Stubborn extremism as a potential pathway to group polarization

Matthew A. Turner and Paul E. Smaldino

Abstract

Group polarization is the widely-observed phenomenon in which the opinions held by members of a small group become more extreme after the group discusses a topic. For example, conservative individuals become even more conservative, while liberal individuals become even more liberal. Social psychologists have offered competing explanations for this phenomenon. These typically require questionable assumptions about human psychology. Here, we posit a more parsimonious explanation: the stubbornness of extreme opinions. Using agent-based modeling, we demonstrate that such "stubborn extremism" gives rise to group polarization, as well as other trends observed across the literature on polarization. Our study revealed a further methodological problem for the study of group polarization: reporting opinions as categories (e.g. on a Likert scale) inflates the observed increase in opinion extremity. We conclude with a call for deeper integration of opinion dynamics modeling with the cognitive science of communication and influence.

Keywords: opinion dynamics; polarization; social influence; agent-based modeling

Introduction

Group polarization is a phenomenon in which the opinions held by members of a small group become more extreme after the group discusses a topic (Myers, 1982; Brown, 1986; Isenberg, 1986; Sunstein, 2002; Sieber & Ziegler, 2019). This phenomenon is socially important for many reasons. First, small groups of advisers often influence executive decisions in government and business. At the "grass roots" level in politics, individuals discuss important issues first in small groups before they vote. Second, group polarization at the local level increases overall polarization at the societal level. Polarization, commonly understood, increases whenever either of two opposed groups moves to a greater extreme, whatever the polarization measure (Bramson et al., 2016). Most studies of political polarization frame the issue in terms of intergroup conflict (Mason, 2018; Klein, 2020). However, we also must understand how group polarization can exacerbate political polarization through increased in-group extremism. Understanding the cognitive mechanisms supporting group polarization is therefore a matter of concern.

Social psychologists have typically offered one of two explanations for group polarization, sometimes combining the two (Sieber & Ziegler, 2019). *Social comparison theory* posits that individuals' privately-held opinions tend be more extreme than those they express publicly, and exposure to consonant opinions gives them confidence to express their true opinions openly (Myers, 1982). *Persuasive arguments theory* posits that when individuals discuss a topic within an already-biased group, they accumulate more persuasive arguments supporting those biases, leading to a more extreme version (Bishop & Myers, 1974; Vinokur & Burstein, 1974). These explanations may explain the empirical phenomenon of group polarization, though more formal modeling will be required to bring precision to the underlying theories (Smaldino, 2017, 2019). Nevertheless, each presumes either that opinions are intrinsically extreme (social comparisons theory) or that moderate opinions exist only due to uncertainty concerning the state of the world (persuasive arguments theory).

We present an alternative explanation for group polarization that, while not mutually exclusive with the other theories discussed, manages to explain the phenomenon of group polarization without assuming anything about the intrinsic distribution of extreme opinions in human groups. We do so by appealing to a property of human psychology we call *stubborn extremism*: as a person's opinion on some topic becomes more extreme, that opinion also becomes more stubborn, i.e. less susceptible to social influence. We support this explanation using a computational model of group polarization. Our model was originally developed for explaining how polarization emerges where two groups become more extremely opposed (Flache & Macy, 2011; Turner & Smaldino, 2018)—it incorporates both negative, repulsive social influence (Cikara & Van Bavel, 2014), assimilative influence, and the stubborn extremism assumption, though repulsive influence is not at work in group polarization because all opinions start out similarly valenced.

Group polarization can emerge computationally by simply assuming agents hold discrete opinions on a multitude of topics (Mueller & Tan, 2018; Banisch & Olbrich, 2019). However, most group polarization studies do not measure participants' binary opinions (e.g., for vs. opposed) on a multitude of topics, but rather measure opinions as falling on a range between strongly for and strongly opposed. Furthermore, the assumption of discrete opinions is problematic from a psychological perspective, since it is rare for quantum leaps in opinion to occur—more often we are influenced gradually over the course of many interactions (Baldassarri & Bearman, 2007, p.793). Our model is most similar to that of Baldassarri and Bearman (2007) in that stubbornness is a function of opinion extremity directly. Martins and Galam (2013) allow for agents to become more or less stubborn, but assume discrete opinions and a separate, continuous measure of open-mindedness/stubbornness. Most other opinion dynamics models that link stubbornness to extremism assume infinitely stubborn extreme agents (sometimes called "zealots") whose opinions are static and whose existence is specified *a priori* by the modeler (Galam & Jacobs, 2007; Mobilia, Petersen, & Redner, 2007; Arendt & Blaha, 2015; Mueller & Tan, 2018). Baldassarri and Bearman (2007) nearly make the connection between stubborn extremism and group polarization, but they mischaracterize group polarization and discuss it in terms of negative influence, saying "interaction with dissimilar others may increase distance, leading to group polarization" (p. 792). Group polarization experiments are designed so that this never occurs. Instead, it is only interaction among relatively likeminded individuals that leads to the group polarization opinion shift.

Current empirical support for the stubborn extremism explanation is positive, though not uniformly so. Zaller (1992) and Converse (2006) established that, at least at the time of their studies, most of the United States electorate, for example, were relatively ignorant of real political issues and easily swayed by momentary predilections and the framing of questions. Guazzini, Cini, Bagnoli, and Ramasco (2015) found that stubborn extremists drove the opinions in groups discussing the use of animals in laboratory experiments, and Lewandowsky, Pilditch, Madsen, Oreskes, and Risbey (2019) found that stubborn extremists have an outsized influence in the perpetuation of scientific misinformation regarding climate change. Group polarization opinion shifts have been observed to increase with the group's initial extremity (Teger & Pruitt, 1967; Myers & Arenson, 1972; Myers, 1982; Brown, 1986). This has only been tested in detail by Teger and Pruitt (1967) and Myers and Arenson (1972), apparently, and has not been established for political opinions. This could cause acceleration of political polarization. Some researchers have suggested that stubbornness is an attribute found generally among people, and is not limited to those with extreme opinions. However, support for this view often comes from studies in which opinions are operationalized as answers to general knowledge tests (such as found in a pub quiz), and not on opinions with political or ethical components in which subjective judgment plays a larger role (Moussaïd, Kämmer, Analytis, & Neth, 2013; Chacoma & Zanette, 2015). More direct empirical tests of the stubborn extremism explanation for group polarization are needed.

Here we investigate whether assuming stubborn extremism can explain and predict observed empirical patterns of group polarization. We do this with an eye towards future empirical experiments. In doing so, we also pay close attention to the scales used to measure opinion extremity. There has recently been scrutiny of the use and analysis of Likert-scaled data in social psychology, which indicates that prematurely sorting continuous data into discrete bins can distort effect sizes (Liddell & Kruschke, 2018). We therefore also examine the effect of using a Likert scale in which simulated agents bin their continuous opinions.

The rest of the paper is organized as follows. We first briefly review the empirical evidence for group polarization upon which we will based our analyses. We will then introduce an agent-based model of opinion dynamics with stubborn extremists, which is adapted from previous work by Flache and Macy (2011), and we will demonstrate how the model supports the stubborn extremism hypothesis. We will then compare our model to the persuasive arguments model of Mäs and Flache (2013), and show how our model can yield a fit to the empirical dataset they test that is at least as congruent. We conclude with limitations of our model's assumptions, and suggestions for future work.

Group polarization studies

A prototypical experiment involving group polarization works as follows. First, participants are pre-screened for their opinions about some issue or set of issues. They give their opinions on an ordinal scale, such as a 7-point Likert scale that ranges from -3, which would indicate "strongly disagree" to +3 which would indicate "strongly agree." Then the participants are sorted into groups with similarly valenced views. The participants are then asked to publicly give their initial opinion on some issue, discuss the issue within their group for a time, and then report their opinion again to the experiments. It is regularly observed that individuals' opinions shift towards greater extremity following group discussion.

Moscovici and Zavalloni (1969) studied group polarization in the context of national (French) and global politics. In their study, they first asked participants the degree to which they agreed or disagreed with the claim that Charles de Gaulle, then president of France, was "too old to carry out such a difficult political task." Second, participants were asked the degree to which they agreed or disagreed that "American economic aid is always used for political pressure." Participants responded on a 7-point Likert scale where -3 indicated total disagreement, 0 represented a neutral psychological position of neither agree nor disagree, and 3 indicated total agreement. Forty individuals in ten groups answered, discussed, then answered again the de Gaulle question, and twenty individuals in five groups did the same for the Americans question, with four per group in each case. In reporting the shifts in opinions, only the mean shifts for the entire subject pool were reported. For the de Gaulle question, a shift was observed from a mean pre-discussion opinion of 1.36 to a mean of 1.82 post-discussion, a shift of 0.46. On the Americans question, the initial pre-discussion mean opinion was 0.88 and the post-discussion mean opinion was 1.69, a shift of 0.81.

Myers and Bishop (1970) asked participants about their attitudes on eight items, where responses varied from -9 to 9, an 18-point scale. -9 indicated maximal racial prejudice and 9 indicated minimal prejudice. Participants were grouped into low, medium, and high prejudice supergroups, from which discussion groups of 4-7 members were formed. The average shift of the low prejudice group was 0.47, while the average shifts of the medium and high prejudice groups were -0.64 and -1.30, respectively. In other words, individuals in the low prejudice group decreased their expressed prejudice levels, while prejudiced groups became more prejudiced.

Myers and Lamm (1975) binned attitudes about the role of women in society from -3 (conservative) to +3 (liberal), again with 0 a neutral opinion. Groups of "chauvinists" showed no significant opinion shift, while "feminists" showed a significant average shift of 0.95. Discussions occurred in groups with four or five members from 95 total participants.

Mäs and Flache (2013) developed a model of opinion change based on persuasive arguments theory and conducted a laboratory study of opinion change similar to the ones described above, concerning the better of two locations (town A or town B) to build a new leisure center. Participants were asked their preference among the two hypothetical towns ("A" or "B"). Participants were sorted into groups "A" and "B," and given a number of pro-A and pro-B arguments they could exchange with other agents. Members of the "A" group (A-Type) were provided with two pro-A arguments and one pro-B argument, while members of the "B" group (B-Type) were provided with one pro-A argument and two pro-B arguments. All members of the "A" group all received the same pro-B argument, and all members of the "B" group received the same pro-A argument.

All participants participated in seven interaction rounds. In each round, participants interacted through a computer interface with a single interaction partner, wherein each participant selected one of their pre-written arguments and sent it to their partner. Participants first interacted with the three other members of their own group, and then interacted with the four members of the other group. After all seven rounds of interaction, participants again reported their opinions.

Mäs and Flache (2013) observed that following within-group interactions, the group's average opinion became more extreme in accordance with the general prediction of group polarization. Their explanation was based on persuasive arguments theory, which also predicts—on the basis of a computational model analyzed by the authors—that when *opposing* groups interact they will become more similar in opinion. This happens because when an individual from one group interacts with a member of the opposing group, they will be exposed to new counterarguments that reduce their extremism towards a more moderate opinion. During the four rounds of between-group interactions in Mäs and Flache's study, the average opinion in both groups converged towards zero. This is predicted by both persuasive arguments and stubborn extremism models, which we demonstrate in computational experiments below.

The model

We developed an agent-based model to demonstrate the stubborn extremism model predicts the empirical results reviewed above. Our goal is to demonstrate that assuming stubborn extremism can lead to group polarization opinion shifts as reliably as social comparisons or persuasive arguments models. The model is similar to a Hopfield network model (Hopfield, 1982), in which node values change based on neighboring node values, and mediated by network weights. Here, these weights are determined by the distance in opinion space between two agents. This model allows for both positive and negative influence, wherein initially similar agents become more similar after interacting, while initially dissimilar agents become more polarized. The model is identical to that studied previously in Flache and Macy (2011) and Turner and Smaldino (2018), but is analyzed here with a different focus than was used in those studies.

We consider a population of N agents, who each have opinions on K different cultural topics. Agent i's opinion on topic k at time t is written $o_{ik,t} \in (-1, 1)$ and changes after i has interacted with its N_i network neighbors. The weight of social influence with each neighbor j is $w_{ij,t}$, with zero direct influence over non-neighbors. Weights depend on the Manhattan distance between agents i and j, normalized over cultural topics: $d_{ij,t} = \frac{1}{K}\sum_{k=1}^{K} |o_{ik,t} - o_{jk,t}|$. The specific operation of these social influence mechanisms is defined by the following dynamical equation

$$o_{ik,t} = o_{ik,t-1} + \Delta o_{ik,t}(1 - |o_{ik,t-1}|^{\alpha}) \qquad (1)$$

where

$$\Delta o_{ik,t} = \frac{1}{2N_i}\sum_{j} w_{ij,t}(o_{jk,t} - o_{ik,t}) \qquad (2)$$

and

$$w_{ij,t} = 1 - d_{ij,t}. \qquad (3)$$

Our model includes both positive and negative influence. Positive influence is when agents become increasingly similar to their dyad partner if the pair are sufficiently similar to begin with ($d_{ij} < 1$). Negative influence is when interaction causes a dyad to become more different, to be repulsed away from one another toward more extreme regions of opinion space if the pair are sufficiently dissimilar to begin with ($d_{ij} > 1$). The parameter α determines the degree to which extreme opinions are stubborn. In the analyses presented here, we use $\alpha = 1$. Stubborn extremism emerges in our model due to the smoothing factor $(1 - |o_{ik,t-1}|)$, which is smaller when $|o_{ik,t-1}|$ is larger. Therefore, more extreme opinions (larger $|o_{ik,t-1}|$) are less susceptible to social influence than less extreme opinions (smaller $|o_{ik,t-1}|$).

Our model generates a number of empirically-observed outcomes. First, we show that our model yields group polarization in an idealized generic case that resembles the studies of Moscovici and Zavalloni (1969), Myers and Bishop (1970), and Myers and Lamm (1975). For our computational experiments, we set the number of agents in the population to $N = 25$ and the number of relevant opinions $K = 1$. The social network for this first experiment was fully connected, meaning all agents could potentially influence all other agents. Second, we represent the Mäs and Flache (2013) empirical experiment with our model and show our model predicts their empirical observations as accurately as their computational model of persuasive arguments theory.

Computational experiments

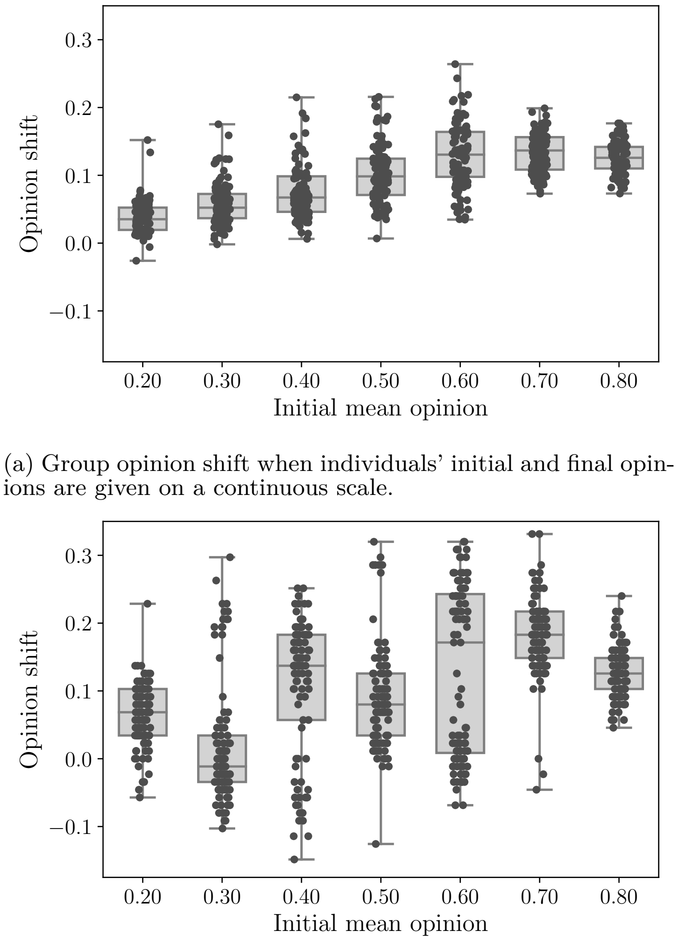
Our first experiment examined the correlation between initial mean opinion and shift magnitude. This also establishes that our model generates group polarization. We set $N = 25$ and $K = 1$. Initial agent opinions were drawn from a normal distribution with $\sigma = 0.25$. In order to demonstrate that our model predicts the correlation between opinion shift and initial opinion extremity, we ran the model with seven different experimental conditions. Each of the seven conditions specified a different mean for the normal distribution from which initial opinions were drawn, $\mu \in \{0.2, 0.3, \ldots, 0.8\}$. For each condition we ran 100 trials. Since opinions are bounded between ± 1 and group polarization experiments force group members to have opinions of the same valence, we re-mapped any drawn opinions greater than 1 to be $+1$ if the drawn opinion was greater than 1, and 0 if the drawn value was less than 0. Each model run consisted of 100 rounds of agent interactions. In one round of agent interaction, N agents are selected at random to update their opinions according to Equation 1. To model a typical group polarization experiment with open discussion, we assume a fully-connected network, so all agents influence one another.

Our second experiment was designed to examine how opinion shifts and the pattern of shift versus initial extremity may be distorted when agents are forced to report their opinions on a seven-point Likert scale. This was done by first transforming the continuous opinions on $(-1,\ 1)$ to a continuous opinion on $(-3.5,\ 3.5)$, representing seven equally-sized bins of width 1, and then rounding each agent's continuous opinion to the closest integer, e.g. 2.9 becomes 3 and 1.49 becomes 1.

Our third and final experiment was designed to generate the results of Mäs and Flache (2013). Here we utilized the multidimensionality of opinions to represent different "persuasive arguments" that participants held. To do this, we set $K = 12$, the total number of persuasive arguments available to each agent in Mäs and Flache's study, and initialized three of the twelve opinions to be non-zero. Recall that in their study, Mäs and Flache provided individuals with one of twelve pre-defined "arguments" they were to share with others to advocate for their opinion. Six of the twelve were chosen as pro-A arguments and six of the twelve were chosen as pro-B arguments. The pro-A arguments were given initial values of $-1/3$ and pro-B arguments given initial values of $1/3$. In our adaptation of this experimental setup, we are using each of K elements of agent i's opinion vector to represent the presence or absence of an argument. As in the Mäs and Flache study, group "A" members all received the same initial pro-B argument, and vice versa. To calculate each agent's scalar opinion based on its $K = 12$ "persuasive argument" components, we first normalize opinions so their absolute values sum to 1, and then averaged over all opinions. This is similar to the persuasive argument model that assumes an individual's opinion is an aggregate of the arguments they know for their position. This computational experiment mirrors Mäs and Flache's persuasive arguments model, but includes stubborn extremism. Furthermore, in our formulation, agents can partially agree or disagree with a given argument, unlike persuasive arguments which assumes an agent either knows an argument or not. For our computational experiment's outcome measure, we calculated the average over all agent opinions in each group at each timestep, and then averaged those averages across 100 trials at each timestep, identical to Mäs and Flache's procedure for obtaining their results (Figures 5 and 6 of their paper).

Implementation

The model was implemented as an agent-based model written in plain Python with user-defined `Agent`, `Model`, and `Experiment` classes. We use NumPy and SciPy for numerical and scientific routines and functions. For full implementation details including instructions for installing and running model code and reproducing our results, please visit the GitHub repository, `https://github.com/mt-digital/group-polarization`. Our computational experiments easily run on a laptop.



(a) Group opinion shift when individuals' initial and final opinions are given on a continuous scale.

(b) Group opinion shift when individuals' initial and final opinions are given on a 7-point Likert scale.

Figure 1: Demonstration of the trend that opinion shift is positively correlated with the mean initial group opinion. The trend is distored by binning into Likert scale responses. Boxes enclose the first and third quartile of the data. 100 trials shown for each condition.

Analysis

Our model predicts that more extreme initial group opinion results in larger shifts up to a certain extremity where the trend reverses (Figure 1a). In terms of stubborn extremism, this general trend is expected because there will be more extremists when the initial mean is greater. These initial extremists exert a greater pull towards extremism when they are more numerous. However, when many agents are extreme and there are few neutral agents to be shifted to more extreme views, the shift begins to decrease in magnitude compared to the maximum shift over initial mean (occurs at initial mean of 0.8 in Figure 1a).

Binning disrupts the positive linear relationship between opinion shift and initial mean group opinion (Figure 1b). This is because, in our model, if enough agents are neutral and not too many agents are extreme, then some agents with an opinion of +3 will shift to +2, and enough agents with opinions of +1 or less do not shift their opinions, the sign of the shift may be negative, and group polarization will not emerge.

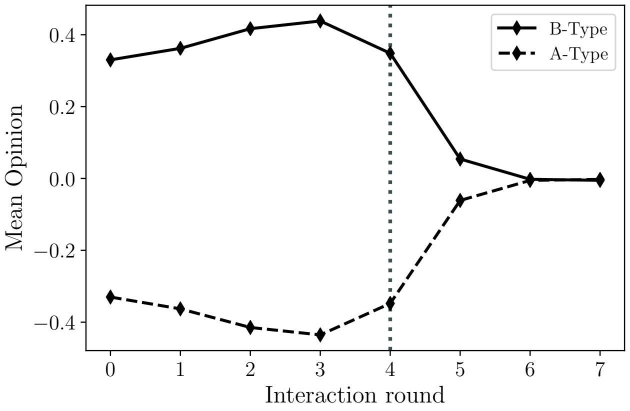


Figure 2: Our model's prediction of group opinions in the Mäs and Flache (2013) study. Within-group interactions are rounds 1-3, intergroup interactions are rounds 4-7.

Our model predicts group polarization as observed by Mäs and Flache (2013), but via the assumption of stubborn extremists instead of persuasive arguments. Our model predicts the same initial increase in the extremity of the average group opinion for both A- and B-Type agents as predicted and observed in Mäs and Flache (2013). Then when A-Types and B-Types interact with one another, our model predicts consensus emerges, as was observed by Mäs and Flache's experiments and predicted by their model (Figure 2 below; compare with Figure 6 Mäs and Flache (2013)). Note that, in our model, no explicit persuasive arguments are exchanged. Instead, each argument is represented as an opinion on a certain cultural topic. Influence occurs on all cultural topics, and similar group members draw one another closer in hypothetical 12-dimensional opinion space through attractive social influence and stubborn extremism, resulting in group polarization.

Discussion

We have shown that stubborn extremists are a feasible explanation for group polarization. Our model that incorporates this simple mechanism predicts behavior observed in a number of empirical studies. These empirical studies have often considered two alternative pathways to group polarization: *persuasive arguments* and *social comparisons*. The persuasive argument theory explains that group polarization occurs because individuals are exposed to more arguments supporting their initial position in contrast with the opposing opinions, thereby strengthening that opinion. At the group level, this leads the average opinion to shift towards an extreme. Alternatively, social comparison theory posits that group polarization is due to group members calculating some optimal opinion to express publicly that takes into account both their private opinion and the perceived social consequences of expressing that opinion. The theory posits that, following group discussion, this optimal public opinion is usually judged to be more extreme than individuals' initially stated opinions.

Neither of these theories are necessarily wrong; indeed, they appear to be both psychologically and sociologically plausible. What we have done is to identify another mechanism, independent of these, which is at least as plausible and which generates known empirical results when formalized in a computational model. Moreover, our model is the only one to assume neither that extreme opinions are the default state (as social comparison theory does), nor that moderate opinions result only from lack of arguments for more extreme positions (as persuasive arguments theory does). Our theory requires only that extreme opinions, once reached, are more stubbornly held than more moderate opinions.

We also demonstrated that, if we assume individuals' opinions are continuous, forced binning of opinions into Likert-style point values leads to distortions in opinion shifts compared to the underlying continuous opinion shifts. Depending on the final distribution of agent opinions, these distortions could be either an over- or under-estimate of the opinion shift. In extreme cases the sign of the opinion shift reversed, making the shift appear to be towards a more moderate group opinion, not more extreme. This puts existing group polarization results in question since most empirical group polarization studies used a Likert scale. This is further compounded by similar effects that could be caused by using metric statistical models on ordinal data, which apparently all existing group polarization studies have done (Liddell & Kruschke, 2018).

Although there is evidence supporting the hypothesis that extreme opinions are more stubbornly held, we are aware of no research specifically investigating the relationship between stubbornly held opinions and group polarization. Future empirical work should evaluate the stubborn extremism hypothesis using a statistical model to detect correlation between opinion extremity and stubbornness. Such work should use both Likert and continuous scales due to the distortions in opinion shift we identified. In the Likert condition, one must use the appropriate statistical model (Liddell & Kruschke, 2018). We advocate the use of both scales for two reasons. First, it is not clear that a continuous model of opinions really is superior to a discrete model of opinions. The empirical situation may determine which opinion model is appropriate. The second reason is that, in practice, responses given on Likert-type scales may be more reliable than those given on continuous scales (Toepoel & Funke, 2018).

Models of opinion dynamics should be able to explain a number of empirical phenomena, including but not limited to group polarization. Another program of future work could be to perform similar computational experiments shown here using alternative, influential models of political polarization, such as Bayesian/information-theoretic models (e.g. Dixit and Weibull (2007)) or algorithmic models (e.g. Dandekar, Goel, and Lee (2013)).

References

Arendt, D. L., & Blaha, L. M. (2015). Opinions, influence, and zealotry: a computational study on stubbornness. *Computational and Mathematical Organization Theory*, *21*(2), 184–209. doi: 10.1007/s10588-015-9181-1

Baldassarri, D., & Bearman, P. (2007). Dynamics of Political Polarization. *American Sociological Review*, *72*(5), 784–811. doi: 10.1177/000312240707200507

Banisch, S., & Olbrich, E. (2019, apr). Opinion polarization by learning from social feedback. *Journal of Mathematical Sociology*, *43*(2), 76–103. doi: 10.1080/0022250X.2018.1517761

Bishop, G. D., & Myers, D. G. (1974). Informational influence in group discussion. *Organizational Behavior and Human Performance*, *12*(1), 92–104. doi: 10.1016/0030-5073(74)90039-7

Bramson, A., Grim, P., Singer, D. J., Fisher, S., Berger, W., Sack, G., & Flocken, C. (2016). Disambiguation of social polarization concepts and measures. *Journal of Mathematical Sociology*, *40*(2), 80–111. doi: 10.1080/0022250X.2016.1147443

Brown, R. (1986). Group polarization. In *Social psychology* (2nd ed., pp. 200–248). New York: Free Press.

Chacoma, A., & Zanette, D. H. (2015). Opinion formation by social influence: From experiments to modeling. *PLoS ONE*, *10*(10), 1–16. doi: 10.1371/journal.pone.0140406

Cikara, M., & Van Bavel, J. J. (2014). The Neuroscience of Intergroup Relations: An Integrative Review. *Perspectives on Psychological Science*, *9*(3), 245–274. doi: 10.1177/1745691614527464

Converse, P. E. (2006). The Nature of Belief Systems in Mass Publics (1964). *Critical Review*, *18*(1-3), 1–74. doi: 10.4324/9780203505984-10

Dandekar, P., Goel, A., & Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5791–6. doi: 10.1073/pnas.1217220110

Dixit, A. K., & Weibull, J. W. (2007). Political Polarization. *Proceedings of the National Academy of*

Sciences of the United States of America, *104*(2), 7351–7356. doi: 10.1073/pnas.0702071104

Flache, A., & Macy, M. W. (2011). Small Worlds and Cultural Polarization. *The Journal of Mathematical Sociology*, *35*(1-3), 146–176.

Galam, S., & Jacobs, F. (2007). The role of inflexible minorities in the breaking of democratic opinion dynamics. *Physica A: Statistical Mechanics and its Applications*, *381*(1-2), 366–376. doi: 10.1016/j.physa.2007.03.034

Guazzini, A., Cini, A., Bagnoli, F., & Ramasco, J. J. (2015). Opinion dynamics within a virtual small group: the stubbornness effect. *Frontiers in Physics*, *3*(September), 1–9. doi: 10.3389/fphy.2015.00065

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(8), 2554–2558. doi: 10.1073/pnas.79.8.2554

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, *50*(6), 1141–1151.

Klein, E. (2020). *Why We're Polarized.* New York: Simon and Schuster.

Lewandowsky, S., Pilditch, T. D., Madsen, J. K., Oreskes, N., & Risbey, J. S. (2019). Influence and seepage: An evidence-resistant minority can affect public opinion and scientific belief formation. *Cognition*, *188*(June 2018), 124–139. doi: 10.1016/j.cognition.2019.01.011

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. doi: 10.1016/j.jesp.2018.08.009

Martins, A. C., & Galam, S. (2013). Building up of individual inflexibility in opinion dynamics. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *87*(4), 1–8. doi: 10.1103/PhysRevE.87.042807

Mäs, M., & Flache, A. (2013). Differentiation without distancing. explaining bi-polarization of opinions without negative influence. *PLoS ONE*, *8*(11).

Mason, L. (2018). *Uncivil Agreement.* Chicago: University of Chicago Press.

Mobilia, M., Petersen, A., & Redner, S. (2007). On the role of zealotry in the voter model. *Journal of Statistical Mechanics: Theory and Experiment*, *2007*(08), P08029–P08029. doi: 10.1088/1742-5468/2007/08/p08029

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, *12*(2), 125–135.

Moussaïd, M., Kämmer, J. E., Analytis, P. P., & Neth, H. (2013). Social influence and the collective dynamics of opinion formation. *PLoS ONE*, *8*(11). doi: 10.1371/journal.pone.0078433

Mueller, S. T., & Tan, Y.-Y. S. (2018). Cognitive perspectives on opinion dynamics: the role of knowledge in consensus formation, opinion divergence, and group polarization. *Journal of Computational Social Science*, *1*(1), 15–48. doi: 10.1007/s42001-017-0004-7

Myers, D. G. (1982). Polarizing Effects of Social Interaction. In H. Brandstätter, J. H. Davis, & G. Stocker-Kreichgauer (Eds.), *Group decision making.* London: Academic Press.

Myers, D. G., & Arenson, S. J. (1972). Enhancement of Dominant Risk Tendencies in Group Discussion. *Psychological Reports*, *30*(2), 615–623. doi: 10.2466/pr0.1972.30.2.615

Myers, D. G., & Bishop, G. D. (1970). Discussion Effects on Racial Attitudes. *Science*, *169*, 778–779.

Myers, D. G., & Lamm, H. (1975). The polarizing effect of group discussion. *American Scientist*, *63*(3), 297–303.

Sieber, J., & Ziegler, R. (2019). Group Polarization Revisited: A Processing Effort Account. *Personality and Social Psychology Bulletin*, *45*(10), 1482–1498. doi: 10.1177/0146167219833389

Smaldino, P. E. (2017). Models Are Stupid, and We Need More of Them. In R. R. Vallacher, A. Nowak, & S. J. Read (Eds.), *Computational models in social psychology.* Psychology Press.

Smaldino, P. E. (2019). Better methods can't make up for mediocre theory. *Nature*, *575*(7781), 9. doi: 10.1038/d41586-019-03350-5

Sunstein, C. (2002). The Law of Group Polarization. *Journal of Political Philosophy*, *10*(2), 175–195.

Teger, A. I., & Pruitt, D. G. (1967). Components of group risk taking. *Journal of Experimental Social Psychology*, *3*(2), 189–205. doi: 10.1016/0022-1031(67)90022-4

Toepoel, V., & Funke, F. (2018). Sliders, visual analogue scales, or buttons: Influence of formats and scales in mobile and desktop surveys. *Mathematical Population Studies*, *25*(2), 112–122. doi: 10.1080/08898480.2018.1439245

Turner, M. A., & Smaldino, P. E. (2018). Paths to Polarization: How Extreme Views, Miscommunication, and Random Chance Drive Opinion Dynamics. *Complexity.*

Vinokur, A., & Burstein, E. (1974). Effects of partially shared persuasive arguments on group-induced shifts: A group-problem-solving approach. *Journal of Personality and Social Psychology*, *29*(3), 305–315. doi: 10.1037/h0036010

Zaller, J. R. (1992). *The Nature and Origins of Mass Opinion.* New York: Cambridge University Press.