking insights into the deliberation outcomes of simulated groups with diverging preferences: when many group members initially misperceive an option as preferred, advocating for this option regardless of others' perceptions will be a very effective way to emerge with a—likewise ill-informed—consensus. But when group members have initial perceptions that closely align with their true preferences, consensus requires an argumentation style that makes members bring forth arguments in favor of a second-best alternative. Additional analyses in the Appendix show that this finding persists when arguments are initially distributed in a lopsided fashion. Results are robust to the level of strategy adherence $\tau$, do not depend on random as opposed to homophilous interaction preferences, and remain similar when a decision is made at only four or five out of six group members perceiving to prefer the same option. Further analyses reported in the Appendix reveal that results are not artifacts of perfectly homogenous groups either; but persist in mixed groups where minorities of agents using the opposite argumentation style are introduced to a population.

## 6 Discussion

Considering the complex relationship between discussion outcomes, argumentation styles and preferences perceptions, our results offer a new perspective on existing theoretical narratives. The simulations support the overall conclusion that the two argumentation styles under consideration have their own merits and defects relative to context. As indicated before, we do not believe that our models fully match real-life deliberations but we believe them to be sufficiently descriptively adequate to lend normative force to the conclusions, in the sense that they can back up tentative claims about the desirability, relative to a context, of one or other argumentation style.

Importantly, in this new perspective we cannot and do not take a stand on the desirability of the outcomes of social deliberations: the merits and defects of ill-informed consensus and truthful disagreement are context-dependent to the extent that choosing an overall favorite would be nonsensical. Our main result is that relative to what is deemed beneficial in a certain context, and on the assumption that group members harbor a latent disagreement but start off with incomplete information about their actual preferences, there are further specifics of the deliberation setting that will make a diplomatic or an advocating argumentation style the more effective one.

This relatively modest and qualified conclusion challenges overly simplistic applications of ideal-type deliberative formats. In particular, assuming that the context makes a consensus desirable, a simplistic reading of the agonistic approach of Mouffe (1999) might suggest that groups in which individuals advocate according to their best evidence will arrive at superior collective conclusions. However, our results support this idea only when initial evidence is fragmented and noisy, leading agents away from their true preferences from the onset, and then only if the differences in preferences are not too pronounced. Conversely, echoing the perspective of Habermas (1985a, b), it may have seemed that groups where members prioritize avoiding offensive engagement and maintain a common ground often exhibit superior deliberative performance. Yet again, our findings only partially align with this expectation: diplomatic argumentation enables consensus only when agents have sufficiently accurate initial convictions, or when preference divergence is high enough for disagreement to be obvious under more abrasive argument sharing.

We hope that our results, apart from revealing the context-sensitivity of styles of deliberation, stimulate a broader engagement of philosophical debate with insights from computational sociology and formal social epistemology. It bears repeating that we do not view our results, or indeed other results from these formal and computational disciplines, as providing stand-alone support for some or other communication style or format for public debate. Rather, we believe that studying the consequences of such styles and formats from up close will help us in our attempts to make public debate and social deliberations more fruitful, transparent, and fair, through an understanding of the dynamics that drives them.

With our emphasis on the fact that our results depend on context, we do not mean to suggest that contextual considerations cannot be accommodated, at least partially, in further model developments. Argued from a utilitarian point of view, the optimal outcome depends on the post factum consequences of the group decision: disagreement can incur costs for both individuals and create unforeseen externalities, but so can an ill-informed consensus. In our model, we assume that the consequences of the group decision—beyond what can be learned from the arguments raised in the deliberation process—are unknown to group members, and accordingly that they do not feature in the deliberation. However, relaxing this assumption is certainly possible, and this would open a new array of strategic considerations to be done by agents accompanied by further theorizing that exceeds the scope of the paper.

Like in any simulation model, simplifying assumptions were made for reasons of parsimony and to the benefit of understandability. This involves, for example, that the relatively basic, additive argument structure used here enabled us to intuitively trace and understand agents' preference formation process. Bayesian preference updating (Assaad et al., 2023; Madsen et al., 2018) is regarded as more 'rational' from the agents' point of view, suggesting that exploring this updating approach is a valuable avenue for future modeling work. Similarly, recent advances in the development of large-language models make it possible to formally represent argument communication in terms of realistic human language rather than the abstract representation of arguments applied here (Betz, 2022; Du et al., 2023). However, while LLMs are powerful tools to generate meaningful text, it is unclear whether they also reliably represent human behavior in complex settings like a debate in which individuals deliberate despite having competing preferences. Another simplifying assumption concerns the focus on a dyadic interaction setting. Restricting argument communication to only one receiver per interaction facilitates an easy, straightforward formalization of an argumentation style that takes receiver preferences into account. Multilateral communication, on the other hand, would involve weighing a multiplicity of receiver preferences against specific persuasion goals, introducing additional complexity and exceeding the scope of this study.

The two argumentation styles elucidated here served as prototypes of self-serving versus cautious communication. Obviously, alternative ways to implement these styles exist, and many alternative argumentation styles can be thought of that should be considered by future research. Having studied groups that only consisted of either advocates or diplomats, the question arises if new patterns of discussion outcomes emerge if groups consist of a mix of individuals with different argumentation styles: for example, whether groups with just one or two diplomats are enough to steer the group towards consensus under high preference divergence. Additional analyses reported in the Appendix explore this possibility and suggest that the deliberation outcomes do not depend on the assumption that groups are homogenous in argumentation style: instead, mixed groups simply produce discussion outcomes that resemble a linear combination of outcomes in homogenous groups.

Next to alternative argumentation styles and mixed groups, another factor that can potentially affect discussion outcomes is whether groups will only make decisions after having reached consensus or whether they rely on voting procedures instead (Priem et al., 1995). Of course, investigating the full spectrum of possible voting rules exceeds the scope of any sensitivity analysis, but we show in the Appendix that results are at least robust under a majority-based voting rule: low preference divergence still facilitates decision-making among advocates when only four or five out of six members must align in their convictions, while high divergence fosters decision-making among diplomats.

In our model, subgroups impose a preference structure that is symmetric and straightforward. But extensions of our model could weaken this symmetry and allow for a multiplicity of individual stakes next to subgroup membership. Likewise, additional biases and heuristics in the interaction between group members can be implemented. A first step in this direction is investigated in the appendix, where we show that discussion outcomes remain similar if agents interact in homophilous (McPherson et al., 2001) instead of random encounters. Although group-wide consensus becomes less likely the higher the homophily level, diplomats will scontinue to find consensus more often than advocates under high preference divergence, while the opposite is the case at low divergence.

Next to homophily, literature on affective polarization (Iyengar et al., 2019) and bounded confidence (Hegselmann & Krause, 2002) suggest that individuals may reject information that comes from disliked or dissimilar sources. While not part of our model in a theoretical sense, our robustness analyses on homophily formally include such behavior: whether individuals encounter outgroup members at a lower probability or accept their arguments with lower likelihood is mathematically interchangeable.

Besides advancing theoretical development to refine the model and validate the robustness of our findings, empirical investigation holds promise in advancing the discourse. Our model's key innovation, the incorporation of agents' conflicting preferences, presents a fertile ground for empirical exploration. Preferences can be rigorously quantified and experimentally manipulated in experiments along the paradigm of behavioral game-theory (Camerer, 2011; Fehr & Gächter, 2000), making it possible to create laboratory settings where human participants operate with preferences akin to those in our model. An important empirical question to answer is, for instance, what argumentation styles individuals are using and whether there are conditions under which humans adopt different styles. Experimental work in social psychology, for instance, suggests that individuals may strategically misrepresent their positions when discussing with members of outgroups holding different positions (Hogg et al., 1990).

Likewise, theoretical work would profit from empirical work on consensus formation in groups. We focused our analyses on the first time a group experienced consensus in that all members perceive to prefer the same option. This consensus, however, is ill-informed since a continuation of the discussion to the point of full information would reintroduce disagreement. An important empirical question is whether individuals notice that they have reached consensus and stop the debate or whether and under what conditions, they continue the exchange of arguments.

## 7 Conclusion

Our simulation analyses suggest that in groups with diverging preferences, deliberation is shaped by the way members raise arguments as well as their initial preference perceptions. Advocating for what one finds personally beneficial only led to truthful disagreement when group members started the discussion with accurate perceptions about their preferences already. Conversely, when initial perceptions were noisy and inaccurate, random initial majorities often convinced the rest of the group of an option they disfavored had full information been present. We compared the behavior of such 'advocates' with that of 'diplomatic' agents who avoid disagreement at the cost of speaking one's actual mind. In these groups, initially accurate (divergent) preference perceptions led to an ill-informed consensus. Inaccurate initial perceptions, on the other hand, eventually resulted in truthful disagreement. Here, the avoidance of disagreement made consensus hard because majorities failed to convince other group members of an option they found least preferable.

## Appendix

### Lopsided Argument Assignment

The main results pertained to groups where arguments were distributed at random prior to deliberation. Here we investigate discussion outcomes of simulated groups where arguments are initially assigned in a selective fashion. We introduce an additional parameter, σ, regulating the probability by which agents draw arguments in support of their most preferred option. They take turns at choosing from the set of available arguments *A'*, one at a time without replacement according to

$$p(a_i) = exp(\sigma * w_{i,o_{max},g}) / \sum\nolimits_{i \in \{A\prime\}} exp(\sigma * w_{i,o_{max},g}) \tag{3}$$

where $o_{max}$ represents the option members of a subgroup $g$ prefer the most. At $\sigma = 0$, arguments are drawn at random, mirroring the setup of discussion groups in the main results. At $\sigma = 2$, argument distribution is highly lopsided, with high chances of $A$ arguments being drawn by $a$ members, $B$ arguments drawn by $\beta$ members, and $C$ arguments being drawn by members of both subgroups with equal probability. To avoid that initial preference perceptions strongly correlate with group members' true preferences independent of the level of σ, the simulation experiments presented here use a low divergence value of $d = 0.1$.

Figure 5 shows that higher σ leads to discussion outcomes similar to higher preference divergence (compare Fig. 2). This is explained by the fact that both parameters tighten the correlation between initial preference perceptions and true preferences (Panel B), either directly through lopsided allocation (σ) or indirectly through differences in argument weights (*d,* see Fig. 2B). Variations of the σ parameter always produce more consensus on option $A$ or $B$ than on $C$ among advocates, with the reason being the low value of preference divergence: at low $d$,

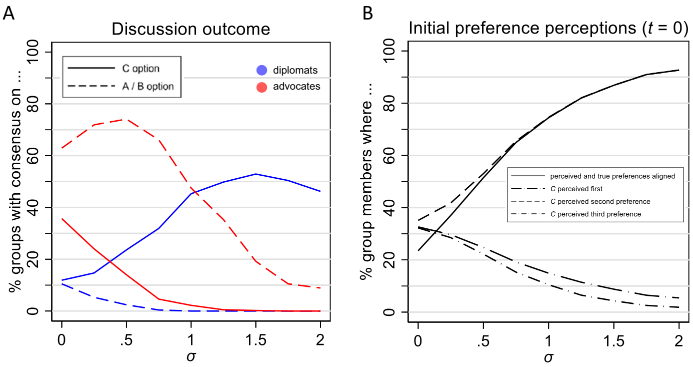


**Fig. 5** Discussion outcomes and initial preference perceptions by sigma at $d=0.1$

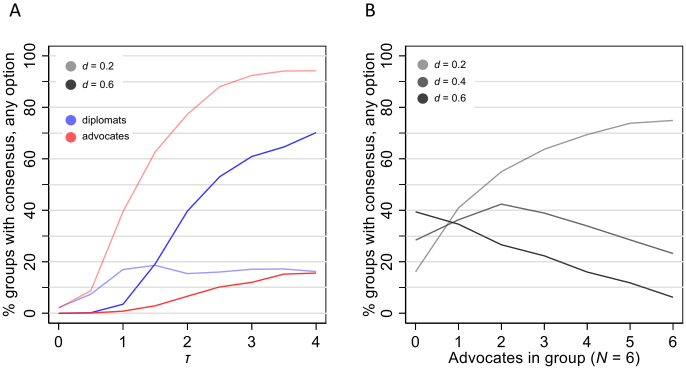


**Fig. 6** Discussion outcomes by adherence $\tau$ (**A**) and number of advocates in group (**B**)

options *A* and *B* are more similar to members of both subgroups, making it easier to align on either of these options.

## Strategy Adherence and Group Composition

Figure 6, Panel A reveals that across the range of the adherence parameter τ and consistent with the main results, diplomats are more likely to form consensus than advocates at high divergence. At low divergence, consensus is more likely among advocates than diplomats. Differences in the effects of the two argumentation styles become bigger in τ. This is explained by the fact higher adherence implies less randomness and a more deterministic selection of arguments according to agents' argumentation styles. At very low adherence ($\tau < 0.5$), argument selection for both styles approximates randomness, resulting in similar probabilities to find consensus.

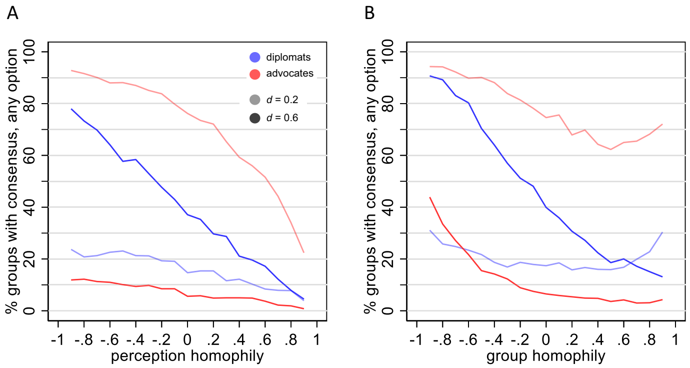


**Fig. 7** Discussion outcomes by perception-based (**A**) and subgroup-based homophily (**B**)

Figure 6, Panel B elucidates discussion outcomes at different divergence levels for mixed groups of advocates and diplomats. At high divergence, the probability of consensus sinks almost linearly in a greater fraction of advocates and rises monotonously at low divergence. Both results are consistent with the main results, showing that consensus occurs more often among diplomats than among advocates at high divergence, while the opposite occurs at low divergence. Only at moderate divergence, a peak in consensus propensity appears at two diplomats and four advocates, exhibiting a curvilinear relationship between consensus propensity and the fraction of advocates in the group. Here, frequent consensus results from diplomats' tendency to raise $C$ arguments, combined with advocates' ability to raise arguments even if they go against their opponents' preference perceptions.

## Preferential Interaction

The simulations that underlie the main results of the paper assume that agents select any interaction partner with equal probability. Here, we test if results are robust to two types of preferential interaction, namely, similarity in perceptions and similarity based on subgroup membership. We regulate interactions between agents by introducing a homophily parameter $h$, ranging from $-1$ to 1. The greater $h$, the more likely a sending agent is to choose a receiving agent with the same trait. A sending agent chooses a receiving agent according to

$$p_y = s_y / \sum\nolimits_{y \in \{Y\}} s_y \tag{4}$$

where $y$ denotes the individual receiver and $Y$ the set of potential receiving agents. $s_y$ represents the trait similarity between sender and receiver, taking on the value of $h/2+0.5$ if sending and receiving agent share the same trait and $1-(h/2+0.5)$ otherwise.

Consistent with the main results, Fig. 7 shows that advocates are more likely than diplomats to form consensus at low levels of preference divergence ($d=0.2$),

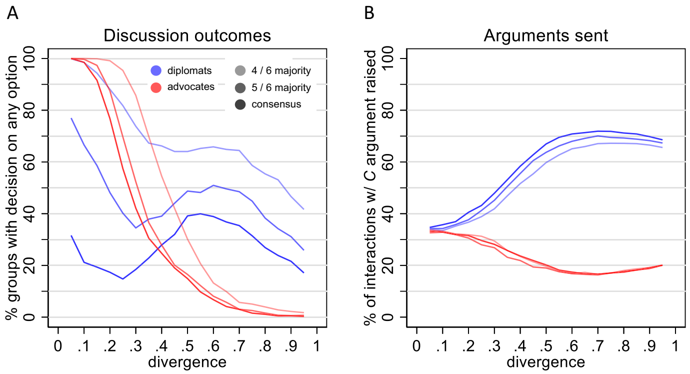


**Fig. 8** Majoritarian decision-making, by divergence

while diplomats are more likely to form consensus at high preference divergence ($d = 0.6$). This is true for the entire parameter range of perception-based homophily (Panel A) and group-based homophily (Panel B). Generally, higher homophily levels indicate lower probabilities of finding consensus. This is because preferential interactions between similar individuals, either in perceptions or subgroup membership, solidify convictions in line with true preferences and result in disagreement as the final outcome of the discussion. Interesting to note is the increase in consensus at low divergence as subgroup homophily reaches higher levels, pointing to potential 'transient diversity' effects (see Stein et al., 2024).

## Decision-Making Without Consensus

Figure 8 reveals that groups would reach similar discussion outcomes if decisions were not made according to full consensus, but according to a 5/6 or 4/6 majority rule instead. Advocates are less likely to converge around any option as divergence levels rise, regardless of whether 4 or 5 group members perceive to prefer the same option. Diplomats tend to disagree more often as divergence levels rise as well, but this tendency is offset by a local peak at around $d = 0.6$, regardless of the underlying decision-making rule. The latter is, again, explained by diplomats' ability to raise arguments in favor of option $C$ as divergence levels rise (Fig. 8, Panel B).

**Acknowledgements** We thank Andreas Flache, Vincenz Frey, Vincent Buskens, the Norms and Networks Cluster at the University of Groningen, participants of the ICS Forum Day, two anonymous reviewers, and the editor Leon Horsten for helpful comments and feedback.

**Funding** Open Access funding enabled and organized by Projekt DEAL. This study is part of the research program Sustainable Cooperation—Roadmaps to Resilient Societies (SCOOP). The authors are grateful to the Netherlands Organization for Scientific Research (NWO) and the Dutch Ministry of Education, Culture and Science (OCW) for generously funding this research in the context of its 2017 Gravitation Program (Grant No. 024.003.025).

**Data Availability** The model is written and built in NetLogo. The data was analyzed using Stata. The simulation model, data output and Stata syntax to reproduce the results of this study are available on the Open Science Framework (OSF) via the following URL: https://osf.io/4tyrz/.

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied, 70*(9), 1–70. https://doi.org/10.1037/h0093718

Assaad, L., Fuchs, R., Jalalimanesh, A., Phillips, K., Schoeppl, L., & Hahn, U. (2023). *A Bayesian agent-based framework for argument exchange across networks* (arXiv:2311.09254). arXiv. https://doi.org/10.48550/arXiv.2311.09254

Baldassarri, D., & Bearman, P. (2007). Dynamics of political polarization. *American Sociological Review, 72*(5), 784–811. https://doi.org/10.1177/000312240707200507

Bernstein, E., Shore, J., & Lazer, D. (2018). How intermittent breaks in interaction improve collective intelligence. *Proceedings of the National Academy of Sciences, 115*(35), 8734–8739. https://doi.org/10.1073/pnas.1802407115

Betz, G. (2022). Natural-language multi-agent simulations of argumentative opinion dynamics. *Journal of Artificial Societies and Social Simulation, 25*(1), 2.

Blume, L. E., Brock, W. A., Durlauf, S. N., & Ioannides, Y. M. (2011). Identification of social interactions. In J. Benhabib, A. Bisin & M. O. Jackson (Eds.), *Handbook of social economics* (Vol. 1, pp. 853–964). North-Holland. https://doi.org/10.1016/B978-0-444-53707-2.00001-3

Brandom, R. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press.

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton: Princeton University Press.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441*(7095), 876–879. https://doi.org/10.1038/nature04766

Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems, 03*(01–04), 87–98. https://doi.org/10.1142/S0219525900000078

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). *Improving factuality and reasoning in language models through multiagent debate* (arXiv:2305.14325). arXiv. https://doi.org/10.48550/arXiv.2305.14325

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90*(4), 980–994. https://doi.org/10.1257/aer.90.4.980

Flache, A., & Macy, M. W. (2011). Small worlds and cultural polarization. *The Journal of Mathematical Sociology, 35*(1–3), 146–176.

Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation, 20*(4), 2.

Greene, W. (2009). Discrete choice modeling. In T. C. Mills & K. Patterson (Eds.), *Palgrave handbook of econometrics: Volume 2: Applied econometrics* (pp. 473–556). Palgrave Macmillan. https://doi.org/10.1057/9780230244405_11

Guo, D., & Yu, A. J. (2019). Human learning and decision-making in the bandit task: Three wrongs make a right. Conference on Cognitive Computational Neuroscience. https://pdfs.semanticscholar.org/decc/a47db4e33ad70ffadeadf622108aa9ec69f8.pdf

Habermas, J. (1985a). *The theory of communicative action: Volume 1: Reason and the rationalization of society* (Vol. 1). Beacon press.

Habermas, J. (1985b). *The theory of communicative action: Volume 2: Lifeword and system: A critique of functionalist reason* (Vol. 2). Beacon Press.

Hahn, U., & Harris, A. J. (2014). What does it mean to be biased: Motivated reasoning and rationality. In B. Ross (Ed.) *Psychology of learning and motivation* (Vol. 61, pp. 41–102). Elsevier. https://www.sciencedirect.com/science/article/pii/B9780128002834000022

Hall, E. T. (1976). *Beyond culture*. Doubleday.

Harlé, K. M., Zhang, S., Schiff, M., Mackey, S., Paulus, M. P., & Yu, A. J. (2015). Altered statistical learning and decision-making in methamphetamine dependence: Evidence from a two-armed bandit task. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01910

Hausman, D. M. (1995). The impossibility of interpersonal utility comparisons. *Mind, 104*(415), 473–490. https://doi.org/10.1093/mind/104.415.473

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, *5*(3). 2–34

Hegselmann, R., & Krause, U. (2006). Truth and cognitive division of labor: First steps towards a computer aided social epistemology. *Journal of Artificial Societies and Social Simulation, 9*(3), 10.

Hogg, M. A., Turner, J. C., & Davidson, B. (1990). Polarized norms and social frames of reference: A test of the self-categorization theory of group polarization. *Basic and Applied Social Psychology, 11*(1), 77–100. https://doi.org/10.1207/s15324834basp1101_6

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science, 22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Keijzer, M. A., Mäs, M., & Flache, A. (2018). Communication in online social networks fosters cultural isolation. *Complexity, 2018*, e9502872. https://doi.org/10.1155/2018/9502872

Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly, 52*(4), 667–694. https://doi.org/10.2189/asqu.52.4.667

Levinthal, D. A. (1997). Adaptation on rugged landscapes. *Management Science, 43*(7), 934–950. https://doi.org/10.1287/mnsc.43.7.934

Lu, L., Yuan, Y. C., & McLeod, P. L. (2012). Twenty-five years of hidden profiles in group decision making: A meta-analysis. *Personality and Social Psychology Review, 16*(1), 54–75. https://doi.org/10.1177/1088868311417243

Madsen, J. K., Bailey, R. M., & Pilditch, T. D. (2018). Large networks of rational agents form persistent echo chambers. *Scientific Reports*, *8*(1), Article 1. https://doi.org/10.1038/s41598-018-25558-7

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*(1), 71–87. https://doi.org/10.1287/orsc.2.1.71

Mark, N. P. (2003). Culture and competition: Homophily and distancing explanations for cultural niches. *American Sociological Review, 68*(3), 319–345. https://doi.org/10.2307/1519727

Mason, W., & Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences, 109*(3), 764–769. https://doi.org/10.1073/pnas.1110069108

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*(1), 415–444.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968. **discussion 74-111**.

Mouffe, C. (1999). Deliberative democracy or agonistic pluralism? *Social Research, 66*, 745–758.

Olsson, E. J. (2013). A Bayesian simulation model of group deliberation and polarization. In F. Zenker (Ed.), *Bayesian argumentation: The practical side of probability* (pp. 113–133). Springer. https://doi.org/10.1007/978-94-007-5357-0_6

Priem, R. L., Harrison, D. A., & Muir, N. K. (1995). Structured conflict and consensus outcomes in group decision making. *Journal of Management, 21*(4), 691–710. https://doi.org/10.1177/014920639502100406

Reverdy, P., & Leonard, N. E. (2015). Parameter estimation in softmax decision-making models with linear objective functions. *IEEE Transactions on Automation Science and Engineering, 13*(1), 54–67.

Robbins, L. (1938). Interpersonal comparisons of utility: A comment. *The Economic Journal, 48*(192), 635. https://doi.org/10.2307/2225051

Sherif, M., & Hovland, C. I. (1961). *Social judgment: Assimilation and contrast effects in communication and attitude change*. Yale University.

Shore, J., Bernstein, E., & Lazer, D. (2015). Facts and figuring: An experimental investigation of network structure and performance in information and solution spaces. *Organization Science, 26*(5), 1432–1446. https://doi.org/10.1287/orsc.2015.0980

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology, 48*, 1467–1478. https://doi.org/10.1037/0022-3514.48.6.1467

Stasser, G., & Titus, W. (2003). Hidden profiles: A brief history. *Psychological Inquiry, 14*(3–4), 304–313. https://doi.org/10.1080/1047840X.2003.9682897

Stein, J., Frey, V., & Flache, A. (2024). Talk less to strangers: How homophily can improve collective decision-making in diverse teams. *Journal of Artificial Societies and Social Simulation*. https://doi.org/10.18564/jasss.5224

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

van Veen, D.-J., Kudesia, R. S., & Heinimann, H. R. (2020). An agent-based model of collective decision-making: How information sharing strategies scale with information overload. *IEEE Transactions on Computational Social Systems, 7*(3), 751–767. https://doi.org/10.1109/TCSS.2020.2986161

Witt, A., Toyokawa, W., Lala, K. N., Gaissmaier, W., & Wu, C. M. (2024). Humans flexibly integrate social information despite interindividual differences in reward. *Proceedings of the National Academy of Sciences, 121*(39), e2404928121. https://doi.org/10.1073/pnas.2404928121

Wittenbaum, G. M., Hollingshead, A. B., & Botero, I. C. (2004). From cooperative to motivated information sharing in groups: Moving beyond the hidden profile paradigm. *Communication Monographs, 71*(3), 286–310. https://doi.org/10.1080/0363452042000299894

Wu, C. M., Deffner, D., Kahl, B., Meder, B., Ho, M. H., & Kurvers, R. H. J. M. (2024). Visual-spatial dynamics drive adaptive social learning in immersive environments (p. 2023.06.28.546887). bioRxiv. https://doi.org/10.1101/2023.06.28.546887

Zollman, K. J. (2010). The epistemic benefit of transient diversity. *Erkenntnis, 72*(1), 17–35.