What interventions can decrease or increase belief polarisation in a population of rational agents?

Piers D. L. Howe (pdhowe@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne

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

Keith Ransom (keith.ransom@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne

Abstract

In many situations where people communicate (e.g., Twitter, Facebook etc), people self-organise into 'echo chambers' of like-minded individuals, with different echo chambers espousing very different beliefs. Why does this occur? Previous work has demonstrated that such belief polarisation can emerge even when all agents are completely rational, as long as their initial beliefs are heterogeneous and they do not automatically know who to trust. In this work, we used agent-based simulations to further investigate the mechanisms for belief polarisation. Our work extended previous work by using a more realistic scenario. In this scenario, we found that previously proposed methods for reducing belief polarisation did not work but we were able to find a new method that did. However, this same method could be reversed by adversarial entities to increase belief polarisation. We discuss how this danger can be best mitigated and what theoretical conclusions be drawn from our findings.

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

Introduction

In today's complex world, a great deal of what we know was learned from other people, often without the ability to verify the information directly. For example, if a doctor gives you a diagnosis, you probably do not have the ability to personally determine whether they were correct. Instead, you might seek a second opinion from a different doctor or even consult the medical literature. Ultimately, however, even these options require you to rely on other people; likely, there is no way for you to independently determine the 'ground truth' of the matter.

But how do you know who to trust? Some people might be misinformed or even be outright deceitful. In an area that is regulated by the government or other regulatory body, you can check that the people you are seeking advice from are qualified and knowledgeable, but this requires you to trust the regulatory body. In others areas, you may be able to use your own direct experience of the matter to determine whose opinions are likely to be correct. However, in many situations even this is not possible.

In such situations, one option might be to communicate with everyone and take the average (i.e. mean) opinion. Such an approach does not require that you determine a priori who is trustworthy and can often yield surprisingly accurate information (Galton, 1907). However, the 'wisdom of crowds' is reduced when people are influenced by each other or they share the same systematic biases (Surowiecki, 2004). It can also fail if some members of the population have extreme beliefs (Navarro, Perfors, Kary, Brown, & Donkin, 2018) and it requires that every person trusts and respects the opinion of everyone else, which is not realistic.

Another option, which many people employ, is to trust those who make claims that are consistent with one's own beliefs (Collins, Hahn, & von Gerber, 2018). Unfortunately, this tends to reinforce whatever belief a person started with, which can lead to the formation of echo chambers, where people only speak to and trust a small group of other people who have similar beliefs to themselves (Axelrod, 2018; Hegselmann & Krause, 2002; Ngampruetikorn & Stephens, 2016). As different echo chambers typically have starkly divergent beliefs, the formation of echo chambers often leads to substantial belief polarisation. Simulations reveal that belief polarisation emerges in a heterogeneous population even when all of the learners involved are rational Bayesian agents (Olsson, 2013; O'Connor & Weatherall, 2018; Madsen, Bailey, & Pilditch, 2018; Perfors & Navarro, 2019).

How can belief polarisation be prevented? Simulation work is particularly valuable for questions like this because it allows researchers to systematically manipulate possible factors, thus investigating the direct causal mechanisms and improving our understanding of the *why* and *how* behind a phenomenon (Gilbert, 2008; Bandini, Manzoni, & Vizzari, 2009; Miller & Page, 2007). For instance, the simulations of Perfors and Navarro (2019) demonstrated that belief polarisation can be almost entirely eliminated in the presence of even a small amount of information from a universally trusted source whose signals are based on the ground truth. Unfortunately, this result not only requires the ground truth to be known but it also requires all agents to start by trusting the ground truth source and to not lose trust if the source tells them something that seems improbable. Both assumption seem unrealistic. Are there interventions to reduce belief polarisation that will work in a more realistic scenario where it is not assumed that the ground truth is known and that people will blindly trust designated sources?

To answer this question we extended the model of Perfors and Navarro (2019) in four ways: (1) we assumed that it is not possible to force an agent to continue to trust a designated source if source's signals differ from the agents' beliefs, (2) we assumed that agents do not have perfect memory, (3) we

assumed that the ground truth is not known, and (4) in keeping with real world findings (Lorenz & Helbing, 2011), we assumed that the initialization of the agents' beliefs is biased, so that the mean belief is not always equal to the ground truth.

The assumption that the ground truth is not known, while restrictive, is a necessary reflection of real life. Despite this assumption, one can still estimate the ground truth by taking the median of the initial beliefs of all the agents (Ugander & Guestrin, 2015). While this estimate may not perfectly correspond to the ground truth, it will typically be closer to the ground truth than the estimates of most of the agents (Surowiecki, 2004). We therefore considered a scenario where all agents had access to a source that provided a signal based on this estimate. In keeping with with Perfors and Navarro (2019), we found that when agents were forced to trust this source, polarisation was greatly reduced. However, when we relaxed this assumption, this method was no longer effective at reducing belief polarisation. Consequently, we investigated two other methods, one of which proved to be quite effective. Unfortunately, in our second study we found that this same method, when reversed, was highly effective at increasing belief polarisation. We show how this danger can be best mitigated and discuss the theoretical consequences of our results. Our work makes three important contributions: (1) we demonstrate how belief polarisation can be reduced in a realistic scenario, (2) we illustrate how this scheme may be exploited to increase polarisation, and (3) and we highlight safeguards to prevent such exploitation. As such, our results have both practical and theoretical significance.

Study 1: Reducing Belief Polarisation

The purpose of the first study was to investigate potential ways of reducing belief polarisation. Perfors and Navarro (2019) found that access to even a small amount of information from a universally trusted source that based its signal on the ground truth is sufficient to greatly reduce belief polarisation. Unfortunately, the real world contains very few (if any) information sources that are universally trusted. In most situations where there are conflicting views, the ground truth cannot be unambiguously determined. We therefore relaxed both of these assumptions. In addition, we assumed that agents do not have a perfect memory and that their beliefs are initialised in a biased manner (Lorenz & Helbing, 2011). Using this model, we investigated both the solution proposed by Perfors and Navarro (2019) and two alternatives: how successful are they for reducing belief polarisation?

Method

In order to enable an appropriate comparison to Perfors and Navarro (2019), we based our model on theirs, altering it in only the four ways specified above. Their model is briefly described here, but we refer the reader to the original article for a detailed justification. Our simulations involve populations of n optimal Bayesian agents who each learnt a belief or hypothesis h by receiving data from other agents. Whereas Perfors and Navarro (2019) allowed n to vary from 6 to 18, for our simulations n was fixed at 18, as these produced the most informative findings. Our agents simultaneously inferred both the trustworthiness t of other agents and which hypothesis h best described the data x; this reflects the observation that people do simultaneously update both their beliefs and their trust in others in light of new data (Petty & Briñol, 2008; Shafto, Eaves, Navarro, & Perfors, 2012; Ransom, Voorspoels, Perfors, & Navarro, 2017; Perfors, Navarro, & Shafto, 2018). Trust was assumed to vary from 0 (no trust) to 1 (full trust) while each agent's beliefs h were represented by a 2D Gaussian parameterised by a mean μ and a symmetric covariance $\mathbf{\Sigma_0}$.

Conditions The purpose of our simulations was to investigate different ways of reducing belief polarisation. In our first study, we ran five conditions. In the first condition, the baseline condition, there was no intervention. The remaining four conditions each employed a different intervention for reducing belief polarisation. Due to the stochastic nature of our simulations, it was necessary to test each intervention multiple times in order to accurately evaluate its effectiveness. Consequently, for each condition we ran 50 runs, with the agents initialized independently at the start of each each run. Code for all simulations is available at `https://osf.io/48rab/`. Below, we describe how each run was initialised and how the updating of beliefs and trust occurred on an iteration by iteration basis.

Initialisation At the start of each run, each agent was individually initialised with the same prior about the covariance, $\mathbf{\Sigma_0}$, as well as a prior about the mean, μ_i, that was drawn separately for each agent from the distribution $\mu \sim N([0,0],\mathbf{\Sigma})$, where $\mathbf{\Sigma} = 0.5\mathbf{I}$, and then transformed by the function $f(x)$ to introduce a positive bias (Lorenz & Helbing, 2011): for $x > 0, f(x) = 1.5x$, else $f(x) = x$. The larger $\mathbf{\Sigma}$ is relative to $\mathbf{\Sigma_0}$, the more heterogenous the population is and the more likely each individual is initialised with beliefs μ that are considered implausible by other individuals (i.e. are not consistent with the belief distributions of other individuals). For our simulations, $\mathbf{\Sigma_0} = 0.05\mathbf{I}$, which is the same value as that used in the HETEROGENOUS condition of Perfors and Navarro (2019). Like Perfors and Navarro (2019), we assumed that the actual truth corresponded to the point $[0,0]$. However, unlike Perfors and Navarro (2019), we assumed that agents did not have access to an information source that knew the ground truth. Instead, in our simulations, whenever an information source needed to know the ground truth in order to determine what signal to broadcast, it had to estimate the ground truth by taking the median value of the initial beliefs of the agents. We used the median, as opposed to the mean, as the median is more robust to bias (Ugander & Guestrin, 2015). Even so, because there were relatively few agents and because the initial beliefs of the agents were biased, this meant that the estimate of the ground truth did not always accurately reflect the actual ground truth. Each agent i was also separately initialised with a trust vector t_i of length n, such that each ele-

ment t_{ij} represented the trust that agent i had in agent j. The initial trust values were drawn from a uniform distribution $t \sim U(0,1)$. Consequently, trust was not symmetric in general, i.e., $t_{ij} \neq t_{ji}$.

Iterations For each run, we conducted 500 iterations. On each iteration, we looped through all n agents in turn. Each agent i selected a single agent j to learn from with a probability proportional to the degree of trust agent i had in agent j. Upon being selected, agent j sampled a single data point x from its belief distribution $x \sim N(\mu_j, \mathbf{\Sigma_0})$. Agent i then listed this data point as received from agent j. Unlike Perfors and Navarro (2019), we assumed that each agent did not have a perfect memory and consequently remembered only the most recent data point received from each agent j and included this information in a list, X_i. Using the list X_i, agent i then updated its trust in *all* agents. This updating was performed using a single step of the Metropolis-Hastings algorithm. Agent i then considered all other agents one agent at a time and, using one step of the Metropolis-Hastings algorithm, updated its current estimate of the true belief, μ_i, using the list of the most recent data point received from every other agent, X_i.

The NO INTERVENTION condition was a replication of the heterogeneous condition of Study 1 of Perfors and Navarro (2019). Its purpose was to show that, when there are no interventions, significant belief polarisation will occur in a heterogeneous population of Bayesian agents, using our more realistic scenario.

The PERFECT TRUST condition models a situation analagous to the ground truth condition of Perfors and Navarro (2019). However, unlike Perfors and Navarro (2019), we assumed the ground truth was not known and instead, on each run, was estimated as the median of the initial beliefs. An information source s was created with mean μ_g centred on this estimate, and was questioned by each agent on 10% of the iterations. When questioned, the source drew a data point from its belief distribution $x_g \sim N(\mu_g, \mathbf{\Sigma_0})$. As this condition was designed to be analogous to the *perfect trust* condition of Perfors and Navarro (2019), it was assumed that all agents perfectly trusted this source for the duration of the run. The purpose of this condition was to attempt to replicate the finding that the presence of a universally-trusted source that approximates the ground truth greatly reduces belief polarisation (Perfors & Navarro, 2019).

The REALISTIC TRUST condition was a more realistic version of the previous condition. In this condition, agents were not forced to trust the source. At the start of each run, we randomly initialised each agent i with a trust in the source s equal to a value given by $t_{is} \sim U(0,1)$, and updated this trust on each iteration in the same manner as for other sources. As before, each agent accessed this source on 10% of the iterations and the data point provided by the source was always drawn from a belief distribution centered on the median of the initial beliefs of the agents, $x_s \sim N(\mu_g, \mathbf{\Sigma_0})$.

The PERSONALISED SIGNAL condition was identical to the REALISTIC TRUST condition except that instead of receiving a signal from the source drawn from a belief distribution centred on μ_g, each agent instead received a signal drawn from a belief distribution centered on a point equal to the sum of 75% of the agent's belief, μ_i, and 25% of μ_g.[1] As before, μ_g was estimated as the median of the initial beliefs of all agents. As the beliefs μ_i of each agent i changed over the course of the run, so too did the signal they received. The intuition is to mimic a personal information source (e.g. a bot) that would over time pull each agent towards what the source believes to be the ground truth but meets them closer to where they are, so as to maintain trust. The goal of this condition was to determine whether having a signal closer to the agent's beliefs would cause the agent to trust it, thus facilitating a gradual shift towards the ground truth.

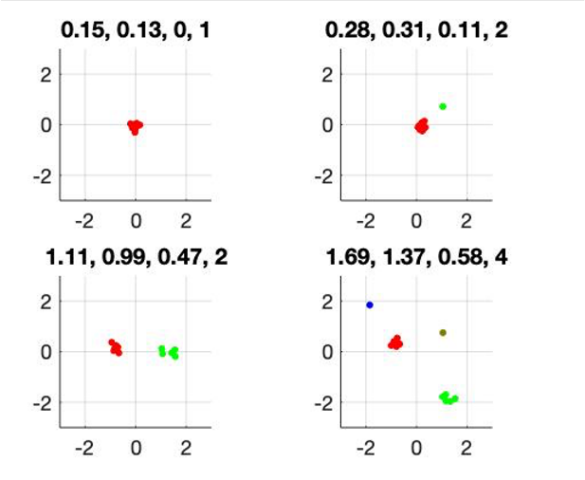

Figure 1: Four example 2D belief distributions. The degree of belief polarisation increases from top-left to bottom-right. In each distribution, each dot represents the 2D belief μ_i of one of the 18 agents. For each distribution, each cluster was given a different colour and the degree of belief polarisation was quantified in four ways: the mean pairwise Euclidean distance, the mean bias, the ratio of the number of internal links to the number of total links, and the number of clusters. These four values are listed in that order above each distribution.

The BETTER INFORMED condition explored the impact of a very different type of strategy. So far, all interventions involved presenting the agent with an external signal on 10% of the iterations, with the hope that this external signal would guide the agent towards the ground truth. As a theoretical exploration, this strategy was interesting, but as a practical strategy it had the limitation of needing to estimate the ground truth as well as possibly needing to present people with messages intermediate between it and their current beliefs. In the BETTER INFORMED condition, we instead affected information flow by making it possible for each agent to listen to only a subset of the other agents. Specifically, each agent i was permitted to listen only to agents whose beliefs μ_j were closer to the estimate of the ground truth, μ_g, than that agent's own belief μ_i. If no such agent existed, the agent listened to

[1] We chose the number 75% because it seemed intuitively sensible; systematically evaluating other values is a goal for future work.

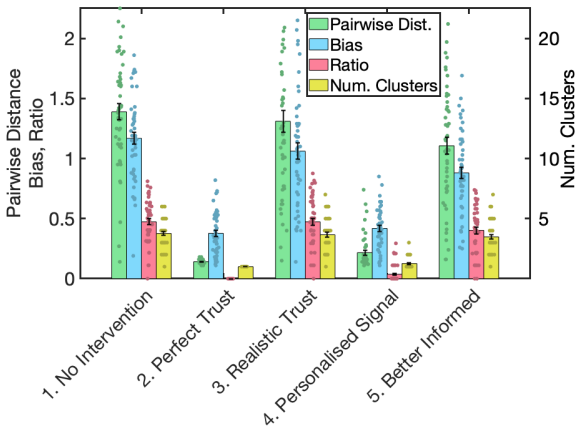

Figure 2: The results from Study 1, which tested interventions for *decreasing* polarisation. For each condition, belief polarisation was measured in four ways: mean pairwise distance, mean bias, ratio of the number of links within clusters to the total number of links and the average number of clusters. Error bars represent the standard error. Dots represent individual data points.

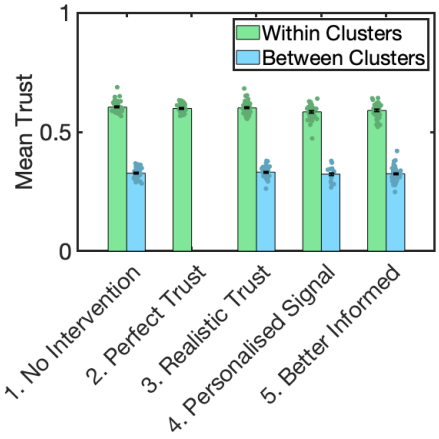

Figure 3: The average trust within clusters and between clusters for each condition in Study 1.

the agent whose mean was closest to μ_g.

Quantifying Belief Polarisation So we could compare our results to those of Perfors and Navarro (2019), we quantified the degree of belief polarisation using the same metric as them: the mean pairwise Euclidean distance between the 2D beliefs of every pair of agents. In addition, we measured the mean bias (i.e. average distance of each agent's belief from the actual ground truth, assumed to be [0, 0]), the ratio of the number of links within a cluster to the total number of links, and the total number of clusters, as determined by the clustering algorithm *dbscan*. While there are additional ways to quantify belief polarisation, these four cover a range of different approaches. As we will demonstrate later, our findings hold, regardless of which one of these four methods of quantifying belief polarisation is used.

Figure 1 shows four representative distributions. Across simulations, belief polarisation increases from left to right and from top to bottom. Consistent with this, all four metrics increase from top-left to bottom-right. Plots of final distributions of the μ values from all our simulations can be found here `https://osf.io/48rab/`

Results

The results are shown in Figure 2. The NO INTERVENTION condition replicated the finding that when nothing is done, there is considerable belief polarisation in a population of heterogeneous Bayesian agents. The PERFECT TRUST condition replicated the finding that when all agents have access to a perfectly trusted information source whose signal approximates the ground truth, belief polarisation is greatly reduced. The REALISTIC TRUST condition demonstrated that when agents are not forced to trust this signal, belief polarisation still occurs. The reason for this is that agents whose beliefs μ_i are not initially close to the estimate of the ground truth gradually decrease their trust in the source over time. This suggests that just having access to a source that approximates the ground truth is not likely to be fully effective at eliminating belief polarisation; agents whose beliefs are furthest from the source's signal will trust it the least and therefore not update their beliefs in response to its signals.

The PERSONALISED SIGNAL condition attempted to get around this problem by personalising the signal that each agent received, adapting it to be close to their current beliefs, so as to maintain trust while nudging the agent towards the source's estimate of the ground truth. This intervention proved to be highly effective at reducing belief polarisation, reducing it almost to the same levels as that in the PERFECT TRUST condition.

For the BETTER INFORMED condition we changed tactics. Instead of presenting agents with information from a knowledgeable source, we arranged for each agent to hear only from other agents that were better informed than itself on 50% of the iterations. Unfortunately, this intervention proved to highly ineffective, mainly because each agent often did not trust the better informed agents because their beliefs were too different from its belief.

Figure 3 shows the average trust between agents for all five conditions for both within clusters and between clusters. The reason why there was no measure of between-cluster trust for the PERFECT TRUST condition was because only a single cluster formed whenever this condition was run. While within-cluster trust was significantly greater than between-cluster trust, there are no substantial differences in trust across conditions.

So, why was within-cluster trust not higher than it was? The answer is that whenever an agent was interrogated it drew a data point from its belief *distribution*. Consequently, even when the mean beliefs of two agents were very similar, the data point produced by one agent would typically be somewhat different from the mean belief of the other agent. Consequently, this would reduce the degree to which the receiver would trust the sender. So, while agents within a cluster

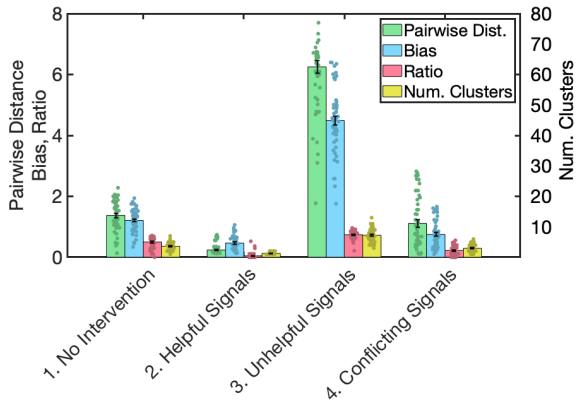

Figure 4: The results from Study 2, which tested both interventions for *decreasing* and interventions for *increasing* belief polarisation. As before, for each condition, belief polarisation was measured in four ways: mean pairwise distance, mean bias, ratio of the number of links within clusters to the total number of links and the average number of clusters. Error bars represent the standard error. Dots represent individual data points.

would mostly trust each other, the trust between them would not be perfect.

So why was between-cluster trust not lower than it was? Agents in different clusters had very different beliefs. So, if an agent in one cluster sent a data point to another agent in a second cluster, the receiver would typically respond by reducing their trust in the sender. As such, one might have expected trust between clusters to have dropped to zero. This didn't happen because agents in different clusters stop talking to each other. At each iteration, each agent selected another agent to receive a data point from, with a probability proportional to the trust the first agent had in the second agent. Consequently, as trust between clusters decreased, eventually agents stopped talking to agents outside of their cluster. This prevented the between-cluster trust from decreasing further.

Study 2: Increasing Belief Polarisation

The flip side of the issue of how to decrease belief polarisation and to improve population-level belief in the truth is the question of what factors *increase* belief polarisation and decrease belief in the truth. This is of interest not only for better understanding the strategy of entities (e.g., foreign agents) who desire to increase social polarisation; it is also important for determining whether characteristics of today's social media environments inadvertently exacerbate polarisation. To investigate this, our second study had four conditions, including a condition where personalised signals were designed to increase belief polarisation and a condition where we pitted against each other two types of personalised signal: those designed to increase belief polarisation versus those designed to decrease belief polarisation.

Conditions As Figure 4 shows, the NO INTERVENTION condition replicated the finding of Study 1. Without any interventions, belief polarisation spontaneously occurred. Similarly, the HELPFUL SIGNALS condition replicated the PERSONALISED SIGNALS condition of Study 1 and showed that introducing personalised signals that nudged each agent towards the estimated ground truth decreased belief polarisation. Whereas in the HELPFUL SIGNALS condition the personalised signals were drawn from a distribution based on a point 75% of the distance from μ_g towards μ_i, in the UNHELPFUL SIGNALS condition the personalised signals were drawn from a distribution based on a point 125% of the distance from μ_g towards μ_i, so would nudge agent i directly *away* from the estimated ground truth. As shown by Figure 4, this was highly effective at increasing belief polarisation: the increase in belief polarisation in the UNHELPFUL SIGNALS condition was much greater than the decrease in belief polarisation in the HELPFUL SIGNALS condition. Presumably, this was due to floor effects in the latter condition - there was only a limited amount of belief polarisation for the helpful personalised signals to remove. Pitting the helpful personalised signals against the unhelpful personalised signals resulted in a stalemate. The resultant degree of belief polarisation was approximately the same as it was in the NO INTERVENTION condition. These results demonstrate the effectiveness of personalised signals for both increasing and decreasing belief polarisation. The reason why this technique works is that having a personal signal close to the agent's existing belief causes the agent to trust it. This allows the signal to continue to influence the agent's belief, for better or for worse, for an extended duration, thereby increasing its effectiveness. While these results give hope that there are potentially effective ways of combating misinformation in a realistic environment, the issue is that helpful signals can be quite easily combated by unhelpful signals. What is needed is a technique that can continue to reduce belief polarisation even in an adversarial situation. Currently, we don't have such a technique, but the personalised signal approach appears to be step in the right direction.

Figure 5 shows the average within-cluster trust and the average between-cluster trust for all four conditions. As in Study 1, within-cluster trust was significantly greater than between-cluster trust, but there were no major differences between conditions.

Discussion

The rise of echo chambers is a growing international problem (Jamieson & Cappella, 2010). It is often seen as primarily an issue for social media, but the tendency to only trust those who have similar beliefs is far more general (Collins et al., 2018). In part, this tendency derives from the fact that, as society gets more scientifically and socially complex, it becomes increasingly hard to independently determine the validity of any given claim. Instead, we must rely on secondary sources, many of which contradict each other or make reference to 'facts' that we have no way of verifying. This is a very difficult epistemological problem: without direct access to a ground truth, it is hard to assess how accurate a source is. Lacking that, it is tempting to judge accuracy by how closely

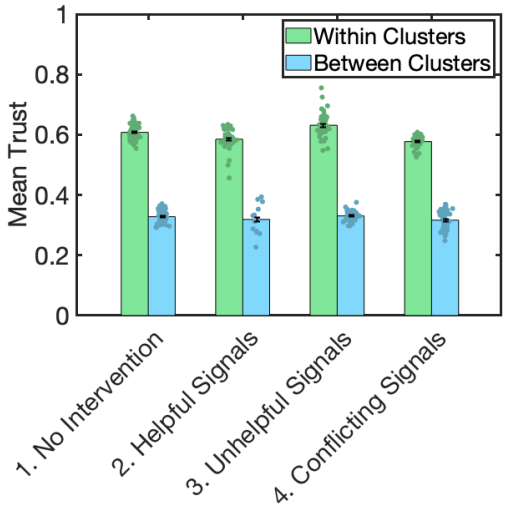

Figure 5: The average trust within clusters and between clusters for each condition in Study 2.

the information provided by a source matches one's own beliefs. Unfortunately, this will often lead to the formation of echo chambers, even when in all other respects the agents act rationally (Madsen et al., 2018; Perfors & Navarro, 2019).

Because belief polarisation often leads to social division, it would be beneficial to society if belief polarisation could be reduced. In our first study we found that reducing it was quite hard. While we found that it was greatly reduced if agents had direct access to a source that approximates the ground truth and were forced to keep trusting this source, regardless of how different its signals were from their current belief, when we relaxed this assumption, those agents whose initial beliefs were furthest from the ground truth became distrustful of the source and consequently didn't update their belief based on it. In other words, the agents who most needed to update their beliefs were the least likely to do so.

Our attempt to reduce belief polarisation by restricting who agents could talk to was similarly unsuccessful. In the BETTER INFORMED condition, on 50% of the iterations, each agent could receive signals only from other agents who were better informed than it was. However, each agent would still only trust those agents that had similar beliefs to it, so was influenced only by the better informed agents that were close to it in belief space. Because the agent would attempt to update its belief in the direction of these other agents, clusters of agents still formed, with all agents in each cluster converging on the μ of the agent in the cluster that was closest to the ground truth. So, while this method did slightly reduce the mean bias, it did not reduce the number of clusters and, consequently, reduced the mean pairwise distance by only a small amount. As such, it was not an effective way of reducing belief polarisation.

Our other attempt to overcome this problem was more successful. In the PERSONALISED SIGNAL condition we introduced a personal signal for each agent that was close to what the agent already believed but closer to the ground truth than the agent's current belief. Importantly, this signal was continuously updated to continue to be close to the agent's current belief. In this way we hoped to maintain trust and to 'lead' each agent progressively towards the truth. Although belief polarisation was not entirely eliminated, it was reduced to a level comparable to that in the PERFECT TRUST condition.

Our second study extended our first study by investigating to what extent personalised signals could *increase* belief polarisation. We found that personalised signals are more effective at increasing belief polarisation than decreasing it. We suspect that this occurred because without any interventions there was a relatively small amount of belief polarisation, as shown by the NO INTERVENTION condition. Consequently, the HELPFUL SIGNALS condition, which used personalised signals to decrease belief polarisation, had relatively little scope to do so. Conversely, the unhelpful signals in the UNHELPFUL SIGNALS condition had not such limits.

Malign actors wishing to increase belief polarisation therefore have an innate advantage in that it is far easier to increase belief polarisation than to decrease it. Of course, the actor creating the unhelpful personalised signals may not be purposely attempting to increase belief polarisation. On social media there is an incentive to increase the number of followers one has and one typically does this by generating 'interesting' posts. While it is not entirely clear what makes a post interesting, there are least two important factors. The first is that the post has to be somewhat believable to the reader. The second is that it needs to be novel and exciting. The first aim can be achieved by creating posts that are similar to what the target audience already believes (Collins et al., 2018). Given that fake news seems to propagate faster and further than real news (Vosoughi & Aral, 2018), the second aim could be achieved by posting false information. In other words, social media users are encouraged by the nature of the system to create false posts that are similar to what their target audience already believes. Such personalised signals would tend to nudge people away from the truth and, as shown in Study 2, are a particularly effective way of increasing belief polarisation. It is therefore possible that some of the belief polarisation present in social media is caused by structural factors that encourage users to create unhelpful personalised signals.

Our findings are preliminary: we did not consider all possible interventions and we could have done more to optimise the parameter values for the interventions that we did study. For instance, it is possible that performance might improve in the PERSONALISED SIGNAL condition if the agents received signals closer or further than 75% of the way between them and the sources estimate of the ground truth. In future work we will perform this systematic search, but in some sense it is not necessary for the take-home message. Even without optimising this intervention, it was already the most successful and *whatever* specific values perform best depend on the specific assumptions about heterogeneity, belief updating, and so

forth, so would not be generalisable.

Our take-home message is essentially a qualitative one: our findings suggest that belief polarisation and the rise of extreme viewpoints like conspiracy theories may be a natural consequence of social media algorithms and the economics of the media ecosystem that reward novelty and engagement, without performing checks on the accuracy of the information that is posted. The agents in our simulations are rationally responding to the impoverished epistemological system and biased information sources; perhaps, at least to some extent, so are humans. Our results also suggest that if an external entity wished to increase polarisation within society, an effective way to achieve this goal would be to send personalised signals to users, perhaps by using a botnet. Individuals are likely to be persuaded only by messages that are close to what they already believe, so it is important to target people individually. Any process for reducing belief polarisation must maintain trust with people whose beliefs are polar opposites. Achieving consensus and maintaining trust is placed in tension: a signal designed to shift people from one pole may seem implausible (and untrustworthy) to those at the other. It is this tension that our simulations capture and that drives at least part of the difficulty of mitigating polarisation in the real world. It is ironic that the best way to reduce people's false beliefs is to share false beliefs, albeit ones that are consistently a little closer to the truth than what the people currently believe. Combating misinformation with misinformation raises ethical issues that would need to be considered before this approach was adopted.

A surprising aspect of our simulations was the unexpectedly high degree of trust between agents in different clusters. Although between-cluster trust was less than within-cluster trust, it wasn't particularly low. The reason for this was that when between-cluster trust began to drop, agents stopped communicating with agents in other clusters and this prevented between-cluster trust falling further. This is surprising because there is increasing animosity in American politics (Levitsky & Ziblatt, 2018) with both Democrats and Republicans increasing distrusting members of the opposite party (S. Iyengar & Westwood, 2019). How can we reconcile these real-world findings with our simulations?

We suspect that part of the reason is that in the real world it is very hard to completely escape from politics and, in particular, it is hard to remain ignorant of what the opposing political party thinks about a range of viewpoints. So while our simulations suggest that both Democrats and Republicans would stop following members of the other party, in reality this may not be possible. It follows that if you continue to be exposed the viewpoints of the other party, then your trust in the other party should continue to fall. This could explain the disconnect between our simulations and the real world findings cited above.

In conclusion, our simulations have produced some interesting findings and have provided some novel insights into the causes of belief polarisation and how it can best be prevented. The work here is preliminary, but we are optimistic about the power of simulations such as these to better understand the mechanisms of belief polarisation and how it arises as a by-product of biases built into human cognition.

References

Axelrod, R. (2018). The dissemination of culture: A model with local convergence and global polarization. *Jn. of Conflict Resolution*, *41*, 203-226.

Bandini, S., Manzoni, S., & Vizzari, G. (2009). Agent based modeling and simulation: An informatics perspective. *Jn. Art. Soc. and Soc. Simulation*, *12*, 1–16.

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

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.

Gilbert, N. (Ed.). (2008). *Agent-based models*. CA, USA: SAGE Publications.

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

Jamieson, K. H., & Cappella, J. N. (2010). *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford, UK: Oxford University Press.

Levitsky, S., & Ziblatt, D. (2018). *How democaries die*. Crown.

Lorenz, R. H. S. F., J., & Helbing, D. (2011). How social influence undermine the wisdom of the crowd effect. *Proceedings of the National Academic of Sciences*, *108*(22), 9020-9025.

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

Miller, J. H., & Page, S. E. (Eds.). (2007). *Complex adaptive systems: An introduction to computational models of social life*. NJ, USA: Princeton University Press.

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

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

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

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

Perfors, A., & Navarro, D. (2019). Why do echo chambers form? The role of trust, population heterogeneity, and objective truth. In *41st Conf. Cog. Sci. Soc.* (pp. 918–5923).

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. Cog. Sci. Soc.*.

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

Ransom, K. J., Voorspoels, W., Perfors, A., & Navarro, D. J. (2017). A cognitive analysis of deception without lying. In *39th Conf. CogSci. Soc.* (pp. 992–997).

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

S. Iyengar, M. L. N. M., Y. Lelkes, & Westwood, S. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, *22*, 129–146.

Surowiecki, J. (2004). *The wisdom of crowds*. New York, NY: Doubleday.

Ugander, D. R., J., & Guestrin, C. (2015). The wisdom of multiple guesses. *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 643-660.

Vosoughi, R. D., S., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151.