Diversity and homophily in group inquiry

Sina Fazelpour (s.fazel-pour@northeastern.edu)
Department of Philosophy and Religion & Khoury College of Computer Sciences
Northeastern University, Boston, MA USA

Hannah Rubin (hannahmrubin@gmail.com)
Department of Philosophy University of Missouri, Columbia, MO USA

Abstract

How do social factors affect group learning in diverse populations? Evidence from cognitive science gives us some insight into this question, but is generally limited to showing how social factors play out in small groups over short time periods. To study larger groups and longer time periods, we argue that we can combine evidence about social factors from cognitive science with agent-based models of group learning. In this vein, we demonstrate the usefulness of idealized models of inquiry, in which the assumption of Bayesian agents is used to isolate and explore the impact of social factors. We show that whether a certain social factor is beneficial to the community's epistemic aims depends on its particular manifestation by focusing on the impacts of homophily – the tendency of individuals to associate with similar others – on group inquiry.

1 Introduction

Diversity, broadly construed, is important to successful inquiry within groups, ranging from juries and deliberative mini-publics to scientific communities. Diversity of social identities can improve the performance of groups through varied cognitive and communicative pathways (Phillips, 2017; Page, 2017; Sulik, Bahrami, & Deroy, 2021). However, most research to date has studied small groups, over short time periods, and there is a difficulty scaling up to larger groups and longer time scales. This leaves important aspects of inquiry in communities of interest, e.g. scientific communities, unexplored.

In this paper, we develop an agent-based model which incorporates personal- and interpersonal-level phenomena studied in cognitive science, as well as empirical evidence regarding the structure of group interactions. We use an idealized model of group inquiry, in which Bayesian agents gather and share evidence, to isolate the impact of social factors on ultimate group success in inquiry. We show that factors that may be beneficial to group inquiry in restricted experimental settings (e.g., where everyone is talking to everyone) may be detrimental when we consider them in the context of how larger groups interact and share information. Thus, we argue, one reason for the gap between potential and realized benefits of demographic diversity (Sulik et al., 2021) is limitations to the inferences we can draw based on experimental results.

In particular, we will investigate when *homophily*, the tendency of individuals to associate with similar others within diverse communities, can be beneficial to inquiry. Homophily can be driven by many different dimensions of similarity (e.g., social identities, attitudes and beliefs, or values) and it can manifest in different structural and behavioral effects (e.g., forming connections, desire to conform, or trust relations). We review some relevant literature regarding important manifestations of homophily in section 2. In section 3, we discuss the usefulness of assuming Bayesian agents in these and other related models of group inquiry.

Then, we present two simulations to show how these various manifestations can impact group inquiry when we scale up to larger groups over longer time periods. First, in section 4 we examine how homophily impacts collective performance by modulating trust relations and identity-based network formation. We consider the effects on inquiry of both identity-based trust and opinion-based trust, for which we develop a novel formalization. We find that homophilic networks are generally more successful than non-homophilic random networks in this context. In section 5, we then consider what happens when we add pressure to conform with others in your social identity group. We find that conformity generally impedes inquiry, and that the effect is more pronounced in homophilic networks. Thus, whether network homophily improves inquiry depends on which other manifestations of homophily are present. Overall, we our findings and conclusions are consonant with a recent push in cognitive science to shift our focus from whether, why, and how diversity is beneficial to *when* it is beneficial (Sulik et al., 2021). Finally, in section 6, we end by discussing the relevance of these findings to arguments regarding proper evaluation diversity initiatives.

2 Manifestations of homophily

In this section, we will provide empirical evidence for each of the manifestations of homophily we model, followed by relevant previous simulation work. Homophily is a particularly important feature to look at when considering the features diverse communities might have, as its been consistently found to be important in many aspects of life. As mentioned, homophily can be driven by many different dimensions of similarity and it can manifest in different structural and behavioral effects. To focus our analysis we consider three structural/behavioral effects of homophily and two dimensions of similarity:

1. Formation of links, influenced by social identity
2. Trust relations, influenced by both social identity and opinion similarity
3. Normative conformity, influenced by social identity

Of course, there are many factors which might be relevant to when we might expect to see benefits of diversity in group inquiry, including other ways homophily may manifest in diverse communities. We focus on a few important aspects of diverse communities, comprised of larger groups and over longer timescales than typically considered in empirical studies, in the hopes that we might eventually build up to a more thorough understanding through future study.

First, we consider network homophily, where links are more likely to be formed between people who belong to the same social identity group. There is good reason to think this will be a feature of a diverse group of inquirers. It is well known that network homophily is a pervasive feature of diverse communities in general (Jackson, 2010). For instance, and of relevance to our models of successful group inquiry, scientific communities are homophilic, especially when it comes to co-authorship patterns (Ferber & Teiman, 1980; McDowell & Smith, 1992; Boschini & Sjögren, 2007; del Carmen & Bing, 2000; West, Jacquet, King, Correll, & Bergstrom, 2013; Wang, Lee, West, Bergstrom, & Erosheva, 2019) and citation patterns (Wardle, 1995; Paris, De Leo, Menozzi, & Gatto, 1998; Ghiasi, Mongeon, Sugimoto, & Larivière, 2018).

Because of homophily's impacts, groups with the same demographic composition can have different levels of "local diversity" (Gomez & Lazer, 2019) and behave in radically different ways. There is motivation for thinking network homophily can affect the spread of knowledge. It has been shown, using a 'belief averaging' model of opinion formation, that homophily can slow the spread of information in a network, such that it takes the network longer to converge to shared opinion (Golub & Jackson, 2012b).

The fact that network homophily has been shown to slow information spread means that one might expect homophily to be beneficial to group inquiry. Limiting information flow can contain misleading results and prevent a community of Bayesian agents from erroneously converging to a false belief (Zollman, 2007, 2010). More specifically, these results indicate that it is sometimes better to have fewer connections between group members, which is a way of limiting information flow. The 'transient diversity' of opinions that persists due to limited information means that people investigate multiple hypotheses for a longer time, increasing the likelihood that they arrive at the truth. Similar results have been found in cases where ultimate success in a complex problem depends on building upon multiple previous innovations; empirical results show that lack of information sharing means less copying from others, leading to a greater number of independent innovations upon which to ultimately build (Derex & Boyd, 2016).

Second, we consider both identity-based and opinion-based trust. In these cases, judgements of similarity go hand-in-hand with judgements of a person's reliability as a source of evidence. Thus Bayesian agents may discount some evidence when updating their beliefs. Studies have found that people are more likely to trust information or arguments from people who are in the same social-identity group, because those people are judged to be more competent or their views are taken to be more worthy of consideration (Turner, Wetherell, & Hogg, 1989; Ahmad, Ahmed, Srivastava, & Poole, 2011; Warkentin, Sharma, Gefen, Rose, & Pavlou, 2018).

Previous models have considered the effect of (identity-based) trust on successful inquiry in groups of Bayesian agents. For instance, since trust is often higher within social identity groups compared to the community as a whole, in certain cases demographic diversity can prevent erroneous convergence to false beliefs by slowing information flow (Fazelpour & Steel, 2021). However, when trust relations are asymmetric, where one group consistently mistrusts the other, the mistrusted group can end up in a privileged epistemic position, though inquiry overall is harmed by the lack of trust (Wu, 2022). While we, like Fazelpour and Steel (2021), consider symmetric trust relationships, we show that when identity-based trust is considered in the context of (demographically) homophilic networks, this same factor may further entrench polarization across social identity groups, preventing the community from converging on the truth.

We also examine the intuitive idea that agents may put more trust in those they perceive to be "like-minded". Many models include a mechanism for having trust depending on similarity of opinion, particularly models of polarization. For instance, Hegselmann, Krause, et al. (2002) provide a model where agents form opinions by taking a weighted average of other agents' opinions who are close enough to their own. Subsequent models allow influence to be continuous; the closer in opinion one is to you, the greater the influence they have on you (Deffuant, Amblard, Weisbuch, & Faure, 2002; Meadows & Cliff, 2012).[1] These models, though, only consider opinion dynamics, where people choose who to listen to (and how much) in absence of individual learning or deliberation. Other models have allowed for the possibility that one might trust another less in the context of deliberation (Angere, 2010; Olsson, 2013) or that people might be more uncertain about evidence provided by others (O'Connor & Weatherall, 2018) the further away their beliefs are from one's own. In our simulations, agents estimate others' like-mindedness, and so calibrate their trust in their opinions, by tracking the (mis)matches between others' opinions—assumed by agents to be signalled by their actions—and their own. Insofar as individuals change their opinions and actions as they gather or receive more evidence over time, this type of opinion-based trust involves a dynamic updating process that is shaped by prior trust as well as current actions.

Finally, we consider what happens when identity-based conformity pressures are also at play. Such (normative) conformity pressures are prevalent in real-world settings, where people tend to feel increased pressure to behave like others in their own social identity group (Cialdini & Goldstein, 2004; Deutsch & Gerard, 1955). In small group settings, it has been found that introducing diversity is beneficial to inquiry: the pressure to conform to one's own social group is reduced and dissenting views relevant to the task are more likely to be elicited (Phillips & Loyd, 2006; Phillips, Liljenquist, & Neale, 2009). Fazelpour and Steel (2021) incorporate these findings into their mathematical models. They show that while conformity is in general detrimental to inquiry, its negative effects are partially counteracted by introducing diversity into a community, when pressure to conform is small and identity-based. Their Bayesian agents update beliefs according to evidence, but perform actions based on a convex combination of their beliefs about the true value of the actions and conformist social pressures. However, we will see that when we consider identity-based conformity in conjunction with other feature of diverse communities (e.g. homophilic network formation), introducing diversity may no longer bring such benefits. Thus, results demonstrating the benefits of these psychological factors may not apply when we consider how people interact and share information in larger group settings.

[1] See (Bramson et al., 2017) for a further review of the literature on polarization.

3 Bayesian agents

We use idealized models of inquiry, in which the assumption of Bayesian agents is used to isolate and explore the impact of social factors. Using Bayesian agents in these sorts of models of inquiry is useful because it allows us to isolate particular social factors from other ways inquiry can go wrong to see the effects of those social factors. This might be thought of as a particular type of strategy in these agent-based models: we idealize away the various mistakes in reasoning real people make in order to focus on the casual effects of other relevant factors.[2]

Previous models of group inquiry take this same strategy as well: assuming Bayesian agents to isolate and explore the impact of social factors. (Note: we do not claim that any of the authors discussed here will agree with the our conceptualization or ascription of this strategy to their work. This is merely our understanding of one way their models licence the conclusions they draw.) To take an example from those models referenced above, Wu (2022)'s use of Bayesian agents allows her to conclude that (asymmetric) lack of trust in the testimony from members of marginalized groups, not a mistake in updating on their evidence, that leads members of dominant groups to come to incorrect conclusions.

This assumption of Bayesian agents is not only useful for studying diversity-relevant social factors. For instance, Holman and Bruner (2017) model the impact of industry funding on scientific research. They show that, if researchers whose (legitimate, but industry-favorable) methods tend to receive more industry funding, the resulting increase in productivity will allow them to train more future researchers in their methods, leading to their methods becoming over-represented in the field. This skews scientific research in favor of industry interests. However, industry-favorable research is not due to any corruption of the scientists; they are performing scientific research and updating properly on the evidence. Instead, this skewing of scientific research is due to the way industry funding interacts with the social structure of the community is set up in training new scientists.

Bayesian agents are generally used, and are seen as helpful, in understanding a kind of group inquiry where beliefs about different propositions matter. This is the kind of inquiry we will consider here, and is often captured by having agents face a bandit problem (which will be described further section 4.1.1). There are other ways modelers conceptualize group inquiry. Some types of inquiry are better conceived of as exchanging arguments or reasons, rather than exhanging evidence which Bayesian agents can update on (A. M. Borg, Frey, Šešelja, & Straßer, 2018; A. Borg, Frey, Šešelja, & Straßer, 2019; Singer et al., 2019). These models do not generally include Bayesian agents, though, some models of argumentation may include Bayesian agents updating on various components of an argument (Assaad et al., 2023). Some models capture exploration through a problem-space (often called an epistemic landscape), where individuals within a group attempt to find the best solutions to a problem, best methods to employ, or something similar (Weisberg & Muldoon, 2009; Grim, 2009;

[2] Of course, there is a substantial literature in philosophy on epistemic functions of simplified/idealized models, which we will not have room to discuss here. For some examples see: Frey and Šešelja (2018); Aydinonat, Reijula, and Ylikoski (2021); Šešelja (2022); Martini and Fernández Pinto (2017); Thicke (2020); Feiten (2023)

Thoma, 2015; Pöyhönen, 2017; Harnagel, 2019). Here, too, this way of modeling can be combined with Bayesian agents updating on evidence regarding the quality of solution, method, etc. (Huang, 2023). On the other hand, some models are aimed at arational group processes, such as opinion sharing (Hegselmann et al., 2002; Golub & Jackson, 2012b; Lassiter, 2021) and there doesn't seem to be much motivation for including Bayesian agents for these kinds of models.

Of course, real people are not perfect Bayesians, so there is a question of the applicability of the results found using a Bayesian agent modeling strategy. However, a qualitative match with experimental evidence where possible gives some confidence that we are not misrepresenting the deliberative process too much or in such a way that the results are undermined. For instance, Fazelpour and Steel (2021) compare their simulation results (found using a model that assumes Bayesian agents) to empirical evidence in cognitive science. They find similar qualitative results to experimental studies: diversity benefits inquiry when trust is higher within social identity groups or when there is more pressure to conform to one's own social group. These social factors (trust and conformity) and effects (the benefits of diversity) have been found in empirical studies. This gives us some confidence in the results of their model, while still leaving an important role for the idealized Bayesian agents to play: In any real group of people empirically studied, there is always the chance that demographically diverse groups benefit from diverse ways of reasoning (or other forms of so-called "cognitive diversity"), while in the model every agent updates in exactly the same (Bayesian) way.

4 Study 1

Study 1 examines the impact of two types of homophily—*identity-driven* and *opinion-driven*—on collective performance in sequential decision-making tasks. We investigated this influence along two specific pathways by which homophilic tendencies can impact group performance: (1) *preferential association* with similar others in network formation and (2) *higher trust* in the testimony of similar others. We explored the impacts of both types of homophily—identity and opinion—along the trust pathway, in addition to the network formation effects of identity-induced homophily.[3]

4.1 Method

4.1.1 Basic computational model.

In broad terms, group inquiry proceeds in our model as Bayesian agents gather and share evidence. Each instance of gathering evidence is an agent getting some data about one of two possible options, A or B. They perform whichever option they believe to be better multiple times and see how successful it is. The success rate is determined by how good that option actually is, but there is noise in the data, so it is not a perfect indicator. Once agents gather their data, they share with all the people

[3] Since network formation in our model is exogenous and opinions change throughout the simulations, we did not consider link formation based on opinion similarity. Future research could incorporate endogenous link formation in order to capture this structural effect of opinion homophily.

they are connected to, i.e. their neighbors on the network. These network connections might be influenced by homophily, indicating that agents are more likely to share and receive evidence from someone like them. After evidence is shared, all agents update their beliefs about the success rates of the options, and therefore which option they think is better. In updating, they take into account their own evidence and evidence of those they are connected to on the network. Here, homophily might also affect how much people take into account others' evidence, as they might weight a neighbor's evidence by some number less than one when updating their beliefs. Once agents have updated their beliefs, they begin a new round of inquiry where each agent experiments with option they now believe to be better, shares evidence with neighbors, and updates beliefs.

More specifically, the agents in our model face a two-armed bandit task. This is a standard formalization of a key type of sequential decision-making task facing research and innovation communities (Sutton & Barto, 2018; Daw, O'doherty, Dayan, Seymour, & Dolan, 2006). At each time point, an agent must decide between one of two options (e.g., a doctor choosing between two choices of treatment). While the agent is unaware of the objective payoffs of the options, it has subjective beliefs about these payoffs. By choosing to experiment with an option, the agent can be thought of as conducting a number of trials and observing the number of successes and failures that ensue. The task is to learn from this feedback at each time point (e.g., number of patients recovered) to find the superior alternative. We model the successes of experimenting with option (or arm), k, at a given time point as a random draw from a binomial distribution, $B(n, \pi_k)$, where n is the number of trials and π_k is k's objective probability of success (Zollman, 2010). The subjective beliefs about the successes of arm, k, is modeled as a beta distribution, $Beta(\alpha_k, \beta_k)$.[4]

Each agent in our studies can belong to one of two identity groups—a membership that can influence patterns of network formation, trust relation, or both (as described below). To model social relations, the agents in our studies are placed on networks of various types (again, described below). A connection between two agents in the network indicates a direct line of influence between them—e.g., in terms of receiving testimonial evidence from one another or observing each others' behavior. In addition to their direct observations, then, each agent also receives evidence from their neighbors in the social network. In this way, depending on the choice of their neighbors, the agents might also receive evidence about an option they themselves did not choose. At each time point, the agents update their beliefs about the payoff of options by incorporating the *weighted* sum of evidence (i.e., successes and failures observed over n trials) collected by themselves and their neighbors.[5] The weighting on a piece of evidence received from a neighbor depends on the focal agent's trust in that neighbor. Finally, given these belief distributions, agents always choose the option that currently has the highest estimated mean (or, as in the next study, highest overall perceived epistemic and non-epistemic value).[6] Figure 1 provides a schematic overview of the model.

[4] Beta distribution is the conjugate prior for binomial distribution (used here to model observed successes), which makes belief updating easier (Blitzstein & Hwang, 2015).

[5] When the evidence is drawn from a binomial distributions and beliefs are modeled as beta distributions, Bayesian updating proceeds as described in Figure 1.

[6] In other words, the agents are greedy and never explore seemingly inferior options. Exploration thus

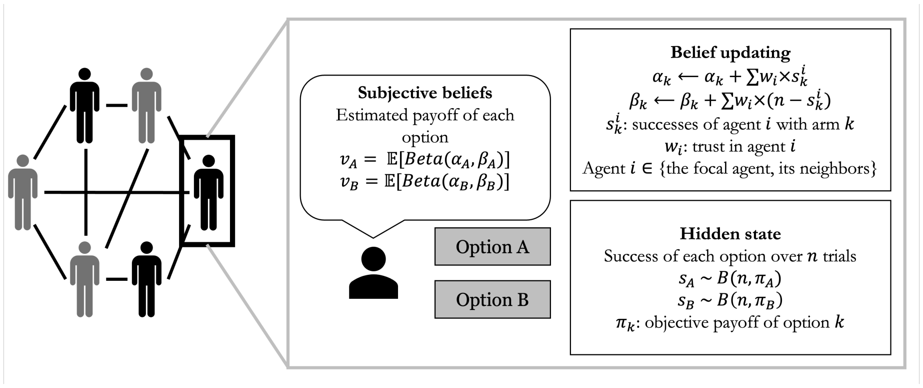

Subjective beliefs
Estimated payoff of each option
$v_A = \mathbb{E}[Beta(\alpha_A, \beta_A)]$
$v_B = \mathbb{E}[Beta(\alpha_B, \beta_B)]$
Option A
Option B
Belief updating
$\alpha_k \longleftarrow \alpha_k + \sum w_i \times s_k^i$
$\beta_k \longleftarrow \beta_k + \sum w_i \times (n - s_k^i)$
s_k^i: successes of agent i with arm k
w_i: trust in agent i
Agent $i \in$ {the focal agent, its neighbors}
Hidden state
Success of each option over n trials
$s_A \sim B(n, \pi_A)$
$s_B \sim B(n, \pi_B)$
π_k: objective payoff of option k

Figure 1: A schematic overview of the agent-based model used in our studies.

Network formation. We examine collective performance across three general types of (bidirectional) network structure: (1) *complete networks*: a fully connected network in which there is a direct link between any two agents, (2) *homophilic networks*: a topology where the pattern of connections between agents is shaped by their *identities*, and (3) *random* where connections are formed independent of identity.[7] Specifically, to construct homophilic networks, we use a variation of Erdős–Rényi random graphs (used to construct our random graphs) called multi-type random graphs, which are often used to model populations with multiple social identity groups (Golub & Jackson, 2012a; Rubin & O'Connor, 2018). In these networks, the probability a link is formed between two agents depends on whether they are in the same social identity group. In-group links (those where the agents are in same social identity group) are formed with a probability p_{in} and out-groups links (those where the agents are from different social identity groups) are formed with a potentially different probability p_{out}.

Of course, varying p_{in} and p_{out} changes both (i) the likelihood of identity-based clustering and (ii) a network's overall sparsity. We can disaggregate the impact of these two factors, and focus specifically on (i), by comparing homophilic networks with *different ratios* $\frac{p_{in}}{p_{out}}$, allowing p_{in} and p_{out} to vary such that, in all such networks, regardless of the ratio, the probability of a connection between any two agents, whatever their identities, remains *invariant*. The random networks that we considered correspond to structures where $\frac{p_{in}}{p_{out}} = 1$ and have the same overall connectivity as the homophilic network we compare them to.[8]

depends on the evidence from neighboring agents.

[7] Importantly, our sampling of potential homophilic and random networks is biased, in the sense that we only explore collective behavior in *connected* network structures—that is, topologies in which there exists a path between any two agents. This ensures that the entire group is involved in inquiry, and that our performance measures are not impacted by outliers (i.e. networks where disconnected agents are acting independently from the rest of the group).

[8] Given a (family of) homophilic network(s) with p_{in} and p_{out}, the probability that there is a connection between any two randomly selected nodes i and j depends on i's identity, j's identity, and the probability that there is a connection between i and j given their identities. In the case of a network of N agents with two

Trust relations between agents. As mentioned above, in this study, we considered two determinants of trust based on (1) similarity of group *identity* and (2) similarity of *opinion-based* like-mindedness. In general, we model trust as a weighting factor, w, in integrating information—successes and failures—arriving to an agent from its neighbors (see Figure 1). Weighting evidence by some measure of its perceived reliability in uncertain environments has been discussed in Bayesian literature, for example by Jeffery (Jeffrey, 1983). More relevant to us, Toelch and Dolan (2015) discuss weighting evidence by some measure of reliability in the context of social learning. Importantly, in the case of identity-based trust w remains fixed, insofar as we keep group identities static in our model. In contrast, since opinions can change as agents gather more observations and update their beliefs, w values for opinion-based trust are dynamic.[9]

Identity-based trust. We follow Fazelpour and Steel (2021) in modeling the impact of identity-based trust on information integration. According to this model, while the evidence from in-group neighbors is treated as if it was directly observed (i.e., $w = 1$ for in-groups), agents give relatively less weight to evidence arriving from out-groups. That is, the successes and failures reported about an option by an out-group neighbor are weighted by a fixed factor, $0 \leq w \leq 1$.

Opinion-based trust. In order to formalize dynamic trust, we need to specify how agents can estimate like-mindedness given the information available to them. We develop a novel measure for dynamic trust here, which we argue captures, intuitively, how trust changes over time in this sort of social situation (described below). A natural way of doing so is for agents to simply track behavioral similarities with others. Specifically, let $A_n^i = \{a_1^i, a_2^i, ..., a_n^i\}$ be agent i's action sequence up to and including experiment n. Agent i can estimate a neighbor j's like-mindedness by comparing their action sequences, such that i's trust in j at n is given by

$$w_n^{ij} = \frac{\sum_{t=1}^{n} \mathbb{1}[a_t^i = a_t^j]}{n}$$

where $\mathbb{1}[.]$ is an indicator function that takes the value of 1 when the agents perform the same action and is 0 otherwise. Accordingly, agents' trust in others ranges between 0 and 1: an agent fully trusts a neighbor when they have taken exactly the same action at each time point. Conversely, agents will have no trust in a neighbor, if they have always chosen different options.

We can get a better intuition about the trust relation by examining an incremental formulation of the equation above

$$w_n^{ij} = w_{n-1}^{ij} + \frac{1}{n} \times (\mathbb{1}[a_n^i = a_n^j] - w_{n-1}^{ij})$$

groups of size N_A and $N - N_A$, this probability can be computed as:

$$p_{any} = \frac{N_A}{N} \times [\frac{N_A - 1}{N - 1} \times p_{in} + \frac{N - N_A}{N - 1} \times p_{out}] + \frac{N - N_A}{N} \times [\frac{N_A}{N - 1} \times p_{out} + \frac{N - N_A - 1}{N - 1} \times p_{in}]$$

It is not difficult to see how p_{in} and p_{out} can in turn be calculated from p_{any} and a specified ratio $\frac{p_{in}}{p_{out}}$. While we focus on the case involving two groups, it is straightforward to extend this beyond the binary case.

[9] In both cases, we assume agents fully trust themselves, i.e., adopt a $w = 1$ in weighting their own observations.

Viewed in this way, i's trust in j at a given time point depends on i's trust in j at the previous time point as well as whether and how j's latest action deviates from (or aligns with) that prior trust. If the deviation from prior trust is positive,[10] then i's trust in j will increase. And trust will decrease, if the deviation is negative. Note that the amount with which trust changes in cases of deviation depends on the particular time point, in the sense that deviations from prior trust are more influential at earlier time points (i.e., smaller n).[11] Intuitively, if an agent is accustomed to listening to another agent, and has been for a long time, it will take more to dissuade then from listening to that agent in the future, whereas if an agent is just starting to talk to another, their behavior as the agent still getting a sense of what they believe in will matter a lot for trust relations down the line.

Experimental design and procedures. We explored the impact of the two types of homophily along network formation and trust pathways across a wide range of parameter settings. Specifically, we varied the range of agents between 10 and 80 (increments of 2). To examine the impact of group size disparity, we varied the size of one of the two groups from 5 agents to 50% of the total network size. In terms of network topologies, in addition to complete network structures, we examined homophilic networks with ratios $\{1, 2, 4, 8\}$ (constructed as described above). While a network with ratio 1 amounts to a random network with no homophilic tendencies, a ratio of 8 indicates that agents are 8 times more likely to connect to in-group others. For identity-based trust, we varied w between 0.05 and 1 (increments of 0.05).[12] Note, however, that below we often refer to the case of $w = 0.1$ as identity-based trust. Finally, throughout, we kept the objective probability of payoff for the two options fixed with $\pi_A = 0.499$ and $\pi_B = 0.5$.

In examining the impact of homophily, we consider two aspects of group performance. We look at *reliability*, the percentage of simulation runs ending in correct, unanimous consensus. In order to gain some insight about the speed at which these communities are arriving at correct (or incorrect) beliefs, we also look at *efficiency*, which is evaluated by comparing reliability after a different time horizons (or number of experiments).

4.2 Results and discussion

We find that sparser networks (random as well as various degrees of homophily) to be more conducive to successful inquiry at longer time horizons when compared to the complete network. This is in line with previous studies (e.g, Zollman (2007)) and can be explained by the fact that in complete networks the transmission of misleading

[10] That is, when j acts in the same way as i, in contrast to i's prior expectation.

[11] This also provides a more general framing of dynamic trust as: $w_n^{ij} = w_{n-1}^{ij} + \tau \times (\mathbb{1}[a_n^i = a_n^j] - w_{n-1}^{ij})$, where τ can be thought of as a factor that determines how much agents tend to adjust their trust in cases of deviations. While in the current model, with $\tau = \frac{1}{n}$, the tendency to adjust will decrease with time, we can also have constant τ in ways that would place more weight on *recent* deviations or (dis)agreements (e.g., when τ is close to 1). This general form occurs in many different learning situations (see, e.g., (Sutton & Barto, 2018)). One might imagine an even broader version with asymmetric change of trust in light of agreement versus disagreement.

[12] Only a subset of these parameters are shown below to focus attention on a few key results.

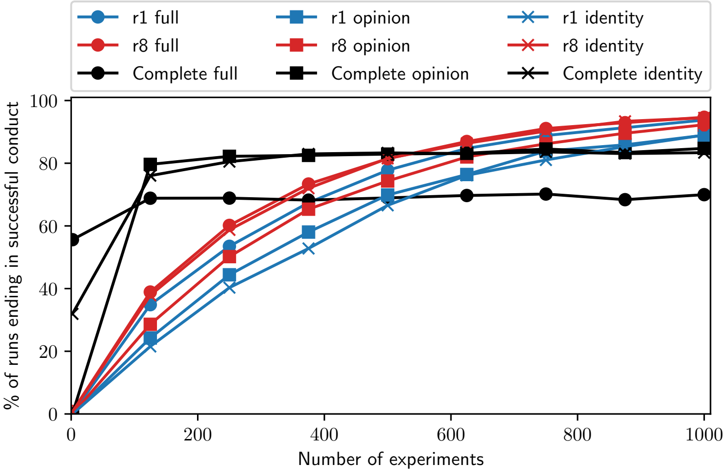
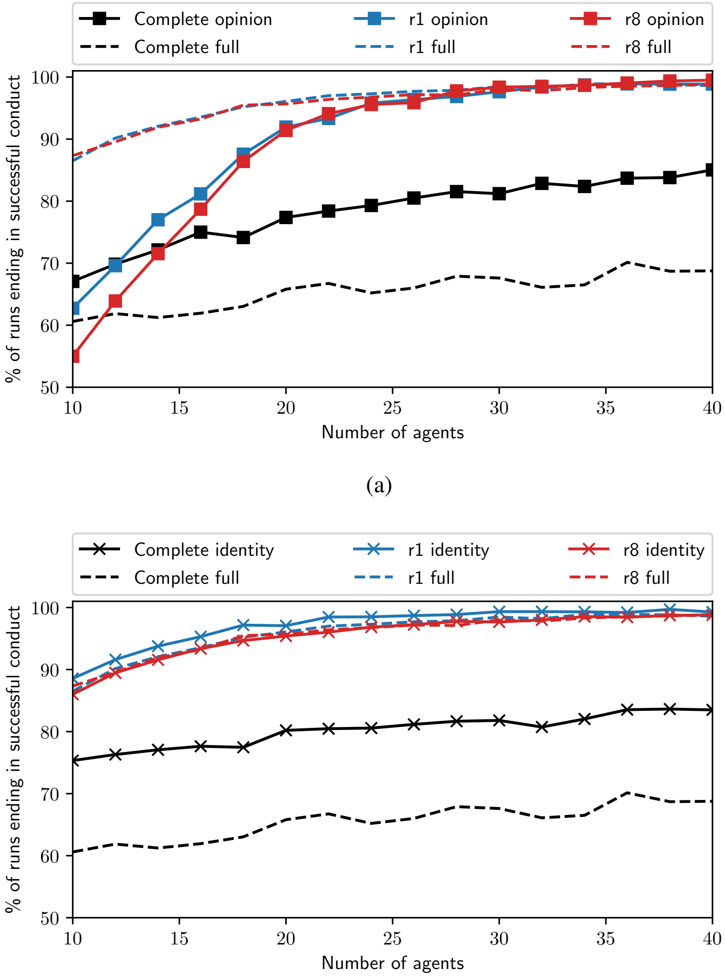


Figure 2: The impact of homophily on efficiency via network formation and trust [Full $w = 1$, opinion-driven, identity-driven $w = 0.1$]. rk refers to homophilic networks where $p_{in} = k \times p_{out}$, while keeping the overall connectivity fixed. Specifically, $r2$ corresponds to a network with $[p_{in}, p_{out}] = [0.4, 0.2]$, $r1$ is the corresponding random network, and $r8$ is a network with $p_{in} = 8p_{out}$. The complete network is included for contrast purposes. All networks of 40 agents with parity of representation. The first data point corresponds to 2 experiments.

results is swift and widespread. As a result, agents in complete networks are particularly susceptible to reaching premature consensus on the wrong option. This is made less likely by sparsity in network connection, as can be seen by comparing the longer term performances of networks with full inter-agent trust in Figure 2 ("Complete full" vs. "r1 full" and "r8 full"). Importantly, the transmission of (mis)information can also be slowed down by lowering levels of trust. Hence, as the figure shows, the performance of complete networks improves when agents' trust behavior is governed by either identity-based or opinion-based considerations (as opposed to full trust). This is similar to findings by Fazelpour and Steel (2021), though they only examine the identity-driven case.

We did not find any appreciable differences in reliability, or success at longer time horizon between networks with varying degrees of homophily (when keeping the type of trust fixed). We did, however, find that across all types of trust, increased identity-driven associations *increased* the efficiency of learning, which was particularly salient in the case of identity-driven trust (see Figure 2).

Increasing the size of the network improved reliability across all network and trust types. Figure 3a shows results for opinion-based trust and figure 3b shows results for identity-based trust. For each of these conditions, the benefits introducing homophilic trust in the complete network are observed for all group sizes. The effects of intro-



(a)

(b)

Figure 3: The impact of homophily on reliability across networks of different size. Homophily's impact on performance (a) via opinion-driven trust and (b) via identity-driven trust ($w = 0.1$). Networks with full inter-group trust ($w = 1$) are shown for contrastive purposes. All networks of 40 agents with parity of representation.

ducing homophilic trust in random and homophilic networks are less straightforward. Looking at figure 3b, it appears there might be some small benefit to introducing identity based trust in random networks when group size is small, but the effects of introducing this type of homophilic trust seem to be negligible overall. That is, identity-driven (as opposed to full) trust had no appreciable impact on successful performance in these networks.

The relationship between opinion-driven trust and network structure is more interesting ("$r1$ opinion" and "$r8$ opinion" in Figure 3a). Specifically, opinion-driven trust was highly detrimental in smaller group sizes. A possible explanation is that the reduction of trust in neighbors who are not seen as like-minded is particularly problematic when the number of neighbors is small to begin with. In such cases agents can quickly end up receiving only evidence that confirms their beliefs. A closer look at our findings supports this explanation: a substantial portion of simulation runs in these networks end up with general polarization (i.e., cases where the collective ends with clusters of opposing, stable opinions that do not fall along identity lines). In "$r1$ opinion" networks, for example, such outcomes constitute 32% of runs in groups of size 10, 26% of runs in groups of size 12 (compared to 4% and 3% respectively, in "$r1$ identity", where trust is driven by identity).

5 Study 2

In study 2, we consider how the presence of identity-based conformity can impact the results from the previous section. Such (normative) conformity pressures are prevalent in real-world settings (Cialdini & Goldstein, 2004; Deutsch & Gerard, 1955), and they are critical to incorporate, because in real-world settings, they are likely to co-present with other manifestations of homophily.

5.1 Method

When conformity pressures are present, the behavior of agents is no longer a faithful reflection of their beliefs. Thought they may be Bayesian agents updating their beliefs properly, their behavior may not accurately reflect this. As a result, in addition to formalizing how conformity pressures shape agent behavior, we also need to consider the impact of this effect on other aspects of behavior (e.g., opinion-based trust) and relevant outcomes (e.g., polarization).

Conformity in diverse networks. While conformity in general is a much studied topic in social psychology, the study of conformity's impact in identity diverse groups is relatively recent (Phillips, Mannix, Neale, & Gruenfeld, 2004; Phillips & Loyd, 2006; Gaither, Apfelbaum, Birnbaum, Babbitt, & Sommers, 2018). Here we use the formalization of conformity's impact on individual decision-making in diverse groups in Fazelpour and Steel (2021):

$$u_j^i = (1 - \kappa) \times v_j^i + \kappa \times \frac{\mathcal{N}_{in}^i(j)}{\mathcal{N}^i}$$

Where u_j^i represents the total perceived value of pursuing option j for agent i. v_j^i is agent i's perceived expected payoff of option j (see Figure 1). $\mathcal{N}^i$ is the total number of i's neighbors and $\mathcal{N}_{in}^i(j)$ are the subset of neighbors who share the same group identity with i (i.e., are considered in-group by i) that pursued option j in the previous time point. Finally, κ represents i's conformist tendency. When $\kappa = 0$, agents simply follow their personal beliefs, but when $\kappa = 1$, agents just follow the majority decision from in-group majority.[13]

This formulation is a modification of the "Other-Total Ratio" (Stasser & Davis, 1981) that is meant to capture two key findings about conformity's impact in identity diverse groups: (1) individuals primarily feel the (normative) pressure to conform to in-groups (reflected in the numerator of the fraction) (Antonio et al., 2004); and (2) the *mere presence* of out-group individuals, regardless of their views, reduces conformity pressure (reflected in the fixed denominator of the fraction) (Phillips, 2017).

Opinion-based trust in the presence of conformity. The presence of conformity complicates our formulation of opinion-based trust as perceived "like-mindedness", since agents might act contrary to their beliefs because of in-group conformity pressure. In this case, agent i observing neighbor j acting in the same way will not necessarily convey j's "like-mindedness" to i. In fact, one could imagine this surface agreement to *increase* i's distrust in j, if i is choosing an action they privately disagree with. To deal with this type of scenario, we assume that, instead of considering what it actually did, the focal agent i compares what it would have done had there been no conformity pressure with j's actual actions. The agents thus adopt an asymmetric attitude towards their own versus others' conduct, downplaying the influence of situational factors (i.e., conformity pressure) in the case of others, but not in their own case. While this is clearly a simplification[14], as the literature on fundamental attribution error in social psychology shows, in many circumstances people do seem to act in similar ways (Ross, 1977).

5.1.1 Experimental design and procedures.

We explored how the presence of identity-based conformity might influence the outcomes of the previous section by varying the extent of conformity pressure between 0 and 0.02 (with increments of 0.002).[15] The presence of conformity requires that we adopt a more fine-grained lens on dependent outcomes. We introduce six new categories of dependent outcomes:

- *Correct all:* Simulation runs that end with all agents pursuing the superior option and believing in their choice.
- *Correct but:* Simulation runs that end with all agents pursuing the superior option, despite the fact that some agents do so as a result of conformity and against their beliefs.
- *Incorrect all:* Simulation runs that end with all agents pursuing the inferior option and believing in their choice.

[13] Throughout, we use the same κ for all agents.

[14] For example, the model ignores the more complicated case where agents might doubt each other's sincerity, thus not taking their actions to be directly reflective of their underlying beliefs. See Mohseni and Williams (2019) for a consideration of this more complicated case.

[15] Given the small different between the objective payoff of the two options, anything outside this range simply amounts to purely conformist behavior.

- *Incorrect but:* Simulation runs that end with all agents pursuing the inferior option, despite the fact that some agents do so as a result of conformity and against their beliefs.
- *Inter-group polarization:* Simulation runs that end with (belief) consensus within identity groups and opposing views between groups.
- *General polarization:* Simulations runs that end with no consensus (in general or within groups).

5.2 Results and discussion

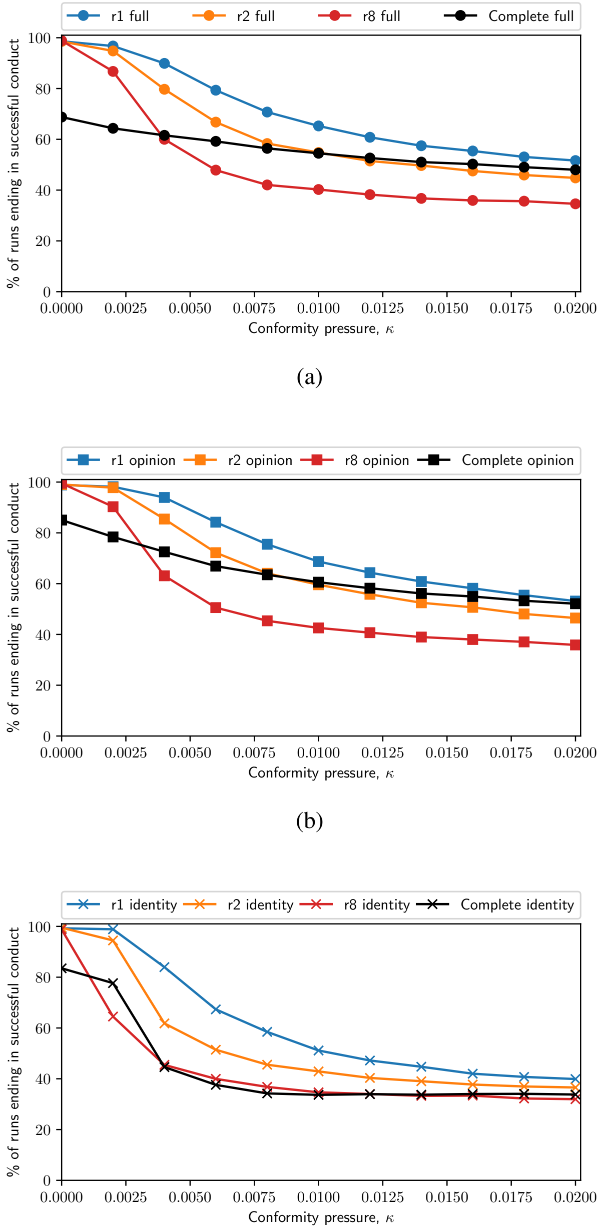
As shown in Figure 4 and consistent with previous studies (O'Connor & Weatherall, 2018; Fazelpour & Steel, 2021), we find normative conformity to be detrimental to successful performance across all network and trust types. Importantly, our results go beyond previous findings that were mainly focused on the impact of conformity in *complete* networks. In particular, we find that the extent of conformity's detrimental impact critically depends on network structure (e.g., random vs. homophilic) and trust type (e.g., full vs. identity- vs. opinion-driven).[16]

Specifically, conformity is particularly detrimental with increased structural homophily and identity-driven trust (and worse still when these are combined). In each figures 4a, 4b, and 4c, increasing homophily (from $r1$ to $r2$, then to $r8$) leads reliability to decrease faster as we increase conformity, κ. As seen in figure 4c, identity based trust leads to worse overall outcomes than opinion-based trust for all network types. Further, this decrease in reliability is particularly sharp for homophilic networks. For instance, once even the smallest amount of conformity is introduced, $r8$ networks decrease from near 100% reliability to less than 70%. This is compared to both the full and opinion-driven trust conditions, where reliability remains above 80%.

That conformity is particularly detrimental with both structural homophily and identity-driven trust (and even worse when those are combined) is to be expected, since homophilic networks in effect decrease diversity in the neighborhood of an agent, thus increasing the conformity pressure on that agent. Identity-based trust exacerbates this situation by preventing agents to learn about the potential superiority of alternative courses of action from the testimony of out-groups.

Perhaps surprisingly, opinion-based trust *curtails* the negative influence of conformity even in homophilic network structures, as figure 5 shows. A possible explanation is that while opinion-driven trust cannot decrease the normative influence of in-group conformity, it can decrease the *epistemic* impact of in-groups when the agent disagrees with them. At the same time, it can lead agents to trust like-minded out-group members. Accordingly, agents are more likely to form correct beliefs or at least beliefs that are in line with certain out-group members. This can in turn result in agents pursuing the superior option and an overall decrease in inter-group polarization. A comparison of the extent of inter-group polarization at lower levels of conformist tendency κ provides support this explanation. Looking at figure 6, we can see that for $\kappa = .0002$, there

[16] Some impacts of network structure also depend on conformity. For instance, in each figures 4a, 4b, and 4c, increasing homophily (from $r1$ to $r2$, then to $r8$) decreases reliability for all $\kappa > 0$ but not for $\kappa = 0$.

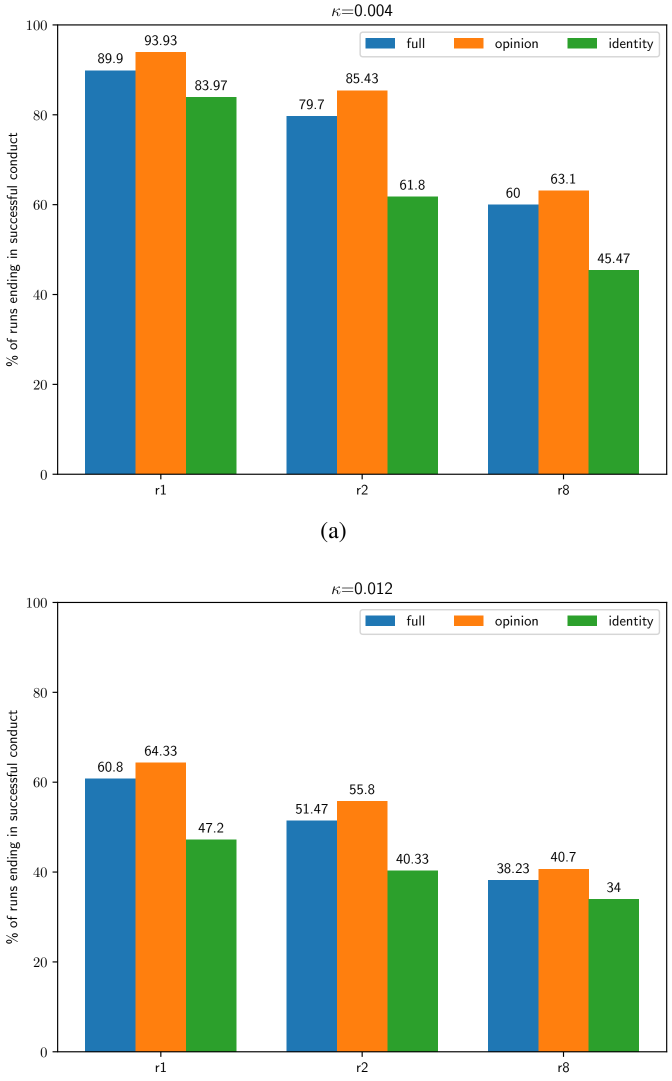


(a)

(b)

(c)

Figure 4: The impact of identity-induced conformity on reliability in (a) full trust (w = 0.1), (b) opinion-driven trust, and (c) identity-driven trust (w = 0.1). All networks of 40 agents with parity of representation.



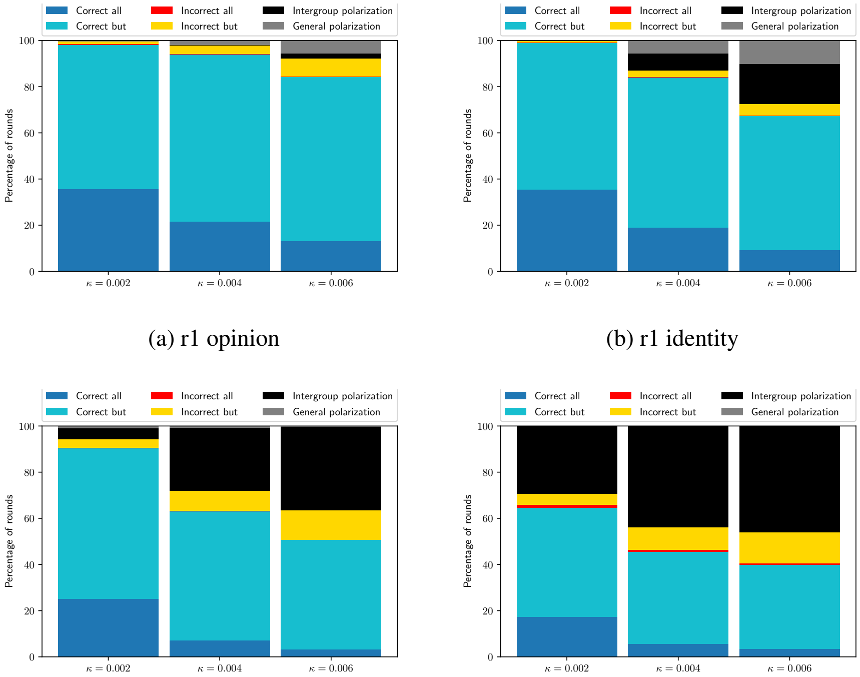
(a)

(b)

Figure 5: The impact of type of trust on reliability for different levels of homophily in (a) a lower conformity condition and (b) a higher conformity condition. All networks of 40 agents with parity of representation.

is much more inter-group polarization in the identity based trust condition (6d) than the opinion-based trust condition (6c).

Figure 6 also helps us conceptualize and explain the general negative effects of homophily. For instance, we can see that the blue areas of the graphs (representing outcomes where everyone performs the correct action) decrease more quickly in ho-



(a) r1 opinion

(b) r1 identity

(c) r8 opinion

(d) r8 identity

Figure 6: The impact of identity-induced conformity on performance disaggregate by outcome types in random networks (r1) with (a) opinion-driven, (b) identity-driven trust; and homophilic networks (r8) with (c) opinion-driven and (d) identity-driven trust. All networks of 40 agents with parity of representation.

mophilic (figures 6c and 6d) versus random graphs (6a and 6b). This is because intergroup polarization growing is much more rapidly as κ increases in homophilic (figures 6c and 6d) versus random (6a and 6b) graphs. Intergroup polarization is particularly common in homophilic networks with identity-driven trust. Because people are both more likely to be connected to and feel pressure to conform to people of their same social identity, we often end up with each social identity group converging on a different action.

6 Discussion

We find that the relationship between homophily and collective performance is complicated. Whether homophily is beneficial – in terms of reliability and/or efficiency – depends both on its particular manifestation. Additionally, whether a certain factor is beneficial depends on the presence of other mediating factors. For example, interestingly, we find that opinion-driven trust impedes convergence to truth in Study 1, but the effect flips in the presence of identity-driven conformity where opinion-driven trust is beneficial. Taking these factors into account has implications for how we make inferences based on research in cognitive science regarding the epistemic benefits of demographic diversity. While lack of trust and differential conformity may be beneficial in experimental settings with small groups, we have shown that these same features of diverse communities can be detrimental to inquiry in larger groups. This makes it difficult to draw inferences from these empirical studies regarding the expected effects of demographic diversity on, e.g., inquiry in scientific communities.

Our results also complicate conclusions regarding how to evaluate policy proposals intending to rectify matters of inequality in collaborative organizations. Diversity according to social identity has been argued to be important to inquiry, resulting in many arguments that we ought to promote demographic diversity because of the ensuing gains in effective inquiry or performance of groups. Arguments of this sort – referred to as "the business case for diversity" or "instrumental diversity rationale" – presume that promoting equity of a certain kind will go hand-in-hand with receiving the benefits of diversity.[17] Often these sorts of arguments have a "private sins as public goods" character, as they aim to convince those only interested in epistemic gains to incidentally promote socially beneficial policies (Schneider, Rubin, & O'Connor, 2021).

There are known issues with the business case for diversity. First, it is morally dubious. In treating marginalized or underrepresented groups as means to achieve an end, it is dehumanizing and justifies exploitative practices (Prescod-Weinstein, 2021; Fehr, 2011; Fehr & Jones, 2022). In fact, the business case often focuses on benefits to members of privileged groups, is associated with increased disparities in academic communities (Starck et al., 2021), and is detrimental to diversity in organizations more generally (Georgeac & Rattan, 2020). Further, many of the benefits of demographic diversity may in fact rest on problematic aspects of social interactions, e.g. lack of trust or devaluation of testimony from marginalized social identity groups (Steel & Bolduc, 2020; Fazelpour & Steel, 2021; Wu, 2022), and attempts to promote diversity by intervening on the structure of communities may backfire and further entrench inequity (Schneider et al., 2021). We add to this list a further complication: whether a certain feature of socially diverse communities is epistemically beneficial at the group level may depend on its particular manifestation.

Taking these factors into account has implications for how we think about implementing and how we justify policy proposals aimed at increasing diversity. We cannot attempt to achieve our epistemic aims through diversity initiatives while ignoring underlying social structures or cultures – this will likely lead to ineffective policies which fail to achieve the desired epistemic benefits. (See also Fehr and Jones (2022)). This is not to argue we ought not to care about epistemic benefits, but policies justified on the basis of a business case for diversity generally ignore background social structures (like homophily) that undercut the ability for diversity to generate the epistemic benefits the policy aims to produce. Rather, we might attempt to foster diversity with policies that simultaneously foster both a demographic and cultural shift within scientific communities (Fehr & Jones, 2022). For example, findings from Study 2 may be

[17] See Steel and Bolduc (2020) and Starck, Sinclair, and Shelton (2021) for an overview of this literature.

seen as suggesting that making opinion similarities salient can decrease the detrimental effects of group-based conformity in interdisciplinary teams. Whether these sorts of results hold in a model suitably adapted to capture interdisciplinary research is the subject of future study.

In this paper, we idealize group inquiry in that we assume groups are comprised of Bayesian agents. This allows us to isolate particular social factors from other ways inquiry can go wrong to see the effects of those social factors. Future work can use this same modeling strategy to investigate the impacts of other social factors. For instance, the model presented here could be extended to include asymmetries between the two groups – such as representation, status, or power differences – which are often present in both demographically diverse groups and interdisciplinary collaborations. Such asymmetries may lead to one group's ideas being both better (in some sense) and less likely to be taken into account, similar to Wu (2022)'s simulation findings that there is a connection between marginalized group members tending to be both epistemically privileged and ignored, or Hofstra et al. (2020)'s empirical results that minority group members are both more likely to produce innovative ideas but are less like to have their ideas taken up by the community.

Focusing on *when* certain aspects of diverse communities are beneficial, which has been useful for thinking about empirical studies (Sulik et al., 2021), is also useful for simulation studies aimed at studying diversity and is important in evaluating proposals to increase diversity. Since different dimensions and impacts of homophily can be co-present, disentangling the sources and consequences of homophily in real-world communities is key. The simulations presented here isololate these important social factors and offer an important first step in this direction, providing theoretical insight into which aspects of homophily impede or promote successful inquiry and under what circumstances.

References

Ahmad, M. A., Ahmed, I., Srivastava, J., & Poole, M. S. (2011). Trust me, i'm an expert: Trust, homophily and expertise in mmos. In *2011 IEEE third international conference on social computing* (pp. 882–887).

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

Antonio, A. L., Chang, M. J., Hakuta, K., Kenny, D. A., Levin, S., & Milem, J. F. (2004). Effects of racial diversity on complex thinking in college students. *Psychological Science*, *15*(8), 507–510. doi: 10.1111/j.0956-7976.2004.00710.x

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

Aydinonat, N. E., Reijula, S., & Ylikoski, P. (2021). Argumentative landscapes: the function of models in social epistemology. *Synthese*, *199*(1), 369–395.

Blitzstein, J. K., & Hwang, J. (2015). *Introduction to probability*. Boca Raton, Florida: CRC Press.

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*, 1–25.

Borg, A. M., Frey, D., Šešelja, D., & Straßer, C. (2018). Epistemic effects of scientific interaction: Approaching the question with an argumentative agent-based model. *Historical Social Research/Historische Sozialforschung*, *43*(1 (163), 285–307.

Boschini, A., & Sjögren, A. (2007). Is team formation gender neutral? evidence from coauthorship patterns. *Journal of Labor Economics*, *25*(2), 325–365.

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.

Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, *55*(1), 591–621. doi: 10.1146/annurev.psych.55.090902.142015

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Deffuant, G., Amblard, F., Weisbuch, G., & Faure, T. (2002). How can extremism prevail? a study based on the relative agreement interaction model. *Journal of artificial societies and social simulation*, *5*(4).

del Carmen, A., & Bing, R. L. (2000). Academic productivity of African Americans in criminology and criminal justice. *Journal of Criminal Justice Education*, *11*(2), 237–249.

Derex, M., & Boyd, R. (2016). Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences*, *113*(11), 2982–2987.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, *51*(3), 629–636. doi: 10.1037/h0046408

Fazelpour, S., & Steel, D. (2021). Diversity, trust and conformity: a simulation study. *Philosophy of Science*.

Fehr, C. (2011). What is in it for me? the benefits of diversity in scientific communities. In *Feminist epistemology and philosophy of science* (pp. 133–155). Springer.

Fehr, C., & Jones, J. (2022). Culture, exploitation, and the epistemic approach to diversity.

Feiten, T. E. (2023). The map/territory relationship in game-theoretic modeling of cultural evolution. *Philosophy of Science*, *90*(5), 1427–1436.

Ferber, M. A., & Teiman, M. (1980). Are women economists at a disadvantage in publishing journal articles? *Eastern Economic Journal*, *6*(3/4), 189–193.

Frey, D., & Šešelja, D. (2018). What is the epistemic function of highly idealized agent-based models of scientific inquiry? *Philosophy of the Social Sciences*, *48*(4), 407–433.

Gaither, S. E., Apfelbaum, E. P., Birnbaum, H. J., Babbitt, L. G., & Sommers, S. R. (2018). Mere Membership in Racially Diverse Groups Reduces Conformity. *Social Psychological and Personality Science*, *9*(4), 402–410. doi: 10.1177/1948550617708013

Georgeac, O., & Rattan, A. (2020). The business case for diversity backfires: detrimental effects of organizations' instrumental diversity rhetoric for underrepresented group members' sense of belonging and performance. *Journal of personality and social psychology*.

Ghiasi, G., Mongeon, P., Sugimoto, C., & Larivière, V. (2018). Gender homophily in citations. In *23rd international conference on science and technology indicators (sti 2018)(september 2018)* (pp. 1519–1525).

Golub, B., & Jackson, M. O. (2012a). How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, *127*(3), 1287–1338.

Golub, B., & Jackson, M. O. (2012b). Network structure and the speed of learning measuring homophily based on its consequences. *Annals of Economics and Statistics/ANNALES D'ÉCONOMIE ET DE STATISTIQUE*, 33–48.

Gomez, C. J., & Lazer, D. M. (2019). Clustering knowledge and dispersing abilities enhances collective problem solving in a network. *Nature communications*, *10*(1), 1–11.

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

Harnagel, A. (2019). A mid-level approach to modeling scientific communities. *Studies in History and Philosophy of Science Part A*, *76*, 49–59.

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).

Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., & McFarland, D. A. (2020). The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, *117*(17), 9284–9291.

Holman, B., & Bruner, J. (2017). Experimentation by industrial selection. *Philosophy of Science*, *84*(5), 1008–1019.

Huang, A. C. (2023). Landscapes and bandits: A unified model of functional and demographic diversity. *Philosophy of Science*, 1–16.

Jackson, M. O. (2010). *Social and economic networks*. Princeton university press.

Jeffrey, R. C. (1983). Bayesianism with a human face.

Lassiter, C. (2021). Arational belief convergence. *Synthese*, *198*(7), 6329–6350.

Martini, C., & Fernández Pinto, M. (2017). Modeling the social organization of science: Chasing complexity through simulations. *European Journal for Philosophy of Science*, *7*(2), 221–238.

McDowell, J. M., & Smith, J. K. (1992). The effect of gender-sorting on propensity to coauthor: Implications for academic promotion. *Economic Inquiry*, *30*(1), 68–82.

Meadows, M., & Cliff, D. (2012). Reexamining the relative agreement model of opinion dynamics. *Journal of Artificial Societies and Social Simulation*, *15*(4), 4.

Mohseni, A., & Williams, C. R. (2019). Truth and conformity on networks. *Erkenntnis*, 1–22.

Olsson, E. J. (2013). A bayesian simulation model of group deliberation and polarization. In *Bayesian argumentation* (pp. 113–133). Springer.

O'Connor, C., & Weatherall, J. O. (2018). Scientific polarization. *European Journal for Philosophy of Science*, *8*(3), 855–875.

Page, S. E. (2017). *The diversity bonus*. Princeton University Press.

Paris, G., De Leo, G., Menozzi, P., & Gatto, M. (1998). Region-based citation bias in science. *Nature*, *396*(6708), 210–210.

Phillips, K. W. (2017). Commentary. what is the real value of diversity in organizations? questioning our assumptions. In *The diversity bonus* (pp. 223–246). Princeton University Press.

Phillips, K. W., Liljenquist, K. A., & Neale, M. A. (2009). Is the pain worth the gain? the advantages and liabilities of agreeing with socially distinct newcomers. *Personality and Social Psychology Bulletin*, *35*(3), 336–350.

Phillips, K. W., & Loyd, D. L. (2006). When surface and deep-level diversity collide: The effects on dissenting group members. *Organizational Behavior and Human Decision Processes*, *99*, 143–160. doi: 10.1016/j.obhdp.2005.12.001

Phillips, K. W., Mannix, E. A., Neale, M. A., & Gruenfeld, D. H. (2004). Diverse groups and information sharing: The effects of congruent ties. *Journal of Experimental Social Psychology*, *40*, 497–510. doi: 10.1016/j.jesp.2003.10.003

Pöyhönen, S. (2017). Value of cognitive diversity in science. *Synthese*, *194*(11), 4519–4540.

Prescod-Weinstein, C. (2021). *The disordered cosmos: A journey into dark matter, spacetime, and dreams deferred*. Hachette UK.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology* (Vol. 10, pp. 173–220). Elsevier.

Rubin, H., & O'Connor, C. (2018). Discrimination and collaboration in science. *Philosophy of Science*, *85*(3), 380–402.

Schneider, M. D., Rubin, H., & O'Connor, C. (2021). Promoting diverse collaborations. In G. Ramsey & A. De Block (Eds.), *The dynamics of science: Computational frontiers in history and philosophy of science.* Pittsburgh University Press.

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

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

Starck, J. G., Sinclair, S., & Shelton, J. N. (2021). How university diversity rationales inform student preferences and outcomes. *Proceedings of the National Academy of Sciences*, *118*(16).

Stasser, G., & Davis, J. H. (1981). Group decision making and social influence: A social interaction sequence model. *Psychological Review*, *88*(6), 523–551. doi: 10.1037/0033-295X.88.6.523

Steel, D., & Bolduc, N. (2020). A closer look at the business case for diversity: The tangled web of equity and epistemic benefits. *Philosophy of the Social Sciences*, *50*(5), 418–443.

Sulik, J., Bahrami, B., & Deroy, O. (2021). The diversity gap: when diversity matters for knowledge.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Thicke, M. (2020). Evaluating formal models of science. *Journal for General Philosophy of Science*, *51*(2), 315–335.

Thoma, J. (2015). The epistemic division of labor revisited. *Philosophy of Science*, *82*(3), 454–472.

Toelch, U., & Dolan, R. J. (2015). Informational and normative influences in conformity from a neurocomputational perspective. *Trends in cognitive sciences*, *19*(10), 579–589.

Turner, J. C., Wetherell, M. S., & Hogg, M. A. (1989). Referent informational influence and group polarization. *British journal of social psychology*, *28*(2), 135–147.

Wang, Y. S., Lee, C. J., West, J. D., Bergstrom, C. T., & Erosheva, E. A. (2019). *Gender-based homophily in collaborations across a heterogeneous scholarly landscape.*

Wardle, D. A. (1995). Journal citation impact factors and parochial citation practices. *Bulletin of the Ecological Society of America*, *76*(2), 102–104.

Warkentin, M., Sharma, S., Gefen, D., Rose, G. M., & Pavlou, P. (2018). Social identity and trust in internet-based voting adoption. *Government Information Quarterly*, *35*(2), 195–209.

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

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PloS one*, *8*(7), e66212.

Wu, J. (2022). Epistemic advantage on the margin. *Philosophy and Phenomenological Research*.

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.