Rational Polarization: Sharing Only One's Best Evidence Can Lead to Group Polarization

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

Ulrike Hahn (u.hahn@bbk.ac.uk)
Centre for Cognition, Computation, and Modelling, Birkbeck, Univ. of London
Malet Street, London, WC1E 7HX, UK

Abstract

Contemporary formal models aim to capture group polarization as the result of deliberation between rational agents. Paradigmatic models do, however, rely on rather limited agents, casting doubt on the conclusion that group polarization can be rationally reconstructed. In this paper, we use a recently developed Bayesian agent-based model of deliberation to investigate this conclusion. This model avoids problems we identify in a group of influential Bayesian polarization models. Our case study shows that a simple mechanism produces realistic patterns of group polarization: limited exchange of evidence across a sparse social network. We reflect on what our results mean for our formal understanding of rational group polarization.

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

Why do groups polarize?

It seems to be a ubiquitous fact that groups may fail to reach consensus through deliberation. Rather, they may diverge and polarize. A remarkably interdisciplinary field has emerged to explain group polarization. This is unsurprising: reaching consensus is the main goal of deliberation in early theories of deliberative democracy (Landemore & Page, 2015), while extreme polarization is seen as a danger to modern democracies today (Sunstein, 2018).

Why would deliberation, the exchange of arguments and evidence, lead to group polarization? Influential theories explain polarization as the result of epistemically irrational behavior, such as motivated reasoning, cognitive biases, and the formation of secluded epistemic bubbles (Kahan, 2013; Taber, Cann, & Kucsova, 2009; Anderson, 2021). Philosophers and computational sociologists have, in the mean time, attempted to develop rational reconstructions of polarization: that is, understanding polarization without assuming irrationality on the part of the group's agents. To do so, computational (simulation) models have emerged, exploring artificial groups of boundedly rational agents engaged in deliberation about a target hypothesis.

It has, however, proven surprisingly hard to produce models in which "simple and intuitive mechanisms produce patterns that even roughly resemble familiar patterns of polarization" (Bramson et al., 2017, p. 115). This naturally casts doubt on the explanation of group polarization as a rational phenomenon. Further, paradigmatic models resort to rather limited agents to produce divergence, rendering questionable whether they are truly rational (for a recent discussion, see Kopecky, 2023).

This paper uses a recently developed simulation framework, NormAN—short for Normative Argument Exchange across Networks (Assaad et al., 2023)—to pursue a simple reconstruction of group polarization as rational. This framework circumvents important problems we identify in a group of influential Bayesian polarization models. We conduct a case study of the NormAN framework, exploring the roles of social network structure and communication rules. We find that groups in NormAN produce realistic belief dispersion patterns. Polarization emerges and increases as the network structure becomes sparser, and (importantly) if agents do not share the entirety of their evidence, but only what they consider to be the best, most truth-conducive evidence. Conversely, if the network is dense and agents share their entire evidence, polarization disappears. This result holds for a range of plausible evidence distributions representing the "topic" under discussion.

Our case study presents a simple and intuitive mechanism for group polarization among rational agents: limited information exchange across a sparse social network. We discuss the implications of our findings on our formal understanding of group polarization.

Models of group polarization

Consider the following generic situation: a group of agents collect and exchange information (in the form of evidence, arguments or testimony) to deliberate whether a certain hypothesis H is true or false. Each forms a degree of belief $P(H) \in [0,1]$ about the hypothesis based on the received information. This scenario, so the modelling literature assumes, captures the essential form of real group deliberation.

Following the formal literature, we understand polarization as belief dispersion induced by deliberation (cf. Bramson et al., 2017, p. 120). If the group converges on one degree of belief, we speak of convergence on consensus. Any other belief distribution can be expressed as belief dispersion of varying degrees. For instance, perfect bi-polarization (i.e., one half of the group believing 1, the other 0) is extreme polarization, consensus means that there is no dispersion, and everything in between we denote as polarization of varying degrees. How do computational models reconstruct polarization among rational agents? There are a plethora of models (e.g., Axelrod,

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

1997; Hegselmann & Krause, 2002; Singer et al., 2019), but in this brief paper, we will focus on Bayesian agent-based models in social epistemology. Bayesian reasoning has long been developed as a standard of rational reasoning, and therefore provides a fitting theoretical framework for modelling rational agents (Hartmann, 2021). Two influential model families have emerged in the Bayesian literature: bandit models and source reliability simulation models.

Bandit models, developed by economists Bala and Goyal (1998) and popularized in philosophy by Zollman (2007), use one or multi-armed bandits where each "arm" represents a theory. Agents test these theories' success rates, aiming to identify the arm with the highest expected payoff. They compute beliefs about the best theory through Bayesian conditionalization based on their own outcomes and those of their neighbors. The standard two-armed bandit models used in Zollman (2007, 2010) do not lead to polarization (correct and incorrect consensus are the only stable states). However, Weatherall and O'Connor introduce modifications facilitating polarization. They include a conformity bias parameter (Weatherall & O'Connor, 2021); another extension has agents perceive evidence from disagreeing neighbors as increasingly uncertain (O'Connor & Weatherall, 2018). Neither extension seems particularly rational (cf. 2018, p. 857).

In bandit models, agents exchange evidence directly, while another class of Bayesian models involves agents exchanging testimony about their beliefs. Testimony's impact is governed by source reliability models, which estimate a source's trustworthiness based on the receiving agent's current belief. If a source's report aligns with the agent's belief (i.e., their expectation), trust increases; if they are contradicted, trust decreases. Trusted sources have more influence on the recipient's degree of belief. This expectation-based updating strategy is supported by empirical research (Collins, Hahn, Von Gerber, & Olsson, 2018), and source reliability models, simple Bayesian networks, implement this strategy. In Olsson and Angere's Laputa simulation (2010; 2011), the first of its kind, agents receive information in the form binary testimony about the hypothesis ("yes"/"no" reports). Both social exchange and non-testimonial inquiry (regulated by self-trust) are modeled using embedded source reliability models. Many subsequent studies have explored this approach (Hahn, Merdes, & von Sydow, 2024; Fränken & Pilditch, 2021; Pallavicini, Hallsson, & Kappel, 2021; Merdes, Von Sydow, & Hahn, 2021). A recently identified problem with these models (and testimony in general) is the difficulty of recognizing dependencies between related sources of information (Hahn, 2023). Two trusted sources (e.g., peers) may both provide a positive report, leading the receiving agent to update twice. However, these two sources' reported beliefs might be based on the same set of evidence, leading the receiving agent to "double-count" this evidence. Double counting is hard to avoid when testimony, rather than evidence exchange carries deliberation in a complex social world. In contrast, explicitly shared evidence can be indexed and recognized, preventing double-counting. In Laputa, double-counting of evidence is aggravated by the dynamics of expectation-based updating: agents begin to trust only agreeing peers, distrust disagreeing ones, and effectively build isolated communities. Within those communities, they repeatedly reaffirm each other's beliefs, losing track of the underlying evidence. This leads Laputa's agents to quickly move to maximal degrees of belief of $P(H) = 1$ or $P(H) = 0$. As Pallavicini et al. (2021) show, populations often escalate to bi-polarization. This is a problem, since the model neither produces realistic polarization patterns (no intermediate dispersion), nor accounts for the fact that realistic agents differentiate between pieces of evidence.

In sum, the discussed Bayesian models have a hard time capturing realistic polarization patterns while maintaining the rationality of their agents. Polarization models from other traditions also typically rely on limiting the agents' willingness to engage with peers that they strongly disagree with (e.g. Hegselmann & Krause, 2002). Of course, there are many other models, some of which use different mechanisms, such as limiting the agents' memory (e.g. Singer et al., 2019).

The NormAN framework

The Bayesian models discussed suggest that polarization can arise from deliberation among boundedly rational agents. However, their reliance on limited agents raises doubts as to whether they really reconstruct polarization rationally. To attempt such a reconstruction, we use the NormAN framework (short for Normative Argument Exchange across Networks, developed in Assaad et al. (2023)). NormAN follows on from earlier models of argument exchange such as Mäs and Flache (2013), taking inspiration from Olsson (2013) and Zollman (2007). It addresses the previously discussed limitations: agents exchange evidence directly, interpret evidence uniformly (without trust updating), recognize indexed evidence, avoid double-counting and have perfect memory.

NormAN 1.0 is a freely available[1] agent-based simulation environment based on three sub-models: a ground-truth world model, individual agents, and a social network across which these agents communicate. The world model specifies the probabilistic relationship between the central hypothesis and certain evidence propositions and determines the truth values of all propositions (both evidence and hypothesis). Agents become aware of the true values of these pieces of evidence, either through individual inquiry, or through communication. Communication is represented by choosing and directly passing on specific pieces of evidence, i.e., argument exchange. Based on the collected evidence, agents form a degree of belief about the hypothesis, which determines how they communicate with their link-neighbors. As an example: if the hypothesis is that a patient has lung cancer, then she probably suffers from shortness of breath, and receives a positive X-ray test (evidence). These propositions are actually true or false in the world. Deliberating agents discover the true values of the evidence propositions, inform each other

[1] https://github.com/NormAN-framework/base-model

about them and use them to compute a belief in the hypothesis.

The ground truth world model is represented by a Bayesian belief network (BN), which specifies the probabilistic connections between two-valued propositions. The modeler chooses one such proposition as the hypothesis H, and a subset as pieces of evidence $E_1, \ldots, E_n$. In each given simulation run, the world model initializes the hypothesis as true or false ($H, \neg H$), and uses the marginal conditional probability of each piece of evidence to stochastically fix their truth values. The result is a string of evidence pieces that are individually true or false, which are (in principle) available to agents and constrain deliberation (e.g., $E_1, \neg E_2, \ldots, \neg E_n$).

Agents collect pieces of evidence by adding the respective truth values to their memory. They form beliefs by aggregating these received pieces of evidence via Bayes' rule (e.g., $P(H|E_1, \neg E_4)$). To do so, they draw on a BN, which, in version 1.0 of NormAN, is a veridical copy of the world model's BN. Hence, agents are not only calibrated to the world model, they all interpret each piece of evidence in the same way: if two agents receive the same set of evidence, they will form the same belief. Each agent starts with the same prior degree of belief $P(H)$. This simple setup can be interpreted as the agents meeting the uniqueness standard of rationality, the position that any body of evidence justifies at most one doxastic attitude toward the target claim (White, 2019).

NormAN agents receive evidence in two ways: through individual inquiry or communication with link-neighbors. Upon receiving new evidence, agents compute a new posterior belief based on their entire evidence set, avoiding order effects. Agents have no memory limitations and never forget evidence; they only add indexed pieces of evidence they have not received before, avoiding any double counting.

Communication is modeled by agents directly passing on a piece of evidence, making the receiver aware of the evidence node's truth value. Which piece of evidence the agent chooses to communicate varies according to the used communication rule. NormAN 1.0 allows the user to explore three different sharing rules: the "random" rule has agents pass on a randomly chosen piece among the previously collected evidence. The "recency" rule makes agents choose recently encountered pieces with a heightened probability. In particular, they share the most recently encountered piece with a probability of x (which in the base model is 0.9), or choose another random piece from their memory (probability of $1-x = 0.1$) (this rule is inspired by Mäs and Flache (2013)).

Finally, the "impact" rule has agents choose what they consider to be their most impactful evidence in favor of their current position. An agent's position is determined by whether their current belief is greater or smaller than the initial belief (i.e., when $P(H|E_i, \ldots, E_k) > P(H)$, they support H). They assess the impact of their collected pieces of evidence by how much they will sway an agent from their prior belief. Hence, evidence E_i's impact is $P(H|E_i) - P(H)$. If an agent's current position is that the hypothesis is true, they share the piece of evidence with the highest positive impact, if their position is that the hypothesis is false, they share the piece of evidence with the greatest negative impact value (and if their belief is equal to their prior, they do not share). This simple impact rule was designed to reflect the idea that speakers seek to communicate what they consider to be the most relevant (and truth-conducive) piece of information.

Lastly, agents can only communicate with one another across symmetrical links determined by a social network. NormAN 1.0 incorporates paradigmatic, modifiable network structures (Watts & Strogatz, 1998), which we will use in our case study.

Polarization case study

Assaad et al. (2023) merely hint at a polarization study in NormAN. This paper presents a comprehensive study of polarization in the NormAN framework, investigating the effects of different social network connectivities and communication rules based on different sets of evidence. NormAN 1.0 already provides the tools to modify connectivity and different sharing styles; we add a new modifiable world model to create a range of plausible evidence distributions.[2]

The generic network

Certain hypotheses are harder to investigate than others: some provide a very conclusive set of evidence, others inconclusive or even misleading evidence. In this case study, we explore how communication rules and social connectivity interact with evidence sets of varying conclusiveness to produce polarization. To this end, we embed a modifiable Bayesian network into NormAN 1.0, which captures the following generic situation (Figure 1): a hypothesis H has ten pieces of (conditionally independent) evidence $E_1, \ldots, E_{10}$. This Bayesian network will serve as the world model and as the agents' representation thereof. The ten pieces will constitute arguments in favor or against the hypothesis (varying in strength), which determine the set's overall conclusiveness. To measure conclusiveness, we use NormAN's "optimal posterior", $op := P(H|E_1, \ldots, E_{10})$. This is the posterior degree of belief agents would compute if they knew the entire evidence. Hence, given H is true, sets with $op = 0.5$ are maximally inconclusive, and they become gradually more conclusive as their op approaches 1.

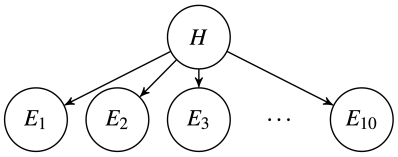

Figure 1: Generic network

[2]Our model code, simulation data and analysis scripts to replicate our results are publicly available on The Open Science Framework (OSF) at https://osf.io/8qz9c/?

Using this generic network, we will assume that the hypothesis is true. Further, we fix $P(H) = 0.5$ (that is, agents are initially completely undecided) and initialize all pieces of evidence as true in the world. To change whether a piece of evidence is interpreted as an argument against/in favor of H, we modify its conditional probabilities. This gives us perfect control of the set's conclusiveness.[3] Generally, whether or not a piece of evidence E_i confirms H will (in Bayesian terms) depend on its likelihood ratio $x_i := P(E_i|H)/P(E_i|\neg H)$. If $x_i > 1$, then E_i confirms the hypothesis, i.e., $P(H|E_i) > P(H)$; if $x_i < 1$, then E_i disconfirms H, and if $x_i = 1$, then E_i is not diagnostic of H.[4] For instance, if $x_k = 4$ and $P(H) = 0.5$, then E_k produces $P(H|E_k) = 0.8$, and is therefore a stronger piece of evidence than one which would only raise an agent's posterior to 0.7.

To create sets of different op (i.e., conclusiveness) we start with a maximally inconclusive set: five evidence nodes confirm the hypothesis, five disconfirm it, and both sides perfectly offset one another, i.e., $op = 0.5$. In particular, the individual posteriors induced by each piece of evidence range from $P(H|E_1) = 0.1$ to $P(H|E_{10}) = 0.9$. This makes for a diverse set of evidence: there are arguments of varying strength both in favor and against H. Building on this baseline, by increasing each piece of evidence's likelihood ratio by a factor γ, we create five more sets of evidence ranging from inconclusive ($op = 0.5$) to very conclusive (op of 0.6, 0.7, 0.8, 0.9 and 0.990). Raising γ makes arguments against the hypothesis less disconfirmatory and those in favor more confirmatory, and as a result, the body evidence becomes more conclusive (Figure 2).[5]

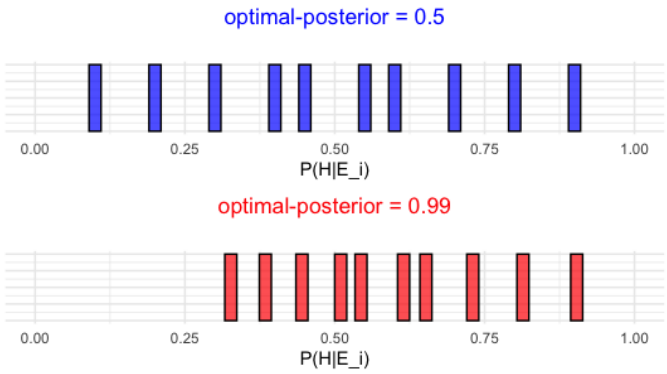

Figure 2: The posteriors induced by single pieces of evidence, $P(H|E_i)$ for E_i. The above graph (blue) shows the perfectly balanced "baseline" evidence set. Below is the most conclusive set (red). Cf. Footnote 5.

Setup

To investigate polarization in NormAN, we tweak three crucial aspects of deliberation: communication styles, social networks and the topic (evidence). The model population consists of 50 agents linked in small-world Watts and Strogatz (1998) networks. By adjusting the mean neighborhood size through varying k (and setting the rewiring-probability to 0), we transition from a disconnected "null" network to nearly complete connectivity (from $k = 0$ to $k = 24$, where every agent has $2 \cdot k$ neighbors). The conclusiveness of evidence is modeled using the six distributions based on the generic network introduced above. Lastly, we explore all sharing styles from NormAN 1.0: random, recency, and impact.

At the beginning of each run, the 10 pieces of evidence are distributed to 10 random agents in the network, while the rest begin with empty memories and the uniform prior belief $P(H) = 0.5$. This setup ensures that each piece of evidence is in principle accessible to the group, and helps track evidence dissemination. Agents are restricted from inquiry, accessing only the initially distributed evidence. Every round, each agent will consider communicating one piece of evidence.

Results

We repeated each parameter configuration 100 times, letting the simulation run until beliefs had stabilized (which happened within 1000 sharing rounds). Polarization was gauged using the Laputa model's measure, the root mean square of the deviation of individual credence from the mean (RMS), which ranges from 0, indicating consensus, to 0.5, indicating perfect bipolarization, i.e., when half of the agents entertain $P(H) = 0$ and the other $P(H) = 1$ (Angere & Olsson, 2017). We also monitored belief spread (i.e., the distance between the highest and lowest belief) and the absolute average deviation from the mean credence (as per Bramson et al. (2017)). The following results are the mean final RMS values of simulation runs (alternative measures are available on OSF).

First, we find that the recency and random sharing rules lead to convergence on the op in every single run: the ten pieces of evidence travel through the entire network, no matter how sparse it is and no matter how conclusive the evidence is, causing the agents to converge on the optimal posterior. Both rules result in the agents fully sharing their entire evidence over the course of the simulation. And since receiving agents neither refuse to communicate due to distrust or disagreement, nor change the interpretation of the evidence, ev-

[3] This also breaks with NormAN 1.0's calibration of agent beliefs and the world model: the likelihoods that the evidence/hypothesis are actually true do not match the agent's perceived likelihoods. As a robustness check, we ran all of our following simulations with perfectly calibrated agents: the agents' beliefs matched the base rate of the hypothesis and the true conditional rates of the pieces of evidence. In these runs, we did not assume H and E_i to be true in each round. This did not change our qualitative findings. The data can be found on OSF (`https://osf.io/8qz9c/?`).

[4] Where disconfirmation is $P(H|E_i) < P(H)$, and lack of diagnosticity is $P(H|E_i) = P(H)$.

[5] On Fig. 2: Generally, if $P(H) = 0.5$, then $P(H|E_i) = x_i/(x_i+1)$. Adding a value γ to x has a smaller effect on the posterior when x is larger. Therefore, adding γ to evidence against H (where $x < 1$) increases the respective posterior more than posteriors created by evidence in favor. E.g., for E_k having $x_k = 4$, $P(H|E_k) = \frac{4}{4+1} = 0.8$. Add $\gamma = 0.1$, then $P'(H|E_k) = \frac{4.1}{4.1+1} \approx 0.804$. On the other hand, E_j having $x_j = 0.25$ creates $P(H|E_j) = \frac{0.25}{0.25+1} = 0.2$. Add $\gamma = 0.1$: $P'(H|E_j) = \frac{0.35}{0.35+1} \approx 0.259$.

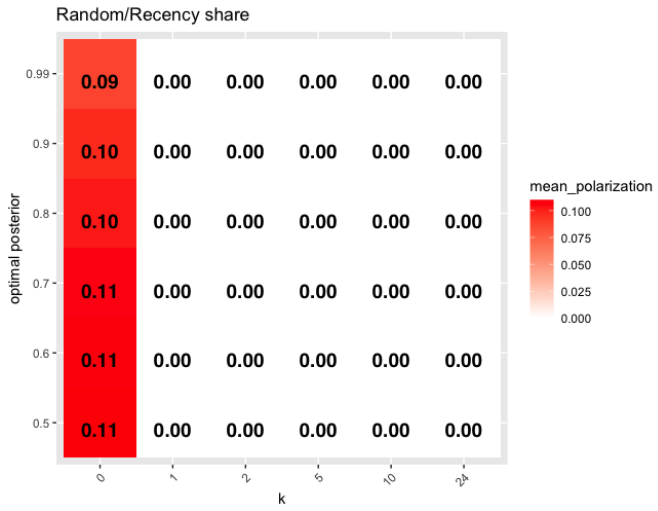

Figure 3: Mean RMS for recency and random share. Polarization disappears through connectivity/communication.

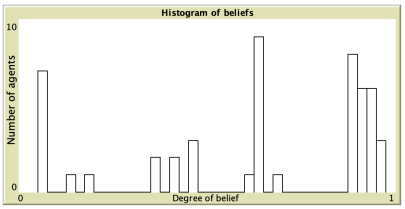

Figure 4: Stable final beliefs in a run using $op = 0.8$, $k = 2$ and the impact sharing rule.

ery agent eventually receives all available evidence and computes the same belief. This result is reminiscent of the convergence observed in the classical bandit model: if the agents exchange the evidence directly and unaltered, they will converge. The resulting consensus belief being determined by the op means that although the agents will converge, the veritistic value of the consensus (i.e., how close this belief is to the truth of H, cf. Goldman (1999)) depends on the evidence. Figure 3 shows this: As long as there is a network connecting all agents ($k > 0$), the RMS drops to 0 via to deliberation.

Polarization merely emerges once agents use impact share: they only communicate the piece of evidence they currently hold to best support their belief. The mechanism is simple: meaningful exchange between two neighbors ceases once the sender has shared what they consider their most impactful evidence. Even if the sender repeats their message, the receiver, having already absorbed the information, disregards it. Communication only resumes if the sender obtains stronger evidence or alters their position—an event that becomes increasingly unlikely: as agents find themselves amidst like-minded peers, the likelihood of encountering transformative evidence diminishes, leading to stagnant communication and stable belief clusters.

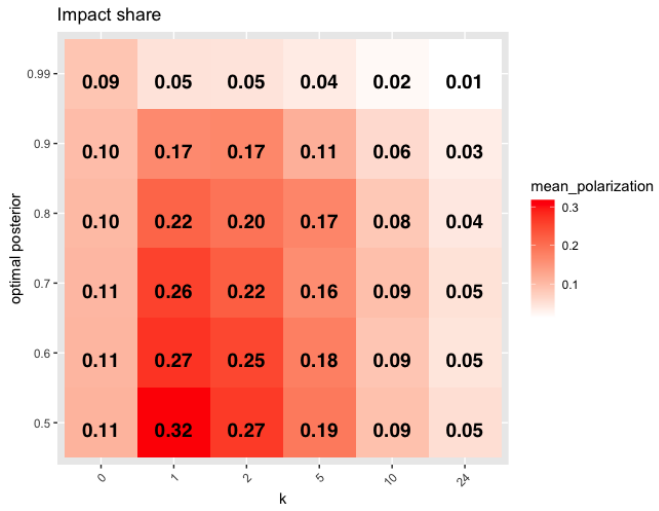

Figure 5: Mean RMS for impact share. Impact sharing produces polarization when the network is sparse and the evidence is inconclusive. To contextualize: with an RMS of 0.32 (highest value), beliefs span a 0.9 interval (using the belief spread measure). With an RMS of 0.01, beliefs cover a narrow 0.09 interval around the optimal posterior.

The emergent polarization stands in contrast to models based on trust dynamics, where agents—perhaps unwittingly—self-select into subpar, closed-off epistemic communities such as filter bubbles or echo chambers (e.g., Fränken & Pilditch, 2021). NormAN simulations uncover a different dynamic: divergence arises from genuine communication efforts, where agents share what they perceive as their most compelling evidence, without further intentional filtering. This goes for both senders and receivers, who never refuse or discount any novel, unexpected information. Furthermore, these simulations produce nuanced, realistic polarization patterns. Rather than splitting into two extreme factions, NormAN agents settle on stable, scattered sets of intermediate beliefs across the unit interval (Figure 4).

As can be seen in Figure 5, impact sharing interacts with network density and evidence conclusiveness. The greatest polarization emerges when each agent has only two neighbors ($k = 1$) and evidence is maximally inconclusive ($op = 0.5$). In these scenarios, there is strong evidence for either position (H and $\neg H$), creating polarization potential: agents exposed to different sets of evidence encompassing strong arguments pulling in opposite directions will draw significantly different conclusions. Sparse networks increase the likelihood of fragmented access to evidence: when all one's neighbors have shared their most impactful evidence, communication halts, blocking the social pathways. With fewer paths, complete blockage becomes more likely for agents, leading different parts of the social network to settle on distinct sets of evidence.

Conversely, when agents are highly connected ($k = 24$) and

the evidence approaches perfect conclusiveness ($op = 0.990$), polarization almost vanishes: Even if agents do not perfectly converge on the optimal posterior due to imperfect exposure to the evidence (through incomplete impact sharing), they will on average obtain more evidence via more link-neighbors. Such growing evidence sets will increasingly overlap and generate similar beliefs. This belief convergence is further facilitated if the available evidence is not strongly divisive, which is the case for very conclusive sets.

Conclusion: Is this rational polarization?

This case study highlights that little is needed to produce realistic group polarization patterns. But are the discussed NormAN agents rational, and can we therefore conclude that group polarization can be rationally reconstructed? We believe that NormAN agents are at least more rational than the discussed Bayesian model agents from the literature: they have perfect memory, do not discount any evidence, never lie or pass on subpar information (cf. Douven & Hegselmann, 2021), and interpret the evidence in the same, unique way, using strict Bayesian updating.

However, the Bayesian framework leaves open completely the rationality of communication rules: should agents always fully share their information? In this case, then, our NormAN simulation suggests that group polarization is not rationally reconstructible (at least not in this framework). As other models, our model, too, relied on limiting agents: they will not share their entire evidence.

We believe, however, that this limitation is reasonable: especially as we increase the size of the evidence set, and navigate a complex topic, it will not be possible for realistic agents to share all they know with all their neighbors. The explored type of impact sharing does not necessarily denote a bad-faith communication strategy. Employing the impact rule implies curating what one shares, prioritizing the communication of what one perceives as the most valuable, truth-conducive evidence. Considering the costliness of both communicating and processing arguments, such a strategy may be pragmatically suitable in many contexts.

Another limitation we have imposed on our agents is that they will not, on their own, collect all pieces of evidence. This, we believe, is also a reasonable limitation: in realistic contexts, different agents will have access to different subests of evidence, both due to practical and cognitive reasons.

Though more model exploration (in NormAN and other frameworks) is necessary to cement the following conjecture, our case study strongly suggests that any type of limited information sharing will, on sparse networks, prevent convergence on a perfectly shared set of information. It is then natural (and rational) for agents to come to different beliefs and polarize—especially if the evidence is inconclusive. This simple mechanism is enough to model realistic polarization, and it is perfectly compatible with standard theories of rationality.

Acknowledgements

The research reported in this paper was supported by the UK's Arts and Humanities Research Council grant AH/V003380/1, and a Mercator Fellowship to U.H. through Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 455912038. L.A. was supported by a Konrad-Adenauer-Stiftung scholarship. Special thanks go to Klee Schöppl, Rafael Fuchs, Kirsty Philips, Borut Trpin and Emelia Stanley for helpful discussions.

References

Anderson, E. (2021). Epistemic bubbles and authoritarian politics. *Political epistemology*, 11–30.

Angere, S. (2010). Knowledge in a social network. *Synthese*, 167–203.

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., Schoeppl, L., & Hahn, U. (2023). A bayesian agent-based framework for argument exchange across networks. *arXiv preprint arXiv:2311.09254*.

Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. *Journal of conflict resolution*, *41*(2), 203–226.

Bala, V., & Goyal, S. (1998). Learning from neighbours. *The review of economic studies*, *65*(3), 595–621.

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*, 18.

Douven, I., & Hegselmann, R. (2021). Mis-and disinformation in a bounded confidence model. *Artificial Intelligence*, *291*, 103415.

Fränken, J.-P., & Pilditch, T. (2021). Cascades across networks are sufficient for the formation of echo chambers: An agent-based model. *Journal of Artificial Societies and Social Simulation*, *24*(3).

Goldman, A. I. (1999). *Knowledge in a social world*. Oxford University Press.

Hahn, U. (2023). Individuals, collectives, and individuals in collectives: The ineliminable role of dependence. *Perspectives on Psychological Science*, 17456916231198479.

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

Hartmann, S. (2021). 4.2 bayes nets and rationality. *The Handbook of Rationality*, 253.

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence: models, analysis and simulation.

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

Kopecky, F. (2023). Argumentation-induced rational issue polarisation. *Philosophical Studies*, 1–25.

Landemore, H., & Page, S. E. (2015). Deliberation and disagreement: Problem solving, prediction, and positive dissensus. *Politics, philosophy & economics*, *14*(3), 229–254.

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

Merdes, C., Von Sydow, M., & Hahn, U. (2021). Formal models of source reliability. *Synthese*, *198*, 5773–5801.

Olsson, E. J. (2011). A simulation approach to veritistic social epistemology. *Episteme*, *8*(2), 127–143.

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

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.

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.

Sunstein, C. (2018). *# republic: Divided democracy in the age of social media*. Princeton university press.

Taber, C. S., Cann, D., & Kucsova, S. (2009). The motivated processing of political arguments. *Political Behavior*, *31*, 137–155.

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

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

White, R. (2019). Epistemic permissiveness. *Contemporary Epistemology: An Anthology*, 267–276.

Zollman, K. J. (2007). The communication structure of epistemic communities. *Philosophy of science*, *74*(5), 574–587.

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