Why Models of Scientific Communication Disagree

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

Abstract

Agent-based models (ABMs) have become a valuable tool in social epistemology for addressing a fundamental question: How should scientists communicate? Yet, ABMs often yield conflicting results—some suggest that high levels of communication decrease group accuracy, while others find it beneficial. Why do models differ so dramatically? We argue that these discrepancies arise from a simple fact: different models use different conceptions of "communication," and model qualitatively different phenomena. To demonstrate the effects of such differences, we integrate three paradigmatic conceptions of communication from the literature—direct evidence sharing, belief averaging, and testimony exchange—into a unifying simulation. This allows us to test them under identical conditions. Our findings suggest that communication is generally beneficial. However, the effects of communication vary significantly by which conception one adopts, even under identical conditions. This lack of robustness highlights a critical issue: outcomes depend heavily on how communication is modeled.

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

Introduction

Science is a fundamentally social activity: discoveries are shared, and results debated, as scientists work collectively to arrive at true beliefs about target hypotheses. This makes scientific inquiry a prime subject for social epistemology, the branch of philosophy examining how groups should interact to successfully converge on correct consensus (Kitcher, 1993). Like any social phenomenon, scientific inquiry is complex and its analysis is challenging. Agent-based models (ABMs) provide a powerful tool, enabling researchers to simulate artificial societies and study how individual behaviors (e.g., communication) shape collective outcomes (e.g., consensus). ABMs of scientific inquiry have been used to address important topics such as diversity, disagreement, and polarization (Šešelja, 2023). One particularly central and general question is: How should scientists communicate to collectively develop true beliefs and maximize accuracy? That is, how often should they communicate, how many peers should they communicate with, and what should they disclose?

Different ABMs offer conflicting answers. Some suggest that high levels of communication can lead to poor consensus (Zollman, 2007; Angere & Olsson, 2017) or polarization (Pallavicini, Hallsson, & Kappel, 2021), while others find that more communication improves outcomes (Hegselmann, Krause, et al., 2006). Hence, not only does the original question——How should scientists communicate?——remain without a definite answer. The divergent results also raise another question: Why do models differ so dramatically?

This paper addresses the latter question. It argues that discrepancies stem from the very different conceptions of "communication" encoded in different models. We identify three paradigmatic communication styles——direct sharing of evidence (Zollman, 2007), averaging of beliefs (Hegselmann et al., 2006), and exchanging testimony (Olsson, 2011)——and integrate them into a single simulation to analyze their effects under identical conditions. This allows us to directly compare previously incomparable modes of communication. In our model, agents collect evidence, communicate, and attempt to determine the truth of a hypothesis across scenarios ranging from simple to challenging, with varying levels of rationality and bias.

Our findings reveal that no single communication style is universally superior; different scenarios favor different approaches. However, one consistent result is that some type of communication generally improves inquiry compared to the absence of communication: either averaging or direct sharing outperform uncommunicative groups, depending on the context. In contrast, testimony generally reduces group performance, even compared to no communication at all.

We interpret our results as follows: Overall, the model highlights the importance of communication. However, testimony, as traditionally modeled in the formal literature, appears too coarse-grained to capture the nuances of efficient scientific communication. Finally, our findings challenge the generality of explicit recommendations derived from individual models: outcomes lack robustness and depend heavily on how communication is conceptualized. Thus, the strength of single ABMs lies not in demonstrating how "communication" simpliciter functions, but in exploring how specific types of communication perform under varying conditions.

What Is Communication? Different Intuitions

Consider a generic scenario: a group of agents deliberate whether a hypothesis H is true (H) or false ($\neg H$). Each agent forms a degree of belief $P(H) \in [0,1]$ based on two sources of information: (i) private inquiry, and (ii) communicated, social information from other agents. A social network, which represents the community's structure, determines who can communicate with whom (cf. Fig. 2). While abstract, this

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

framework serves as a template for sophisticated models of scientific inquiry. Many ABMs adopt this structure, differing in how they model private inquiry (i) and communication (ii) (e.g., Olsson, 2011; Hahn, Hansen, & Olsson, 2018; Hegselmann & Krause, 2015; Michelini, Javier, Houkes, Šešelja, & Straßer, 2023; Assaad et al., 2023).

Dynamically, models evolve as follows: each agent conducts a private inquiry and then communicates with their peers. We can think of private inquiry as performing an experiment, yielding a result with a certain diagnostic value concerning H. How should scientific communication be modeled? While there are many approaches (cf. Šešelja, 2023), we focus on three paradigmatic types that align with intuitive notions of communication.

First, there is simply the direct transmission of one's experimental results. If one agent achieves a result, they will transmit it unfiltered to their neighbors in the social network. This also means that receiving agents reevaluate the finding; it is like sharing "raw data." Second, there is the averaging of beliefs: agents communicate by adopting the weighted average of their own belief and the beliefs of their interlocutors. One particularly influential view in epistemology, the Equal Weight View, prescribes agents to adopt the belief that exactly "splits the difference" (Carey & Matheson, 2013). Lastly, there is testimony: agents communicate by stating their beliefs, such as "I believe that H is true." The sending agent becomes a source of information; knowing their belief can serve as direct evidence in favor of H.

Some of the best known ABM frameworks are based on these intuitions. The next section briefly outlines them, sketching their main results concerning the central question: How should scientists communicate?

Models of Scientific Inquiry

ABMs in the philosophy of science have become a large and fruitful field, making it unfeasible to survey them fully here (c.f. Douven, 2019; Šešelja, 2023; Šešelja, 2022). Three modeling frameworks stand out for both their primacy and widespread appeal: averaging models, multi-armed bandits, and Bayesian models of testimony. Each is based on a different intuition of communication.[1]

Among the best-known averaging models are Hegselmann-Krause (HK) models, or bounded confidence models (Hegselmann, Krause, et al., 2002; Hegselmann et al., 2006), which stem from earlier models of consensus formation (Lehrer, Wagner, Lehrer, & Wagner, 1981). In these models, agents hold opinions (between 0 and 1) about an intermediate value τ that they aim to correctly determine. They communicate with peers whose opinions are within a "confidence interval" of size ε from their own. The average of those beliefs within the interval forms the agent's social evidence component. The agent's final belief is a linear average of this social component and an evidential, private signal from the world.

In another popular framework, agents share evidence directly. Bandit models, first developed by Bala and Goyal (1998) and introduced to philosophy by Zollman (2007, 2010), use one or multi-armed bandits, where each "arm" represents a theory. Agents test the success rates of these theories by performing experiments ("pulling an arm"), aiming to identify the theory with the highest expected payoff. They update their beliefs about the best theory through Bayesian conditionalization, based on their own outcomes and those of their link-neighbors in a social network. Here, communication is direct, and receiving peers reevaluate findings as though they had produced them themselves.

Another class of Bayesian ABMs, originating with Angere and Olsson's (2010; 2011) *Laputa* model, involves agents exchanging testimony about their beliefs: If an agent is sufficiently sure of H, they utter "I believe H." In these models, testimony takes the form of a binary report on the hypothesis ("yes"/"no" reports). The impact of such reports is determined by the receiver's trust in the speaker, which is estimated through source reliability models (cf. Merdes, Von Sydow, & Hahn, 2020). The latter is an important task: perceived reliability determines the informational content of received testimony—one's "trust" in the source. Testimony from a trusted source will be perceived as highly informative evidence pertaining to the hypothesis, while distrusted sources are ignored or potentially even "anti-updated" on (cf. Olsson, 2020). Formal epistemologists have devised many different Bayesian models of source reliability—originating with Bovens and Hartmann (2003),[2] which use Bayesian reasoning to compute perceived reliability, taking testimony and prior beliefs about the hypothesis into account (cf. Hahn, Merdes, & von Sydow, 2018; Collins, Hahn, Von Gerber, & Olsson, 2018).

All of these models have been adapted and modified for various specific applications, many addressing questions of optimal communication (e.g., Douven & Riegler, 2010; Riegler & Douven, 2009, 2010; Weatherall & O'Connor, 2021; Michelini et al., 2023; Huang, 2023). Perhaps the best-known result comes from Zollman's application of the bandit model, suggesting that sparsely connected scientific communities, which communicate less, are more likely to converge on the optimal theory than more densely connected ones (though this result has been questioned by Rosenstock, Bruner, and O'Connor, 2017). The reasoning is that highly communicative groups may prematurely settle on a suboptimal consensus, abandoning the exploration of more fruitful strategies. Applications of the *Laputa* testimony model show similar results: frequent communication can lead not

[1] There are many more frameworks, which use entirely different concepts and representations, such as those using abstract argumentation frameworks (Borg, Frey, Šešelja, & Straßer, 2017, 2019), and epistemic landscape models (Grim, 2009; Weisberg & Muldoon, 2009).

[2] The model in Bovens and Hartmann (2003) does not consider unreliable sources as anti-reliable, in contrast to the ones used in Angere (2010); Olsson (2011). For a discussion of the effects of anti-reliability in ABMs, see Hahn, Merdes, and von Sydow (2023); Olsson (2020).

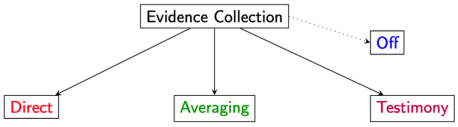

Figure 1: Sketch of the model.

only to wrong consensus but also to pronounced polarization. This prompts Angere and Olsson (2017) to issue the warning, "Publish Late, Publish Rarely!"[3] in scientific contexts, lest the network of agents be "spammed" with recycled, low-quality information (for similar findings, see Hahn, Hansen, & Olsson, 2020).

This general result——that more communication is detrimental——seems, at first glance, surprising. It also contrasts starkly with averaging models: in those models, more communication (i.e., averaging) virtually guarantees consensus (cf. Lehrer, 1976), while testimony models tend to produce more polarization. In the case of HK models, "(t)he interplay of both, seeking the truth (...) and a social exchange process that includes all, may finally lead to a consensus at least fairly close to the truth" (Hegselmann et al., 2006, p. 7). So here, more communication is better, both in terms of consensus and accuracy.

Questions and Aim

Why do these models yield such vastly different recommendations? Answering this question is not straightforward: different frameworks are not easily comparable—partly because they do not model individual inquiry in the same way.[4] A unifying model is needed to directly compare the impacts of different conceptions of communication: we introduce a model where (i) private evidence collection is fixed and (ii) the three discussed styles of communication run in parallel, including a "no communication" baseline ("shadow agents" from Hahn, Merdes, & von Sydow, 2024). Every round, agents first collect one piece of evidence each, and then split up into "counterfactual" groups: an averaging group, a testimony group, a direct-sharing group, and an "off" group (cf. Fig. 1), each trying to determine whether a central hypothesis H is true or false under identical conditions.[5]

With this model, we will explore how each style of communication is affected by different conditions: (a) whether the problem is hard or easy, i.e., whether the underlying evidence is conclusive, (b) whether agents are biased, and (c) whether the network is tightly connected. The goal is to address the questions: why do different models offer different recommendations, and which styles of communication are best suited for modeling scientific interaction?

A disclaimer: while our model is inspired by the mentioned frameworks, it is not intended as a precise reconstruction of any specific one. Nevertheless, our model will reveal commonsensical patterns reflective of key insights from the literature on each framework.

A Unifying Model

Agents, representing scientists, aim to determine whether a binary hypothesis H is true or false (H or $\neg H$). Their degree of belief, $P(H)$, is shaped by two factors: (i) private inquiry and (ii) communication. Communication between agents occurs within a social network of symmetrical links. Without loss of generality, we assume that H is always true.[6]

(i) Collecting Evidence

Each round, agents perform a test that yields either confirming evidence (positive result) or negative evidence. We denote these events as propositions: E_i for "Test i confirms the hypothesis," and $\neg E_i$ for "Test i disconfirms it." The probability of drawing E_i is determined by the frequency $\mathtt{f} \in (0.5, 1)$, and the probability of $\neg E_i$ is $1 - \mathtt{f}$. This mirrors a binomial process where each test outcome is independent.[7]

Starting with $P(H) = 0.5$, agents update their beliefs via Bayesian updating. The impact of evidence E_i on $P(H)$ is determined by its likelihood-ratio (LHR): $LHR_i = \frac{P(E_i|H)}{P(E_i|\neg H)}$.[8] We assume that the agents' perceptions of LHRs match the true frequencies: they interpret the evidence "correctly." In formal terms: $P(E_i|H) = \mathtt{f}$. Using Olsson's Symmetry Assumption (2011), we posit that the agents perceive $P(E|H) = P(\neg E|\neg H)$. Hence, the LHR of a positive piece of evidence is $LHR^+ = \frac{\mathtt{f}}{1-\mathtt{f}}$ and that of negative evidence is $LHR^- = \frac{1-\mathtt{f}}{\mathtt{f}}$. Formally, the agents collect these pieces of evidence in their memory (e.g., $\{E_1, \ldots, E_n\}$); their degree of belief is computed via strict conditionalization on received evidence $P(H|E_1, \ldots, E_n)$ (if they have received n results).

Since H is true, frequency $\mathtt{f}$ determines how easy the problem is: if it is closer to 1, not only are most results correct—they are also weighted more in favor of H. The setup ensures that, as long as $\mathtt{f} > 0.5$, agents will, with increasing evidence, converge to $P(H) = 1$ in expectation. Lower $\mathtt{f}$ (but > 0.5) mean that more evidence is needed to converge on $P(H) = 1$.

To relax the assumption of perfect calibration, we introduce bias: agents' perceptions of LHRs are distorted by a bias parameter δ drawn from a uniform distribution between $[-0.1, 0.1]$ (inspired by Baccini and Hartmann, 2022). Each agent k has their own bias δ_k, which is applied as follows:

[3] "Our study indicates that, as the scientific community becomes increasingly connected, scholars should (...) be increasingly self-critical when deciding whether to publish or not, even if this means publishing late and rarely" Angere and Olsson (2017, p. 24).

[4] *Laputa*'s agents receive binary reports from the world, HK's agents are continuously "drawn" toward the true intermediate value τ, and bandit agents must make a trade-off between strategies.

[5] This model setup takes considerable inspiration from Hahn, Assaad, and Burton (2024), which compares a Bayesian model of argument exchange (Assaad et al., 2023) to averaging models (Hegselmann et al., 2002).

[6] Our model code, simulation data and analysis scripts are available on The Open Science Framework (OSF) at https://osf.io/b2mnj/?view_only=1effb76fda504b329e56e23794a68fc5 (Please copy the link carefully into your browser).

[7] The evidence collection protocol is inspired by Huang (2023).

[8] If $LHR_i > 1$, E_i confirms H; if $LHR_i < 1$, E_i disconfirms H; if $LHR_i = 1$, E_i is non-diagnostic.

$LHR \cdot e^{\delta_k}$. This skews the computed posteriors in a roughly symmetrical manner, so that in biased simulations, some agents systematically undervalue the evidence, while others overvalue it.[9]

(ii) Communication

Agents compute their belief $P(H)$ both via private inquiry and via communication. Each simulation round, they first inquire; then, in four parallel runs, agents use one of the following rules, computing distinct "counterfactual" beliefs (cf. Fig. 1): $P_{off}(H)$, $P_{direct}(H)$, $P_{average}(H)$, $P_{testimony}(H)$.

No Sharing Agents form beliefs based solely on their own evidence: $P_{off}(H) = P(H|E_1, \ldots, E_n)$.

Direct Sharing Agents transmit the evidence they gathered, which the receiver adds to their own, as though they had drawn it themselves. This re-evaluation may be distorted by bias.

Averaging Agents update their $P_{average}$ as a weighted average of their own, evidential belief and the mean belief of their neighbors:

$P_{average} = \alpha \cdot P(H|E_1, \ldots, E_n) + (1-\alpha) \cdot P_{social}(H)$

where $P_{social}(H)$ is simply the average belief of one's neighbors (we take this formula from Hegselmann et al., 2006). In our baseline model, $\alpha = 0.5$: agents trust the combined belief of peers as much as their evidence.

Testimony Testimonial reports are binary and represented as variables: REP^k means "informant k asserts H," and $\neg REP^k$ means "informant k asserts $\neg H$" (cf. Bovens & Hartmann, 2003; Merdes et al., 2020). That is, other agents' reports serve directly as evidence about H. Formally, this means that agents simply add reports to their memory (e.g., $\{E_1, REP_1^k, \ldots, E_n, REP_n^k\}$). Agents conditionalize on this information: $P_{testmony}(H|E_1, REP_1^k, \ldots, E_n, REP_n^k)$. To model literally the idea of testimony as "taking a neighbor's report as evidence," we treat a positive report REP akin to a positive piece of evidence E_i, with LHR^+, and a negative report as a negative piece $\neg E_i$, with LHR^-. To express their position, agents send a positive report if their belief $P(H) > 0.5$, and a negative report if $P(H) < 0.5$. Put differently, agents "trust" a peer's testimony as much as a piece of evidence. This is a simplified version of *Laputa*'s procedure; since agents do not dynamically update their "trust" in their neighbors (i.e., the perceived LHR of a report), they are akin to the fixed-trust agents in Hahn, Merdes, and von Sydow (2024).[10]

Dynamics At each time step of the simulation, agents first collect a piece of evidence (i) and then communicate with all their neighbors (ii) in one of three ways, updating their respective "counterfactual" belief variables. Direct sharing involves agents passing on their most recently collected piece of evidence. Testimony sharing has agents send one report per round. Finally, agents update their beliefs by taking the weighted average of their neighbors' beliefs.

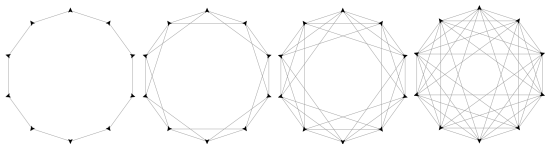

Figure 2: Social Networks (Watts & Strogatz, 1998).

Output evaluation

Accuracy is measured using the Brier score (Brier, 1950), a standard measure used in the literature (cf. Huang, 2023). The Brier score for agent k is defined as $(P^k(H) - V(H))^2$, where $V(H) = 1$ if the hypothesis is true. For enhanced readability, we use the inverted score: $BS^k = 1 - (P^k(H) - V(H))^2$. Thus, the mean Brier score, reflecting collective accuracy, is calculated as $BS^{\text{mean}} = \frac{1}{N}\sum_{i=1}^{n}(1 - (P^i(H) - V(H))^2)$ (for N agents). To contrast the four counterfactual scenarios, we analyze four distinct Brier scores: $BS^{\text{mean}}_{\text{off}}$, $BS^{\text{mean}}_{\text{direct}}$, $BS^{\text{mean}}_{\text{average}}$, and $BS^{\text{mean}}_{\text{testimony}}$. A unanimous belief of $P(H) = 1$ yields a maximum score of $BS^{\text{mean}} = 1$, whereas consensus on $P(H) = 0$ results in $BS^{\text{mean}} = 0$. [11]

Case Study

We simulated groups of $n = 10$ agents, each drawing and communicating for 100 rounds per run. Two key variables were varied: problem difficulty and agent competence. The frequency of positive, correct evidence (`f`) ranged from 0.51 to 0.9. With $\texttt{f} = 0.51$, 100 evidence pieces may not suffice to reach $P(H) = 1$, whereas at $\texttt{f} = 0.9$, agents can reach correct consensus even without communication. That is, `f` models the difficulty of the "research question." Lastly, network connectivity ranged from 2 to 8 neighbors, covering sparse to nearly fully connected networks (cf. Fig 2). For this paper, results are averaged over network densities, leaving the analysis of topology effects for future work.

Unbiased agents

When agents are unbiased and interpret evidence uniformly and correctly, increasing `f` well above 0.5 makes individual inquiry sufficient for convergence on the true belief $P(H) = 1$, even without communication. As shown in Fig. 3, as `f` increases, the accuracy measure of the "off" group approaches 1. However, direct sharing and averaging significantly enhance correct convergence.

Direct sharing better promotes convergence on $P(H)$ even at lower `f` values—a predictable outcome. Since the evidence is truth-conducive and agents are unbiased, more evidence naturally improves accuracy. Instead of collecting

[9] E.g., suppose a prior $P(H) = 0.5$ and $\texttt{f} = 0.6$; then the unbiased $LHR^+ = \frac{0.6}{0.4} = 1.5$. The unbiased posterior of one positive piece of evidence E_i would be $P(H|E_i) = 0.6$. With a positive bias $\delta = 0.1$, the biased posterior is ≈ 0.624. For a negative bias $\delta = -0.1$, it is ≈ 0.576.

[10] We leave explorations of different reliability updates in this context of model comparison for future work.

[11] We also monitored mean beliefs, and the dispersion/polarization of beliefs (cf. Bramson et al., 2017). This data can be found on OSF.

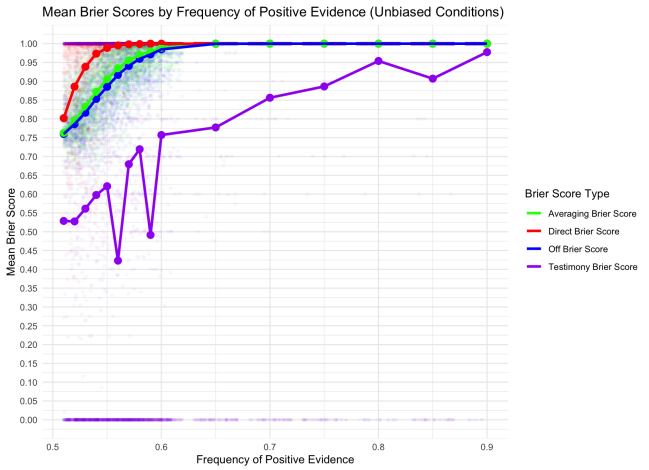

Figure 3: **Unbiased Conditions:** (Inverted) Brier scores. A score of 1 indicates perfect group convergence on the correct conclusion, i.e., $P(H) = 1$. Solid lines represent the mean final scores, averaged across 100 runs for each of four network densities (totaling 400 runs per condition). Each dot represents the outcome of a single run. What may appear as continuous purple lines at the top and bottom of the plot are in fact dense clusters of individual dots, illustrating the extreme outcomes frequently observed in the testimony condition.

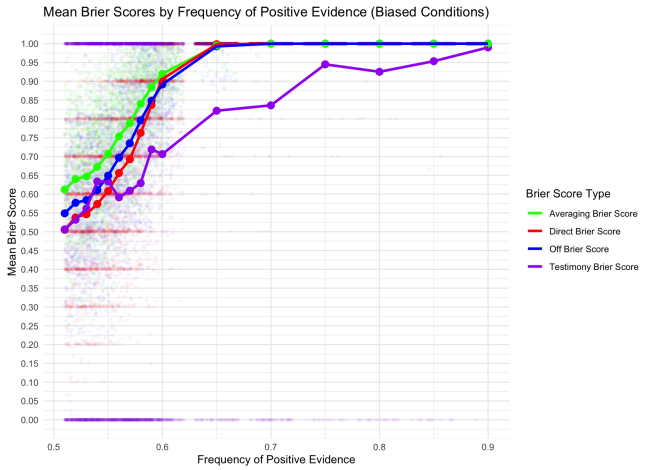

Figure 4: **Biased Conditions:** (Inverted) Brier scores. Plot conventions as in Fig. 3.

100 pieces of evidence individually, agents collectively access $100 + 100 \cdot m$ pieces, where m represents their number of link-neighbors.

Next in line for accuracy is averaging. While agents receive more information through sharing, averaging moderates beliefs away from extremities. By definition, the mean of n beliefs is less extreme than the highest or lowest instance. As a result, averaging slows the collective shift toward a belief of 1, particularly when some agents believe the hypothesis to be false and pull the mean belief downward (cf. Hahn, Assaad, and Burton (2024) for a similar result). This effect is especially pronounced when $\mathtt{f} \approx 0.5$, where the evidence is inconclusive. Nevertheless, communication through averaging still guides agents to $P(H)$ more quickly than no communication at all.

By contrast, testimony performs strikingly poorly. It is the only sharing rule that produces incorrect consensus, frequently converging on $P(H) = 0$ despite truth-conducive evidence. This outcome arises from a self-reinforcing, runaway dynamic reminiscent of an information cascade. Early evidence may lead some agents to believe that $P(H) < 0.5$, prompting negative testimony. This, in turn, convinces neighbors that the hypothesis is false, amplifying the cycle. Once all neighboring agents believe H is false, they reinforce one another through continued negative testimony. Even if agents later encounter positive evidence, it is drowned out by repeated negative testimony. This is particularly likely in well-connected networks, where agents receive much more testimony than evidence collected through inquiry.

Moreover, this runaway effect causes the testimony group to converge collectively on extreme beliefs—$P(H) \approx 0$ or $P(H) \approx 1$—even when the evidence is uncertain ($\mathtt{f} \approx 0.5$). Testimony fosters overconfidence, making extreme outcomes highly sensitive to initial conditions. As these outcomes depend on the first "triggering" rounds, the stochastic nature of the model explains the volatility of testimony's accuracy, particularly in areas of inconclusive evidence (cf. Fig. 3).

Biased agents

Now let us examine biased agents: each agent has a noise parameter that skews their perception of the evidence. Since these bias parameters are drawn from a uniform distribution, we can expect roughly equal numbers of positively and negatively biased agents within the population. When $\mathtt{f}$ is low—meaning the evidence is not highly diagnostic—biases have a greater relative impact on the computed posteriors ($P(H|E_i)$, see footnote 9).

As a result, when agents are biased and $\mathtt{f}$ is low (e.g., within $[0.51, 0.6]$), inquiry alone does not guarantee convergence to $P(H) \approx 1$. Some agents, systematically undervaluing the evidence, are drawn toward $P(H) \approx 0$, as illustrated in Fig. 4. In this parameter space, direct sharing performs worse than the "off" group: when biased agents reinterpret shared evidence, the increased volume of evidence does not necessarily lead to greater accuracy. In fact, more evidence can accelerate convergence toward erroneous beliefs for negatively biased agents, pulling them more quickly toward $P(H) \approx 0$.

Among biased populations, averaging, not the "off" group, performs best. While unbiased agents still interpret evidence correctly (i.e., according to its true likelihood), the biased agents' perceptions act as noisy but independent estimates of an underlying correct quantity. Averaging these estimates mitigates individual errors, resulting in a more accurate collective belief. We have hit on the well-known "wisdom of the crowds" effect (Galton, 1907).

Lastly, testimony still fares significantly worse: The self-reinforcing, runaway effect remains dominant, possibly leading the entire group to an incorrect consensus of $P(H) \approx 0$—even when the evidence is very conclusive and the other groups converge on correct consensus.

Conclusions: What Can We Learn from Models of Scientific Inquiry?

How to interpret these results? We take them to mean threefold: first, they suggest that testimony is an inherently problematic model of scientific communication. Second, our model does show that communication can indeed be an epistemic benefit: it helps groups converge to true consensus quicker. However, these benefits hinge on the particular model design, which determines whether well-known effects of modeled communication can be leveraged.

Drawing on established models, it is no surprise that we recover paradigmatic effects. Most strikingly, the runaway effect of testimony-based communication: as explored in Hahn et al. (2020), our agents, too, can become completely detached from the underlying evidence. As the "off" group shows, even in cases where agents would, through individual inquiry alone, converge on $P(H) = 1$, testimony can pull them towards incorrect beliefs. This effect, being led to believe something due to social influence that runs counter to one's personal information, has been dubbed the "bandwagon" effect in the *Laputa* model (2020, p. 1552). However, in our model, certain conditions hold which do not typically hold in *Laputa*: First, it is never the case that agents testify without having received new evidence; they receive evidence each round. This is something recommended in "Publish Late, Publish Rarely!" (2017) to improve group competence and diminish the "recycling" of evidence. Furthermore, agents do not dynamically update their belief of their neighbors' reliability, a feature that demonstrably causes group polarization (Hahn, Merdes, & von Sydow, 2024). Nevertheless, testimony in this simple form diminishes group performance.

There are different reasons why testimony-based communication can be detrimental: from the perils of trust-updating, to a missing account of dependencies between sources (e.g., Hahn, 2023). Our study suggests simply: disclosing only "yes"/"no" statements is insufficient for cogent communication, especially in scientific discourse. Rather, epistemic agents ought to disclose more fully their beliefs, or the evidence on which they are based. Indeed, this is common practice: open access, pre-registration of studies and data-sharing are essential parts of scientific communication. Abstracting away from these essential aspects renders testimony vulnerable to detachment from underlying evidence—an outcome which seems essentially unscientific.

What about the other discussed styles of communication? Direct sharing does disclose evidence, while averaging fully discloses agents' degrees of belief. It depends on the boundary conditions whether these styles of communication fare better or worse than a "no communication" group: our study shows that under different conditions, these styles leverage well-known effects. If evidence is truth-conducive in expectation, because agents are assumed to interpret it correctly, then direct sharing, simply increasing access to evidence, is good. In contrast, if agents are biased in certain ways, then averaging can help cancel out individual errors in interpretation, and prove beneficial when direct sharing does not.

But of course, whether or not these effects can be leveraged depends on the model setup. This is what we take our study to show: since different styles of communications are being modeled as completely different formal mechanisms, it is unsurprising that they would showcase different effects and strengths in different conditions. Averaging, direct sharing and testimony are qualitatively different mechanisms, not to be easily subsumed under a single concept—not to speak of many styles we have not explicitly modeled here (e.g., the exchange of arguments, cf. Borg et al., 2019). We could have set up our model in a myriad of other ways: we could have drawn the biases differently, have used a multi-armed bandit to model inquiry or chosen a different representation of the research problem altogether (e.g., as landscapes, Grim, 2009, or Bayesian networks, Assaad et al., 2023).

How to proceed? It stands to reason that modelers, who wish to derive explicit recommendations from their simulations, ought to either argue why their model of communication is suited for their particular target system (e.g., scientific communities), or show that their effect is robust across frameworks. However, it is generally accepted that ABMs from social epistemology are meant to provide merely "how-possibly" explanations, and serve as particularly useful thought-experiments (cf. Reutlinger, Hangleiter, & Hartmann, 2018; Mayo-Wilson & Zollman, 2021), rather than for explicit "how-actually" explanations. Furthermore, testing effects across different frameworks is being called for (so-called "structural robustness analysis," cf. Šešelja, 2023), and has, at times, been done (cf. Douven & Hegselmann, 2022).

What, then, can we learn from ABMs? We believe that they yield how-possibly explanations of the following type: "If communication is conceived of as [`concept`], then communication can be detrimental." They help explore how their particular conception of communication fares in different conditions. Our particular model suggests the following explanations: If communication is conceived of as testimony, then it can drown out evidence. If communication is conceived of as averaging, then it can cancel out biased, noisy perceptions. And if communication is viewed as direct sharing, it provides more access to evidence and is likely beneficial when additional evidence is expected to be truth-conducive. In sum, the answer to the question "Why do models of scientific communication disagree?" is straightforward: they study distinct, yet inherently interesting, communication phenomena.

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 especially grateful to Ulrike Hahn, Stephan Hartmann and Alexander Reutlinger for their invaluable advice and support. I also thank Klee Schöppl and Rafael Fuchs and for many insightful conversations on this topic and ABMs in general. Finally, I thank two anonymous reviewers for their helpful comments on an earlier draft of this manuscript.

References

Angere, S. (2010). Knowledge in a social network. *Preprint without journal information*.

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*.

Baccini, E., & Hartmann, S. (2022). The myside bias in argument evaluation: A bayesian model. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

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

Borg, A., Frey, D., Šešelja, D., & Straßer, C. (2017). Examining network effects in an argumentative agent-based model of scientific inquiry. In *International workshop on logic, rationality and interaction* (pp. 391–406).

Borg, A., Frey, D., Šešelja, D., & Straßer, C. (2019). Theory-choice, transient diversity and the efficiency of scientific inquiry. *European Journal for Philosophy of Science*, *9*(2), 26.

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.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3.

Carey, B., & Matheson, J. (2013). How skeptical is the equal weight view? In *Disagreement and skepticism* (pp. 131–149). Routledge.

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.

Douven, I. (2019). Computational models in social epistemology. In *The routledge handbook of social epistemology* (pp. 457–465). Routledge.

Douven, I., & Hegselmann, R. (2022). Network effects in a bounded confidence model. *Studies in History and Philosophy of Science*, *94*, 56–71.

Douven, I., & Riegler, A. (2010). Extending the hegselmann–krause model i. *Logic Journal of IGPL*, *18*(2), 323–335.

Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, *75*(7), 450–451.

Grim, P. (2009). Threshold phenomena in epistemic networks. In *2009 aaai fall symposium series.*

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

Hahn, U., Assaad, L., & Burton, J. W. (2024). Opinion averaging versus argument exchange. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Hahn, U., Hansen, J. U., & Olsson, E. J. (2018). Truth tracking performance of social networks: how connectivity and clustering can make groups less competent. *Synthese*, 1–31.

Hahn, U., Hansen, J. U., & Olsson, E. J. (2020). Truth tracking performance of social networks: How connectivity and clustering can make groups less competent. *Synthese*, *197*, 1511–1541.

Hahn, U., Merdes, C., & von Sydow, M. (2018). How good is your evidence and how would you know? *Topics in Cognitive Science*, *10*(4), 660–678.

Hahn, U., Merdes, C., & von Sydow, M. (2023). Knowledge through social networks: Accuracy, error, and polarisation. *PlOS one*.

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. (2015). Opinion dynamics under the influence of radical groups, charismatic leaders, and other constant signals: A simple unifying model. *NHM*, *10*(3), 477–509.

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).

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

Huang, A. C. (2023). Track records: A cautionary tale.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press, USA.

Lehrer, K. (1976). When rational disagreement is impossible. *Noûs*, 327–332.

Lehrer, K., Wagner, C., Lehrer, K., & Wagner, C. (1981). Consensus and philosophical issues. *Rational Consensus in Science and Society: A Philosophical and Mathematical Study*, 3–18.

Mayo-Wilson, C., & Zollman, K. J. (2021). The computational philosophy: simulation as a core philosophical method. *Synthese*, *199*(1), 3647–3673.

Merdes, C., Von Sydow, M., & Hahn, U. (2020). Formal models of source reliability. *Synthese*, 1–29.

Michelini, M., Javier, O., Houkes, W., Šešelja, D., & Straßer, C. (2023). Scientific disagreements and the diagnosticity of evidence: how too much data may lead to polarization.

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

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

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

Reutlinger, A., Hangleiter, D., & Hartmann, S. (2018). Understanding (with) toy models. *The British Journal for the Philosophy of Science*.

Riegler, A., & Douven, I. (2009). Extending the hegselmann–krause model iii: From single beliefs to complex belief states. *Episteme*, *6*(2), 145–163.

Riegler, A., & Douven, I. (2010). Extending the hegselmann-krause model ii.

Rosenstock, S., Bruner, J., & O'Connor, C. (2017). In epistemic networks, is less really more? *Philosophy of Science*, *84*(2), 234–252.

Šešelja, D. (2022). Agent-based models of scientific interaction. *Philosophy Compass*, *17*(7), e12855.

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.

Weisberg, M., & Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of science*, *76*(2), 225–252.

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.

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.