Philosophy of Science (2024), **91**, 329–350
doi:10.1017/psa.2023.176

ARTICLE

Can Confirmation Bias Improve Group Learning?

Nathan Gabriel and Cailin O'Connor

University of California Irvine
Corresponding author: Nathan Gabriel; Email: nathangabriel@ucmerced.edu

(Received 21 April 2022; revised 02 August 2023; accepted 13 December 2023; first published online 09 January 2024)

Abstract

Confirmation bias has been widely studied for its role in failures of reasoning. Individuals exhibiting confirmation bias fail to engage with information that contradicts their current beliefs, and, as a result, can fail to abandon inaccurate beliefs. But although most investigations of confirmation bias focus on individual learning, human knowledge is typically developed within a social structure. We use network models to show that moderate confirmation bias often improves group learning. However, a downside is that a stronger form of confirmation bias can hurt the knowledge-producing capacity of the community.

1. Introduction

Chaffee and McLeod (1973) offered individuals a choice of pamphlets to read about upcoming elections. They found that individuals tended to choose those pamphlets that fit with their current preferences, rather than those that opposed them. Mynatt et al. (1978) presented subjects with a dynamic system on a computer and asked them to discover the laws governing this system. They found that once subjects generated hypotheses about the system they followed up with tests that would tend to confirm their hypotheses, rather than disconfirm them. Lord et al. (1979) conducted an experiment on individuals with strong views on the death penalty. They found that when these subjects were offered new information regarding the deterrent effect of the death penalty they were very resistant to changing their opinions. Sweeney and Gruber (1984) surveyed members of the public during the Watergate hearings and found that those who had voted for Nixon tended to ignore information about the hearings compared to those who had voted for McGovern.

These studies are just a few of those outlining the pervasive impact of confirmation bias on human learning. Confirmation bias refers to a cluster of related behaviors whereby individuals tend to seek out, to interpret, to favor, and to selectively recall information that confirms beliefs they already hold, while avoiding or ignoring information that disconfirms these beliefs. It has been widely implicated in the prevalence and persistence of false beliefs. Individuals exhibiting this bias often ignore information that might help them develop accurate beliefs about the world. Most notably, they are susceptible to holding on to false beliefs which have been discredited (Festinger et al. 2017; Anderson et al. 1980; Johnson and Seifert 1994; Lewandowsky et al. 2012).

Confirmation bias has mostly been studied at the individual level—i.e., how does it influence individual beliefs and behaviors? Human knowledge and belief, though, are deeply social. Individuals influence the beliefs of those they interact with, and are influenced in turn. Ideas and evidence are shared via social networks in ways that impact further learning and exploration. This leads to a set of questions: How does confirmation bias influence learning and belief in human groups? Is it harmful to groups in the same way it seems to be harmful to individuals? A few authors have considered, in particular, whether confirmation bias might have unexpected or surprising *benefits* to group inquiry. Could confirmation bias actually be epistemically useful in the right contexts?

We use network models to study these questions. In particular, we draw on the network epistemology paradigm first developed in economics by Bala and Goyal (1998) to study learning in groups. Subsequently, this framework has been widely employed in social epistemology and the philosophy of science to study related topics such as the emergence of consensus in scientific communities (Zollman 2007, 2010) and the impacts of social biases on group learning (O'Connor and Weatherall 2018). Unlike some other sorts of network models, in this paradigm agents gather and share data and evidence with each other. This is an important feature in studying confirmation bias since this bias impacts the way individuals deal with evidence they receive.

We find that in models incorporating moderate levels of confirmation bias groups do *better* than in models where individuals do not exhibit confirmation bias. Dogmatic individuals who do not easily change positions force the group to more extensively test their options, and thus avoid pre-emptively settling on a poor one. This result reflects claims from philosophers and psychologists who have argued that tendencies related to irrational stubbornness, such as confirmation bias, might benefit group learning in this way (Kuhn 1977; Popper 1975; Solomon 1992, 2007; Mercier and Sperber 2017; Smart 2018). Our results also echo modeling findings from Zollman (2010), who shows that groups of stubborn individuals sometimes learn better than more individually rational learners.[1] In our case, confirmation bias functions as a sort of stubbornness. It leads individuals to keep exploring theories that might otherwise seem suboptimal, and, in doing so, to sometimes discover that these theories are actually worthwhile.

There is a downside to confirmation bias, though. While moderate levels can promote accurate group-level learning, we find that a more robust type of confirmation bias leads individuals to entirely ignore theories they do not currently favor. In such cases, communities can polarize, and epistemic progress is harmed. This suggests that while our models help confirm a useful function of confirmation bias, worries about its harms are still legitimate even when considered from the group perspective.

[1] See also Xu et al. (2016), Frey and Šešelja (2018, 2020), and Boroomand and Smaldino (2023).

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

The paper will proceed as follows. In section 2 we describe relevant literature, first focusing on empirical work on confirmation bias. We then briefly survey related modeling work. Section 3 outlines our model, which incorporates a form of confirmation bias into epistemic network models. In section 4 we present two sets of results. The first considers models with a moderate level of confirmation bias, and shows how this bias can improve learning in a community. The second considers models where confirmation bias drives polarization, and prevents good group learning. In the conclusion we draw some more general lessons for social epistemology and philosophy of science. One relates to the independence thesis—that irrational individuals can form rational groups, and vice versa (Mayo-Wilson et al. 2011). Our models provide one more vein of support for this claim. Another relates to the rationality or irrationality of ignoring data as a Bayesian learner. And a last point regards what simple models of polarization can tell us.

2. Previous literature

2.1. Confirmation bias

As noted, confirmation bias is a blanket term for a set of behaviors where are unresponsive or resistant to evidence challenging their currently held beliefs (Klayman 1995; Nickerson 1998; Mercier and Sperber 2017). The models we present will not adequately track all forms of confirmation bias. They do, however, reflect behaviors seen by those engaging in what is called selective exposure bias, as well as those who selectively interpret evidence.

Selective exposure occurs when individuals tend to select or seek out information confirming their beliefs. This could involve avoidance of disconsonant information (Hart et al. 2009) or pursuit of consonant information (Garrett 2009; Stroud 2017). The study by Chaffee and McLeod (1973) where participants chose pamphlets to read about an upcoming election is an example of selective exposure bias. While selective exposure has been most frequently studied in the context of politicized information, it need not be. Johnston (1996) observes it in participants seeking to confirm their stereotypes about doctors. Olson and Zanna (1979) find selective exposure in participants' art viewing preferences. Stroud (2017) gives a wider overview of these and related results.

As will become clear, our models can also represent confirmation bias that involves selective interpretation or rejection of evidence. Recall Lord et al. (1979), where subjects received information both supporting and opposing the efficacy of the death penalty as a deterrent to crime. This information did little to change subjects' opinions on the topic, suggesting they selectively rejected information opposing their view. Gadenne and Oswald (1986) demonstrate a similar effect in subject ratings of the importance of information confirming vs. challenging their beliefs about a fictional crime. Taber and Lodge (2006) gave participants pairs of equally strong arguments in favor of and against affirmative action and gun control, and found subjects shifted their beliefs in the direction they already leaned. In each of these cases, individuals seemed to selectively reject only the information challenging their views.

As noted, many previous authors have argued that confirmation bias may be epistemically harmful. Nickerson (1998) writes that "[m]ost commentators, by far, have seen the confirmation bias as a human failing, a tendency that is at once

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

pervasive and irrational" (205). It has been argued that confirmation bias leads to irrational preferences for early information, which grounds or anchors opinions (Baron 2000). In addition, confirmation bias can lead subjects to hold on to beliefs which have been discredited (Festinger et al. 2017; Anderson et al. 1980; Johnson and Seifert 1994; Nickerson 1998; Lewandowsky et al. 2012). Another worry has to do with "attitude polarization," exhibited in Taber and Lodge (2006), where individuals shift their beliefs in different directions when presented with the same evidence.

Further worries about confirmation bias have focused on communities of learners rather than individuals. Attitude polarization, for example, might drive wider societal polarization on important topics (Nickerson 1998; Lilienfeld et al. 2009). For this reason, Lilienfeld et al. (2009) describe confirmation bias as the bias "most pivotal to ideological extremism and inter- and intragroup conflict" (391).

Specific worries focus on both scientific communities and social media sites. Scientific researchers may be irrationally receptive to data consistent with their beliefs, and resistant to data that does not fit. Koehler (1993) and Hergovich et al. (2010), for example, find that scientists rate studies as of higher quality when they confirm prior beliefs. If so, perhaps the scientific process is negatively impacted.

It has also been argued that confirmation bias may harm social media communities. Pariser (2011) argues that "filter bubbles" occur when recommendation algorithms are sensitive to content that users prefer, including information that confirms already held views. "Echo chambers" occur when users seek out digital spaces—news platforms, followees, social media groups, etc.—that mostly confirm the beliefs they already hold. While there is debate about the impact of these effects, researchers have argued that they promote polarization (Conover et al. 2011; Sunstein 2018; Chitra and Musco 2020), harm knowledge (Holone 2016), and lead to worryingly uniform information streams (Sunstein 2018; Nikolov et al. 2015) (but see Flaxman et al. 2016).

While most previous work has focused on harms, some authors argue for potential benefits from confirmation bias. Part of the thinking is that such a pervasive bias would not exist if it was entirely harmful (Evans 1989; Mercier and Sperber 2017; Butera et al. 2018). With respect to individual reasoning, some argue that testing the plausibility of a likely hypothesis is beneficial compared to searching out other, maybe less likely, hypotheses (Klayman and Ha 1987; Klayman 1995; Laughlin et al. 1991; Oaksford and Chater 2003). Lefebvre et al. (2022) show how confirmation bias can lead agents to choose good options even when they are prone to noisy decision making.[2]

Another line of thinking, more relevant to the current paper, suggests that confirmation bias, and other sorts of irrational stubbornness, may be beneficial in group settings.[3] The main idea is that stubborn individuals promote a wider exploration of ideas/options within a group, and prevent premature herding onto one consensus. Kuhn (1977) suggests that disagreement is crucial in science to promote

[2] Rollwage and Fleming (2021) also use a decision-theoretic model to argue that when agents can accurately assess their own confidence the harms of confirmation bias can be reduced.

[3] Some authors also argue that confirmation bias could be beneficial in interpersonal settings, either for reasoning about social partners (Leyens et al. 1999; Snyder and Stukas Jr 1999) or when competence is threatened by social competition (Butera et al. 2018).

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

exploration of a variety of promising theories. Some irrational stubbornness is acceptable in generating this disagreement. Popper (1975) is not too worried about confirmation bias because as, he argues, the critical aspect of science as practised in a group will eliminate poor theories. He argues that "... a limited amount of dogmatism is necessary for progress: without a serious struggle for survival in which the old theories are tenaciously defended, none of the competing theories can show their mettle" (98). Solomon (1992) points out that in the debate over continental drift, tendencies like confirmation bias played a positive role in the persistence and spread of (ultimately correct) theories. (See also Solomon 2007.) All these accounts focus on how irrational intransigence can promote the exploration of diverse theories, and ultimately benefit group learning.

In addition, Mercier and Sperber (2017) argue that when peers disagree, confirmation bias allows them to divide labor by developing good arguments in favor of opposing positions. They are then jointly in a position to consider these arguments and come to a good conclusion. This fits with a larger picture where reasoning evolved in a social setting, and what look like detrimental biases actually have beneficial functions for groups. All these arguments fit with what Smart (2018) calls "Mandevillian Intelligence"—the idea that epistemic vices at the individual level can sometimes be virtues at the collective level. He identifies confirmation bias as such a vice (virtue) for the reasons listed above.

The results we will present are largely in keeping with these arguments for the group benefits of confirmation bias. Before presenting them, though, we take some time to address previous, relevant modeling work.

2.2. Previous models

To this point, there seem to be few models incorporating confirmation bias specifically to study its effects on epistemic groups. Geschke et al. (2019) present a "triple filter-bubble" model, where they consider impacts of (i) confirmation bias, (ii) homophilic friend networks, and (iii) filtering algorithms on attitudes of agents. They find that a combination of confirmation bias and filtering algorithms can lead to segmented "echo chambers" where small, isolated groups with similar attitudes share information. Their model, however, does not attempt to isolate confirmation bias as a causal factor in group learning. In addition, they focus on attitudes or opinions that shift as individuals average with those of others they trust. As will become clear, our model isolates the effects of confirmation bias, and also models learning as belief updating on evidence, thus providing better structure to track something like real-world confirmation bias.

There is a wider set of models originating from the work of Hegselmann and Krause (2002), where agents have "opinions" represented by numbers in a space, such as the interval $[0, 1]$. They update opinions by averaging with others they come in contact with. If agents only average with those in a close "neighborhood" of their beliefs they settle into distinct camps with different opinions. This could perhaps be taken as a representation of confirmation bias, since individuals are only sensitive to opinions near their own. But, again, there is no representation in these models of evidence or of belief revision based on evidence.

As noted, we draw on the network epistemology framework in building our model. While this framework has not been used to model confirmation bias, there have been

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

some relevant previous models where actors devalue or ignore some data for reasons related to irrational biases. O'Connor and Weatherall (2018) develop a model where agents update on evidence less strongly when it is shared by those with different beliefs. This devaluing focuses on the source of information, rather than its content (as occurs in confirmation bias). Reflecting some of our results, though, they find that devaluation at a low level is not harmful, but at a higher level eventually causes polarization. Wu (2023) presents models where a dominant group devalues or ignores information coming from a marginalized group. Wu's model (again) can yield stable polarization under conditions in which this devaluation is very strong.[4] In both cases, and, as will become clear, in our models, polarization emerges only in those cases where agents begin to entirely ignore data coming from some peers.

There is another set of relevant results from epistemic network models. Zollman (2007, 2010) shows that, counterintuitively, communities tend to reach accurate consensus more often when the individuals in them are less connected. In highly connected networks, early strings of misleading evidence can influence the entire group to preemptively reject potentially promising theories. Less-connected networks preserve a diversity of beliefs and practices longer, meaning there is more time to explore the benefits of different theories. A very similar dynamic explains why, in our model, moderate levels of confirmation bias actually benefit a group. Zollman (2010) finds similar benefits to groups composed of "stubborn" individuals, i.e., ones who start with more extreme priors and thus learn less quickly. Frey and Šešelja (2018, 2020) generate similar results for another operationalization of intransigence. And Xu et al. (2016) yield similar results for another type of model. In our model, confirmation bias creates a similar sort of stubbornness.[5]

One last relevant set of models find related results using NK-landscape models, where actors search a problem landscape for solutions. March (1991), Lazer and Friedman (2007), and Fang et al. (2010) show how less-connected groups of agents may be more successful at search because they search the space more widely and avoid getting stuck at local optima. Mason et al. (2008) and Derex and Boyd (2016) confirm this empirically. And Boroomand and Smaldino (2023), in draft work, find that groups searching NK-landscapes adopt better solutions when individuals have preferences for their own, current solutions. This is arguably a form of irrational stubbornness that improves group outcomes. (Their models, though, do not involve actors with preferences for confirmatory data the way ours do.)

3. Model

3.1. Base model

As discussed, our model starts with the network epistemology framework (Bala and Goyal 1998), which has been widely used in recent work on social epistemology and the philosophy of science. Our version of the model builds off that presented in Zollman (2010).

[4] See also Fazelpour and Steel (2022).

[5] See Wu and O'Connor (2023) for an overview of network models considering how mechanisms that thus promote transient diversity of practice improve group outcomes. And see Smart (2018) for a summary of modeling and empirical results showing how individual epistemic vice can promote group exploration.

There are two key features of this framework: a decision problem and a network. The decision problem represents a situation where agents want to develop accurate, action-guiding beliefs about the world, but start off unsure about which actions are the best ones. In particular, we use a *two-armed bandit problem*, which is equivalent to a slot machine with two arms that pay out at different rates.[6] The problem is then to figure out which arm is better. We call the two options A (or "all right") and B (or "better"). For our version of the model, we let the probabilities that each arm pays off be $p_B = 0.5$ and $p_A = p_B - \varepsilon$. In other words, there is always a benefit to taking option B, with the difference between the arms determined by the value of ε.

Agents learn about the options by testing them, and then updating their beliefs on the basis of these tests. Simulations of the model start by randomly assigning beliefs to the agents about the two options. In particular, we use two *beta distributions* to model agent beliefs about the two arms. These are distributions from 0 to 1, tracking how much likelihood the agent assigns to each possible probability of the arm in question. The details of the distribution are not crucial to understand here.[7] What is important is that there are two key parameters for each distribution, α and β. These can be thought of as tracking a history of successes (α) and failures (β) in tests of the arms. When new data is encountered, say n trials of an arm with s successes, posterior beliefs are then represented by a beta distribution with parameters $\alpha + s$ and $\beta + n - s$. It is easy to calculate the expectation of this distribution, which is $\alpha/(\alpha + \beta)$.

Following Zollman (2010), we initialize agents by randomly selecting α and β from $[0, 4]$. The set-up means that at the beginning of a trial, the agents are fairly flexible since their distributions are based on relatively little data. As more trials are performed, expectation becomes more rigid. For example, if $\alpha = \beta = 2$, then the expectation is 0.5. Expectation is flexible in that if the next three pulls are failures, then expectation drops to $2/(2 + 5) \approx 0.286$. However, if a thousand trials resulted in $\alpha = \beta = 500$, three repeated failures would result in an expectation $500/(500 + 503) \approx 0.499$ (which is still close to 0.5). In simulation, if the agents continue to observe data from the arms, their beta distributions tend to become more and more tightly peaked at the correct probability value, and harder to shift with small strings of data.

As a simulation progresses we assume that in each round agents select the option they think more promising, i.e., the one with a higher expectation given their beliefs. This assumption corresponds with a myopic focus on maximizing current expected payoff. While this will not always be a good representation of learning scenarios, it represents the idea that people tend to test those actions and theories they think are

[6] Note that previous investigations of confirmation bias on the individual level have used these and similar decision problems (Rollwage and Fleming 2021; Lefebvre et al. 2022).

[7] The function is defined as follows.

Definition (*Beta distribution*) A function on $[0,\ 1]$, $f(\cdot)$, is a beta distribution iff, for some $\alpha > 0$ and $\beta > 0$,

$$f(x) = \frac{x^{(\alpha-1)}(1-x)^{(\beta-1)}}{B(\alpha, \beta)},$$

where $B(\alpha, \beta) = \int_0^1 u^{(\alpha-1)}(1-u)^{(\beta-1)}\, du$.

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

promising.[8] Each agent gathers some number of data points, n, from their preferred arm. After doing so, they update their beliefs in light of the results they gather, but also in light of data gathered by neighbors. This is where the network aspect of the model becomes relevant. Agents are arrayed as nodes on a network, and it is assumed they see data from all those with whom they share a connection.

To summarize, this model represents a social learning scenario where members of a community (i) attempt to figure out which of two actions/options/beliefs is more successful, (ii) use their current beliefs to guide their data-gathering practices, and (iii) share data with each other. This is often taken as a good model of scientific theory development (Zollman 2010; Holman and Bruner 2015; Kummerfeld and Zollman 2015; Weatherall et al. 2020; Frey and Šešelja 2020) or the emergence of social consensus/beliefs more broadly (Bala and Goyal 1998; O'Connor and Weatherall 2018; Wu 2023; Fazelpour and Steel 2022).

In this base model, networks of agents eventually settle on consensus—either preferring the better option B, or the worse option A. If they settle on A, they stop exploring option B, and fail to learn that it is, in fact, better. This can happen if, for instance, misleading strings of data convince a wide swath of the group that B is worse than it really is.

3.2. Modeling confirmation bias

How do we incorporate confirmation bias into this framework? As noted, confirmation bias is varied and tracks multiple phenomena (Klayman 1995). For this reason, we develop a few basic models of confirmation bias that track the general trend of ignoring or rejecting evidence that does not accord with current beliefs. The goal is to study the ways such a trend may influence group learning in principle, rather than to exactly capture any particular version of confirmation bias.

For each round of simulation, after trial results are shared according to network connections, agents have some probability of accepting and updating their beliefs based on the shared results. This probability is based on how likely they believe those results are given their prior beliefs, λ. This likelihood is a function of the agent's current beta distribution parameters, α and β, as well as the details of the results, successes, s, per number of draws, n.[9] An agent calculates λ separately for each set of results shared via a network connection. Examples of these probabilities as a function of an agent's α and β values are shown in figure 1.

[8] Kummerfeld and Zollman (2015) present models of this sort where agents also explore options that they think are suboptimal.

[9] The likelihood for some agent of some set of results is given by a beta-binomial probability mass function:

$$pmf_X(s, n, \alpha, \beta) = \binom{n}{s} \frac{B(s + \alpha, n - s + \beta)}{B(\alpha, \beta)},$$

where $B(\alpha, \beta) = \int_0^1 u^{(\alpha-1)}(1-u)^{(\beta-1)}\,du$, X is the action (A or B) that generated the results, α and β are the values corresponding to the receiving agent's beliefs about action X, n is the number of pulls, and s is the number of successes in shared results. For further discussion of the beta-binomial probability mass function, see see Johnson et al. (2005, 282) or Gupta and Nadarajah (2004, 425).

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

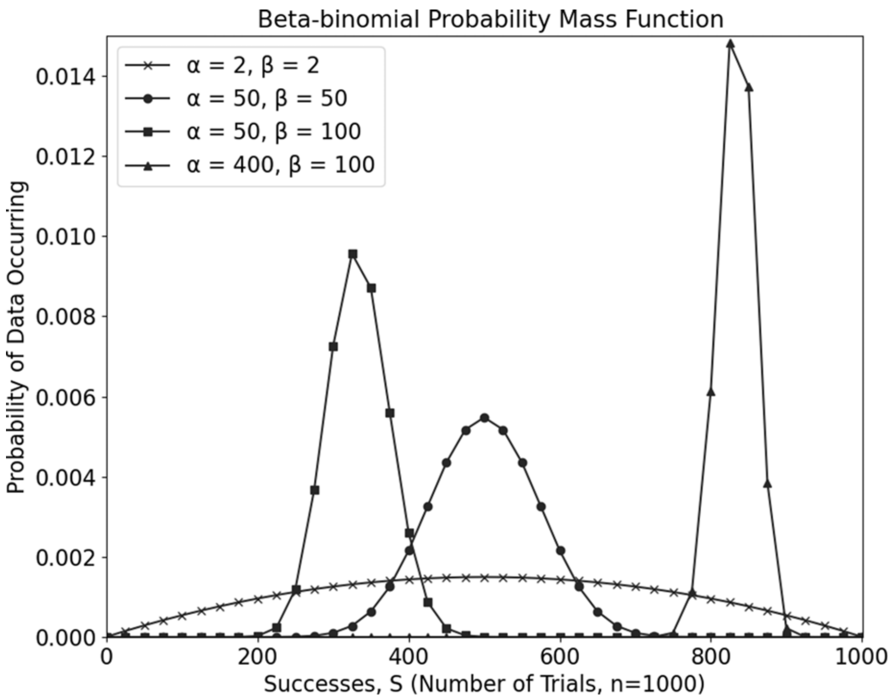

Figure 1. The probability mass functions of beta-binomial distributions for different values of α and β.

Additionally, the model includes an intolerance parameter, t, that impacts how likely agents are to accept or reject results for a given prior probability of those results occurring. The probability of an agent accepting a set of results is $p_{\text{accept}} = \lambda^t$. When t is low, agents are more tolerant of results they consider unlikely, and when t is high they tend to reject such results. For example, suppose an agent thinks some shared results have a 5% chance of occurring given their prior beliefs (i.e., $\lambda = 0.05$). Then, for $t = 1$, the agent has a probability of accepting $p_{\text{accept}} = 0.05$. For $t = 2$, the agent is extremely intolerant with $p_{\text{accept}} = 0.05^2 = 0.0025$.[10] For $t = 0.5$, the agent is more tolerant and $p_{\text{accept}} = 0.05^{0.5} = 0.22$. And when $t = 0$ the probability of acceptance is always 1, i.e., our model reverts to the base model with no confirmation bias. Whenever evidence is accepted, agents update their beliefs using Bayes' rule as described. Agents never reject evidence they generated themselves.[11] This feature mimics confirmation bias by representing either (i) a situation in which agents are selectively avoiding data that does not fit with their priors, or (ii) engaging with, but rejecting, this data and thus failing to update on it.

[10] We do not actually consider values of $t > 1$ in our simulations because generally prior probabilities of evidence are fairly small to begin with.

[11] This is true across our models, and we take it to be psychologically realistic. We ran limited simulations to confirm that this choice did not significantly impact results. In all cases, results were very similar in models where agents also applied confirmation bias to their own results.

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

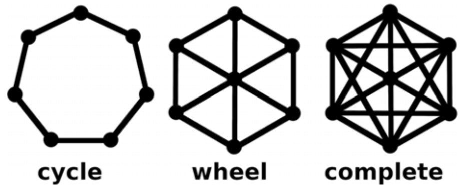

Figure 2. Several network structures.

Notice that, for a given tolerance t, agents with the same expectation do not typically have the same probability of accepting evidence. For example, $\alpha = \beta = 2$ gives the same 0.5 expectation as $\alpha = \beta = 50$, but for any $t \neq 0$, an agent with the former beliefs will be more likely to accept a 1000-test trial with 650 successes. The latter agent finds this data less likely because of the relative strength of their beliefs (see figure 1). In general, stronger beliefs in this model will be associated with a higher likelihood of rejecting disconsonant data. This aspect of the model neatly dovetails with empirical findings suggesting that confirmation bias is stronger for beliefs that individuals are more confident in (Rollwage et al. 2020).

We consider several different simple network structures, including the cycle, wheel, and complete networks (see figure 2). We also consider Erdös–Rényi random networks, which are generated by taking some parameter b, and connecting any two nodes in the network with that probability (Erdős and Rényi 1960). In general, we find qualitatively robust results across network structures. For each simulation run we initialize agents as described, and let them engage in learning until the community reaches a stable state.

4. Results

4.1. Moderate confirmation bias

In the model just described, notice that actors can be very unlikely to update on some data sets. But the structure of the beta distribution and our rule for rejecting evidence means that they always accept data they encounter with some probability. Whenever agents continue to test different theories, their data continues to reach network neighbors and shape the beliefs of these neighbors. This mutual influence means that, as in previous versions of the model without confirmation bias, actors in our model always reach consensus eventually: either correct consensus that B is better, or incorrect consensus on A. The question is, how does the introduction of confirmation bias influence the frequency with which correct vs. incorrect consensus emerges?

Surprisingly, we find that confirmation bias improves the knowledge-producing capacity of epistemic networks, in that it increases the likelihood a particular network will reach correct consensus. This finding is robust across network structures, and variations in other parameters (network size, number of pulls per round n, difference between the arms ε).[12] Figure 3 shows this result for the wheel network with different

[12] In all the results presented we hold $\varepsilon = 0.001$ and $n = 1000$. These choices follow previous authors. They also keep the difficulty of the bandit problem in a range where it is at least somewhat challenging to identify the better option. This reflects the fact that we wish to model the sort of problem that might actually pose a challenge to a community trying to solve it. If ε is larger, or n larger, the problem is easier and more communities reach accurate consensus in this sort of model.

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

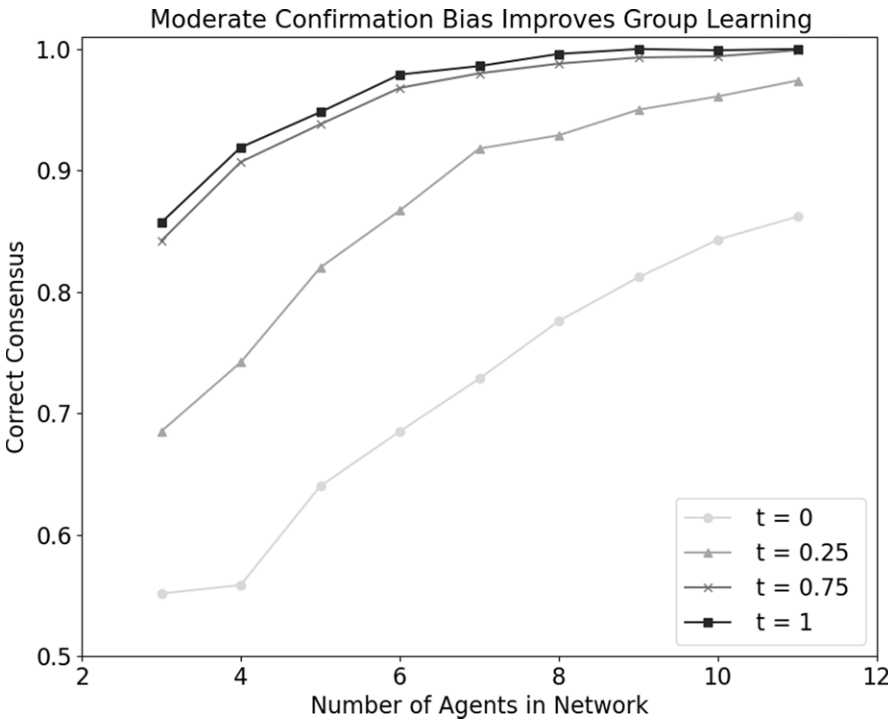

Figure 3. When agents use moderate levels of confirmation bias, groups tend to reach accurate consensus more often. This figure shows results for small wheel networks. Qualitative results are robust across parameter values. $\varepsilon = 0.001$, $n = 1000$.

numbers of agents. The results are averages over 1000 runs of simulation for each parameter value. Each trace tracks a different amount of confirmation bias, as modulated by t. As is clear, the larger t is, i.e., the more confirmation bias, the more often the network of agents correctly concludes that B is the better option.[13,14]

As noted, this trend is robust across parameter values. In figure 4 we show similar results for larger graphs randomly generated using the Erdős–Rényi (ER) algorithm described above. Again, higher levels of confirmation bias correspond to better group learning.

[13] For all results displayed, we ran simulations long enough to reach stable consensus. To check replicability, many of our models were coded independently by two separate team members. Results were all highly similar, with some small variations based on exact details of algorithm implementation.

[14] In one variation, we drop the assumption that agents always accept their own data and instead allow agents to reject their own information according to the same dynamics with which they accept or reject other's data. Results were similar, and qualitative findings were robust. For example, with $b = 0.5$ and $t = 0.25$, correct consensus rates, for 4, 6, 9, 12, 15, and 25 agents respectively, shifted from 0.757, 0.875, 0.952, 0.971, 0.990, 0.997 as shown in figure 3 to 0.819, 0.908, 0.970, 0.988, 0.995, 1.000 in the variation in which agents can reject their own data. This variation had similar results for the model of strong confirmation bias reported in section 4.2. Code is available at https://github.com/nathanlgabriel/confirmation_bias_illusory_truth.

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

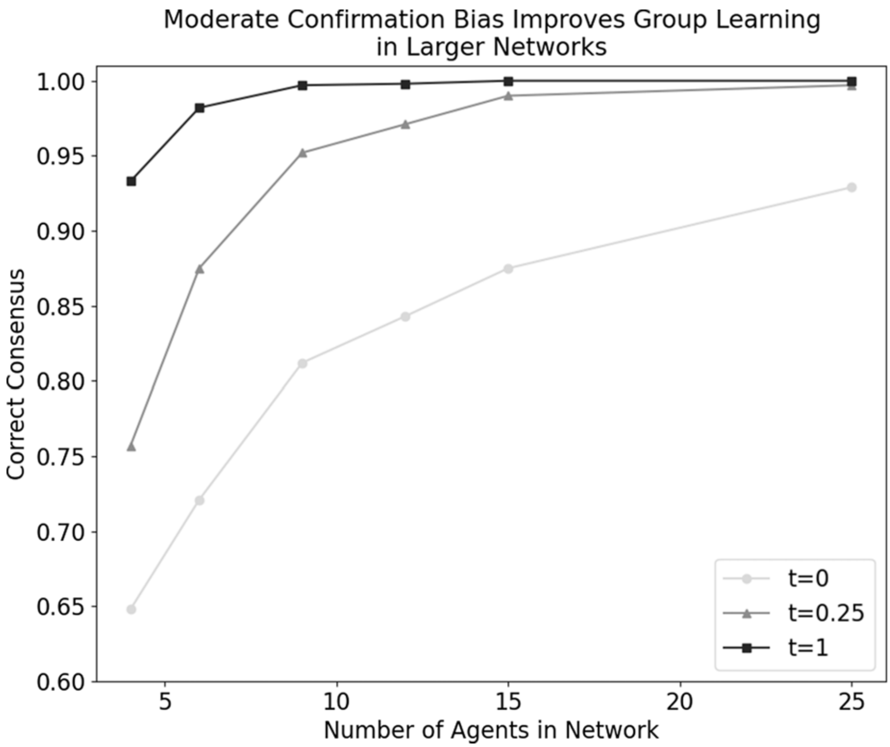

Figure 4. When agents use moderate levels of confirmation bias, groups tend to reach accurate consensus more often. This figure shows results for moderate-sized ER random networks with the probability of connection between any two nodes $b = 0.5$. Qualitative results are robust across parameter values. $\varepsilon = 0.001$, $n = 1000$.

As previously mentioned, this finding relates to results from Zollman (2007, 2010) showing that both lowering connectivity and increasing stubbornness can improve outcomes in this sort of model. This "Zollman effect" occurs because individuals can influence each other too strongly, and, as a result, incorrectly settle on option A as a result of early strings of misleading data. By making agents less willing to accept data that might change their mind, confirmation bias decreases social influence in a similar way to decreasing connectivity or stubbornness and leads to longer periods of exploration for both theories. This, in turn, increases the chances that the entire group selects the better option B in the end. While it is surprising that a reasoning bias which is usually treated as worrisome can actually improve the performance of a group, this result, as noted, reflects previous claims from philosophers and psychologists. The mechanism we identify—where confirmation bias leads to continued exploration and data gathering about multiple theories or actions—is very similar to that described by Kuhn (1977), Popper (1975), Solomon (1992, 2007), and Smart (2018).

To test the robustness of our general finding, we implement another version of the model. Confirmation bias in the first version responds to the likelihood of some data set given current beliefs. But confirmation bias often occurs in the context of fairly coarse-grained information. What if we suppose individuals ignore details of the data and simply ask, which theory does this data support? And, do I think that theory is the better one? In deciding to accept or reject a set of data in this version of the model, the actor calculates their probability that B is better than A, or vice versa, and scales with an intolerance parameter as before.[15] Actors accept any data set supporting B (or A) with probability P_{accept}.

The qualitative results of this "coarse-grained" model are similar to the previous one. Across parameters, increasing confirmation bias leads to improved group outcomes. Figure 5 shows results for ER random networks with different numbers of agents. As is clear, a higher value of t is again associated with a greater probability that the group adopts a consensus on the better option, B.

Our results to this point seem to suggest that confirmation bias is an unmitigated good in a group setting. It is true that the sort of confirmation bias modeled so far always improves group consensus formation in our models. There are a few caveats, though. First, for parameter settings where the decision problem is relatively easy—where the network (N) is large, agents draw more data (n is large), and/or the two arms are relatively easy to disambiguate (ε is large)—most groups successfully learn to choose the correct arm. In these cases confirmation bias does little to improve learning.[16] On the other hand, confirmation bias as we model it always slows down consensus formation, sometimes very dramatically. This creates a trade-off between speed of learning and accuracy of consensus formation (Zollman 2007, 2010). In cases where it is important for a group to quickly reach consensus, then, confirmation bias might cause problems. Second, as will become clear in the next section, stronger assumptions about what confirmation bias entails will shift this narrative.

4.2. Strong confirmation bias

To this point, we have only considered models where agents always have some probability of updating on data they encounter, though this probability may be small. This means that all agents continue to exert influence on each other, regardless of what they believe and what sorts of data they gather. This influence might be small, but it ensures that, given enough time, the community will eventually reach consensus on one of the two options.

But what if agents sometimes entirely discount data that does not fit their prior beliefs? We now look at a much simpler version of confirmation bias. Agents calculate how likely some data set is given their current beliefs, as before. If that probability is below some threshold, h, they discard the data. If it is above that threshold, they update on it.

[15] That is, we calculate P_{accept} as

$$P_{\text{accept}} = \left[\sum_{i=0}^{999} \left(pmf_A(i, n, \alpha_A, \beta_A) * \sum_{j=i+1}^{1000} pmf_B(j, n, \alpha_B, \beta_B) \right) \right]^t,$$

where $pmf_X(s, n, \alpha, \beta))$ is the same as before.

[16] See also Rosenstock et al. (2017), who point out that the benefits of network connectivity shown in Zollman (2010) are only relevant to difficult problems.

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

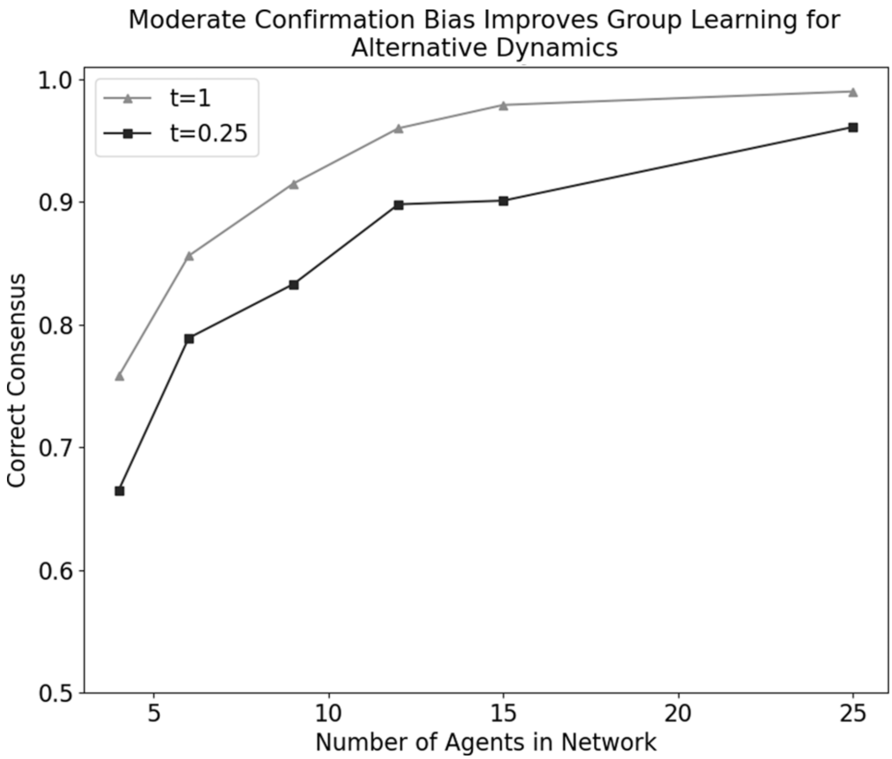

Figure 5. Moderate confirmation bias increases epistemic success under a different operationalization of confirmation bias. This figure shows results for moderate-sized ER random networks with the probability of connection between any two nodes $b = 0.5$. Qualitative results are robust across parameter values. $\varepsilon = 0.001$, $n = 1000$.

In this version of the model, we now observe outcomes where groups do not settle on consensus. It is possible for subgroups to emerge which favor different options, and where data supporting the alternative position is unpersuasive to each group. This can be understood as a form of polarization—agents within the same community settle on stable, mutually exclusive beliefs, and do not come to consensus even in the face of continued interaction and sharing of evidence.[17]

Figure 6 shows results for Erdős–Rényi random networks with different thresholds for ignoring discordant data, h. As is clear, as the cutoff becomes more stringent, fewer simulations end up adopting an accurate consensus.

As noted, much of the reason that communities fail to reach accurate consensus in these models is because they polarize. When this happens, some actors adopt accurate beliefs, but others do not. Because actors with inaccurate beliefs develop credences where the accurate belief looks very unlikely to them, they become entirely

[17] There are many ways the term polarization is used. Here we operationalize it as any outcome where the community fails to reach consensus, and where this lack of consensus is stable. This approximately tracks notions of polarization that have to do with failure of a community to agree on matters of fact.

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

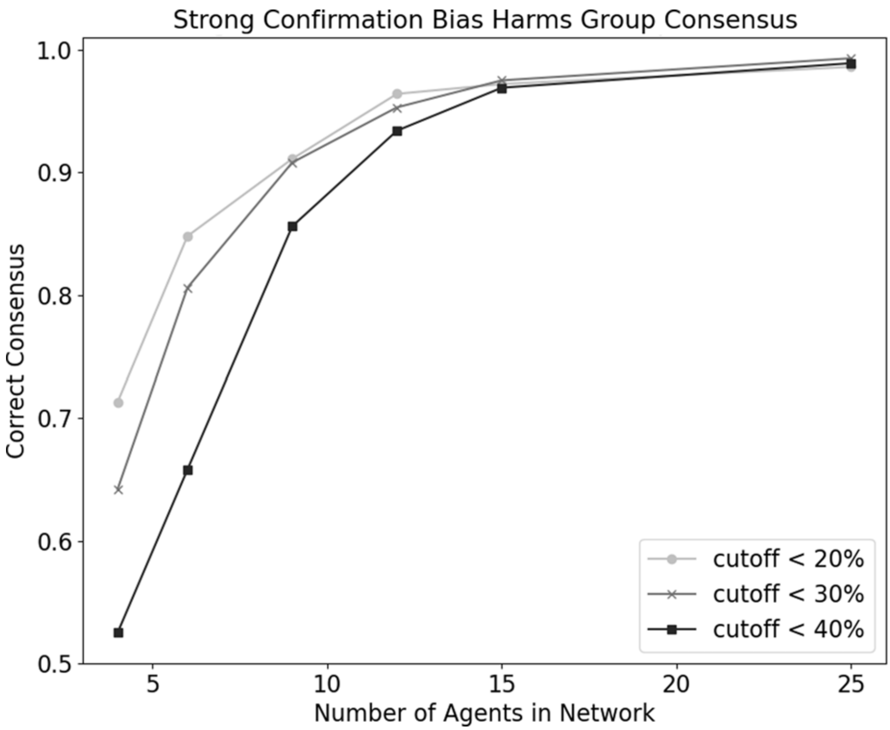

Figure 6. Strong confirmation bias hurts group learning. This figure shows results for moderate-sized ER random networks with the probability of connection between any two nodes $b = 0.5$. Qualitative results are robust across parameter values. $\varepsilon = 0.001$, $n = 1000$.

insensitive to data that might improve their epistemic state. As figure 7 shows, polarization occurs more often the stronger the agents' confirmation bias. Both accurate and inaccurate consensus become less common. For parameter values where only very likely data is accepted, polarization almost always emerges.

Another question we might ask is: how does this stronger form of confirmation bias impact the general epistemic success of agents in the network? Note that since polarization occurs in these models this is a slightly different question than how strong confirmation bias impacts correct group consensus. Given that confirmation bias leads to an increase in polarization, and a decrease in both correct and incorrect consensus formation, it is not immediately clear whether it is epistemically harmful on average.

In general, we find that this stronger form of confirmation bias leads fewer individual actors, on average, to hold correct beliefs. As is evident in figure 8 for high levels of strong confirmation bias, fewer individuals hold true beliefs. In this figure notice that for lower levels of confirmation bias there is relatively little impact on average true belief. In fact, given details of network size, we find that there is often a slight advantage to a little confirmation bias for the reasons outlined in the last section—it prevents premature lock-in on false consensus.[18] This slight advantage is

[18] In the simulations pictured here, the 20%–30% cutoff range does the best by a hair.

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

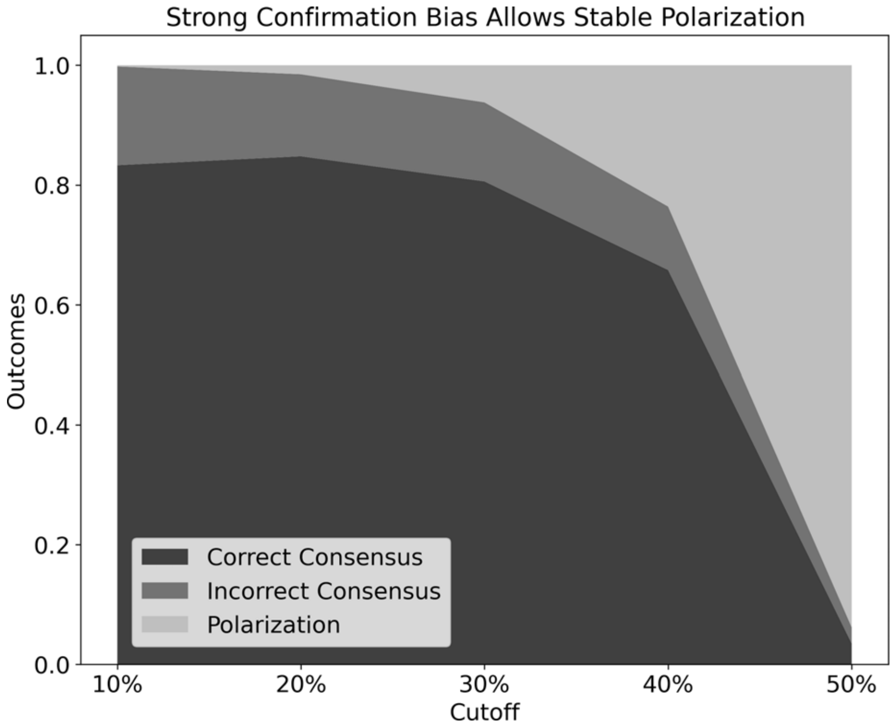

Figure 7. Strong confirmation bias leads to polarization. This figure shows results for ER random networks with the probability of connection between any two nodes $b = 0.5$. Qualitative results are robust across parameter values. $N = 6$, $\varepsilon = 0.001$, $n = 1000$.

eventually outweighed by the negative impacts of too much distrust. As confirmation bias increases, eventually too many agents adopt false beliefs, and fail to engage with disconfirmatory evidence.

At this point, it may seem that small differences in how confirmation bias is modeled have large impacts on how it influences group learning. As long as agents continue to have some influence on each other, no matter how small, confirmation bias improves consensus formation (and thus average true beliefs). Once this is no longer true, it generally harms average true beliefs. This picture is not quite right. Recall from the previous section that moderate confirmation bias always slows consensus formation, sometimes dramatically. When this happens, a network can remain in a state of transient polarization for a long period of time. If we stopped our models at some arbitrary time period, rather than always running them to a stable state, the two sorts of confirmation bias would look more similar. In both cases confirmation bias leads to polarization, but in one case that polarization eventually resolves, and this process improves community learning. The take-away is thus a complex one—confirmation bias can have surprising benefits on group learning, and for the very reasons supposed by some previous authors, but these benefits are neither simple, nor unmitigated.

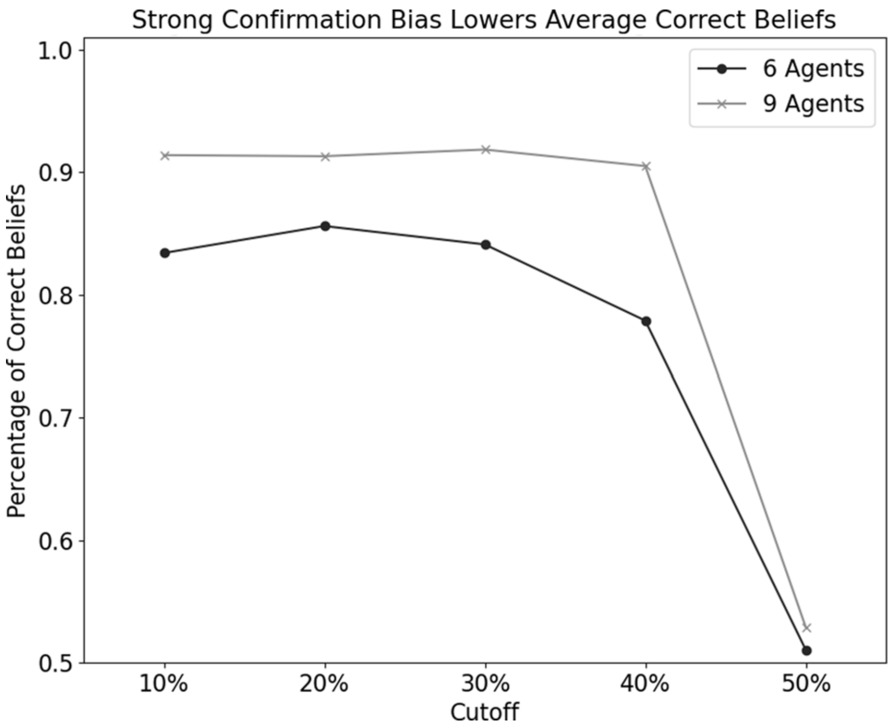

Figure 8. Avegerage correct beliefs under strong confirmation bias. This figure shows results for ER random networks of size 6 and 9, with the probability of connection between any two nodes $b = 0.5$. Qualitative results are robust across parameter values. $\varepsilon = 0.001$, $n = 1000$.

5. Conclusion

We find that confirmation bias, in a moderate form, improves the epistemic performance of agents in a networked community. This is perhaps surprising given that previous work mostly emphasizes epistemic harms of confirmation bias. By decreasing the chances that a group pre-emptively settles on a promising option, confirmation bias can improve the likelihood the group chooses optimal options in the long run. In this, it can play a similar role to decreased network connectivity or stubbornness (Zollman 2007, 2010; Xu et al. 2016; Wu 2023). The downside is that more robust confirmation bias, where agents entirely ignore data that is too disconsonant with their current beliefs, can lead to polarization, and harm the epistemic success of a community. Our modeling results thus provide potential support for arguments from previous scholars regarding the benefits of confirmation bias to groups, but also a caution. Too much confirmation bias does not provide such benefits.

There are several discussions in philosophy and social sciences where our results are relevant. Mayo-Wilson et al. (2011) argue for the *independence thesis*—that rationality of individual agents and rationality of the groups they form sometimes come apart. Our results lend support to this claim. While there is a great deal of evidence suggesting that confirmation bias is not ideal for individual reasoners, our models suggest it can nonetheless improve group reasoning under the right

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

conditions. This, as noted, relates to the notion of Mandevillian intelligence from Smart (2018).

This argument about the independence thesis connects up with debates about whether it is ever rational to ignore free evidence.[19] According to Good's theorem, it is always rational to update in such cases (Good 1967). The proof relies on the idea that an individual who wishes to maximize their expected utility will not do worse, and will often do better, by updating on available, free information. But in our models agents sometimes choose to ignore evidence, and thus increase their chances of eventually holding true beliefs. Of course, in the meantime they ignore good evidence that should, on average, improve the success of their actions. Whether or not they "should" ignore evidence in this case arguably depends on what their goals are. But if the central goal is to eventually settle on the truth, we show that ignoring some data can help in a group learning setting.

As noted, our results are consonant with previous argumentation regarding the value of stubbornness or dogmatism to science. There is a question, though, about whether confirmation bias, or other forms of arguably irrational stubbornness, are the best mechanisms by which to improve group learning. Santana (2021) argues that stubbornness in science can have negative consequences, such as hurting public trust. Wu and O'Connor (2023) give an overview of the literature on transient diversity of beliefs in network models, and argue that in scientific communities there are better ways to ensure this diversity than to encourage actors to be stubborn. For example, centralized funding bodies can promote exploration across topics instead. By doing so, they allow scientists to learn about data rationally, but still prevent premature adoption of suboptimal theories. But Wu and O'Connor's conclusions are specific to scientific disciplines where there are levers for coordinating exploration across a group. When it comes to more general epistemic groups, especially outside of science, such coordination may not be possible. If so, confirmation bias may provide benefits that are not available via more efficient routes.

One larger discussion this paper contributes to regards the mechanisms that can lead to polarization in real communities. Such mechanisms often include feedback loops wherein similarity of opinion/belief leads to increased influence between individuals, and vice versa. Individuals whose beliefs diverge end up failing to influence each other, and their divergent beliefs become stable. But under this general heading, theorists have identified a number of different such mechanisms. Hegselmann and Krause (2002) show how this can happen if individuals fail to update on the opinions of those who do not share their opinions. Weatherall and O'Connor (2020) find polarization emerges when individuals conform with those in their social cliques, and thus ignore data from those outside. Pariser (2011) argues that algorithms can drive polarization by supplying only information that users like in the face of confirmation bias. Echo chambers function when individuals seek out and connect to friends and peers who share their beliefs (see also modeling work by Baldassarri and Bearman 2007). Wu (2023) finds polarization arises when entire groups mistrust other groups based on social identity. O'Connor and Weatherall (2018) find that polarization emerges when actors do not trust data from peers who hold different beliefs. And in

[19] Of course, if data is costly, a rational agent might not be willing to pay the costs to update on it. But in our modeling set-up, we assume that data may be shared cost-free.

https://doi.org/10.1017/psa.2023.176 Published online by Cambridge University Press

our models polarization can follow from confirmation bias because subgroups ignore different sets of disconfirmatory data.

This suggests that identifying sufficient causes of polarization is very different from identifying necessary, or even likely, causes of polarization. It also suggests that, in considering real instances of polarization, researchers should be sensitive to many possible causes. Thus, experimental/empirical research and modeling are both necessary in figuring out just what real causes are at work in producing social polarization.

As a last note before concluding, we would like to discuss limitations of our models. Of course, the models we present are highly simplified compared to real social networks. This means that the results should, of course, be taken with a grain of salt. In particular, we only consider one type of learning problem—the one-armed bandit model. The question remains whether and to what degree these results will be robust. We suspect that models with other problems might yield similar results. The general benefit of slowing group learning, and promoting a period of exploration, has been established across a number of models with different problems and mechanisms. We leave this for future research.

We conclude with one last note about why models are especially useful to this project. Psychological traits like confirmation bias are widespread and deeply ingrained. It is not easy to intervene on them in experimental settings. This means that it is hard to devise an experiment where one group learns with confirmation bias, and one without. Models allow us to gain causal control on the ways confirmation bias can impact group learning, even if we do so for a simplified system.

Acknowledgements. This material is based upon work supported by the National Science Foundation under Grant No. 1922424. Many thanks to the members of our NSF working group for discussion and feedback on drafts of this proposal—Clara Bradley, Matthew Coates, Carolina Flores, David Freeborn, Yu Fu, Ben Genta, Daniel Herrmann, Aydin Mohseni, Ainsley Pullen, Jim Weatherall, and Jingyi Wu. Thanks to the anonymous reviewers for comments. And thanks to members of the DFG Network for comments, especially Audrey Harnagel, Patrick Grim, Christoph Merdes, Matteo Michelini, and Dunja Seselja.

References

Anderson, Craig A., Mark R. Lepper, and Lee Ross. 1980. "Perseverance of Social Theories: The Role of Explanation in the Persistence of Discredited Information". *Journal of Personality and Social Psychology* 39 (6):1037–49. https://psycnet.apa.org/doi/10.1037/h0077720

Bala, Venkatesh and Sanjeev Goyal. 1998. "Learning from Neighbors". *Review of Economic Studies* 65 (3):595–621.

Baldassarri, Delia and Peter Bearman. 2007. "Dynamics of Political Polarization". *American Sociological Review* 72 (5):784–811. https://doi.org/10.1177/000312240707200507

Baron, Jonathan. 2000. *Thinking and Deciding*. Cambridge: Cambridge University Press.

Boroomand, Amin and Paul E. Smaldino. 2023. "Superiority Bias and Communication Noise Can Enhance Collective Problem Solving". *Journal of Artificial Societies and Social Simulation* 23 (3):14. https://doi.org/10.18564/jasss.5154

Butera, Fabrizio, Nicolas Sommet, and Claudia Toma. 2018. "Confirmation as Coping with Competition". *European Review of Social Psychology* 29 (1):299–339. https://doi.org/10.1080/10463283.2018.1539908

Chaffee, Steven H. and Jack M. McLeod. 1973. "Individual vs. Social Predictors of Information Seeking". *Journalism Quarterly* 50 (2):237–45.

Chitra, Uthsav and Christopher Musco. 2020. "Analyzing the Impact of Filter Bubbles on Social Network Polarization". In *Proceedings of the 13th International Conference on Web Search and Data Mining*, 115–123. New York: Association for Computing Machinery. https://doi.org/10.1145/3336191.3371825

Conover, Michael, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. "Political Polarization on Twitter". *Proceedings of the International AAAI Conference on Web and Social Media* 5 (1):89–96. https://doi.org/10.1609/icwsm.v5i1.14126

Derex, Maxime and Robert Boyd. 2016. "Partial Connectivity Increases Cultural Accumulation Within Groups". *Proceedings of the National Academy of Sciences* 113 (11):2982–7. https://doi.org/10.1073/pnas.1518798113

Erdös, Paul and Alfréd Rényi. 1960. "On the Evolution of Random Graphs". *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5 (1):17–60.

Evans, Jonathan St. B. T. 1989. *Bias in Human Reasoning: Causes and Consequences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fang, Christina, Jeho Lee, and Melissa A. Schilling. 2010. "Balancing Exploration and Exploitation through Structural Design: The Isolation of Subgroups and Organizational Learning". *Organization Science* 21 (3):625–42. https://doi.org/10.1287/orsc.1090.0468

Fazelpour, Sina and Daniel Steel. 2022. "Diversity, Trust, and Conformity: A Simulation Study". *Philosophy of Science* 89 (2):209–31. https://doi.org/10.1017/psa.2021.25

Festinger, Leon, Henry Riecken, and Stanley Schachter. 2017. *When Prophecy Fails: A Social and Psychological Study of a Modern Group that Predicted the Destruction of the World*. Morrisville, NC: Lulu Press, Inc.

Flaxman, Seth, Sharad Goel, and Justin M. Rao. 2016. "Filter Bubbles, Echo Chambers, and Online News Consumption". *Public Opinion Quarterly* 80 (S1):298–320. https://doi.org/10.1093/poq/nfw006

Frey, Daniel and Dunja Šešelja. 2018. "What Is the Epistemic Function of Highly Idealized Agent-Based Models of Scientific Inquiry?" *Philosophy of the Social Sciences* 48 (4):407–33. https://doi.org/10.1177/0048393118767085

Frey, Daniel and Dunja Šešelja. 2020. "Robustness and Idealizations in Agent-Based Models of Scientific Interaction". *The British Journal for the Philosophy of Science* 71 (4):1411–37. https://doi.org/10.1093/bjps/axy039

Gadenne, Volker and Margit Oswald. 1986. "Entstehung und Veränderung von Bestätigungstendenzen beim Testen von Hypothesen". *Zeitschrift für experimentelle und angewandte Psychologie* 33:360–74.

Garrett, R. Kelly. 2009. "Echo Chambers Online?: Politically Motivated Selective Exposure among Internet News Users". *Journal of Computer-Mediated Communication* 14 (2):265–85. https://doi.org/10.1111/j.1083-6101.2009.01440.x

Geschke, Daniel, Jan Lorenz, and Peter Holtz. 2019. "The Triple-Filter Bubble: Using Agent-Based Modelling to Test a Meta-Theoretical Framework for the Emergence of Filter Bubbles and Echo Chambers". *British Journal of Social Psychology* 58 (1):129–49. https://doi.org/10.1111/bjso.12286

Good, Irving J. 1967. "On the Principle of Total Evidence". *The British Journal for the Philosophy of Science* 17 (4):319–21. https://doi.org/10.1093/bjps/17.4.319

Gupta, Arjun K. and Saralees Nadarajah. 2004. *Handbook of Beta Distribution and its Applications*. Boca Raton, FL: CRC Press.

Hart, William, Dolores Albarracín, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, and Lisa Merrill". 2009. "Feeling Validated versus Being Correct: A Meta-Analysis of Selective Exposure to Information. *Psychological Bulletin* 135 (4):555–88. https://doi.org/10.1037/a0015701

Hegselmann, Rainer and Ulrich Krause. 2002. "Opinion Dynamics and Bounded Confidence Models, Analysis, and Simulation". *Journal of Artificial Societies and Social Simulation* 5 (3).

Hergovich, Andreas, Reinhard Schott, and Christoph Burger. 2010. "Biased Evaluation of Abstracts Depending on Topic and Conclusion: Further Evidence of a Confirmation Bias within Scientific Psychology". *Current Psychology* 29 (3):188–209. https://doi.org/10.1007/s12144-010-9087-5

Holman, Bennett and Justin P. Bruner. 2015. "The Problem of Intransigently Biased Agents". *Philosophy of Science* 82 (5):956–68. https://doi.org/10.1086/683344

Holone, Harald. 2016. "The Filter Bubble and its Effect on Online Personal Health Information". *Croatian Medical Journal* 57 (3):298–301. https://doi.org/10.3325

Johnson, Hollyn M. and Colleen M. Seifert. 1994. "Sources of the Continued Influence Effect: When Misinformation in Memory Affects Later Inferences." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20 (6):1420. https://psycnet.apa.org/doi/10.1037/0278-7393.20.6.1420

Johnson, Norman Lloyd, Samuel Kotz, and Adrienne W. Kemp. 2005. *Univariate Discrete Distributions*. New York: Wiley.

Johnston, Lucy". 1996. "Resisting Change: Information-Seeking and Stereotype Change. *European Journal of Social Psychology* 26 (5):799–825. https://doi.org/10.1002/(SICI)1099-0992(199609)26:5

Klayman, Joshua. 1995. "Varieties of Confirmation Bias". *Psychology of Learning and Motivation* 32:385–418. https://doi.org/10.1016/S0079-7421(08)60315-1

Klayman, Joshua and Young-Won Ha. 1987. "Confirmation, Disconfirmation, and Information in Hypothesis Testing". *Psychological Review* 94 (2):211–28. https://doi.org/10.1037/0033-295X.94.2.211

Koehler, Jonathan J. 1993. "The Influence of Prior Beliefs on Scientific Judgments of Evidence Quality". *Organizational Behavior and Human Decision Processes* 56 (1):28–55. https://doi.org/10.1006/obhd.1993.1044

Kuhn, Thomas S. 1977. "Objectivity, Value Judgment, and Theory Choice". In *The Essential Tension: Selected Studies in Scientific Tradition and Change*, 320–39. Chicago, IL: University of Chicago Press.

Kummerfeld, Erich and Kevin J. S. Zollman. 2015. "Conservatism and the Scientific State of Nature". *The British Journal for the Philosophy of Science* 67 (4):1057–76. https://doi.org/10.1093/bjps/axv013

Laughlin, Patrick R., Scott W. VanderStoep, and Andrea B. Hollingshead. 1991. "Collective versus Individual Induction: Recognition of Truth, Rejection of Error, and Collective Information Processing". *Journal of Personality and Social Psychology* 61 (1):50–67. https://doi.org/10.1037/0022-3514.61.1.50

Lazer, David and Allan Friedman. 2007. "The Network Structure of Exploration and Exploitation". *Administrative Science Quarterly* 52 (4):667–94. https://doi.org/10.2189/asqu.52.4.667

Lefebvre, Germain, Christopher Summerfield, and Rafal Bogacz. 2022. "A Normative Account of Confirmation Bias during Reinforcement Learning". *Neural Computation* 34 (2):307–37. https://doi.org/10.1162/neco_a_01455

Lewandowsky, Stephan, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. "Misinformation and its Correction: Continued Influence and Successful Debiasing". *Psychological Science in the Public Interest* 13 (3):106–31. https://doi.org/10.1177/1529100612451018

Leyens, Jacques-Philippe, Benoit Dardenne, Vincent Yzerbyt, Nathalie Scaillet, and Mark Snyder. 1999. "Confirmation and Disconfirmation: Their Social Advantages". *European Review of Social Psychology* 10 (1):199–230. https://doi.org/10.1080/14792779943000062

Lilienfeld, Scott O., Rachel Ammirati, and Kristin Landfield. 2009. "Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare?" *Perspectives on Psychological Science* 4 (4):390–8. https://doi.org/10.1111/j.1745-6924.2009.01144.x

Lord, Charles G., Lee Ross, and Mark R. Lepper. 1979. "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence". *Journal of Personality and Social Psychology* 37 (11):2098–109. https://doi.org/10.1037/0022-3514.37.11.2098

March, James G. 1991. "Exploration and Exploitation in Organizational Learning". *Organization Science* 2 (1):71–87.

Mason, Winter A., Andy Jones, and Robert L. Goldstone. 2008. "Propagation of Innovations in Networked Groups". *Journal of Experimental Psychology: General* 137 (3):422–33. https://doi.org/10.1037/a0012798

Mayo-Wilson, Conor, Kevin J. S. Zollman, and David Danks. 2011. "The Independence Thesis: When Individual and Social Epistemology Diverge". *Philosophy of Science* 78 (4):653–77. https://doi.org/10.1086/661777

Mercier, Hugo and Dan Sperber. 2017. *The Enigma of Reason*. Cambridge, MA: Harvard University Press.

Mynatt, Clifford R., Michael E. Doherty, and Ryan D. Tweney. 1978. "Consequences of Confirmation and Disconfirmation in a Simulated Research Environment". *Quarterly Journal of Experimental Psychology* 30 (3):395–406. https://doi.org/10.1080/00335557843000007

Nickerson, Raymond S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises". *Review of General Psychology* 2 (2):175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nikolov, Dimitar, Diego F. M. Oliveira, Alessandro Flammini, and Filippo Menczer. 2015. "Measuring Online Social Bubbles". *PeerJ Computer Science* 1:e38. https://doi.org/10.7717/peerj-cs.38

Oaksford, Mike and Nick Chater. 2003. "Optimal Data Selection: Revision, Review, and Reevaluation". *Psychonomic Bulletin & Review* 10:289–318. https://doi.org/10.3758/BF03196492

O'Connor, Cailin and James Owen Weatherall. 2018. "Scientific Polarization". *European Journal for Philosophy of Science* 8 (3):855–75. https://doi.org/10.1007/s13194-018-0213-9

Olson, James M. and Mark P. Zanna. 1979. "A New Look at Selective Exposure". *Journal of Experimental Social Psychology* 15 (1):1–15. https://doi.org/10.1016/0022-1031(79)90014-3

Pariser, Eli. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. London: Penguin.

Popper, Karl. 1975. "The Rationality of Scientific Revolutions". In *Problems of Scientific Revolution: Progress and Obstacles to Progress*, edited by Rom Harré, 320–39. Oxford: Clarendon Press.

Rollwage, Max and Stephen M. Fleming. 2021. "Confirmation Bias Is Adaptive when Coupled with Efficient Metacognition". *Philosophical Transactions of the Royal Society B* 376 (1822):20200131. https://doi.org/10.1098/rstb.2020.0131

Rollwage, Max, Alisa Loosen, Tobias U. Hauser, Rani Moran, Raymond J. Dolan, and Stephen M. Fleming. 2020. "Confidence Drives a Neural Confirmation Bias". *Nature Communications* 11:2634. https://doi.org/10.1038/s41467-020-16278-6

Rosenstock, Sarita, Justin Bruner, and Cailin O'Connor. 2017. "In Epistemic Networks, Is Less Really More?" *Philosophy of Science* 84 (2):234–52. https://doi.org/10.1086/690717

Santana, Carlos. 2021. "Let's Not Agree To Disagree: The Role of Strategic Disagreement in Science". *Synthese* 198 (25):6159–77. https://doi.org/10.1007/s11229-019-02202-z

Smart, Paul R. 2018. "Mandevillian Intelligence". *Synthese* 195:4169–200. https://doi.org/10.1007/s11229-017-1414-z

Snyder, Mark and Arthur A. Stukas Jr. 1999. "Interpersonal Processes: The Interplay of Cognitive, Motivational, and Behavioral Activities in Social Interaction". *Annual Review of Psychology* 50 (1): 273–303. https://doi.org/10.1146/annurev.psych.50.1.273

Solomon, Miriam. 1992. "Scientific Rationality and Human Reasoning". *Philosophy of Science* 59 (3):439–55.

Solomon, Miriam. 2007. *Social Empiricism*. Cambridge, MA: MIT Press.

Stroud, Natalie Jomini. 2017. "Selective Exposure Theories". In *The Oxford Handbook of Political Communication*, edited by Kate Kenski and Kathleen Hall Jamieson. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199793471.013.009_update_001

Sunstein, Cass R. 2018. *#Republic: Divided Democracy in the Age of Social Media*. Princeton, NJ: Princeton University Press.

Sweeney, Paul D. and Kathy L. Gruber. 1984. "Selective Exposure: Voter Information Preferences and the Watergate Affair." *Journal of Personality and Social Psychology* 46 (6):1208–21. https://doi.org/10.1037/0022-3514.46.6.1208

Taber, Charles S. and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs". *American Journal of Political Science* 50 (3):755–69. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Weatherall, James Owen and Cailin O'Connor. 2020. "Conformity in Scientific Networks". *Synthese* 198:7257–78. https://doi.org/10.1007/s11229-019-02520-2

Weatherall, James Owen, Cailin O'Connor, and Justin P. Bruner. 2020. "How To Beat Science and Influence People: Policymakers and Propaganda in Epistemic Networks". *The British Journal for the Philosophy of Science* 71 (4):1157–86. https://doi.org/10.1093/bjps/axy062

Wu, Jingyi. 2023. "Epistemic Advantage on the Margin: A Network Standpoint Epistemology". *Philosophy and Phenomenological Research* 106 (3):755–77. https://doi.org/10.1111/phpr.12895

Wu, Jingyi and Cailin O'Connor". 2023. "How Should We Promote Transient Diversity in Science? *Synthese* 201:37. https://doi.org/10.1007/s11229-023-04037-1

Xu, Bo, Renjing Liu, and Zhengwen He. 2016. "Individual Irrationality, Network Structure, and Collective Intelligence: An Agent-Based Simulation Approach". *Complexity* 21 (S1):44–54. https://doi.org/10.1002/cplx.21709

Zollman, Kevin. 2007. "The Communication Structure of Epistemic Communities". *Philosophy of Science* 74 (5):574–87. https://doi.org/10.1086/525605

Zollman, Kevin. 2010. "The Epistemic Benefit of Transient Diversity". *Erkenntnis* 72 (1):17–35. https://doi.org/10.1007/s10670-009-9194-6

Cite this article: Gabriel, Nathan and Cailin O'Connor. 2024. "Can Confirmation Bias Improve Group Learning?" *Philosophy of Science* 91 (2):329–350. https://doi.org/10.1017/psa.2023.176