Exploring Effects of Self-Censoring through Agent-Based Simulation

Klee Schöppl (l.u.schoppl@rug.nl)
Department of Theoretical Philosophy, University of Groningen
Groningen, 9712 GL NL

Ulrike Hahn (u.hahn@bbk.ac.uk)
Centre for Cognition, Computation, and Modelling, Birkbeck College, University of London
London, WC1E 7HX U.K.

Abstract

Recent years have seen an explosion of theoretical interest, as well as increasingly fraught real-world debate, around issues to do with discourse participation. For example, marginalised groups may find themselves excluded or may exclude themselves from discourse contexts that are hostile. This not only has ethical implications, but likely impacts epistemic outcomes. The nature and scale of such outcomes remain difficult to estimate in practice. In this paper, we use agent-based modelling to explore the implications of a tendency toward 'agreeableness' whereby agents might shape their communication so as to reduce direct conflict. Our simulations show that even mild tendencies to avoid disagreement can have significant consequences for information exchange and the resultant beliefs within a population.

Keywords: Argumentation; Agent-based modelling; Communication; Normative reasoning; Polarization;

Introduction

Dotson (2011) identifies the practice of testimonial smothering, wherein speakers are coerced or pressured by epistemic violence into tailoring their own testimony to the biases and identities of those listening. But even in the absence of being motivated by epistemic injustice (Fricker, 2007), individuals may see mere disagreement as uncomfortable, and as potentially harmful to their social relationships, and choose to adapt their behaviour accordingly (Vraga, Thorson, Kligler-Vilenchik, & Gee, 2015). Possible strategies to avoid conflict might range from strategically choosing what to share to self-censoring through non-participation (Hayes, Scheufele, & Huge, 2006). According to a recent publication (Roos, Utz, Koudenburg, & Postmes, 2022), pragmatic attempts to avoid disagreement in direct communication may lead to lessened perceptions of polarization, and instead convey an inflated impression of agreement.

In this paper, we investigate the possibility of adverse effects of agreeableness. Counterintuitively, widespread agreeable behaviour in argument exchange that aims to minimize disagreement might, in fact, increase actual polarization. The possibility that agents individually and truthfully assert those arguments to their interlocutors which they assume to pose the lowest risk of disagreement may thereby hinder the effective diffusion of truth across a social network on a macro level.

To examine epistemic consequences of agreeableness we conduct agent-based simulations using a recently proposed framework for argument exchange by Bayesian agents (Assaad et al., 2023). ABMs allow insight into emergent consequences of multiply interacting agents (Klein, Marx, & Fischbach, 2018). Consideration specifically of Bayesian agents, who are optimal in their reasoning and evidence aggregation, provides a useful baseline in as much as adverse effects that are observable even here are good candidates for actual real-world problems.

We implement a communication rule for agents wanting to avoid disagreement by selecting for communication, amongst all the facts known to them, only those they expect to be agreeable to their interlocutors. Given that our agreeable agents will only ever assert truths, this is a rather moderate deviation from optimal communication. We consider it compatible with both epistemically idealized versions of testimonial smothering on the one hand, and a mere urge for interpersonal politeness on the other.

Our study of the impacts of this agreeableness targets two different communication contexts: First, we are interested in local, pairwise communication between individuals, engaged in repeated personal dialogues across their social network. Second, we study a context in which agents communicate to all of their interlocutors at once, a form of 'broadcast' found in communication on online social media where messages become available to all followers or all readers of a particular exchange.

The Model

NormAN, short for normative argument exchange across networks, is a recent Bayesian modelling framework introduced to study opinion dynamics in the context of complex argument domains (Assaad et al., 2023). Building on its base version (NormAN version 1.0, referred to as 'base model' in the following), our extension is implemented using *NetLogo* (Wilensky, 1999), its *R*-extension (Thiele & Grimm, 2010), and the R-packages bnLearn (Scutari, 2009) and gRain (Højsgaard, 2012).

The World

The world in NormAN is governed by a causal structure encoded in a Bayes' net[1]: a directed, acyclical graph (DAG) with a conditional probability distribution. Nodes in the objective, world DAG represent random variables, and arrows

[1]For a brief introduction to Bayes' nets and their use in formal epistemology, see Bovens and Hartmann (2004), Chapter 3.

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.*

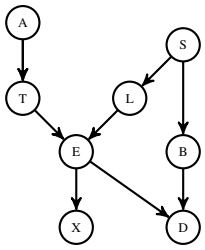

Figure 1: The 'Asia' network (Lauritzen & Spiegelhalter, 1988).

represent causal dependencies between them. One variable in the world-DAG becomes the 'hypothesis' under discussion, and in each run of NormAN, the hypothesis is either true or false, with a prior probability chosen by the modeller.

A number of remaining variables become evidence about the hypothesis. In each run of the model, their truth values are stochastically determined by conditionalizing the entire Bayes' net on the actual value of the hypothesis in this run. As a result, each run of NormAN features a set of diverse pieces of evidence which, in expectation, will allow approximation of the hypothesis' truth value.

For a simple example, consider the DAG in Figure 1: in the so-called 'Asia' network, the L-node represents the variable of whether one specific patient is ill with lung cancer. If this is the hypothesis under discussion, a variety of other, causally relevant variables represented in the DAG become valuable pieces of evidence: whether the patient is a smoker (S), suffering from bronchitis (B) or dyspnoea (D), and whether the patient has recently visited Asia (A), which might have caused them to develop tuberculosis (T), a potential alternative explanation for if their chest x-ray (X) is showing problematic results. On any given run of NormAN, the patient either does or does not suffer from lung cancer, and based on that value, the model generates plausible evidence distributions using the Asia network; in other words, for each run, NormAN uses the conditional probability distribution of the Bayes' net to determine a list of evidence values (such as smoker=true, bronchitis=false etc.) that are true in that world and subsequently form the basis of evidence acquisition and exchange.

The Inhabitants

The model world is inhabited by agents, Bayesian reasoners concerned with finding out the actual truth value of the hypothesis. They entertain a 'subjective mirror' of the objective Bayes' net, meaning they are aware of the actual causal structure of the world, and whether—and to which extent—the various pieces of evidence and the hypothesis hang together. Agents may gain access to the actual values of individual evidence nodes through inquiry, or start a run having pre-drawn any specified number of pieces of evidence.

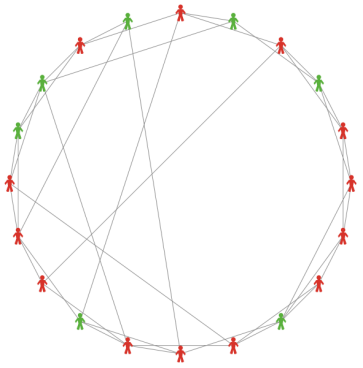

Figure 2: Example population of 20 agents assembled in a Watts-Strogatz (Watts & Strogatz, 1998) 'small world' network with a mean degree $k = 2$ and a rewiring probability $p = 0.2$.

Agents are connected via a network of communication links, which allows the study of a variety of network topology types. See Figure 2 for an example of the kinds of networks we use in the simulations for this paper.

Across their communication links, agents may assert pieces of evidence as arguments for or against the hypothesis under consideration. Crucially, in doing so, they are effectively pointing their neighbours directly at the truth values of evidence nodes in the world, rather than, for example, merely asserting testimony about what they themselves believe to be the case (unlike, for example, in Olsson (2011) and Angere's (2010) *Laputa* model). Following our previous example, a group of agents deliberating whether a patient has lung cancer may assert the observable fact that the patient is a smoker as an argument to support the hypothesis, or the fact that the x-ray results are unproblematic as an argument to attack it.

Given constraints on the amount of information transmittable per communicative exchange, agents must choose which particular argument to assert in a given exchange. In the base model of NormAN, agents communicate at most a single piece of evidence at a time, and we follow this setup here. To determine agent choice of what to communicate at a given time step, the base model implements so-called communication rules, of which we will use two: 'random share' and 'impact share'. Each time agents performing 'random'-sharing wish to communicate with their neighbours, they simply select a random argument known to them. Agents that perform 'impact'-sharing, on the other hand, will select arguments to assert based on their own beliefs about the hypothesis. If they believe the hypothesis to a higher degree than their initial base rate expectation (what we will call a 'positive polarity belief'), they will assert—amongst the arguments known to them—the strongest one that supports ('positive polarity argument') the hypothesis. Likewise, when they hold a belief of negative polarity, they will assert the strongest argument attacking the hypothesis. Impact-sharing encodes agents' attempts to share what they consider most relevant to

the discussion at hand, the best argument in favour of their own, current belief about the hypothesis.

Any time an agent in NormAN receives access to a piece of evidence they had not previously encountered, they update their beliefs about the hypothesis by the use of Bayesian conditionalization on the new evidence. To do so, they make use of their causal conception of the world, which is encoded in their subjective Bayes' nets. Agents that gain access to all evidence available in the current run of the model thus adopt as their new degree of belief in the hypothesis the 'optimal posterior'. The optimal posterior is the best anyone can do given access to all pieces of evidence and knowledge of the causal relationships between the evidence and the hypothesis and provides a good benchmark for the epistemic performance of a population. Given the stochasticity involved in initializing each run's unique distribution of evidence, the optimal posterior may of course fluctuate between two runs sharing the same world-DAG and hypothesis truth value.

Our Extensions

In order to study the impacts of agreeableness, we implemented three changes to the base model.

First, an option for agents to communicate pair-wise, to satisfy their preference for agreeableness in local, one-on-one communication. In the base model, agents always communicate to all of their link-neighbors at once. Our extension allows us to contrast scenarios where agents communicate with each of their neighbours individually with a scenario in which they indiscriminately broadcast to their entire network.

Second, we added heterogeneity to the base model, in the form of implicitly typed agents. To allow the study of parts of the agent population having a preference for agreeableness in their communication, we extended the model to allow mixing of two agent types making use of any two communication rules.

Third, we implemented 'sample'-sharing, an agreeable communication rule: agents using sample-sharing keep track of the polarities of each of their neighbours last-asserted argument. When communicating to these neighbours, agreeable agents will then assert arguments of matching polarity back to them, in an attempt to minimize tension. This process is straightforward in pairwise communication, and handled via majority rule when broadcasting to multiple neighbours at once. Note that even our agreeable NormAN agents will only ever point their interlocutors to true pieces of evidence as arguments about the hypothesis. In this, they differ, for instance, from the Bayesian agents in Mohseni and Williams's (2021) model which, under the influence of conformity-bias, may assert public hypothesis-opinions which they privately expect to be false.

Results

Method

Each data point is based on the results from 100 model runs of 25 time steps each, with each run featuring a small-world network (Watts-Strogatz, 20 agents, $k = 2$ and rewiring probability of 0.2). For better comparison with the polarization case study in the original NormAN paper (Assaad et al., 2023), our world uses the 'Vole' net (Fenton & Neil, 2018), instead of the simpler Asia net we used for explanatory purposes above.[2] Each run is initialized with a true hypothesis, all agents communicate once per time step (requiring a chattiness of 1 and conviction-threshold of 0) and start each run knowing exactly one piece of evidence (pre-draws and max-draws of 1).[3]

In each of the runs, we measured two key values, polarization and mean error. For polarization,

$$\sqrt{\frac{1}{n}\sum(p_n(HYP) - p_{mean}(HYP))^2}$$

the model first calculates the mean belief $p_{mean}(HYP)$ of all n agents in the population. From there, it determines, for each agent, the squared distance between their own belief and $p_{mean}(HYP)$. Finally, it takes the square root of the mean of these distances (Angere & Olsson, 2017). Intuitively, this measures how far the average agent in this run is from the mean belief. Error is measured analogously, as the square root of the mean squared distance between each agent's belief about the hypothesis and the optimal posterior p_{opt} reachable given perfect evidence:

$$\sqrt{\frac{1}{n}\sum(p_n(HYP) - p_{opt}(HYP))^2}$$

We take our results to be indicative of more general trends that hold for a variety of, but by far not for all, combinations of social networks, causal structures and evidence distributions. In particular, effect sizes will depend on the epistemic situation being sufficiently challenging to properly discriminate between the effectiveness of different sharing rules: Neither entirely obvious hypotheses, nor those that cannot be answered by the population either way, constitute the kinds of tipping-point environments where these dynamics can meaningfully affect polarization or error.

To see this, consider the relationship between the base rate with which the hypothesis is true across the many model runs analyzed in our experiments, and the agents' initial expectations (priors) of that hypothesis: In the NormAN model, agents' prior beliefs in the hypothesis are given by its marginal probability in their (subjective) causal model of the world absent any other evidence. Return to the Asia net as an example: When attempting to diagnose whether a patient has lung cancer, for instance, the agents would go into deliberation with an expectation of lung cancer that matches the actual base rate thereof. The closer the match between the truth rate across model runs, and the marginal expectation entertained

[2] Vole is an artificial Bayes' net modelled after the Agatha Christie play 'Witness for the Prosecution', and features 22 nodes, of which 6 count as evidence for the purpose of our simulation. For a full description of this network see (Assaad et al., 2023).

[3] See this link for access to our model code and simulation data.

by the agents, the less they polarize and the less they end up being wrong: If accurate, this expectation of the base rate is effectively 'free information' that is encoded in the agents' priors separately from inquiry and communication, and better priors make for better outcomes! Still, whether agents harbour in an epistemic environment that resembles their expectation or not, the relative trend of how sample-sharing affects populations of impact- or random-sharers remains stable.

For another relevant factor, take the total amount of evidence available to the population relative to the amount of evidence required for the calculation of the optimal posterior, which is largely determined by the number of agents, and the probability and amount of signals they receive from the world. If many agents each have initial access to many pieces of evidence, and there are only a relatively small number of such pieces necessary to arrive at the optimal posterior, then clearly polarization and erroneous agent beliefs become very unlikely, even if communication across the network happens to be imperfect.

Pairwise Communication

For the first case, study, consider the results shown in Figure 3: We start on the left with a population of agents that are engaged in pairwise communication using random-share. At every time step, they have individual conversations with each of their link neighbours, in which they assert an argument they know at random.

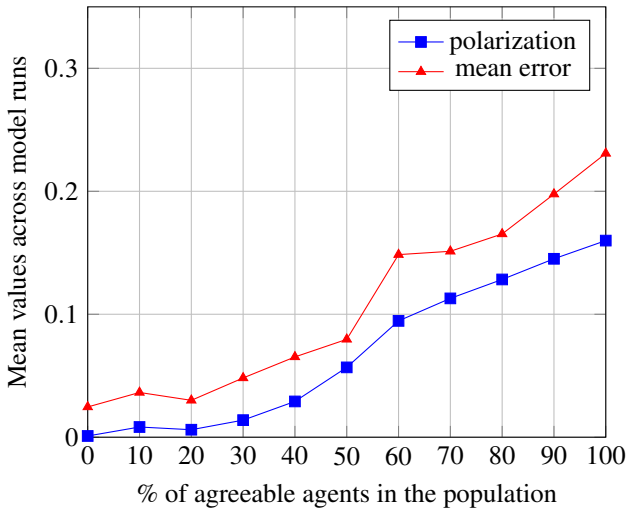

Figure 3: Impact of the presence of agreeable agents on the mean error and mean polarization of a population of random-sharers engaged in pairwise communication.

As the original NormAN paper already notes, these conditions will lead to low rates of polarization, and low rates of mean error. When randomly sharing arguments, all agents in a network will eventually converge on the same set of arguments known almost certainly. Unless one or more pieces of evidence are never drawn by any of the agents, all agents will also arrive at the optimal posterior. From this rather ideal starting point, we increased the percentage of agents with a strict preference for being agreeable in steps of 10%. These agents try to match the polarity of the last argument they received from each of their neighbors on a pairwise basis, by telling them a random known argument of said polarity. Agreeableness in this case study serves to partly cement initial inequalities in evidence distribution across social networks: Agents that start by asserting whichever argument is known to them initially, will subsequently not be confronted by agreeable neighbours with any arguments of contrasting polarity. As such, agreeable agents may come to act as buffers between parts of the social network that happen to be initially polarized due to chance, maintaining this polarization until the end of a simulation run.

Figure 4 shows what happened when we introduced a tendency to be agreeable to a population of agents that use the communication rule impact-sharing: Agreeable agents now determine whether their interlocutors have last shared an argument that confirms or disconfirms the hypothesis and will respond in kind. However, among the suitable arguments available to them, agreeable impact-sharers communicate those with their preferred impact. They are happy to share the strongest arguments they know to confirm or disconfirm hypotheses in accordance with their beliefs. However, when needing to share an argument that confirms a hypothesis they believe to be unlikely true, they will share the weakest such argument, and vice versa when asked to disconfirm a hypothesis that they believe in.

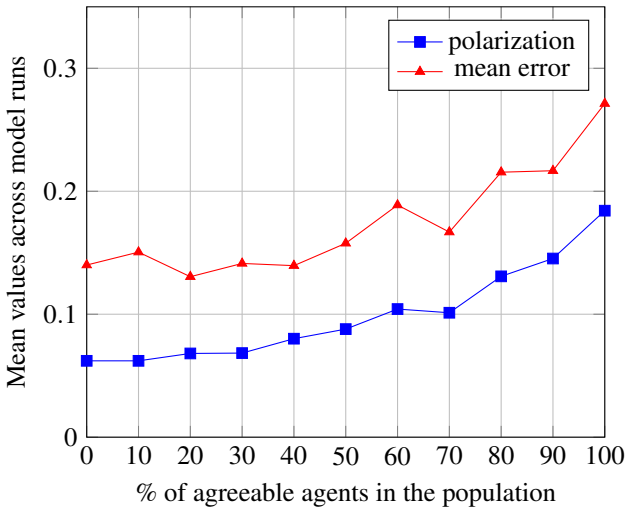

Figure 4: Impact of the presence of agreeable agents on the mean error and mean polarization of a population of impact-sharers engaged in pairwise communication.

Starting at 0% agreeable agents, our findings match those reported by Assaad et al. (2023): in a population of purely impact-sharing agents, error and polarization are present to a much higher degree than among those that randomly assert known arguments. While such a communication rule

makes sense from the perspective of individual agents, it can hinder the diffusion of arguments across the network and lead to hidden profiles, as agents prefer repeating previous assertions over sharing all arguments known to them. However, as we increase the percentage of agreeable agents, things get considerably worse: Previously, while communication links between two impact-sharers of matching polarity served to obscure subsets of their respectively known arguments, at least when two impact-sharers of mismatched polarity communicated, they presented each other with contradicting arguments, thereby counteracting polarization. Now, however, agreeable agents might find themselves slotted between agents of different polarities, and they will tell each side something agreeable, effectively blocking off crucial communication links.

Broadcasting

Let us next take a look at how these populations behave when engaged in communication that more closely resembles social media use.

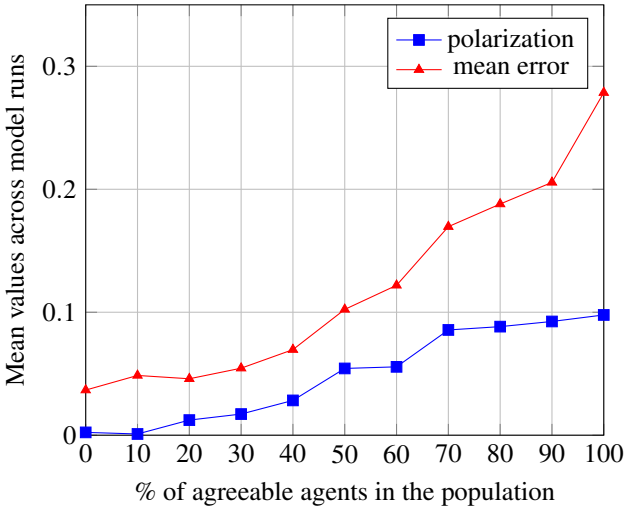

Figure 5: Impact of the presence of agreeable agents on the mean error and polarization of a population of random-sharers engaged in broadcast-style communication.

Figure 5 displays the results from introducing agreeableness to a population of agents that randomly share arguments known with all of their link neighbours simultaneously. Agreeable agents in this scenario keep track of whether the majority of their neighbours last asserted arguments in support of, or against the truth of the hypothesis, and will then try to match this polarity with the next argument they broadcast into their network. Among the suitable known arguments, they will pick at random.

As the percentage of agreeable agents increases, mean error rates climb to even greater heights than in the local, pairwise setting seen above. The mechanism behind this is as follows: whichever polarity of arguments happens to be dominant at the beginning of any run influences future assertions, as agreeable agents seek to avoid being out of line with their neighbours. Rather than just happening in local, pairwise interactions, this may now cascade across the network, as more and more agreeable agents start joining in with an initial cluster of matched-polarity beliefs, helping to extensively distribute all (and only) known arguments with that polarity.

Notice, however, that polarization—while still increasing with the percentage of agreeable agents—remains much lower than in the pairwise version. Previously, an initial mismatch in the polarity of arguments asserted between any two agents led agreeable agents to consistently share what they take to be most agreeable to their interlocutor, potentially causing or maintaining local, pairwise polarization. Now, however, there exists the possibility for a unidirectional override, where agents that stand out from the crowd will be presented with arguments of contradicting polarity even by their agreeable interlocutors if those are motivated by suitable majority ratios among their own link neighbours. These dynamics of high error rates with comparatively low polarization make this case study reminiscent of filter bubbles.

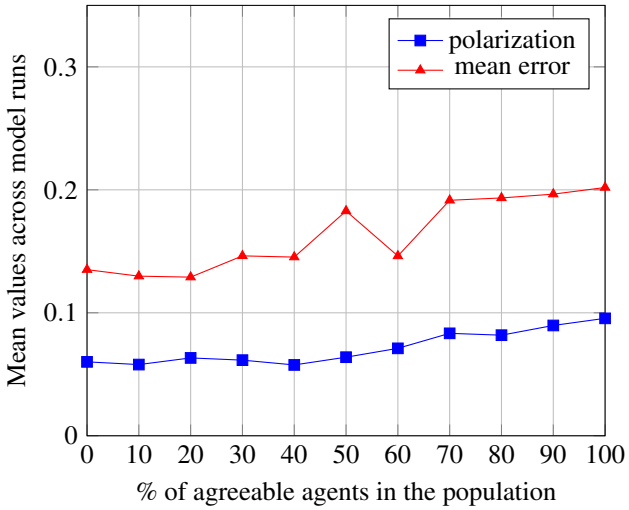

Figure 6: Impact of the presence of agreeable agents on the mean error and polarization of a population of impact-sharers engaged in broadcast-style communication.

Lastly, Figure 6 contains the results of introducing agreeableness to a population of agents performing broadcast-style impact-sharing. Agreeable agents in this case study will assert only arguments that fit those last received from the majority of their neighbours, but—within these constraints—will still share the argument with their preferred impact.

Unlike the pairwise version, the effect of increasing the percentage of agreeable agents is much weaker in this scenario. This is because previously, differences in beliefs between two such agents quickly led to stable, local, pairwise polarization, effectively blocking off many communication links all over the network. Now, this too can be overwritten, as agreeable agents that are backed by a majority of their

own neighbours will begin to assert disagreeable arguments across formerly blocked-off links. In fact, given that in expectation, each agent's individual, initial evidence is more likely than not to be of correct polarity, it is relatively unlikely for agents to find themselves in situations where the majority of their impact-sharing neighbours initially assert arguments of incorrect polarity, which serves to counteract the otherwise harmful impacts of agreeableness.

Conclusion

We extended the recently introduced agent-based model NormAN to study the effects of agreeable communication on polarization and accuracy across a population of Bayesian reasoners. We find that epistemic outcomes are consistently worsened with respect to both measures, but especially for populations engaged in pairwise, strategic communication of arguments. Equally vulnerable in terms of mean error, even if not of polarization, are populations engaged in broadcast style, random-sharing. Despite the fact that all agents in our model communicate exclusively true facts as arguments and are perfectly Bayesian reasoners possessing accurate understandings of the causal structure governing their world, the presence of agreeable communication nevertheless served to undermine effective diffusion of arguments across the network.

Our results thus illustrate the challenges of trying to promote and sustain healthy discourse. Even modest self-imposed information filtering can have adverse consequences on accuracy and polarization. The bulk of research on self-censoring has involved so-called 'spirals of silence' that involve individuals suppressing their contributions on morally laden issues for fear of social isolation in circumstances where they perceive themselves at odds with majority views (Scheufle & Moy, 2000). Our simulations illustrate how readily much 'softer' constraints on what is communicated can have significant aggregate-level effects.

At the same time, however, a failure to maintain a discussion context that feels comfortable will inevitably lead people to withdraw from the discourse (Powers, Koliska, & Guha, 2019), thus achieving the same potentially problematic filtering by a different means. Understanding how to create environments in which disagreement feels tolerable thus seems a top priority for beneficial deliberation.

Acknowledgments

U.H. was supported by Arts and Humanities Research Council grant AH/V003380/1, and a Mercator Fellowship through Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 455912038.

K.S. was supported by the research program Sustainable Cooperation – Roadmaps to Resilient Societies (SCOOP). They are grateful for funding from the Netherlands Organization for Scientific Research (NWO) and the Dutch Ministry of Education, Culture and Science (OCW) in the context of its 2017 Gravitation Program (grant number 024.003.025).

References

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*.

Bovens, L., & Hartmann, S. (2004). *Bayesian epistemology*. OUP Oxford.

Dotson, K. (2011). Tracking epistemic violence, tracking practices of silencing. *Hypatia*, *26*(2), 236-257. doi: 10 .1111/j.1527-2001.2011.01177.x

Fenton, N., & Neil, M. (2018). *Risk assessment and decision analysis with bayesian networks*. Crc Press.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Hayes, A. F., Scheufele, D. A., & Huge, M. E. (2006). Nonparticipation as self-censorship: Publicly observable political activity in a polarized opinion climate. *Political Behavior*, *28*, 259–283.

Højsgaard, S. (2012). Graphical independence networks with the grain package for r. *Journal of Statistical Software*, *46*, 1–26.

Klein, D., Marx, J., & Fischbach, K. (2018). Agent-based modeling in social science, history, and philosophy. an introduction. *Historical Social Research/Historische Sozialforschung*, *43*(1 (163), 7–27.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computation with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, *50*(2), 157–224. Retrieved from `https://www.jstor.org/stable/2345762`

Mohseni, A., & Williams, C. R. (2021). Truth and conformity on networks. *Erkenntnis*, *86*, 1509–1530.

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

Powers, E., Koliska, M., & Guha, P. (2019). "shouting matches and echo chambers": perceived identity threats and political self-censorship on social media. *International Journal of Communication*, *13*, 20.

Roos, C. A., Utz, S., Koudenburg, N., & Postmes, T. (2022). Diplomacy online: A case of mistaking broadcasting for dialogue.

Scheufle, D. A., & Moy, P. (2000). Twenty-five years of the spiral of silence: A conceptual review and empirical outlook. *International journal of public opinion research*, *12*(1), 3–28.

Scutari, M. (2009). Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*.

Thiele, J. C., & Grimm, V. (2010). Netlogo meets r: Linking agent-based models with a toolbox for their analysis. *Environmental Modelling & Software*, *25*(8), 972–974.

Vraga, E. K., Thorson, K., Kligler-Vilenchik, N., & Gee, E. (2015). How individual sensitivities to disagreement shape youth political expression on facebook. *Computers in Human Behavior*, *45*, 281–289.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, *393*(6684), 440–442.

Wilensky, U. (1999). Netlogo. Retrieved from `https://ccl.northwestern.edu/netlogo/`