UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Multi-Option Polarization: How Deliberating More Options Both Increases and Decreases Polarization

Permalink

https://escholarship.org/uc/item/5dt8g8c1

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Author

Assaad, Leon

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Multi-Option Polarization: How Deliberating More Options Both Increases and Decreases Polarization

Leon Assaad (L.Assaad@campus.lmu.de)
Munich Center for Mathematical Philosophy, LMU Munich
Ludwigstraße 31, 80539 Munich, Germany

Abstract

Formal models in social epistemology explore why rational agents might polarize. While paradigmatic models focus on binary topics, e.g., "Is H true or false?", many real-world issues involve multi-option topics: "Which of $n > 2$ options is true/best?" This paper introduces a model of rational deliberation on multi-option topics to address the following question: As a group discusses more options, should we expect their beliefs to polarize more or less? We find a dual effect: as the number of options increases, agents are more likely to disagree on which option is most likely correct. This makes it harder to reach consensus on a single position. At the same time, their beliefs—and thus their disagreements—become less extreme. Hence, while agents are more likely to disagree, these disagreements are less intense. Since each trend aligns with a familiar concept of polarization, more options can increase or decrease polarization, depending on one's measurement.

Keywords: Agent-Based Model; Polarization; Bayesian Inference; Deliberation; Social Epistemology

Introduction

Polarization seems omnipresent—both as a political issue and as a research topic across disciplines. While the term "polarization" has many meanings (Bramson et al., 2017), one type has garnered particular attention from social epistemologists: group polarization, the tendency for group members' beliefs about a topic to diverge through deliberation. Group polarization, whether in political (Singer et al., 2019) or scientific debates (O'Connor & Weatherall, 2018), appears undesirable: deliberation is often aimed at achieving (correct) consensus, so belief divergence constitutes a failure of the process.

While influential theories attribute polarization to irrational behaviors (e.g., Taber & Lodge, 2006; Kahan, 2013), philosophers seek to understand it as a rational phenomenon. Formal models have played a central role in this effort by simulating artificial groups of boundedly rational agents deliberating on a target hypothesis. Such "rational reconstructions" uncover potential rationales underlying polarization and highlight conditions under which even unbiased agents would struggle to converge.

Paradigmatic models have focused on agents debating questions of the form "Is H true or false?" or "Which of two options is best?" However, many real-world questions, from scientific inquiries to democratic voting, take the form of multi-option topics: "Which of $n > 2$ options is true/best?" Cases where agents deliberate more than two candidate options are underexplored in the formal polarization literature.

Addressing this gap, this paper tackles two challenges. First, it introduces a new model of rational deliberation for multi-option topics. Second, it investigates the effects of increasing the number of options: as we move beyond binary topics, should we expect polarization to increase or decrease?

The proposed model extends Assaad and Hahn's (2024) simulation of deliberation (and polarization) via evidence exchange. It models the topic as a variable with n values, representing the considered options. Does increasing n exacerbate polarization? The answer depends on how multi-option polarization is defined. This paper adapts two measures from the binary polarization literature: (1) the diversity of positions, where a position represents the option an agent considers most likely true, and (2) belief dispersion, quantifying the intensity of disagreement about each option. The findings reveal a dual effect: as n increases, position diversity (1) rises, making consensus on a single preferred option harder to achieve. However, average belief dispersion (2) decreases—disagreements about individual options become less extreme. Hence, while agents are more likely to disagree about which option is correct, these disagreements are less intense. But is this polarization? Increasing the number of discussed options either amplifies or mitigates polarization, depending on the chosen measure. While our result is general and intuitive, it highlights that multi-option polarization is highly measure-sensitive, subject to different intuitions and interpretations.

Modeling Polarization

Consider the following scenario: a group deliberates whether a hypothesis H is true (H) or false ($\neg H$). Agents gather and exchange evidence for and against H, each forming a degree of belief that H is true, denoted as $P(H) \in [0,1]$, based on the information received. Over time, the group may either reach consensus on a common belief or diverge, with some agents believing H is likely true while others believe it is false.

This latter outcome is often referred to as group polarization in the socio-epistemic literature. As a result of social exchange or exposure to evidence, agents' beliefs can diverge—one agent's belief in $P(H)$ may decrease while another's increases (Henderson & Gebharter, 2021). In group settings, polarization is more challenging to quantify,[1] but it

[1] Bramson et al. (2017) distinguish nine different measurements.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

typically refers to the dispersion of group beliefs over time. In extreme cases, the group splits into completely opposing camps—one believing $P(H) \approx 0$, the other $P(H) \approx 1$ (Pallavicini, Hallsson, & Kappel, 2021).

Using formal models of deliberation, philosophers have explored whether and why rational agents might polarize (e.g., Hegselmann, Krause, et al., 2002; O'Connor & Weatherall, 2018; Olsson, 2013). These studies highlight general mechanisms that are (arguably) compatible with rationality. One mechanism leading to polarization is that agents interpret the same evidence differently. Due to differing background beliefs, agents assign different likelihoods to the evidence, leading some to see it as supporting H and others as supporting $\neg H$ (e.g., Jern, Chang, & Kemp, 2009, 2014; Henderson & Gebharter, 2021; Olsson, 2020; Freeborn, 2024).

However, even if agents interpret evidence uniformly, they may still polarize. Another mechanism involves agents receiving different subsets of evidence, with some subsets supporting H and others supporting $\neg H$. In fact, it is quite likely that agents of a group will end up knowing different portions of the available information, especially when there is a large body of evidence. This can lead to divergent beliefs, particularly when strong evidence exists for both sides. Then, incomplete and unequal access to evidence can result in significant polarization, with some agents believing $P(H) \approx 1$ and others believing $P(H) \approx 0$.

Assaad and Hahn (2024) show that this type of polarization may arise through group deliberation when agents restrict their evidence-sharing to what they consider their "best," most truth-conducive evidence. In their simulation, communication is modeled as agents strategically passing on pieces of evidence. If agents share only their strongest evidence, polarization emerges—and it is exacerbated when (i) agents have few interlocutors (i.e., the social network is sparse) and (ii) strong evidence exists for both H and $\neg H$. When agents do not share all their evidence, information does not travel across the entire network. As a result, subgroups emerge, each exposed to different sets of evidence. This fragmentation is more likely in sparse networks (i). And when strong evidence exists for both sides (ii), different evidence subsets may pull agents' beliefs further apart.

There are other models of polarization based on unequal access to evidence. Mäs and Flache (2013) show that agents may settle on different arguments due to homophilic network structures, while in Singer et al. (2019), agents consider different evidence due to limited memory.[2] The general mechanism can take many forms: when agents receive different evidence, they may polarize. This paper investigates this broader mechanism in a multi-option context.

Though not a comprehensive survey of the literature (cf. Šešelja, 2023), this short introduction highlights that paradigmatic models of polarization focus on binary questions: agents must choose between H and $\neg H$ (e.g., Pallavicini et al., 2021; Olsson, 2013) or between two options, A and B (e.g., Zollman, 2010; O'Connor & Weatherall, 2018). We believe this to be a significant limitation.

[2]See also Kopecky (2024); Schöppl and Hahn (2024).

Multi-Option Topics

Important decision problems may not "boil down to two options" (List & Goodin, 2001). This is evident in most democracies, where voters choose among multiple parties. It is also true in science, where more than two candidate theories often compete as the best explanation (Stegenga, 2013), and in judicial cases, where there may be more than two suspects (Hahn & Hartmann, 2020). Whenever agents ask, "Which of $n > 2$ options is true/best?" they are deliberating a multi-option topic (MOT).

To begin modeling MOTs, we draw on social choice theory. List and Goodin (2001) extend the Condorcet Jury Theorem to cases where there are n options, "where precisely one option (...) is supposed to be the epistemically 'correct' outcome" (2001, p. 9). Following this description, we make this assumption about MOTs: the n options are mutually exclusive and jointly exhaustive (MEJE), meaning exactly one option is correct. This assumption excludes cases where no option is correct, multiple options are correct, or new options emerge during deliberation.[3] While real-world discussions may involve such complexities, MEJE provides a useful starting point for analysis and applies broadly—it covers all cases where a group seeks a single "best" option.[4]

How can we embed MOTs in a model of group deliberation? In the scenario above, agents deliberate a hypothesis by collecting and exchanging evidence. Thus, any model of deliberation must be based on a model of evidence—one that specifies how evidence relates to the hypothesis and how (rational) agents interpret it.

A Bayesian Model of MOT

Here we develop a model of MOT—comprising a hypothesis and a set of evidence—before using it in a simulation of group deliberation. To do so, we adopt a Bayesian approach. Bayesian epistemology is a well-established standard of rationality in philosophy (Bovens & Hartmann, 2003) and is widely used to model human reasoning (Hahn, Harris, & Corner, 2009; Collins, Hahn, Von Gerber, & Olsson, 2018). Bayesianism posits that rational agents' beliefs about propositions behave like probabilities and should be updated according to Bayes' rule (Sprenger & Hartmann, 2019).

The NormAN framework (*Normative Argument Exchange across Networks*, Assaad et al., 2023) takes the Bayesian machinery to create deliberative scenarios in a straightforward manner. It uses Bayesian networks (BNs) to model the probabilistic relationships between a central hypothesis and its

[3]The addition of new propositions (or options) to an agent's awareness is an important topic in epistemology. In this context, "catch-all" propositions, such as a "none of the above" choice, can help ensure exhaustiveness (de Canson, 2024).

[4]See Freeborn (2024); Weatherall and O'Connor (2021b) for models where agents deliberate and "factionalize" around sets of binary topics. These cases do not satisfy MEJE.

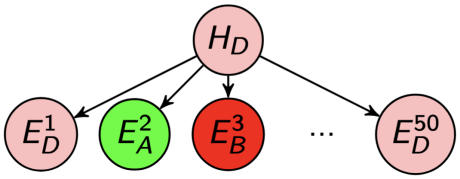


Figure 1: A generic BN with 50 independent pieces of evidence. Colors indicate the value proposition each takes (in the subscript), which are generated via NormAN's stochastic initialization. This is an example distribution.

	H_A	H_B	H_C	...	H_n
E_A	x	$\frac{1-x}{n-1}$	$\frac{1-x}{n-1}$	...	$\frac{1-x}{n-1}$
E_B	$\frac{1-x}{n-1}$	x	$\frac{1-x}{n-1}$	...	$\frac{1-x}{n-1}$
E_C	$\frac{1-x}{n-1}$	$\frac{1-x}{n-1}$	x	...	$\frac{1-x}{n-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
E_n	$\frac{1-x}{n-1}$	$\frac{1-x}{n-1}$	$\frac{1-x}{n-1}$	...	x

Figure 2: A generic CPT for independent pieces of evidence.

evidence—thus specifying the topic under discussion. BNs are directed graphs combined with probability distributions, where each node represents a proposition and the edges indicate their dependencies. In NormAN, a BN generates a "ground truth" for each deliberation scenario by stochastically initializing the hypothesis and evidence with specific values. Consider a binary hypothesis H, which can be either true (H) or false ($\neg H$),[5] with a base rate of $P(H) = 0.5$. This means that in a given simulation, there is a 50% chance that H will be true. The probabilities of evidence E (e.g., a test result) are determined by their true positive and false positive rates, $P(E|H)$ and $P(E|\neg H)$. For instance, if $P(E|H) = 0.6$, then when H is true, there is a 60% chance that E will be positive. This represents the simplest case, where the evidence depends solely on the hypothesis. Assaad and Hahn (2024) encode this setup in their "generic" BN, where a binary hypothesis has k independent pieces of evidence (cf. Fig. 1). Each simulation, therefore, generates a set such as $E^1, \neg E^2, \ldots, E^k$, with evidence both supporting and opposing H, varying in diagnosticity. Given the quality of the evidence, its bulk will likely point to the true hypothesis state.

Agents reason about the hypothesis by receiving and interpreting evidence, updating their beliefs via Bayes' Theorem ($P(H|E)$). In NormAN, agents' perceived probabilities align with the world model's BN; they update using the correct probabilities. This makes them ideally rational reasoners: they interpret evidence uniformly and correctly. As the hypothesis in Assaad and Hahn (2024) is binary, agents maintain two beliefs: $P(H)$ and $P(\neg H) = 1 - P(H)$.

Let us extend the world model's BN to capture a MOT. First, we modify the hypothesis proposition: H can now take one of n values, meaning $\mathrm{H} \in H_A, H_B, \ldots, H_n$, where H_i denotes that i is the correct option. MEJE is built into this hypothesis model: just as a binary hypothesis is either true or false (and neither both nor neither), only one option can be correct. Hence, the probabilities assigned by agents to each option must sum to 1: $P(H_A) + P(H_B) + \ldots + P(H_n) = 1$. An agent's beliefs can be represented as an array of length n. For example, if $n = 3$, $\texttt{beliefs} = [0.2, 0.5, 0.3]$ represents $P(H_A) = 0.2$, $P(H_B) = 0.5$, and $P(H_C) = 0.3$.

We now turn to the evidence. We adopt Assaad and Hahn's simple causal structure (Fig. 1), where each piece of evidence is conditionally independent. However, in our model, evidence E can take n values. Formally, $\mathrm{E} \in E_A, E_B, \ldots, E_n$, where E_i means "Evidence E supports the truth of H_i." This added flexibility comes with increased complexity: defining the corresponding probability distribution requires constructing an $n \times n$ conditional probability table (CPT) for each piece of evidence, specifying the likelihood of each evidence state E_i given hypothesis H_j, $P(E_i|H_j)$. To simplify, we assume that evidence has a true positive rate, or more generally, a fixed hit-probability of $P(E_i|H_i) = x$, while the probability of an erroneous indication is evenly distributed across the other options: $P(E_i|H_{j\neq i}) = \frac{1-x}{n-1}$ (cf. Fig. 2). For now, each piece of evidence follows this same CPT,[6] though we can relax this assumption. However, it helps us address the question: What is the status of a newly added option?

This will depend on the prior probability of the new option, $P(H_{n+1})$, and the likelihood of evidence supporting it. If we move from n to $n+1$ options while keeping the prior probabilities of the original n options fixed, the new option's prior must be zero, meaning it will not be considered in the model. For a new option to be genuinely viable, it must be initially plausible, and there must be a chance of evidence supporting it. Hence, we assume that the new option has the same prior as all others: $P(H_i) = \frac{1}{n}$ for all i. This goes both for the probability that an option is correct in the world model and for the agents' prior beliefs, which are calibrated in NormAN.

What about the evidence, characterized by its hit-probability x? As we add more options, a fixed x makes the evidence relatively stronger. Intuitively, as n increases while x remains fixed, the evidence becomes more effective at distinguishing the correct option from each incorrect one. For instance, with two options and a hit-probability of $x = 0.6$, a piece of evidence is relatively weak. But with 10 options, $x = 0.6$ makes evidence far more diagnostic than in the two-option case, and therefore significantly stronger. This is reflected in a rational agent's updates: stronger evidence leads to greater belief updates. Hence, keeping x fixed as n increases makes it easier to determine the correct option.

To keep evidence strength constant, we normalize $x = \frac{2}{n+1}$. This ensures that the evidence remains equally effective in distinguishing the correct option from any incorrect one, re-

[5] We denote propositional variables in roman script and their values in italics.

[6] Though they can and likely will take different values.

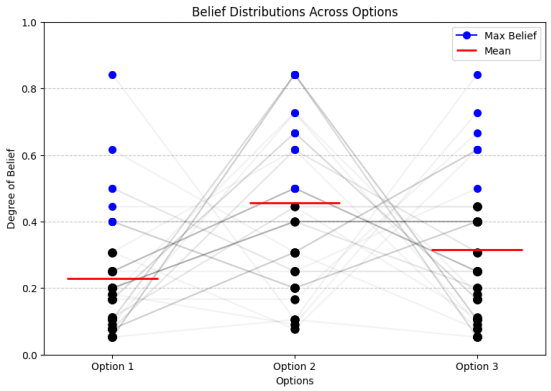


Figure 3: The beliefs of 50 agents about 3 options, based on subsets of five pieces of evidence each.

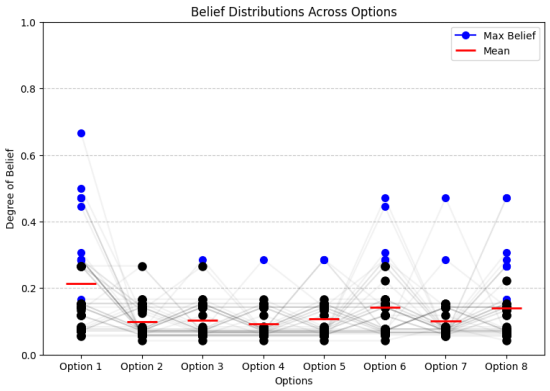


Figure 4: The beliefs of 50 agents about 8 options, based on subsets of five pieces of evidence each.

gardless of the number of options. As a result, the size of the updates to an agent's belief in the indicated option remains consistent. This is achieved by the pairwise likelihood ratio, $\frac{P(E_i|H_i)}{P(E_i|H_{j\neq i})}$, remaining constant as n increases.[7]

By normalizing x in this way, we introduce a viable new option, H_{n+1}, in a principled manner: it has the same prior probability as the existing options and can be supported by evidence of equal strength. Of course, this does not mean that in real scenarios, adding a new option (e.g., in a court of law; Hahn, 2020) necessarily introduces a viable option as we defined it here. Given our idealizations, we note that the general findings in the next subsection have been robustness-checked by randomly varying the CPTs of each piece of evidence within the bounds of probabilistic rules (available on OSF, cf. footnote 7). In these settings, each evidence piece had a different CPT (and therefore varied in strength). Our qualitative results hold, but this is unsurprising: as we will see, what drives our result is that the new option has some prior probability and that there is likely some evidence for it.

Measuring Multi-Option Polarization: A Dual Effect

We now have a model of MOT: a focal hypothesis H takes one of n values[8] and stochastically generates a set of evidence, each supporting one of the n options. When moving from n to $n+1$ options, the new option is assigned the same prior probability and is likely to be supported by some evidence, making H_{n+1} a viable new option.

Recall, the mechanism we explore is how unequal access to evidence can lead to polarization. We can now investigate this in a MOT context. Using the topic model above, we simulate a hypothesis with 50 pieces of evidence. This produces a sequence of 50 "tests," pointing toward different options (e.g., $E_A^1, E_B^2, E_C^3, \ldots, E_A^{50}$). Then, we let 50 agents each draw 5 pieces of evidence at random from this set. That is, each agent likely has a different subset. We repeat this process as we increase n and plot the agents' beliefs.

Figures 3 and 4 display the beliefs of 50 agents from example simulations: the left panel for $n = 3$ and the right panel for $n = 8$. Each dot represents an agent's belief (y-axis) in an option (x-axis); connected dots indicate individual agents' `belief` arrays. The red bar shows the population's mean belief in that option. Finally, blue dots mark each agent's highest belief—their favored option. Let us call an agent's favored option their "position."

Two things are immediately apparent. First, the belief "curves" of agents are flatter in the $n = 8$ case. This is natural: adding a viable option makes it likely that there will be at least some evidence for H_{n+1}. Since agents' beliefs must sum to 1, receiving evidence for multiple options spreads their probability mass "thinner." As a result, the mean beliefs in each option are lower, and each agent's highest belief (blue) also decreases on average. Importantly, this flattening means that agents' beliefs about any particular option are less dispersed in the $n+1$ case than in the n case.

Secondly, moving from n to $n+1$ leads to greater diversity in favored options. In the $n = 3$ case, each of the three options is defended by some agents as their preferred choice, resulting in three distinct positions. Meanwhile, in the $n = 8$ case, there are seven positions, indicating a broader range of preferred candidates. More generally, as n increases, agents' belief curves not only become flatter, but the diversity of their positions—the maxima of their curves—also grows.

Thus, increasing the number of options has a dual effect. First, the growing diversity of positions makes it less likely that agents will agree on a single preferred option. In a social choice scenario, the question "Who will you vote for?" would receive a broader range of answers. Additionally, the size of groups of agreeing agents shrinks, making it more likely that

[7] We prove this in an appendix, available on The Open Science Framework (OSF) along with our model code, simulation data and analysis scripts: `https://osf.io/p7x6d/?view_only=42a9fa92c1c544ecb0f37a38d25427ab` (Please copy the link carefully into your browser).

[8] Each round, there is a designated "ground truth." Each option is equally likely: the probability of H taking any given value is $(1/n)$, and the agents are aware of these "base rate" probabilities.

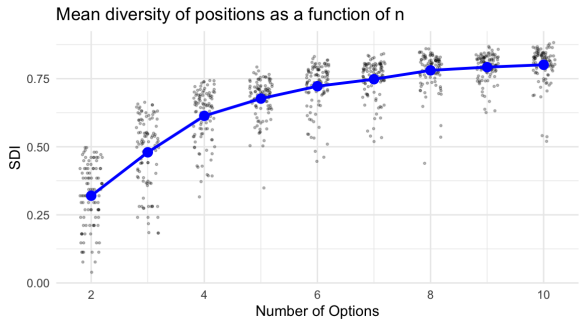


Figure 5: Drawing different subsets: Mean diversity of positions (*SDI*) as a function of the number n. Blue points indicate mean values, gray dots indicate individual simulations

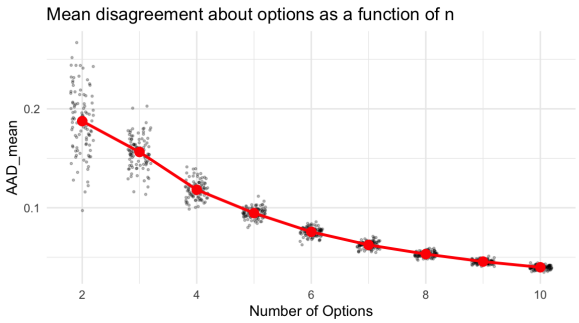


Figure 6: Drawing different subsets: Mean disagreement about options (AAD_{mean}) as a function of the number n. Red points indicate mean values.

agents will disagree about which option is most likely.

On the other hand, as the number of options increases, agents' beliefs become less extreme on average—including in their favored option. While they may still disagree on the most likely option, their degrees of belief are less polarized, as reflected in the narrower dispersion of belief values. The consequence of adding new options is: while it becomes harder for agents to converge on a single favored choice, it also reduces extremism in their beliefs, making disagreements about individual options less pronounced.

Is this polarization? There is no single agreed-upon way to measure it, even in binary cases (cf. Hahn, Merdes, & von Sydow, 2024, Appendix). Different measures capture different aspects of "polarization," and this becomes even more complex in multi-option contexts (MOT), where beliefs are represented as sets of probabilities rather than single values.[9]

However, our discussion clearly highlights two conceptions from the binary literature. First, the idea that a diversity of positions reflects polarization: when agents fail to agree on a single option, they are considered polarized (cf. Weatherall and O'Connor (2021a); O'Connor and Weatherall (2018), models where agents must choose between two options). Greater diversity—i.e., more numerous but smaller, distinct camps—corresponds to less agreement and, therefore, to polarization. One way to measure this notion of polarization is Simpson's Diversity Index (*SDI*), which calculates the probability that two randomly chosen agents have the same position (Simpson, 1949; McDonald & Dimmick, 2003). Formally, $SDI = 1 - \sum_{i=1}^{n} p_i^2$, where n is the number of options and p_i is the proportion of agents who favor option i. A higher value indicates greater diversity.[10]

The second way to measure multi-option polarization—namely, as the extremity of disagreement—is based on common dispersion measures used in binary contexts (e.g., Pallavicini et al., 2021; Angere & Olsson, 2017), applied separately to each option. Measures of statistical dispersion assess how far apart beliefs are. They reach their maximum when half of the population believes $P(H) = 1$ and the other half believes $P(H) = 0$—perfect bi-polarization—and their minimum at perfect consensus. We use the average absolute deviation from the mean (AAD) to measure the dispersion of beliefs for each option i, denoted as AAD_i: "How dispersed are beliefs in H_i?"[11] Overall dispersion is then measured as $AAD_{mean} = \frac{1}{n}\sum_{i=1}^{n} AAD_i$. Higher values mark more extreme disagreements (cf. Bramson et al., 2017, p. 120).

Figures 5 and 6 display the aggregate *SDI* and *AAD* values from our simulations, where 50 agents draw subsets of 5 pieces of evidence each for increasing values of n (each setting was run 100 times). More options lead to greater diversity of positions but less extreme disagreements.

A Social Simulation

We have shown that receiving different subsets of evidence can polarize rational agents in multi-option contexts. However, we have not yet examined how such unequal subsets arise through social deliberation. As discussed earlier, models highlight multiple plausible mechanisms for this, such as agents remembering different sets of evidence (Singer et al., 2019) or preferential attachment in social networks (Mäs & Flache, 2013). Assaad and Hahn's (2024) model showcases another simple mechanism: agents only share their "best," most impactful evidence. We conjecture that any mechanism that yields unequal evidence subsets can also give rise to multi-option polarization and reproduce the effect shown above. While a fuller exploration of social dynamics in MOT settings remains for future work, we illustrate the point by replicating the simple simulation study from Assaad and Hahn's (2024) model using our MOT framework.

Fifty agents are placed on a social network of varying den-

[9] One approach is to measure the overall distance between probability distributions, such as the Jensen-Shannon divergence (a measure of the difference between probability distributions, cf. Nielsen, 2019) or calculating the average pairwise Euclidean distance.While our simulation model (OSF) tracks these distances, we do not use or interpret them in this paper.

[10] The index ranges from 0 (all agents share the same position) to $1 - (1/n)$ (evenly distributed positions).

[11] $AAD_i = \frac{1}{N}\sum_{j=1}^{N} \left| P^j(H_i) - \bar{P}(H_i) \right|$ where $P^j(H_i)$ is agent j's belief in H_i and $\bar{P}(H_i)$ is the population's mean belief in option H_i. AAD_i can take values from 0 to to 0.5.

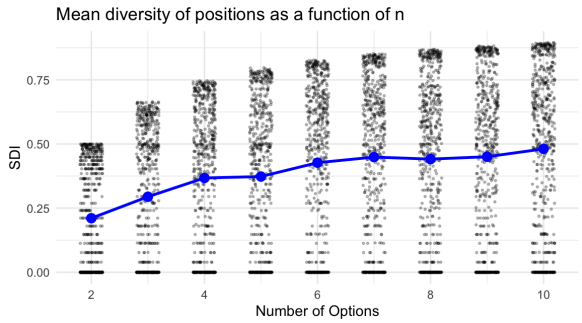


Figure 7: Social deliberation: Mean diversity of positions (*SDI*) as a function of the number n. Blue points indicate mean values, gray dots indicate individual simulations

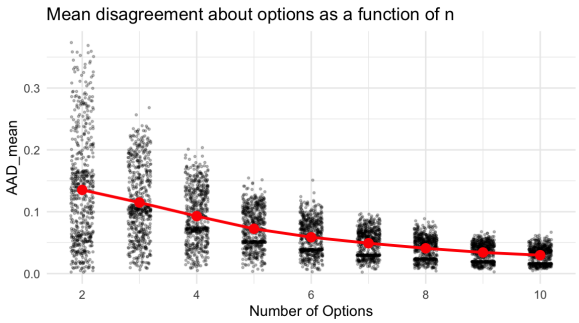


Figure 8: Social deliberation: Mean disagreement about options (AAD_{mean}) as a function of the number n. Red points indicate mean values.

sity. At the start of the simulation, each agent is assigned one of the 50 stochastically initialized pieces of evidence, which can support any of the n options. To deliberate, agents only share the most impactful evidence in favor of their position. Formally, this is the piece of evidence that maximally increases their degree of belief in their preferred option.[12] For robustness, we explored a full range of network connectivity. Using a small-world network (Watts & Strogatz, 1998), we adjusted the mean neighborhood size by varying the parameter k (with a rewiring probability of 0), transitioning from a disconnected "null" network to a nearly complete network.[13] We increased the number of viable options n from $n = 2$ to $n = 10$, running each simulation setting 100 times (6,000 simulations in total), each time until agent beliefs stabilized.

Figures 7 and 8 visualize the effects of increasing n on both types of polarization discussed. Figure 7 shows that the mean diversity of positions (as measured by *SDI*) increases as n increases. Meanwhile, the mean disagreement about any individual option (as measured by AAD_{mean}) decreases (Fig. 8).

The aggregate values of the dual effect are slightly muted by the social dynamics. This is due to our inclusion of highly connected networks: in (almost) full networks (each of the 50 agents has 48 neighbors), even preferential sharing leads to homogeneous evidence subsets. Connected to (almost) everyone, all agents receive the same evidence from the first senders, hence likely converging on the supported position. Once converged, agents preferentially share evidence supporting that position. As a result, all agents come to know the same subset of evidence supporting that option; their evidence subsets become homogeneous, and disagreements decrease. This outcome is less likely in sparse networks, where preferential sharing leads to pronounced position diversity and belief dispersion. We leave a detailed exploration of deliberative dynamics for future work, noting here that the identified effect remains robust in this simple social scenario.

[12]The impact of evidence E_j on option H_i is $P(H_i|E_j) - P(H_i)$. If multiple pieces of evidence have the same impact, agents will only send one (a feature which can be disabled in the simulation).

[13]Specifically, as in the original study, we used $k = 0$, $k = 1$, $k = 2$, $k = 5$, $k = 10$, and $k = 24$. Each agent has $2 \cdot k$ neighbors.

Conclusion: What Really Is Multi-Option Polarization?

Our model shows that adding more options to group deliberation has a dual effect on agent beliefs: while reaching consensus becomes harder, disagreements about individual options become less extreme. This outcome stems from (a) all options being initially probable, (b) each option likely having some supporting evidence, (c) agents receiving different pieces of evidence, and (d) only one option being correct at a time. Such conditions are (arguably) common in decision problems faced by epistemic communities debating the "right thing to do." However, our model has limitations that point to directions for future research. First, it does not account for the qualitative similarity between options, which can influence the importance of disagreements in positions. Additionally, the assumption that only one option can be correct may not always hold. Both aspects should be explored further using more advanced Bayesian networks.

Philosophically, our study highlights how measure-sensitive multi-option polarization is. Using two measures based on binary polarization intuitions, we found that they come apart: depending on one's perspective, adding more options may either reduce polarization or increase it. Alternatively, one might find a completely different measure of polarization more appropriate, as different measures will be better suited to different contexts. Thus, our study shows that further work is needed to better understand multi-option polarization, both in refining models and in deepening our conceptual grasp of the question: "How polarized is a group deliberating more than two options?"

Acknowledgments

I gratefully acknowledge funding provided by the Konrad-Adenauer-Stiftung (PhD scholarship) and the support of the Chair of Philosophy of Science at the Munich Center for Mathematical Philosophy (LMU). I am grateful to Ulrike Hahn, Stephan Hartmann and Alexander Reutlinger for their invaluable feedback and support. I thank Rafael Fuchs and Klee Schöppl for insightful conversations on this topic.

References

Angere, S., & Olsson, E. J. (2017). Publish late, publish rarely!: Network density and group performance in scientific communication. In *Scientific collaboration and collective knowledge* (pp. 34–62). Oxford University Press.

Assaad, L., Fuchs, R., Jalalimanesh, A., Phillips, K., Schöppl, K., & Hahn, U. (2023). A bayesian agent-based framework for argument exchange across networks. *arXiv preprint arXiv:2311.09254*.

Assaad, L., & Hahn, U. (2024). Rational polarization: Sharing only one's best evidence can lead to group polarization. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Bovens, L., & Hartmann, S. (2003). *Bayesian Epistemology*. Oxford University Press.

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

Collins, P. J., Hahn, U., Von Gerber, Y., & Olsson, E. J. (2018). The bi-directional relationship between source characteristics and message content. *Frontiers in psychology*, *9*, 317842.

de Canson, C. (2024). On algebra relativisation. *Mind*, fzae052.

Freeborn, D. P. W. (2024). Rational factionalization for agents with probabilistically related beliefs. *Synthese*, *203*(2), 46.

Hahn, U. (2020). Argument quality in real world argumentation. *Trends in cognitive sciences*, *24*(5), 363–374.

Hahn, U., Harris, A. J., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic*, *29*(4), 337–367.

Hahn, U., & Hartmann, S. (2020). Reasonable doubt and alternative hypotheses: A bayesian analysis.

Hahn, U., Merdes, C., & von Sydow, M. (2024). Knowledge through social networks: Accuracy, error, and polarisation. *Plos one*, *19*(1), e0294815.

Hegselmann, R., Krause, U., et al. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, *5*(3).

Henderson, L., & Gebharter, A. (2021). The role of source reliability in belief polarisation. *Synthese*, *199*(3), 10253–10276.

Jern, A., Chang, K.-m., & Kemp, C. (2009). Bayesian belief polarization. *Advances in neural information processing systems*, *22*.

Jern, A., Chang, K.-M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological review*, *121*(2), 206.

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision making*, *8*(4), 407–424.

Kopecky, F. (2024). Argumentation-induced rational issue polarisation. *Philosophical Studies*, *181*(1), 83–107.

List, C., & Goodin, R. E. (2001). Epistemic democracy: Generalizing the condorcet jury theorem.

Mäs, M., & Flache, A. (2013). Differentiation without distancing. explaining bi-polarization of opinions without negative influence. *PloS one*, *8*(11), e74516.

McDonald, D. G., & Dimmick, J. (2003). The conceptualization and measurement of diversity. *Communication Research*, *30*(1), 60–79.

Nielsen, F. (2019). On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, *21*(5), 485.

Olsson, E. J. (2013). A bayesian simulation model of group deliberation and polarization. *Bayesian argumentation: The practical side of probability*, 113–133.

Olsson, E. J. (2020). Why bayesian agents polarize. In *The epistemology of group disagreement* (pp. 211–229). Routledge.

O'Connor, C., & Weatherall, J. O. (2018). Scientific polarization. *European Journal for Philosophy of Science*, *8*, 855–875.

Pallavicini, J., Hallsson, B., & Kappel, K. (2021). Polarization in groups of bayesian agents. *Synthese*, *198*, 1–55.

Schöppl, K., & Hahn, U. (2024). Exploring effects of self-censoring through agent-based simulation. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Simpson, E. (1949). Measurement of diversity. *Nature*, *163*.

Singer, D. J., Bramson, A., Grim, P., Holman, B., Jung, J., Kovaka, K., ... Berger, W. J. (2019). Rational social and political polarization. *Philosophical Studies*, *176*, 2243–2267.

Sprenger, J., & Hartmann, S. (2019). *Bayesian philosophy of science*. Oxford University Press.

Stegenga, J. (2013). An impossibility theorem for amalgamating evidence. *Synthese*, *190*, 2391–2411.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, *50*(3), 755–769.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature*, *393*(6684), 440.

Weatherall, J. O., & O'Connor, C. (2021a). Conformity in scientific networks. *Synthese*, *198*(8), 7257–7278.

Weatherall, J. O., & O'Connor, C. (2021b). Endogenous epistemic factionalization. *Synthese*, *198*(Suppl 25), 6179–6200.

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

Šešelja, D. (2023). Agent-Based Modeling in the Philosophy of Science. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2023 ed.). Metaphysics Research Lab, Stanford University.