Why do echo chambers form?
The role of trust, population heterogeneity, and objective truth

Amy Perfors (amy.perfors@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne

Danielle J. Navarro (d.navarro@unsw.edu.au)
School of Psychology, University of New South Wales

Abstract

Many real-world situations involve learning entirely or mostly based on the information provided by other people, which creates a thorny epistemological problem: how does one determine which of those people to trust? Previous work has shown that even populations of rational Bayesian agents, faced with this problem, polarise into "echo chambers" characterised by different beliefs and low levels of between-group trust. In this study we show that this general result holds even when the reasoners have a more complex meaning space and can communicate about their beliefs in a more nuanced way. However, even a tiny amount of exposure to a mutually trusted "ground truth" is sufficient to eliminate polarisation. Societal and psychological implications are discussed.

Keywords: Bayesian reasoning; echo chambers; polarisation; social inference; trust; epistemology

Introduction

The real world is full of situations where the vast majority of what we learn comes from other people. In some, like language learning, the "ground truth" of the matter simply *is* whatever people agree that it is. However, many other situations pose a much more challenging epistemological problem: the ground truth is (at least mostly) inaccessible, and the only way to learn about it is to rely on other people. Regardless of why the truth is often inaccessible – due to spatial or temporal distance, or difficulties in interpreting ambiguous data – people are often faced with questions of this character. Did humans evolve or were we created by a superior being? Did Trump assist the Russians to influence the US 2016 elections? Did Bob have an affair with Mindy? In all of these cases, there *is* a truth of the matter, but it is not a truth that is directly accessible to most people. All of the data is mediated through other agents – scientists studying evolution, politicians receiving confidential documents, journalists deciding what to report on, Bob and Mindy – and few have the access or training necessary to make sense of the data on their own.

What is a rational learner to do in this difficult epistemological situation? One option would be to simply try to communicate fully with everybody and update one's beliefs accordingly. When this happens, groups of Bayesian learners will converge to a shared belief system equivalent to the population prior, at least when organised as chains (Griffiths & Kalish, 2007) or fully interconnected (Whalen & Griffiths, 2017). When data are additionally generated from an external ground truth, the convergent distribution is also shaped by that world (Perfors & Navarro, 2014). However, these results only hold when agents cannot select who to talk to and when all share the same prior. When people have heterogeneous priors, the beliefs of the population are systematically distorted towards the beliefs of the most extreme individuals (Navarro, Perfors, Kary, Brown, & Donkin, 2018).

This amplification of extreme priors is concerning because it suggests that the process of information transmission itself can distort belief – and that this occurs even if all agents are fully rational and can share information fully. But our situation in real world is even more difficult. Limited by temporal and cognitive constraints, people cannot exchange information with everyone else. Moreover, the real world includes people who you might not want to learn from – not just because they have different or more extreme priors, but because they might be completely wrong or actively deceptive.

Intuitively, one solution to this dilemma would be for agents to learn who *not* to trust: to lower the weight given to the data from people who are inaccurate or miscalibrated. This is an appealing idea, but raises an important question: in the absence of any direct access to the ground truth, how should a rational learner determine who is to be trusted? One possibility is that agents might favour those who seem to make sense: those who make claims that are consistent with one's own beliefs. Indeed, there is evidence that people do adopt this strategy (Collins, Hahn, & von Gerber, 2018). Unfortunately, trusting people with similar beliefs more often leads to polarisation (e.g., Axelrod, 1997; Hegselmann & Krause, 2002; Olsson, 2013; Ngampruetikorn & Stephens, 2016; O'Connor & Weatherall, 2018; Madsen, Bailey, & Pilditch, 2018). Instead of converging on a shared set of beliefs, populations split into echo chambers: sub-groups characterised by high trust and shared beliefs within groups, but low trust and shared beliefs between groups.

Although this general result is robust and has been shown in a variety of modelling paradigms, in many cases the reasoners in such paradigms are not meant to be optimal (e.g., Axelrod, 1997; Hegselmann & Krause, 2002; Ngampruetikorn & Stephens, 2016). Some studies that *do* use Bayesian agents have established that polarisation arises even when all of the agents reason rationally (Olsson, 2013; O'Connor & Weatherall, 2018; Madsen et al., 2018); however, these studies generally involve fairly impoverished one-dimensional meaning spaces and agents who can only communicate about those spaces in a limited way. For instance, the agents in Olsson (2013) may believe in a proposition to only some degree (e.g., 70%) but are only capable of communicating binary ("yes" or "no") beliefs about the proposition.

The agents in Madsen et al. (2018) are permitted more nuance, being able to communicate their beliefs about the mean of a one-dimensional Gaussian, but have no way to communicate their level of certainty. Would polarisation still arise in groups of Bayesian agents with a richer space belief and the ability to communicate those beliefs in a more nuanced way? We explore this question here.

In Study 1 we present a new modelling paradigm in which agents must learn and communicate about a two-dimensional meaning space by sampling items from their current beliefs, while simultaneously making inferences about which of the other agents are trustworthy. We show that, as long as the distribution of prior beliefs in the population is sufficiently heterogeneous, echo chambers form even in this circumstance. Study 2 investigates whether polarisation can be eliminated and trust built by selectively communicating about only some topics (dimensions). We find that this is not a solution: doing so does build trust but at the cost of never coming into agreement. In Study 3 we explore another potential solution: access to a mutually trusted ground truth. Reassuringly, when agents have access to such a truth – even if it makes up only a tiny fraction of all of the data – polarisation is eliminated.

Study 1: Baseline

Method

Our simulations involve populations of n optimal Bayesian agents who each learn a hypothesis by receiving data from other agents (we vary $n = 6$ or $n = 18$). Agents perform inference over which other agents are trustworthy t at the same time as inferring which hypothesis h best describes the data x seen so far by calculating the joint posterior $P(t,h|x)$. Performing joint inference over trust and beliefs is somewhat different from the typical approach, in which agents directly prefer others who have similar beliefs (Olsson, 2013; Madsen et al., 2018; O'Connor & Weatherall, 2018). We opted for this approach for two reasons. First, people appear to make inferences about trust at the same time that they evaluate beliefs, and use their perceptions of trust to decide whose data to rely on (Petty & Briñol, 2008; Shafto, Eaves, Navarro, & Perfors, 2012; Perfors, Navarro, & Shafto, 2018). More importantly for our purposes, explicitly differentiating inferences about trust from beliefs allows us to explore what happens if agents can change their communication style (but not their beliefs) in order to build trust, as in Study 2.

Trust is a real value between 0.0 (no trust) to 1.0 (perfect trust) while beliefs consist of 2D Gaussians parameterised by an unknown mean μ and a known symmetric covariance Σ_0, as described in more detail below.

Initialisation. Each agent a is initialised with a different prior belief about the mean $\mu_a \sim N(0,\Sigma)$, where $\Sigma = 0.5\mathbf{I}$. All agents share the same prior about the covariance Σ_0. We manipulate population heterogeneity by changing the size of the prior covariance Σ_0 relative to the initial generating covariance Σ. Populations with high heterogeneity are initialised with means that are more "distant" in belief space relative to their beliefs about how wide the category is. There are three conditions, each defined by their covariance matrix Σ_0: HOMOGENEOUS ($\Sigma_0 = 0.25\mathbf{I}$), NEUTRAL ($\Sigma_0 = 0.15\mathbf{I}$), and HETEROGENEOUS $\Sigma_0 = 0.05\mathbf{I}$).

It would have been mathematically equivalent to manipulate heterogeneity by keeping the agents' covariance priors Σ_0 constant and varying the covariance of the generating distribution Σ; the important thing is the ratio of the two. (We chose to do it this way because one of our dependent variables is the average distance between agents in belief space, and this permits all conditions to be initialised with a similar average distance.) Smaller initial covariance matrices imply more heterogeneity because heterogeneous populations contain more individuals who are more likely to initially disagree (by inferring that the data provided by the other was unlikely). The same intuition is captured in other paradigms via the tendency to seek out those who are distant in belief space; agents with less of this tendency are more likely to polarise (Olsson, 2013; O'Connor & Weatherall, 2018; Madsen et al., 2018).

Agents are also initialised with trust vectors t with one cell for each other agent in the population, such that $t \sim \text{Beta}(1,1)$. This prior means that each agent may initially trust any other to any degree. Because the prior is weak, it is easily changed in response to data.

Iterations. During each iteration we loop through our population of n agents. At each iteration, agent i selects another agent j to learn from, proportional to the relative degree of trust i has in j. Upon being selected, agent j samples a single data point x at random from their hypothesis such that $x \sim N(\mu_j,\Sigma_0)$. Agent i then then updates their beliefs about $\mu_\mathbf{i}$ in the direction of x.[1] Thus, each iteration involves agents learning from others, in all cases revising their beliefs in the direction of the data provided, but weighting the data that was provided by trusted agents more.

At each iteration each agent i also updates their trust in all other agents j, based on the data $\mathbf{X_j}$ provided by each. The intuition is that agents will infer trustworthiness based on the extent that the other says sensible things: in this context, that means that agent j will be trusted proportional to the degree to which the data they provide to i is consistent with i's own beliefs. Agent i accomplishes this by computing the probability that they themselves would have generated that data $P(\mathbf{X}_j|N(\mu_i,\Sigma_0))$ and comparing it to the probability that it was generated by an uninformative and unhelpful other $P(\mathbf{X}_j|N(0,\Sigma_u))$.[2] Agents are thus more likely to trust those who provide data that is consistent with their own beliefs.

[1] Technically, agent i performs $n-1$ Metropolis-Hastings steps, one for each of the other agents j, in which the likelihood is calculated for all of the data points $\mathbf{X}_j$ shared by j, including the new data point x. Likelihood is weighted by trust in that agent, so that agents who are more trusted have more of an affect on belief revision.

[2] The reason for comparing against a baseline is that the raw probability of an agent providing any set of datapoints is low in absolute terms, and without the comparison all simulations tend for all agents to trust nobody. Results are qualitatively similar for a wide range of choices for the covariance of the uninformative baseline, as long as it is larger than Σ_0. All simulations here set $\Sigma_u = \mathbf{I}$.

Our approach is most similar to that of Madsen et al. (2018), but there are a few key differences in addition to those already discussed. First, their agents make inferences about both mean *and* variance, and communicate by providing the mean directly rather than sampling from their posterior. Polarisation occurs in their simulations at least in part because the learned variances approach size zero. This was probably facilitated by the fact that agents could not sample from their distributions when providing data and thus could only give point estimates, leading to a severe underestimation of the variance. Here we test whether polarisation still emerges even with agents with constant variance who can also provide more information about the extent of their distribution.

A second difference is that their agents can revise their beliefs *away* from the data they receive, whereas ours cannot. This sort of belief revision is not necessarily irrational (Jern, Chang, & Kemp, 2014), but it is difficult to determine to what extent it drives polarisation in Madsen et al. (2018). In order to explore whether polarisation arises even when the conditions for it are as unfavourable as possible, our agents disregard data they do not trust rather than move away from it.

Results

For each condition and population size, we ran 50 runs (differing only in the initial random distribution of agents in belief space) for 500 iterations each. All of our simulations were characterised by changes in the beliefs of the agents as well as their mutual trust. We consider each in turn.

Trust. We can visualise the distribution of trust across the population using pairwise mutual trust matrices T in which T_{ij} denotes the trust that agent i has toward j. We are specifically interested in the distribution of trust within the population: does it tend to be uniform, or are there clusters of agents who highly trust in each other but distrust anyone else? As the top right panel of Figure 1 shows, this clustering can be quantified using Gini mean difference (GiniMD): the mean absolute difference between all distinct elements in the pairwise trust matrices. A lower GiniMD indicates a higher shared trust, and GiniMD values over 0.3 correspond to highly polarised populations: the pairwise trust matrices show a "block" structure in which agents are in subgroups characterised by high within-group trust and low between-group trust.

As the top of Figure 1 shows, regardless of the population size, populations with HETEROGENEOUS agents were highly likely to become polarised. An ANOVA found a significant effect of condition on GiniMD ($F(2,296) = 29.34, p < 0.0001$) but not number of agents ($F(1,296) = 0.81, p = 0.369$). Initial random differences in beliefs between agents were exacerbated as they grew to trust those with similar beliefs and minimised data from those with dissimilar beliefs. Heterogeneity was the determining factor because it affected how much weight agent i put on data from j. In heterogeneous populations, more agents had initial beliefs that were far from the covariance of other agents; they were thus more apt to be distrusted. Once distrusted, they could not recover.

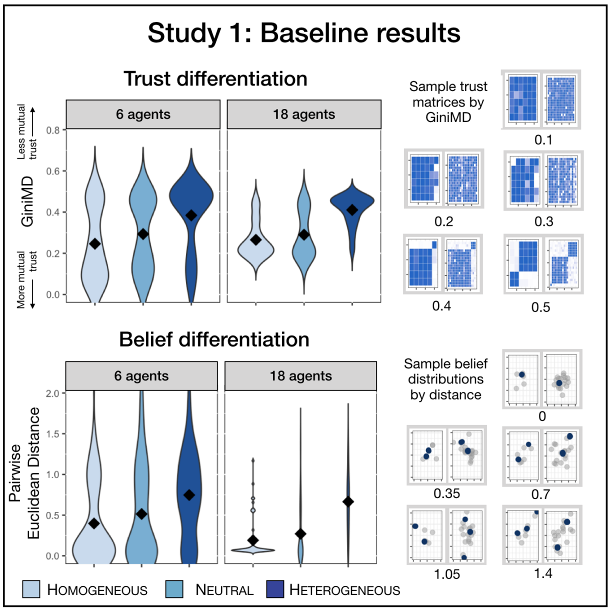

Figure 1: *Study 1: Emergence of polarisation.* Top: Polarisation is evident in the pairwise mutual trust matrices between agents, and quantified using GiniMD (right). Values above 0.3 indicate that agents have formed subgroups characterised by high within-group trust and low between-group trust. Populations of all size become polarised when they are HETEROGENEOUS (left), despite the fact that all agents are optimal Bayesian reasoners. Bottom: More HETEROGENEOUS agents also show a greater divergence in beliefs (left). Sample runs (right) showing the average pairwise Euclidean distance between agents in belief space (grey dots plot the locations of agents' initial hypotheses (μ) and dark blue dots plot the final ones) reveal that larger differences tend to correspond to more than one cluster in belief space.

The bottom of Figure 1 illustrates that these trust-based echo chambers correspond to greater average distance from each other in belief space; agents do not converge on a shared belief. As before, this effect was driven by population heterogeneity ($F(2,296) = 22.11, p < 0.0001$), although population size was also significant ($F(1,296) = 11.24, p = 0.001$). Even though agents in all conditions began the simulations at similar distances in belief space from each other, the HETEROGENEOUS agents tended to form widely-separated clusters while more HOMOGENEOUS agents were more likely to converge on the same belief. Distance in belief space and trust clustering thus both tell the same story: in sufficiently heterogeneous populations, polarisation is highly likely, even when all of the agents involved are optimal Bayesian reasoners. Consistent with this, there is a strong correlation between GiniMD and distance ($r = 0.81, t(298) = 23.7, p < 0.0001$).

How might we disrupt this tendency toward polarisation? Study 2 explores one idea: building trust by communicating tactically. Our agents are always constrained to be honest, but here we make it possible for them to refrain from communicating about topics on which disagreement is likely.

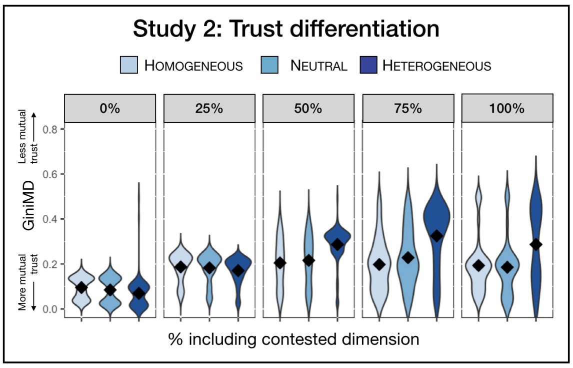

Figure 2: *Emergence of trust when agents can avoid contested subjects.* Average GiniMD as a function of the proportion of time agents included information about the contested dimension. As the dimension is included less, the agents show ever-higher levels of mutual trust. Trust is consistently unpolarised by the time the contested dimension is included 25% of the time, even in the HETEROGENEOUS condition.

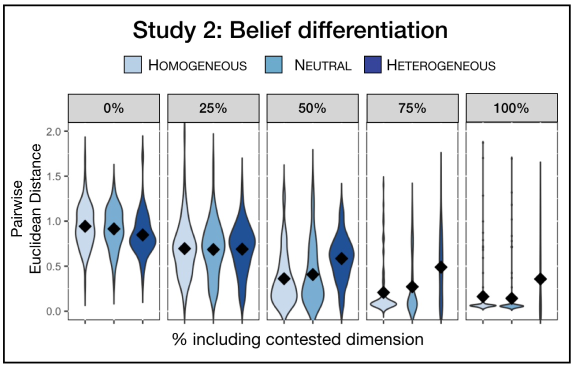

Figure 3: *Evolution of belief when agents can avoid contested subjects.* Average pairwise distance in belief space as a function of the proportion of time agents included information about the contested dimension. As the dimension is included less, the agents show more divergence in beliefs; as trust increases, the divergence in beliefs increases more. Thus, lower polarisation does not reflect more agreement.

Study 2: Tactical topic selection

Method

One of the simplifications we made in Study 1 was to assume that agents were required to communicate fully as well as honestly. In real life, however, people have discretion in what they choose to talk about. If you are visiting an uncle with whom you disagree politically, you might spend the majority of your time talking about something that you agree on, like football. This enables you to grow trust in each other and might give you the space to occasionally talk about politics.

Does adopting this strategy decrease the emergence of polarisation? Key to answering this question is realising that it is important to talk about contested issues at least some of the time: otherwise, you might trust each other, but still have irreconcileable beliefs about the facts of the matter. In these simulations we test whether there are any "sweet spots" in which agents can talk about contested beliefs just enough to come to agreement and maintain trust.

We tested this by initialising the agents differently. Where before the initial means for agents $\mu_\mathbf{a}$ were generated by sampling from a Gaussian with symmetric covariance matrix $0.5\mathbf{I}$, in Study 2 we sampled them from an asymmetric matrix with the same covariance as before along one dimension but four times tighter along the other. This meant that agents *a priori* only disagreed on one dimension, rather than two.

We then systematically varied the proportion of time that agents chose to include the contested dimension that they were more likely to disagree on. If an agent received a data point that did not include that dimension, they "filled it in" themselves by sampling it from their own prior. This was done in order to maximise the probability of eliminating polarisation; if it cannot be avoided even when agents are making the most charitable assumptions about what is going unstated, then it would be even harder to avoid if agents are making less charitable assumptions.

Results

The results suggest that enabling agents to only discuss one dimension and avoid contested dimensions *does* increase mutual trust, but the price of this is that agents no longer form a shared set of beliefs. As Figure 2 shows, communicating less about the contested dimension systematically increases trust ($F(4,1495) = 117.5, p < 0.0001$). If the contested dimension is included only half of the time, GiniMD values are consistently below 0.4, and if it is included 25% of the time or less the level of polarisation is nearly nonexistent.[3]

However, as Figure 3 reveals, that lack of polarisation corresponds to situations where the average distance between beliefs has increased substantially ($F(4,1495) = 137.7, p < 0.0001$). When the contested dimension is included half of the time, the average distance between beliefs is even higher than in the baseline HETEROGENEOUS case, even though the trust levels are still low. By the time polarisation has been eliminated in the trust matrices (when talking about the contested dimension 25% of the time or less), agents radically differ in their beliefs. What appears to be happening is that, unaffected by external data, evolution along that dimension proceeds in a random walk. Thus, although agents agree with each other on the non-contested dimension, they diverge ever more strongly on the contested one.

Thus, the higher levels of trust have not bought more agreement: they just reflect the fact that some topics are not discussed. Most importantly, we could find no "sweet spots" in our simulations where strategically communicating about contested beliefs only part of the time could allow trust to be maintained *and* beliefs to converge. This finding should be interpreted with caution because it depends to some extent on choices we made about values of Σ_0, Σ_u, and Σ. However, it is not reassuring that the divergence in belief occurs *before*

[3] For ease of presentation, we collapse across population size in the figures and analyses but the qualitative effect is identical whether there are 6 or 18 agents in the population.

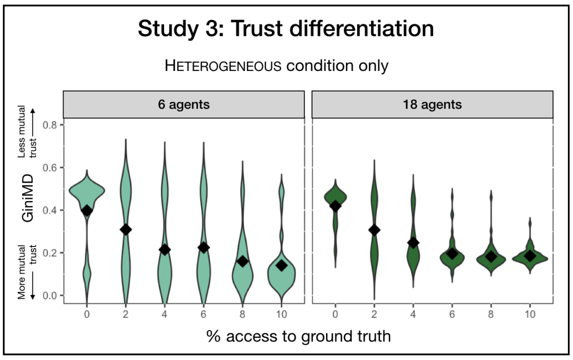

Figure 4: *Evolution of trust when agents have access to the ground truth.* It only takes a little bit of access to a ground truth source that everyone agrees is trustworthy to disrupt the formation of echo chambers, even in a HETEROGENEOUS population. Even 2% of all data points make a big difference, and by 4% or so everyone trusts everyone else.

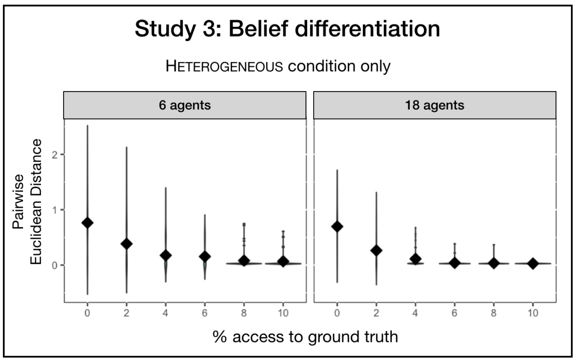

Figure 5: *Evolution of belief when agents have access to the ground truth.* It only takes a little bit of access to a ground truth source that everyone agrees is trustworthy to disrupt the formation of echo chambers, even in a HETEROGENEOUS population. Even 2% of all data points make a big difference, and by 4% or so there are no differences in beliefs.

the emergence of mutual trust, suggesting that even if such a sweet spot exists, it is tiny and highly dependent on a very specific set of parameter choices.

So far we have found that echo chambers persistently form in populations of rational agents, despite making as many charitable assumptions as possible: our agents do not revise beliefs away from those they disagree with and communicate about a rich meaning space in a way that includes their confidence (variance) about the mean rather than the mean alone. Even with these assumptions, as long as the initial beliefs are heterogeneous enough, agents cluster into echo chambers. Allowing them to build trust by communicating more often on less contentious topics does not solve this problem; communicating rarely enough to build trust means not communicating often enough to converge on a set of shared beliefs. Taken together, this appears to support the intuition we began with: this is a very difficult epistemological problem. How can one sensibly learn from others when you have no way to evaluate who to trust aside from the data they provide, and no way to evaluate that data against the state of the world?

These considerations suggest that echo chamber formation might be eliminated by simply giving agents access to some mutually-agreed upon ground truth of the matter. This might be data supplied by the external world directly or information provided by an objective observer; all that is necessary is that everyone has access to it and everyone trusts it. Does access to the ground truth disrupt the formation of echo chambers? If so, how little is required?

Earlier work has investigated these questions and found that access to the ground truth is not sufficient to disrupt echo chamber formation (O'Connor & Weatherall, 2018; Madsen et al., 2018). However, in O'Connor and Weatherall (2018) the agents sought out such evidence in a confirmatory way, testing their current hypothesis only. It is possible that receiving data relevant to all hypotheses might have led to a different result. Furthermore, agents in Madsen et al. (2018) often ended up ignoring the ground truth because it was outside of their inferred variance, which had shrunk to zero. In that sense it was not actually a "ground truth", because although it was available to all, very few people trusted it. In Study 3 we therefore provide a ground truth that all agents have access to and all trust equally.

Study 3: Ground truth

Method

Our method was exactly the same as in Study 1, except that sometimes the agents received a data point x_g sampled from the "ground truth" of the world, $x_g \sim N(0,\Sigma)$. Agents revised their belief based on this data exactly as they did on any other data; the only difference is that they did not perform inference over trust, instead assuming perfect trust in the source. We systematically vary how often agents have access to the ground truth. Because echo chambers only emerged in the HETEROGENEOUS condition in Study 1, we consider only that condition here. As in Study 2, for ease of presentation we combine the runs with 6 and 18 agents.

Results

As Figures 4 and 5 show, even a very small amount of access to ground truth data is sufficient to disrupt the formation of echo chambers. When only 2% of the data comes from the ground truth, a substantial proportion of runs result in high levels of mutual trust and shared beliefs. When 4% of the data is ground truth, polarisation is consistently eliminated: even initially HETEROGENEOUS agents converge on the same set of shared beliefs and trust everybody in the population.

Discussion

This paper is part of a growing literature investigating what happens to populations of rational agents when faced with a difficult epistemological puzzle: how to learn a set of beliefs from other people, without having access to external evidence about those beliefs or knowing *a priori* who to trust.

Consistent with that literature, we find that echo chambers consistently emerge, despite making every effort we could to eliminate them. Even though we provided agents with a richer meaning space and more nuanced communication abilities than other studies, polarisation was still highly likely as long as the population was sufficiently heterogeneous in their initial beliefs. One contribution of our work, therefore, is to further underline the robustness of this effect.

We make several larger contributions as well. First, we show that enabling agents to strategically talk less about topics that they disagree on did not solve the problem. Avoiding those topics did lead to improve trust, but at the expense of *increasing* the distance between beliefs; we found no "sweet spot" where both mutual trust and shared belief were possible. To our knowledge this is the first attempt to simulate the population-level effects that results from agents adopting different communicative tactics. Our framework is rich enough to investigate many other such tactics. What happens if people sample based not just on their own beliefs, but also on their inferences about the beliefs of others? What if people deliberately select more or less extreme beliefs, in an effort to shift the Overton window of acceptable discourse? How vulnerable are these strategies to deceptive or malicious agents?

Our work is also the first, to our knowledge, to show that having access to a trusted "ground truth" is an extremely powerful way to break the echo chamber effect. Previous work found that ground truth did not help that much (Madsen et al., 2018; O'Connor & Weatherall, 2018), but as discussed before, this was probably because of specific modelling choices that resulted in their "ground truth" being neither fully shared nor fully trusted. When it *is* shared and trusted, only a small proportion of data is necessary for even initially heterogeneous populations to develop high trust and converge on shared beliefs. The reason for this is that this common ground breaks the vicious cycle and creates a virtuous one: agents make inferences about their beliefs based in part on the ground truth data, thus trusting agents more who agree with it, and so forth. Our framework is flexible enough to enable further exploration of the robustness of this effect. How important is it that *everyone* have access to it? What if the ground truth is more accessible or less ambiguous to some? Is there any way for agents to identify those people that cannot be "gamed" by malicious agents seeking to mislead?

Our finding about the necessity of the ground truth may have important implications in light of the "post-truth" era that many believe we are now in (Lewandowsky, Ecker, & Cook, 2017). This era is characterised not only by attempts to delegitimise previously trusted sources but, more profoundly, a pervasive denial that a truth exists at all and a persistent belief that no sources are to be trusted (McCright & Dunlap, 2017). Indeed, one of the characteristics of fascism was a denial of the utility of external evidence (Varshizky, 2012), and conspiracy theories are associated with lower levels of trust in external sources (Einstein & Glick, 2015). Our simulations suggest why: shared access to the truth is one of the few things that might rescue agents from an otherwise inescapable epistemic trap. Agents who do not have access or belief in this truth are far easier to confuse, polarise, and manipulate.

Although our work further demonstrates that echo chamber formation is a robust and consistent effect even in populations of perfectly rational learners, it does suggest a key to disrupting them. Perhaps polarisation can be minimised and trust increased not by throwing more evidence toward mistaken beliefs, but by working to persuade people instead that objective truth exists and shoring up their (perceived) capacity to access and evaluate it.

Acknowledgments

This work was supported by the Australian Defence Science Technology Group Strategic Research Initiative scheme. Thanks to Piers Howe, Yoshi Kashima, Nic Fay, Charles Kemp, Alexei Filinkov, and Lucia Falzon for helpful conversations.

References

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

Collins, P., Hahn, U., & von Gerber, Y. (2018). The bi-directional relationship between source characteristics and message content. *Frontiers in Psychology*, *9*.

Einstein, K., & Glick, D. (2015). Do I think BLS data are BS? The consequences of conspiracy theories. *Pol. Beh.*, *37*, 679–701.

Griffiths, T., & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cogn. Sci.*, *31*(3), 441–480.

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Jn Artificial Societies and Social Simulation*, *5*(3).

Jern, A., Chang, K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2).

Lewandowsky, S., Ecker, U., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Jn Applied Res. Mem. Cogn.*, *6*, 353–369.

Madsen, J., Bailey, R., & Pilditch, T. (2018). Large networks of rational agents form persistent echo chambers. *Sci. Reports*, *8*.

McCright, A., & Dunlap, R. (2017). Combatting misinformation requires recognizing its types and the factors that facilitate its spread and resonance. *Jn Applied Res. Mem. Cogn.*, *6*, 389–396.

Navarro, D. J., Perfors, A., Kary, A., Brown, S., & Donkin, C. (2018). When extremists win: Cultural transmission via iterated learning when populations are heterogeneous. *Cognitive Science*, *42*(7), 2108–2149.

Ngampruetikorn, V., & Stephens, G. (2016). Bias, belief, and consensus: Collective opinion formation on fluctuating networks. *Physical Review E*, *94*(5).

O'Connor, C., & Weatherall, J. (2018). Scientific polarization. *European Jn for Phil Science*, *8*, 855–875.

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

Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cogn. Sci*, *38*(4), 775–793.

Perfors, A., Navarro, D. J., & Shafto, P. (2018). Stronger evidence isn't always better: A role for social inference in evidence selection and interpretation. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *40th Conf. Cognitive Science Soc.*

Petty, R., & Briñol, P. (2008). Persuasion: from single to multiple to metacognitive processes. *Persp. Psychol. Sci.*, *3*, 137–147.

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*, *15*, 436–447.

Varshizky, A. (2012). Alfred rosenberg: The nazi weltanschauung as modern gnosis. *Politics, Religion, and Ideology*, *13*, 311–331.

Whalen, D., & Griffiths, T. (2017). Adding population structure to models of language evolution by iterated learning. *Jn of Mathematical Psychology*, *76*, 1–6.