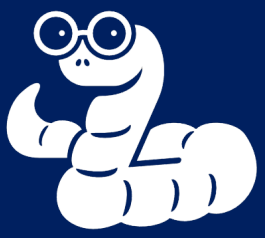

WORKing papers in Management Science

WORMS/25/03

# Depolarizing power of anticonformity

Arkadiusz Lipiecki[1]
Katarzyna Sznajd-Weron[1]

[1] Faculty of Management, Wrocław University of Science and Technology, Poland

WORMS is a joint initiative of the Management Science departments of the Wrocław University of Science and Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

# Depolarizing Power of Anticonformity

Arkadiusz Lipiecki[a], Katarzyna Sznajd-Weron[a]

[a]*Faculty of Management, Wrocław University of Science and Technology, Wyb. Wyspianskiego 27, Wrocław, 50-370, Poland*

**Abstract**

Political polarization hinders collective decision-making across multiple domains, from public health to environmental policy. Therefore, depolarization strategies are crucial and have been increasingly studied. Anticonformity, responding to social influence by opposing the opinions of others, has been associated with increased polarization, while its potential role as a depolarizing force has been largely overlooked. Although psychologists point to different forms of anticonformity, most computational models focus solely on xenophobia, prejudice against outsiders, which radicalizes opinions. Our work addresses this gap by considering another type of anticonformity – asserting uniqueness. We propose the counterintuitive hypothesis that increasing the disagreement by anticonforming to the influence group can reduce issue-based polarization. Within a family of computational models, we show that a depolarizing intervention based on promoting uniqueness may be more effective than traditional interventions, such as decreasing in-group favoritism or enhancing tolerance. We discuss the relevance of our findings through the lens of recent psychological experiments on strategic anticonformity, which demonstrate the potential of applying the proposed depolarizing intervention in real-world settings.

*Keywords:* Political polarization, strategic anticonformity, opinion dynamics, bounded confidence

## 1. Introduction

When a polarized society has to make a collective decision, how can the members of opposing factions reach a compromise? Numerous papers studied how facilitating discussion across the party line and promoting openness to different views can decrease issue-based polarization (Santoro and Broockman, 2022; Bail et al., 2018; Combs et al., 2023; de Jong, 2024; Wojcieszak and Warner, 2020). We approach this problem by considering an overlooked type of anticonformity and propose a counterintuitive solution — instead of directly promoting agreement, we suggest that encouraging people to stand out from the group can act as a significantly more effective and universal depolarizing strategy. In an extensive computational study of a family of opinion dynamics models we demonstrate how anticonformity-based behavioral intervention can help polarized communities reach the middle ground. Finally, we provide an empirical justification for our theoretical framework, based on the recently conducted social experiments.

First, we want to emphasize that the problem of depolarizing a society is distinct from preventing polarization, and for this reason the starting point of our considerations is a society that is already divided into two factions, which represent partisan or ideological identity. This division is the cornerstone of political polarization (Smith et al., 2024; Fasching et al., 2024; Baldassarri and Page, 2021; Mason, 2018) that draws the line between *us* (the in-group) and *them* (the out-group). In polarized societies, positions on other issues sort along this single axis of conflict (McCoy et al., 2018), either in the process of partisan sorting or issue alignment (Baldassarri and Gelman, 2008; Kozlowski and Murphy, 2021). Recent studies highlight the

*Email addresses:* arkadiusz.lipiecki@pwr.edu.pl (Arkadiusz Lipiecki), katarzyna.weron@pwr.edu.pl (Katarzyna Sznajd-Weron)

growing partisan polarization on climate policy issues (Chen et al., 2021; David Caldwell and Vivyan, 2024). Hence, in our study, the agents embedded in this factional social environment face a public conflict over the decision among three ranked choices, and the choice preference of each agent is initially aligned with its camp: individuals within each faction share the same preference, while members of opposing factions hold maximally divergent positions. Since affective polarization is without a doubt an important aspect of the modern public debate (Iyengar et al., 2019; Druckman et al., 2021), we consider scenarios with different levels of in-group favoritism, which allows to capture the structural aspect of polarization — the formation of clusters of like-minded people. This phenomenon, closely related to echo chambers, is particularly prevalent in online social networks (Cota et al., 2019; Ebeling et al., 2023; Falkenberg et al., 2024; Fraxanet et al., 2024; Hohmann et al., 2023; Magdaci et al., 2022). In this context, we explore the effectiveness of various interventions aimed at overcoming the issue-based polarization, which refers to the differences in individuals' opinions. We focus on this type of polarization, because failing to reach a middle ground can hinder the initiatives that require collective action and interfere with the decision-making process, *e.g*. regarding environmental issues (Falkenberg et al., 2022; Judge et al., 2023; David Caldwell and Vivyan, 2024). And with increasingly severe effects of climate change, collective action is more important than ever before. In their academic call to action, Debnath et al. (2023) stress: "The time for substantial climate action is now, and public opinion matters in people-centric system-level change". Hence it comes as no surprise that afters years of intensive research on the issue of polarization, depolarization interventions are becoming the focal point, both within computational models (Axelrod et al., 2021; Górski et al., 2023; Ojer et al., 2023; Borges et al., 2024; Schweighofer and Garcia, 2024) and empirical social studies (Balietti et al., 2021; Wojcieszak and Warner, 2020; Gugushvili, 2024; Doornbosch et al., 2024; Park and Warner, 2024; de Jong, 2024; Flores et al., 2022).

In this paper, we ask whether anticonformity can be used as a depolarizing strategy and, if so, under what conditions it is effective. The novelty of this paper is two-fold. First, to our knowledge, anticonformity has never been used as a depolarizing strategy. In contrast, it has generally been believed to strengthen polarization, mainly because previous studies have focused on xenophobia-driven anticonformity, which is only one of several types of anticonformity (Nyczka et al., 2018). In this paper, we propose the use of anticonformity motivated by the desire to assert uniqueness as a depolarizing intervention strategy, an idea that has not yet been explored in computational or conceptual models. Second, we introduce and analyze an entire family of computational models to rigorously test our hypothesis on the effectiveness of anticonformity as a depolarizing force. All of these models are derived from the three-state $q$-voter model (Lipiecki and Sznajd-Weron, 2022), but only one has been analyzed so far, and in a simpler version that does not distinguish between in-group and out-group interactions, which is a key concept in this paper. We will explain this in more detail after we introduce the model, in the last part of Section 3.

## 2. Related works

In the context of our paper, a particularly interesting approach was suggested by Axelrod et al. (2021). They developed the Attraction–Repulsion Model — an agent-based model in which individuals are more likely to be exposed to and attracted by views similar to their own, while being repelled by views that are overly dissimilar. Based on this model, several interventions were proposed. The first is to increase tolerance (broadening the range of acceptable viewpoints) in order to prevent or slow polarization by fostering a more moderate population. This idea is consistent with the proposition of Van Der Maas et al. who suggested that one intervention to reduce polarization is to introduce agents into the system with more flexible attitudes that allow compromise. He used the metaphor of the meat-eating vegetarian — an agent who acts as a bridge between vegetarians and meat-eaters and can therefore communicate with and influence both groups (Van Der Maas et al., 2020). The logic of these interventions seems intuitive — those holding moderate views attract some of the more radical individuals. Yet, as is usually the case with emergent phenomena, our intu-

ition can fail in predicting the macro effects of micro level interventions. A somewhat contradictory result obtained in other agent-based models shows that sometimes increasing individuals' openness to assimilative social influence leads to more opinion bi-polarization (Liu et al., 2022). In agreement with this finding, another depolarization strategy proposed by Axelrod et al. (2021) focuses on reducing the exposure to dissimilar views. There is conflicting evidence in the literature as to whether increased contact with opponents increases polarization, decreases it, or has little or no effect (Santoro and Broockman, 2022; Bail et al., 2018; Combs et al., 2023; de Jong, 2024; Wojcieszak and Warner, 2020).

These results paint a puzzling picture. On one hand – it seems beneficial to be open to opposing views, on the other hand – being exposed to people that are ideologically different from us can escalate the divide. However, when we consider the rules of social response typically considered in computational models, this seeming contradiction can be resolved — negative interactions between individuals holding opposing views drives them further away, hence decreasing the frequency of such interactions has a positive impact on polarization. This social response, which can be called xenophobia, is to the best of our knowledge the only type of anticonformity considered so far in the context of polarization (Siedlecki et al., 2016; Axelrod et al., 2021; Liu et al., 2022). But this is merely one form of anticonformity among many others (Nail et al., 2013). Another may be to assert uniqueness or to establish one's behavioral freedom. In this case, people tend to distance themselves from similar individuals. Recently, Dvorak et al. (2024) have shown that standing out from the group can be a strategic response in certain situations where people are evaluated. Their study provides much needed experimental insights into the occurrence of anticonformity in group interactions. Yet they conclude that "societal benefits and costs of conformity and anticonformity may vary considerably across situations". Hence in this paper, we pose the question: Can this type of anticonformity depolarize polarized factions? To answer this question we evaluate the interplay among three different depolarizing strategies: enhancing anticonformity through asserting uniqueness, fostering tolerance, and promoting out-group contacts. We show that only anticonformity consistently contributes to depolarization, while the effectiveness of the other two interventions depends on specific conditions.

## 3. Model

In order to thoroughly test the role of strategic anticonformity in the depolarization of an initially polarized system, we will consider several versions of the general model of three-state opinion dynamics. We begin the description of the model with the part that is common to all versions.

We consider the system of $N$ agents, also called voters or individuals. Each agent, indexed by $x \in \{1, 2, \ldots, N\}$, is characterized by two attributes, one static and one dynamic. The static attribute denotes group membership — everyone can belong to one of two groups: A or B. We write that a given agent is a member of group A/B, or alternatively, that a given agent is of type A/B. The second attribute is a dynamic variable $s_x(t)$:

$$s_x(t) = i \in \{-1, 0, 1\}, \tag{1}$$

describing the individual opinion at time $t$.

The voters update their opinions after being exposed to group influence. Groups that the agents interact with can be composed from members of both the same faction (in-group interactions) and the opposing one (out-group interactions). The only distinction between in-group and out-group interactions will rely on their frequency. To describe this, we introduce a parameter $\beta$, defined as the frequency of out-group interaction. Thus $\beta < 1/2$ corresponds to interaction pattern with a homophily bias, which models the in-group favoritism. Noteworthy, our approach does not entail the out-group hostility, which is typically associated with affective polarization (Judge et al., 2023; Smith et al., 2024). However, as reviewed by Hewstone et al. (2002), the in-group favoritism does not have to be accompanied by the out-group antagonism. Furthermore, Iyengar et al. (2019) argue that more research is required to establish which sentiment contributes

more to the affective polarization. They point to evidence that in some situations, partisan bias is driven mostly by in-group love rather than out-group dislike.

Having laid out the interaction patterns, we can proceed to the rules of opinion dynamics in our setup, which are given by the threshold $q$-voter model (Nyczka et al., 2018; Vieira and Anteneodo, 2018; Vieira et al., 2020). This model is particularly interesting in the context of social interactions. Firstly, it belongs to the class of complex contagion models, which means that one agent must be reinforced by multiple others to be influenced, as opposed to simple contagions where one individual is sufficient to induce change. While simple contagion is appropriate for modeling the spread of a disease, complex contagion is much more appropriate when dealing with opinion dynamics (Centola and Macy, 2007; Vasconcelos et al., 2019; Borges et al., 2024). Secondly, it is a reversible model, by which we mean that the states of individual agents can shift back and forth (Kempe et al., 2015). Such reversibility reflects the fact that opinions are relatively volatile (Glynn et al., 2018). Furthermore, the threshold $q$-voter model can describe both the unanimity rule and the majority rule, as well as all intermediate types of influence used in psycho-social experiments, as reviewed briefly in Nyczka et al. (2018).

In our computational model, we adapt the threshold $q$-voter rule to the two-faction structure. Thus, at each time step, a target agent is randomly selected and then a group of $Q$ agents, called the $Q$-panel, is formed. The members of the $Q$-panel are drawn at random either from faction $A$ or $B$, with the selection frequency determined by the parameter $\beta$: with probability $\beta$, a member of the $Q$-panel is chosen from the faction to which the target agent does not belong, and with probability $1-\beta$, from its own faction, as shown in Fig. 1. If there are at least $q$ agents within the $Q$-panel who share the same opinion, they have enough influence to affect the target; we will refer to this influential group as the $q$-source. The value of $q$ must lie within the interval $[\lfloor Q/2+1 \rfloor, Q]$, where the left bound corresponds to the majority rule, and the right bound corresponds to the unanimity rule, considered in the original formulation of the $q$-voter model (Castellano et al., 2009). In terms of the social influence exerted on the target, we extend the model beyond its original formulation. Specifically, we will consider the following types of social responses:

1. **Conformity with bounded confidence**: The target agent will change its opinion only if there are at least $q$ agents in the influence group who (a) share the same opinion and (b) this opinion differs from the target's opinion by exactly 1. The target agent then adopts the opinion signaled by the $q$-source. The rule described by (b) is the so called bounded confidence (BC), which was first studied for continuous opinions (Deffuant et al., 2000; Hegselmann and Krause, 2002) and is by now one of the most common mechanisms employed in opinion dynamics models (Ye et al., 2025). This type of social response, which is related to homophily, is considered to be one of the most important factors in the emergence of polarization (Axelrod, 1997; Mäs and Flache, 2013).
2. **Conformity without bounded confidence**: The target agent will change its opinion only if there are at least $q$ agents in the influence group whose opinions are on the same side with respect to the opinion of the target, i.e. there are either at least $q$ agents $x$ such that $s_x(t) < s_{x'}(t)$, or $q$ agents $x$ such that $s_x(t) > s_{x'}(t)$, where $x'$ is the target agent. The opinion of the target then changes by one in the direction of the opinions of the $q$-sources.
3. **Anticonformity**: We examine the form of anticonformity that occurs when an agent interacts with a group that shares its views and changes its opinion to stand out from the crowd, in the spirit of *asserting uniqueness*. Thus, the $q$-source is formed if at least $q$ agents from the $Q$-panel have exactly the same opinion as the target. If their opinion is $-1$ or 1, the target will set its state to 0; and if the opinion is 0, it will randomly choose between $-1$ and 1.

The first and third types of social responses have been considered in the three-state $q$-voter model (Lipiecki and Sznajd-Weron, 2022), but never within the threshold $q$-voter model. The second type of conformity has never been considered before. It can be understood as a higher tolerance scenario, since agents now also conform to those whose opinions differ from their own by more than 1. Moreover, unanimity in the

$q$-source is not required. Thus, the introduction of this type of social response will allow us to study the role of tolerance enhancement. On the other hand, it can also be treated as a different model of conformity, allowing us to test the universality of our results.

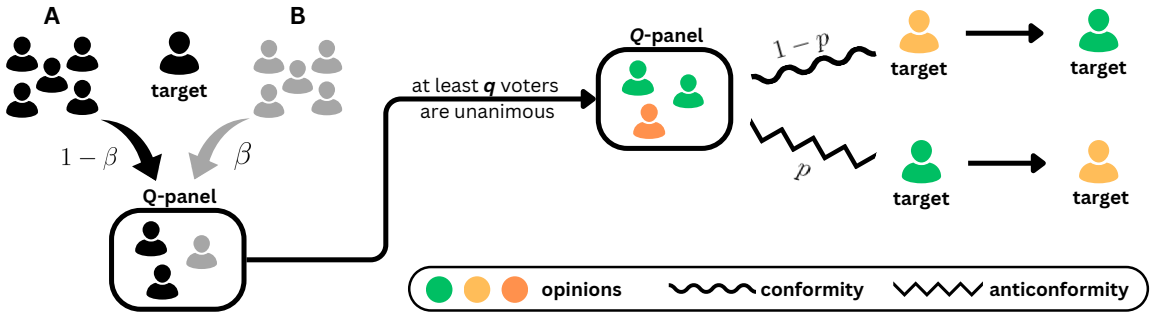


Figure 1: Schematic visualizations of the dynamic model with conformity and anticonformity as a depolarizing intervention. First, a target agent is randomly selected, and the members of the $Q$-panel are randomly drawn from either faction $A$ or $B$, with the proportion controlled by the parameter $\beta$. Second, if at least $q$ agents within the $Q$-panel share the same opinion, they influence the target: with probability $1-p$, the target conforms to the $Q$-panel, and with probability $p$, the target anticonforms.

We will consider all three combinations of the two different types of social responses defined above. As an example, schematic visualization of the model with conformity and anticonformity as a depolarizing intervention is shown in Fig. 1. Each combination will be examined under two approaches: dynamic and static, resulting in a total of six different models. In the dynamic approach, agents randomly switch between two types of social response, for example between anticonformity (with probability $p$) and conformity (with probability $1-p$). In the static approach, agents' social responses remain fixed over time: a fraction $p$ of agents always exhibit anticonformity, while a fraction $1-p$ always conforms. Of course, the same goes for other combinations of social responses.

In statistical physics, the dynamic and static approaches are referred to as annealed and quenched, respectively. In the social context, the annealed (dynamic) approach has been associated with situation-specific responses, whereas the quenched (static) approach has been linked to person-specific responses. However, in this paper, we do not use this terminology because we consider the different types of social responses as intervention strategies, rather than associating them with situational or personality-based factors.

At this point, we would like to clarify the novelty of this paper from a modeling perspective in more detail than was done in the introduction. So far, only one of these six models has been studied, which is a combination of conformity with bounded confidence and anticonformity within the dynamic approach (Lipiecki and Sznajd-Weron, 2022). Moreover, this analysis was performed only within the $q$-voter model and never within the threshold $q$-voter model. Finally, the model was examined in a different setting, where each agent had the same probability of interacting with all other agents, corresponding to $\beta = 0$. Therefore, all results that will be presented in this paper are new.

The model defined in this section can be described by the set of ordinary differential equations (ODEs) representing the temporal change $c_{iX}(t)$ of the concentration of agents with a given opinion in each of the two factions. The ODEs for all six versions of the model are derived in Appendix A, along with the details of the numerical analysis we conducted in Appendix B. For additional explanation of the models, we refer the reader to pseudocodes of Monte Carlo simulations provided in Appendix D, which can be used for numerical analysis of finite-size systems.

## 4. Polarization Index

Since our goal is to investigate how effectively depolarization interventions work for two polarized groups, we start from a state of maximum polarization as the initial condition. However, to evaluate po-

larization strength, we first need to define a suitable measure.

Let $c_{iX}(t)$ denote the concentration of agents with opinion $i \in \{-1, 0, 1\}$ in faction $X \in \{A, B\}$ at time $t$:

$$c_{iX}(t) :- \frac{N_{iX}(t)}{N}, \tag{2}$$

where $N_{iX}(t)$ is the number of agents with opinion $i$ in faction $X$ at time $t$.

We use the well-established *polarization index* $\mu$ (Morales et al., 2015) to quantify polarization and consensus. Although originally defined for continuous opinions in the interval $[-1, 1]$, we note that it can be adapted to our model. By mapping the concentrations of our three discrete opinions onto Dirac delta distributions centered at $-1$, 0 and 1 with respective masses of $c_{-1}$, $c_0$ and $c_1$ (where $c_i = c_{iA} + c_{iB}$), we define the *polarization index* $\mu$ for our three-state opinion system:

$$\mu = \frac{1 - |c_{-1} - c_1|}{2} \left( \frac{c_1}{c_0 + c_1} + \frac{c_{-1}}{c_{-1} + c_0} \right). \tag{3}$$

According to this definition, maximum polarization ($\mu = 1$) occurs if and only if $c_{-1} = c_1 = 1/2$, while it vanishes when a single opinion reaches full support. In our study, all opinion distributions corresponding to polarized states are characterized by $\mu > 1/2$, while consensus states by $\mu < 1/2$, see Appendix C.

## 5. Results

To compare different types of interventions across various models, we will construct phase diagrams that show the polarization index $\mu$ in the stationary state as a function of the probability of out-group interaction, $\beta$, and the level of intervention, $p$. The phases on these diagrams will be identified by looking at the opinion distribution in each of the group, as well as in the entire system, as shown in Fig. 2. The main result comparing different intervention strategies within all six models is shown in Fig. 3. We have studied the models for various values of the parameters $Q$ and $q$, but we first focus on the results for $Q = 4$ and $q = 3$, as this corresponds to a 75% majority in the influence group – a value commonly used in psychological experiments (Goldstein et al., 2008; Nyczka et al., 2018). Let us start with the model of conformity with BC & anticonformity as a depolarizing intervention, for which phase diagrams are shown in the left panels of Fig. 3. We see that the phase diagram is composed of three phases in both the dynamic and static approach. However, these phases are different in each of the two approaches. In the dynamic version, the only possible type of consensus is the middle-ground consensus. It is the strongest just after crossing the boundary, denoted by the white line, between the reddish (polarization) and blueish (consensus) regions. After crossing this line, the dominance of the middle opinion decreases and the systems moves towards disagreement ($\mu = 1/2$), where the concentrations of all three opinions are equal.

For the dynamic approach, increasing the probability of out-group interaction, *i.e.*, the number of contacts between members of two groups, does not depolarize the system, and some level of anticonformity is always necessary. Moreover, promoting out-group interactions in the presence of moderate anticonformity can lead to the transition from between-group to in-group polarization. This is accompanied by an increase in the polarization index (darker red color), achieving the opposite effect. However, there exists an optimal $\beta$, for which the anticonformity level required to reach middle-ground consensus is the smallest. It means that more frequent out-group interactions can be beneficial, but not always and only to a certain extent. In the static approach (lower panels of Fig. 3), however, the system can be depolarized either by encouraging more frequent out-group interactions or by increasing anticonformity. Yet, unless the probability of out-group interaction exceeds a certain threshold, raising anticonformity remains the only effective method for depolarization. Furthermore, when out-group interaction is high but anticonformity is low, the system tends to reach pole consensus—where the extreme option dominates. Such a consensus can deeply dissatisfy minority agents with anti-dominant opinions. Additionally, pole consensus arises through symmetry breaking,

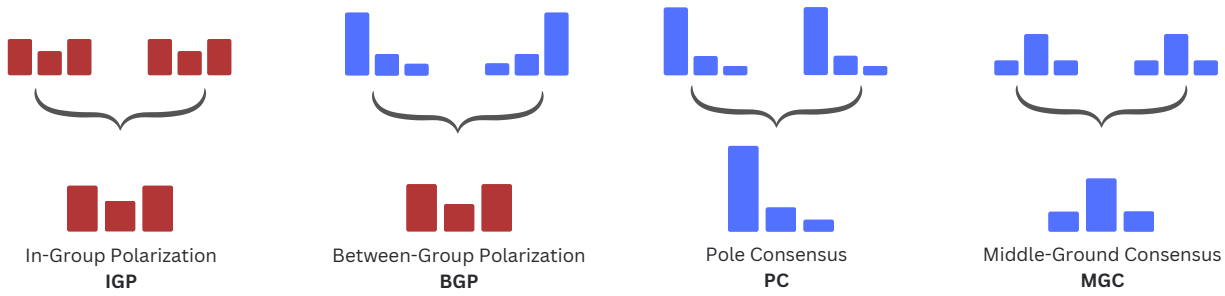


Figure 2: Schematic visualizations of phases corresponding to social structures based on the stationary opinion distributions. The top row shows the final distribution of opinions within each of the two factions, while the bottom row shows the distribution of opinions across the entire system. Note that Pole Consensus may appear as the dominance of either the left or right opinion, due to the symmetry between the extreme opinions. The opinion distributions presented in the scheme are exemplary, formal definitions used for phase classification are provided in Appendix C. We use the same color coding to distinguish between consensus and polarization in all subsequent figures: blue for consensus and red for polarization.

making it impossible to predict which extreme will prevail—a source of uncertainty that complicates managing collective change. To achieve middle-ground consensus, interventions promoting anticonformity are essential, as seen in the dynamic approach.

For the model of conformity without BC and anticonformity as a depolarizing intervention (middle panels of Fig. 3), the results from both the dynamic and static approaches are nearly identical. In this model, there are two methods to depolarize the system: increasing the probability $\beta$ of out-group interaction, or enhancing anticonformity. Below a certain threshold of $\beta$, anticonformity remains necessary to depolarize the system. Thus, even in a tolerant environment where people conform to those who hold significantly different opinions, some level of anticonformity is still required to achieve depolarization. However, for larger values of $\beta$, no anticonformity is needed to depolarize the system. Consequently, we could expect that conformity without BC would serve as an effective depolarizing intervention if the probability $\beta$ of out-group interaction is sufficiently high.

We examine this expectation in the final scenario: conformity with BC and conformity without BC as a depolarizing intervention, as shown in the right panels of Fig. 3. This time, the results for the dynamic and static approaches are not identical, but still similar, with two notable differences: in the static approach, the area representing the pole consensus phase is larger, while the area representing the polarization phase is smaller compared to the dynamic approach. More interestingly, the region of polarization is considerably larger than in the model with anticonformity. Below a certain value of $\beta$, the system remains polarized. Noteworthy, the probability $\beta$ of out-group interaction required to depolarize the system is higher than in the model that includes anticonformity. Furthermore, when the in-group favoritism is strong (low values of $\beta$), it is not possible to overcome polarization with this intervention.

It is worth to note that the transitions between phases are discontinuous, which is indicated by the jump of the polarization index, as shown in Fig. 4. Similar discontinuous transitions related to the existence of tipping points were observed in another computational model of opinion dynamics used to describe partisan polarization (Macy et al., 2021). In their conclusions, the authors wrote *Future research is needed to investigate the possibility that tipping dynamics generalize to other models of polarization*. In this regard, we show that discontinuous phases are indeed somewhat universal in partisan polarization.

So far, we have presented results only for a $Q$-panel size of 4 and a $q$-source size of 3. However, we have also tested other values of $Q$ and $q$, finding that the results change only quantitatively, as shown in Fig. 5 for the model of conformity with BC and anticonformity. We have presented the results for only a few values of $Q$: specifically, $Q = 4$ in Fig. 3 and $Q = 2, 3$ in Fig. 5. This choice of parameters is guided by empirical social studies, which indicate that the size of a freely formed human group rarely exceeds five (Desportes

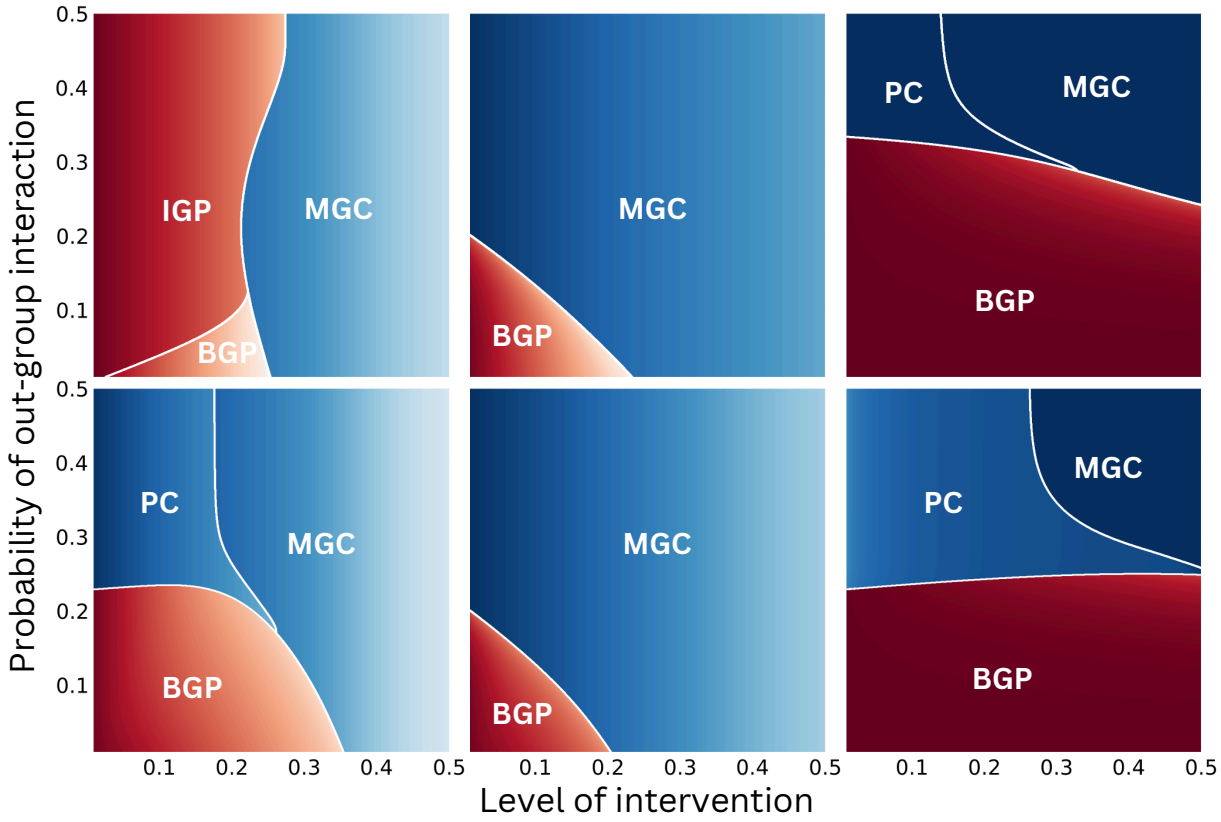


Figure 3: The polarization index, $\mu$, in the stationary state as a function of the probability of out-group interaction, $\beta$, and the level of intervention, $p$, for $Q = 4$ and $q = 3$. Reddish colors indicate $\mu > 1/2$, corresponding to polarized states **BGP** and **IGP**, while bluish colors indicate $\mu < 1/2$, corresponding to consensus states **PC** and **MGC**. The darker the color, the stronger the degree of polarization or consensus. White lines mark the boundaries between phases defined in Fig. 2. The results are presented for: **conformity with BC & anticonformity** as a depolarizing intervention (left panels), **conformity without BC & anticonformity** as a depolarizing intervention (middle panels), and **conformity with BC & conformity without BC** as a depolarizing intervention (right panels). The top panels display results for the dynamic approach, where $p$ represents the probability of intervention, while the bottom panels show results for the static approach, where $p$ indicates the fraction of agents permanently engaging in intervention behavior. The results were obtained by numerical evolution of the dynamic equations, as described in Appendix B. In the presented plots, the level of intervention starts from a small-non zero value $p = 0.01$. For $p = 0$, when only conformity with BC can take effect (left and right panels), the systems would not evolve from their initial state of total polarization, as per the model definition.

and Lemaine, 1988). Moreover, research shows that communication in small groups (up to five people) tends to be more interactive and dialogic (Fay et al., 2000). Additionally, small groups of 3–4 people have recently been used in a social experiment on strategic anticonformity (Dvorak et al., 2024), within which we will discuss our results in Section 6. However, to assess the model's sensitivity, we have conducted calculations over a much larger range of $Q$ and $q$, and the results remain qualitatively similar.

## 6. Conclusions

Social psychologists have identified several types of anticonformity, including strategic self-anticonformity, also known as reverse psychology (MacDonald et al., 2011; Nail et al., 2013). This form of anticonformity involves expressing an opinion contrary to one's true beliefs to persuade others toward a desired outcome, particularly when one expects that others will resist direct suggestions.

In this paper we propose a slightly different form of strategic anticonformity, although it still involves opposing others. In our conceptualization, its strategic aspect does not stem from a prospective reward for

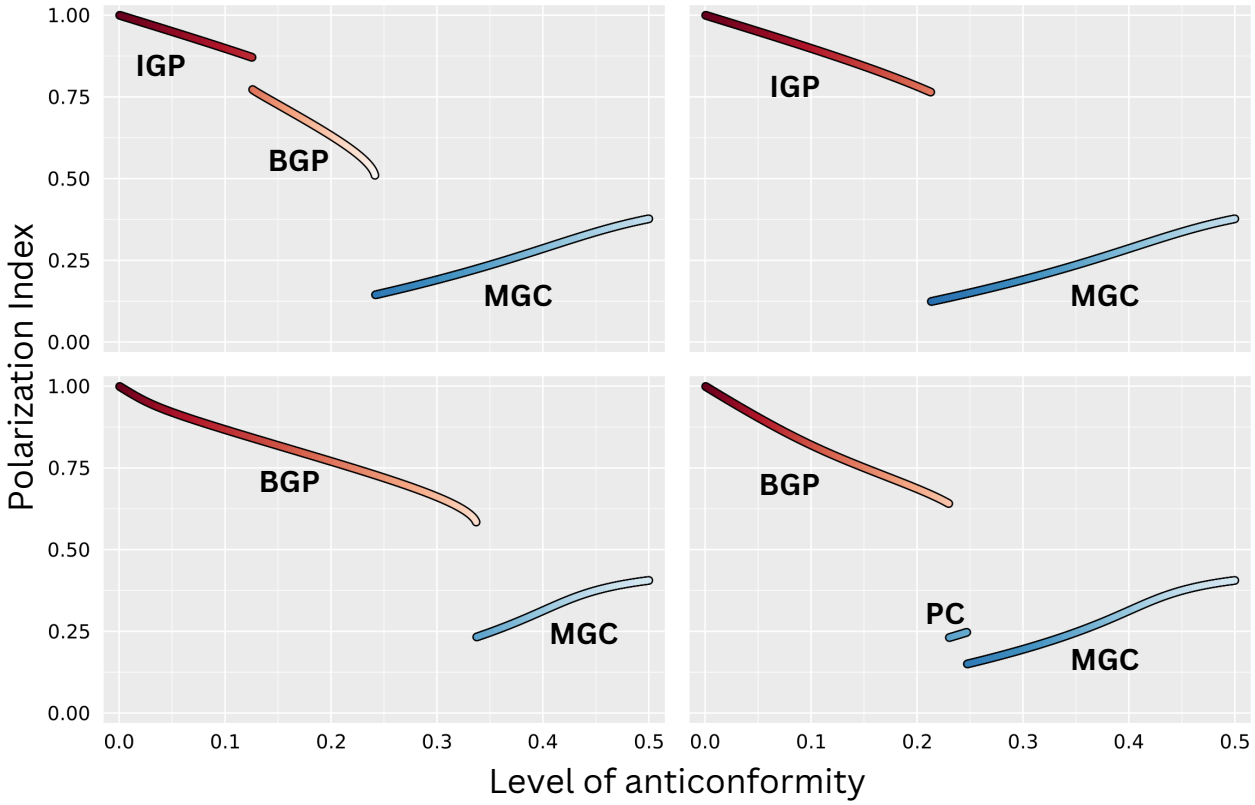


Figure 4: Dependence between the polarization index $\mu$ and the level of anticonformity for the **conformity with BC & anticonformity** model with $Q = 4$ and $q = 3$. Results for the dynamic approach are shown in the top panels, and for the static approach in the bottom panels. The left panels correspond to probability of out-group interaction $\beta = 0.05$, while the right panels to $\beta = 0.2$. The results were obtained by numerical evolution of the dynamic equations, as described in Appendix B.

an individual, but rather for a collective. We propose that promoting asserting uniqueness among polarized communities can act as an effective depolarizing intervention. Imagine that you are a manager of a team that is experiencing polarization. Its members often cannot agree on a solution and become deadlocked, which hampers their ability to achieve objectives. To facilitate reaching a middle-ground, you could promote anticonformity among your subordinates, so that some of them will try to stand out from the group. According to the results of this study, this could help to depolarize the system more effectively than other interventions, such as increasing tolerance (here represented by conformity without BC) or increasing the frequency of contact between polarized groups (here represented by probability $\beta$ of out-group interaction).

In our study, as in several others (Axelrod et al., 2021; Macy et al., 2021), greater tolerance supports depolarization. However, in our case, it can only act as a depolarizing intervention if the number of contacts between groups is sufficiently large. Similarly, a greater number of out-group contacts may promote depolarization, but it can also amplify existing polarization, depending on how people respond to social influence. Interestingly, our results show that the latter can occur even without xenophobia or out-group antagonism, which were examined in related works (Axelrod et al., 2021; Siedlecki et al., 2016).

Comparing the intervention based on strategic anticonformity with increasing the tolerance (conformity without BC) showed that the latter cannot overcome polarization in the presence of strong in-group interaction bias. Hence, if we do not have the means to influence interaction patterns and increase the number of connections across party lines, anticonformity remains the only effective depolarization strategy.

This result seems particularly interesting in light of the recently obtained results of psychological experiments conducted by Dvorak et al. (2024). Firstly, in their experiments conformity and anticonformity were

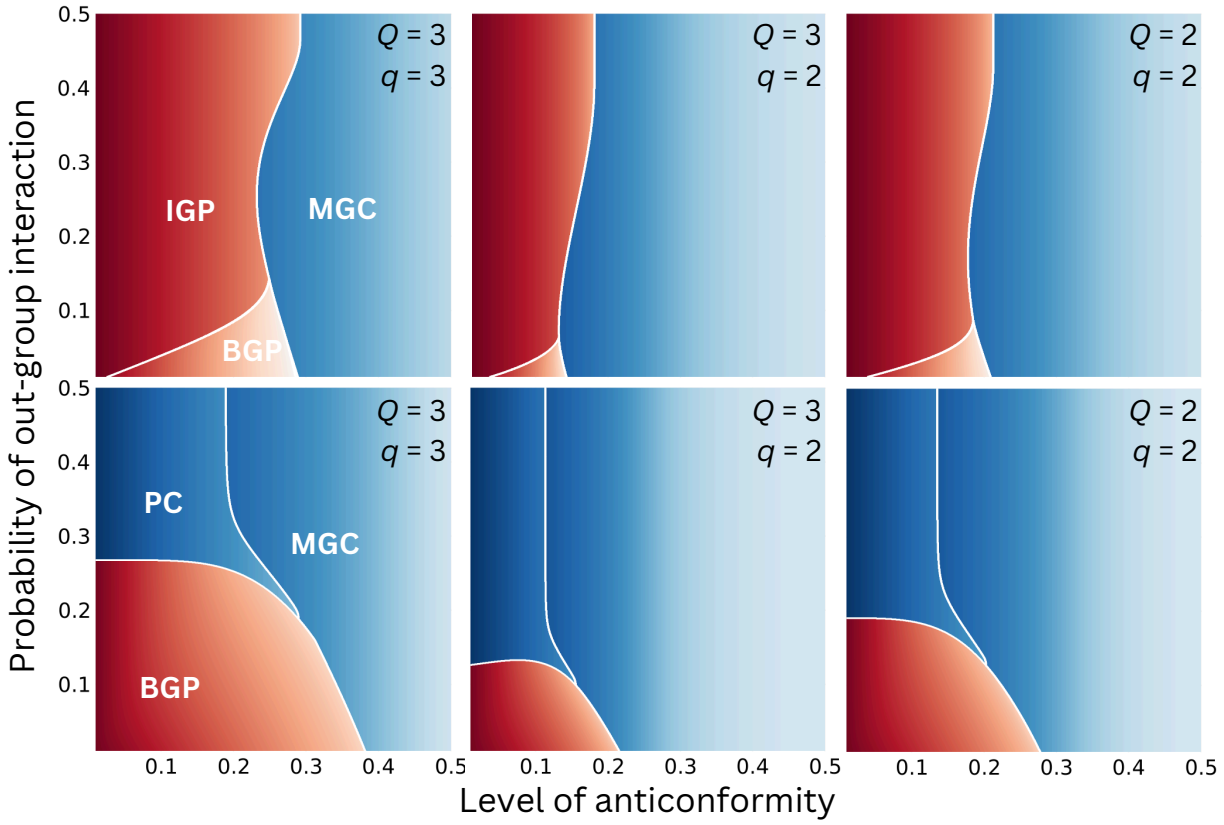


Figure 5: The polarization index, $\mu$, in the stationary state as a function of the probability of out-group interaction, $\beta$, and the level of anticonformity, $p$, the conformity with BC & anticonformity model. Reddish colors indicate $\mu > 1/2$, corresponding to polarized states **BGP** and **IGP**, while bluish colors indicate $\mu < 1/2$, corresponding to consensus states **PC** and **MGC**. The darker the color, the stronger the degree of polarization or consensus. White lines mark the boundaries between phases defined in Fig. 2. Results for the dynamic approach are shown in the top panels, while results for the static approach are shown in the bottom panels. Different sizes of the $Q$-panel (the influence group) and the $q$-source (the threshold required for influence) are indicated in the upper right corner of each panel. The results were obtained by numerical evolution of the dynamic equations, as described in Appendix B.

measured by the choice adjustment of the participant with respect to the majority choice among two other group members. This corresponds exactly to the $q = Q = 2$ scenario in our model. Secondly, they find that reward-based incentives can increase anticonformity, even in the context of factual statements. This leads us to believe that the proposed intervention could be successfully implemented by inducing anticonformity with an appropriate reward prospect. Thirdly, they observe that social responses are heterogeneous, which is captured in our model by considering situation-specific and person-specific behavior. Our results show that the depolarizing effect of anticonformity is universal across these scenarios. Finally, their estimates of the frequency of anticonformity in the prospect of reward start at 15%, which is of the same order as the critical probability of anticonformity required to depolarize our agent system. Furthermore, our research suggests that tipping dynamics is quite universal within the tested models of opinion dynamics. In view of this, we believe that our study, although theoretical, provides significant insights into the interplay of anticonformity and polarization and that the proposed concept of strategic anticonformity as a depolarizing intervention displays strong potential for real-world applications.

### CRediT authorship contribution statement

**Arkadiusz Lipiecki:** Conceptualization, Methodology, Software, Investigation, Visualization, Writing – original draft. **Katarzyna Sznajd-Weron:** Funding acquisition, Supervision, Writing – review & editing.

## Declaration of competing interests

The authors declare no competing interests.


## Acknowledgments

This research was funded by the National Science Centre, Poland under the OPUS call in the Weave programme, project no. 2023/51/I/HS6/02269.

During the preparation of this work, the author(s) used DeepL Write to improve the writing style and correct typographical and grammatical errors. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.


## Code availability

The source code used in this study is publicly available at github.com/lipiecki/DepolarizingAnticonformity.jl as a replication package developed in Julia programming language.

## Appendix A. Evolution equations

### *Appendix A.1. Dynamic approach*

Let $c_{iX}(t)$ denote the concentration of agents with opinion $i \in \{-1, 0, 1\}$ in faction $X \in \{A, B\}$ at time $t$:

$$c_{iX}(t) := \frac{N_{iX}(t)}{N}, \tag{A.1}$$

where $N_{iX}(t)$ is the number of agents with opinion $i$ in faction $X$ at time $t$. The number of agents in each faction is conserved and equal to $N/2$, so the following equations must hold:

$$\begin{aligned} c_{-1A}(t) + c_{0A}(t) + c_{1A}(t) &= \frac{1}{2}, \\ c_{-1B}(t) + c_{0B}(t) + c_{1B}(t) &= \frac{1}{2}. \end{aligned} \tag{A.2}$$

Using the above equations we can reduce the number of independent variables to four. Due to the symmetry of the boundary opinions ($-1$ and 1), we decide to describe the system with $c_{-1A}$, $c_{1A}$, $c_{-1B}$, $c_{1B}$. The evolution of the system can be given in terms of transition rates $\gamma_X^{i\to j}$, representing a probability that an agent will change its opinion from $i$ to $j$, given that it belongs to faction $X$:

$$\begin{cases} \dfrac{dc_{-1A}}{dt} &= c_{0A}\gamma_A^{0\to -1} - c_{-1A}\gamma_A^{-1\to 0}, \\ \dfrac{dc_{1A}}{dt} &= c_{0A}\gamma_A^{0\to 1} - c_{1A}\gamma_A^{1\to 0}, \\ \dfrac{dc_{-1B}}{dt} &= c_{0B}\gamma_B^{0\to -1} - c_{-1B}\gamma_B^{-1\to 0}, \\ \dfrac{dc_{1B}}{dt} &= c_{0B}\gamma_B^{0\to 1} - c_{1B}\gamma_B^{1\to 0}. \end{cases} \tag{A.3}$$

Note that only transition probabilities between neighboring states, i.e. $(i, j) : |i - j| = 1$ are included in the above system of evolution equations. This is due to the fact that within our model only unit changes of opinion are allowed. In the dynamic approach, we consider two behavior rules, $\mathcal{B}_p$ and $\mathcal{B}_{1-p}$, at each time step, a randomly selected target agent behaves according to rule $\mathcal{B}_p$ with probability $p$ or rule $\mathcal{B}_{1-p}$ with probability $(1 - p)$. The explicit formulas for transition probabilities $\gamma_X^{i\to j}$ depend on the selected rules. Below, we consider three rule combinations presented in the paper.

*Conformity with bounded confidence and anticonformity.* In this version of the model, $\mathcal{B}_p$ corresponds to anticonformity, while $\mathcal{B}_{1-p}$ to conformity with bounded confidence. The transition probability takes the form:

$$\gamma_X^{i\to j} = \left[\left(p - \frac{p}{2}\mathbb{1}_{i=0}\right) S_{i|X} + (1-p)\, S_{j|X}\right] \mathbb{1}_{|i-j|=1}. \tag{A.4}$$

In the above equation, $S_{i|X}$ is probability that at least $q$ voters in the $Q$-panel of a voter from faction $X$ hold opinion $i$:

$$\begin{aligned} S_{i|X} &= \sum_{r=q}^{Q} \binom{Q}{r} \left[(1-\beta)\,\tfrac{1}{2} c_{iX} + \beta \tfrac{1}{2} c_{i\bar{X}}\right]^r \left[1 - (1-\beta)\,\tfrac{1}{2} c_{iX} - \beta \tfrac{1}{2} c_{i\bar{X}}\right]^{Q-r} \\ &= I\left[(1-\beta)\,\tfrac{1}{2} c_{iX} + \beta \tfrac{1}{2} c_{i\bar{X}},\, q,\, Q-q-1\right], \end{aligned} \tag{A.5}$$

where $I$ is the regularized incomplete beta function and $\bar{X}$ denotes the faction opposite to $X$.

*Conformity without bounded confidence and anticonformity.* In this version of the model, $\mathcal{B}_p$ corresponds to anticonformity, while $\mathcal{B}_{1-p}$ to conformity without bounded confidence. The transition probability takes the form:

$$\begin{aligned} \gamma_X^{i\to j} = &\left[\left(p - \frac{p}{2}\mathbb{1}_{i=0}\right) S_{i|X} + (1-p)\, S_{\{k:k>i\}|X}\right] \mathbb{1}_{j-i=1} \\ &+ \left[\left(p - \frac{p}{2}\mathbb{1}_{i=0}\right) S_{i|X} + (1-p)\, S_{\{k:k<i\}|X}\right] \mathbb{1}_{i-j=1}, \end{aligned} \tag{A.6}$$

where $S_{\{k:k>i\}|X}$ and $S_{\{k:k<i\}|X}$, are the probabilities that at least $q$ voters in the $Q$-panel of a voter from faction $X$ hold opinion larger than $i$ and, respectively, smaller than $i$. They can be expressed as:

$$\begin{aligned} S_{\{k:k>i\}|X} &= I\left[\sum_{k>i} (1-\beta)\,\tfrac{1}{2} c_{kX} + \beta \tfrac{1}{2} c_{k\bar{X}},\, q,\, Q-q-1\right], \\ S_{\{k:k<i\}|X} &= I\left[\sum_{k<i} (1-\beta)\,\tfrac{1}{2} c_{kX} + \beta \tfrac{1}{2} c_{k\bar{X}},\, q,\, Q-q-1\right]. \end{aligned} \tag{A.7}$$

*Conformity with and without bounded confidence.* In this version of the model, $\mathcal{B}_p$ corresponds to conformity without bounded confidence, while $\mathcal{B}_{1-p}$ to conformity with bounded confidence. The transition probability takes the form:

$$\begin{aligned} \gamma_X^{i\to j} = &\left[p S_{\{k:k>i\}|X} + (1-p)\, S_{j|X}\right] \mathbb{1}_{j-i=1} \\ &+ \left[p S_{\{k:k<i\}|X} + (1-p)\, S_{j|X}\right] \mathbb{1}_{i-j=1}. \end{aligned} \tag{A.8}$$

*Appendix A.2. Static approach*

In the static approach, voters are heterogeneous with respect to the behavior rules. Each voter is randomly assigned a rule, $\mathcal{B}_p$ with probability $p$ or $\mathcal{B}_{1-p}$ with probability $1-p$, and always follows this rule. Therefore, to describe the evolution of the system we need to consider the concentration of agents with opinion $i \in \{-1, 0, 1\}$ in faction $X \in \{A, B\}$ that follow behavior rule $\mathcal{B} \in \{\mathcal{B}_p, \mathcal{B}_{1-p}\}$:

$$c_{iX\mathcal{B}}(t) := \frac{N_{iX\mathcal{B}}(t)}{N}, \tag{A.9}$$

where $N_{iX\mathcal{B}}(t)$ is the number of agents with opinion $i$ in faction $X$ and behavior $\mathcal{B}$ at time $t$. The total number of agents in each faction, as well as the total number of agents assigned with $\mathcal{B}_p$ and $\mathcal{B}_{1-p}$ remain constant. Furthermore, since the behavior rules of each agent are independent and identically distributed

random variables, the concentration of voters, which follow rule $\mathcal{B}$ is the same within faction $A$ and $B$. This can be formalized as

$$\begin{aligned} c_{-1X\mathcal{B}_p}(t) + c_{0X\mathcal{B}_p}(t) + c_{1X\mathcal{B}_p}(t) &= \frac{p}{2}, \\ c_{-1X\mathcal{B}_{1-p}}(t) + c_{0X\mathcal{B}_{1-p}}(t) + c_{1X\mathcal{B}_{1-p}}(t) &= \frac{1-p}{2}. \end{aligned} \tag{A.10}$$

This allows us to reduce the number of independent variables to 8. To retain consistency with the derivation for dynamic approach, we choose $c_{-1A\mathcal{B}_p}$, $c_{1A\mathcal{B}_p}$, $c_{-1A\mathcal{B}_{1-p}}$, $c_{1A\mathcal{B}_{1-p}}$, $c_{-1B\mathcal{B}_p}$, $c_{1B\mathcal{B}_p}$, $c_{-1B\mathcal{B}_{1-p}}$, $c_{1B\mathcal{B}_{1-p}}$. Their evolution is expressed in terms of transition probabilities $\gamma_{X\mathcal{B}}^{i\to j}$ between neighboring states $i$ and $j$:

$$\begin{cases} \dfrac{dc_{-1A\mathcal{B}_p}}{dt} &= c_{0A\mathcal{B}_p}\gamma_{A\mathcal{B}_p}^{0\to -1} - c_{-1A\mathcal{B}_p}\gamma_{A\mathcal{B}_p}^{-1\to 0}, \\ \dfrac{dc_{1A\mathcal{B}_p}}{dt} &= c_{0A\mathcal{B}_p}\gamma_{A\mathcal{B}_p}^{0\to 1} - c_{1A\mathcal{B}_p}\gamma_{A\mathcal{B}_p}^{1\to 0}, \\ \dfrac{dc_{-1A\mathcal{B}_{1-p}}}{dt} &= c_{0A\mathcal{B}_{1-p}}\gamma_{A\mathcal{B}_{1-p}}^{0\to -1} - c_{-1A\mathcal{B}_{1-p}}\gamma_{A\mathcal{B}_{1-p}}^{-1\to 0}, \\ \dfrac{dc_{1A\mathcal{B}_{1-p}}}{dt} &= c_{0A\mathcal{B}_{1-p}}\gamma_{A\mathcal{B}_{1-p}}^{0\to 1} - c_{1A\mathcal{B}_{1-p}}\gamma_{A\mathcal{B}_{1-p}}^{1\to 0}, \\ \dfrac{dc_{-1B\mathcal{B}_p}}{dt} &= c_{0B\mathcal{B}_p}\gamma_{B\mathcal{B}_p}^{0\to -1} - c_{-1B\mathcal{B}_p}\gamma_{B\mathcal{B}_p}^{-1\to 0}, \\ \dfrac{dc_{1B\mathcal{B}_p}}{dt} &= c_{0B\mathcal{B}_p}\gamma_{B\mathcal{B}_p}^{0\to 1} - c_{1B\mathcal{B}_p}\gamma_{B\mathcal{B}_p}^{1\to 0}, \\ \dfrac{dc_{-1B\mathcal{B}_{1-p}}}{dt} &= c_{0B\mathcal{B}_{1-p}}\gamma_{B\mathcal{B}_{1-p}}^{0\to -1} - c_{-1B\mathcal{B}_{1-p}}\gamma_{B\mathcal{B}_{1-p}}^{-1\to 0}, \\ \dfrac{dc_{1B\mathcal{B}_{1-p}}}{dt} &= c_{0B\mathcal{B}_{1-p}}\gamma_{B\mathcal{B}_{1-p}}^{0\to 1} - c_{1B\mathcal{B}_{1-p}}\gamma_{B\mathcal{B}_{1-p}}^{1\to 0}. \end{cases} \tag{A.11}$$

The transition probability $\gamma_{X\mathcal{B}}^{i\to j}$ assumes different forms depending on the behavior rule $\mathcal{B}$. Below, we provide the formulas for three rules that are considered in the paper.

*Conformity with bounded confidence.* If rule $\mathcal{B}$ corresponds to conformity with bounded confidence, the transition probabilities are given by

$$\gamma_{X\mathcal{B}}^{i\to j} = S_{j|X}\mathbb{1}_{|i-j|=1}. \tag{A.12}$$

*Conformity without bounded confidence.* When $\mathcal{B}$ is the conformity without bounded confidence, the transition probabilities take the form:

$$\gamma_{X\mathcal{B}}^{i\to j} = S_{\{k:k>i\}|X}\mathbb{1}_{j-i=1} + S_{\{k:k<i\}|X}\mathbb{1}_{i-j=1}. \tag{A.13}$$

*Anticonformity.* For $\mathcal{B}$ equivalent to anticonformity, we obtain:

$$\gamma_{X\mathcal{B}}^{i\to j} = \left(p - \frac{p}{2}\mathbb{1}_{i=0}\right)S_{i|X}\mathbb{1}_{|i-j|=1}. \tag{A.14}$$

Plugging Eqs. (A.12)-(A.14) into system (A.11), we obtain final equations for the evolution of our model.

## Appendix B. Numerical evolution of the systems and sensitivity analysis

Given the systems of differential equations (A.3) and (A.11), we obtain the trajectories of the model by solving the equations numerically. In this study, we consider the initial condition corresponding to total between-group polarization, i.e., all voters in faction $A$ are assigned with opinion $-1$, while all voters in

faction $B$ hold opinion 1. However, since some of the models considered in the study exhibit symmetry breaking, symmetric initial conditions can correspond to unstable fixed points of the dynamical system. Therefore, we slightly perturb the initial opinion concentrations in faction $B$ with $\varepsilon = 10^{-6}$ so that $c_{1B}(t = 0) = 1/2 - \varepsilon$ and $c_{0B}(t = 0) = \varepsilon$. The dynamic equations are then evolved with an ODE solver until a steady state is reached.

To check the robustness of our results, we perform a sensitivity analysis by conducting the numerical evolution of systems for perturbed initial conditions and comparing the phase classification against the unperturbed system. The perturbations to initial conditions are parameterized by $\Delta$ through the following formula:

$$c_{0B}(0) = \varepsilon + \Delta, \quad c_{1B}(0) = \frac{1}{2} - \varepsilon - \Delta, \tag{B.1}$$

where $\Delta \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. We then color the phase space with the minimal $\Delta$ for which the phase classification of the perturbed system was different from that of the unperturbed one: the darker the color, the higher the sensitivity (smaller $\Delta$) of phase classification to the initial conditions. Figs. B.6 and B.7. present the sensitivity analysis for the results presented in Figs. 3. and 5. In line with expectations, some systems exhibit high sensitivity to initial conditions only close to the boundary between the phases, while others are strongly robust.

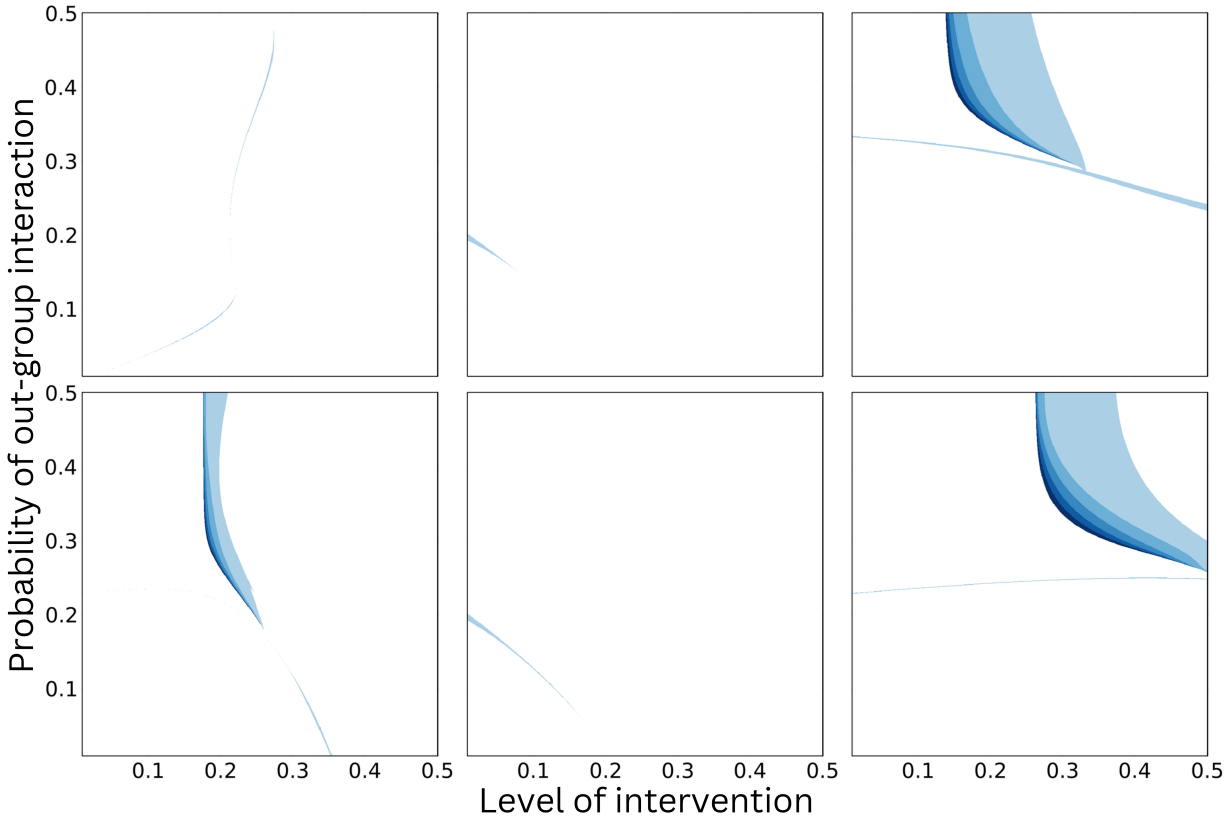


Figure B.6: Sensitivity plots for the results presented in Fig. 3., i.e. for $Q = 4$ and $q = 3$. The darker the color, the more sensitive the phase classification is to the perturbation in initial conditions described by B.1, the most intensive blue corresponds to sensitivity to $\Delta = 10^{-5}$, while white color indicates that the classification does not change even for $\Delta = 10^{-1}$. The results are presented for: **conformity with BC & anticonformity** as a depolarizing intervention (left panels), **conformity without BC & anticonformity** as a depolarizing intervention (middle panels), and **conformity with BC & conformity without BC** as a depolarizing intervention (right panels). The top panels display results for the dynamic approach, where $p$ represents the probability of intervention, while the bottom panels show results for the static approach, where $p$ indicates the fraction of agents permanently engaging in intervention behavior.

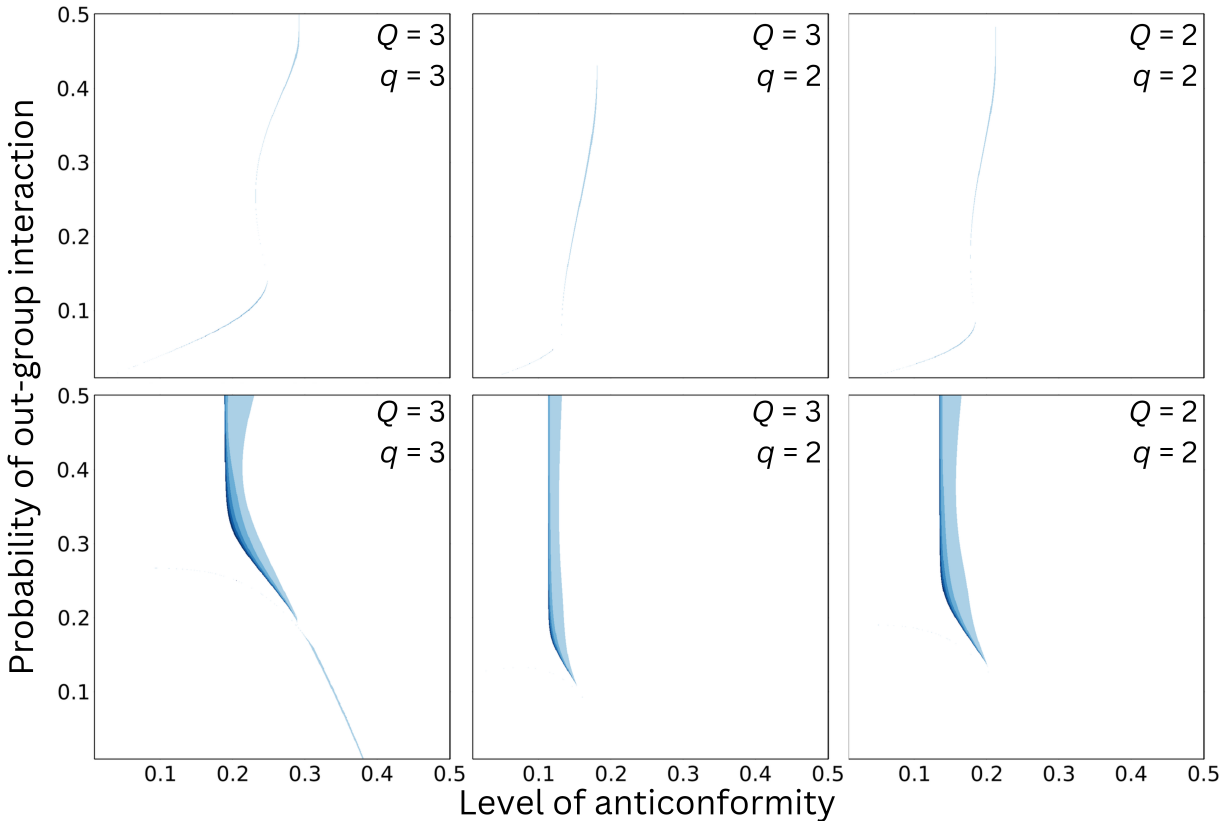


Figure B.7: Sensitivity plots for the results presented in Fig. 5. The darker the color, the more sensitive the phase classification is to the perturbation in initial conditions described by B.1, the most intensive blue corresponds to sensitivity to $\Delta = 10^{-5}$, while white color indicates that the classification does not change even for $\Delta = 10^{-1}$.. Results for the dynamic approach are shown in the top panels, while results for the static approach are shown in the bottom panels. Different sizes of the $Q$-panel (the influence group) and the $q$-source (the threshold required for influence) are indicated in the upper right corner of each panel. The results were obtained by numerical evolution of the dynamic equations, as described in Appendix B.

## Appendix C. Phase classification

In the paper we present a schematic visualization for four distinct phases obtained in the examined models. In the following, we lay out a formal classification according to which the phase diagrams were obtained. Firstly, let us consider how the values of the polarization index $\mu$ change with the distribution of the three-state opinion. In a simple case of symmetric distribution, we get $\mu > 1/2$ if the distribution is bimodal, $\mu = 1/2$ when it is uniform, and $\mu < 1/2$ for a unimodal distribution. Thus the threshold $1/2$ will act as our reference for distinguishing polarized and consensus states. Due to the fact that we are examining two-faction social structure, the phase classification of our system will be based on system-wide polarization index $\mu$, as well as the faction-level indices $\mu_A$ and $\mu_B$, defined as

$$\mu_X = \frac{1 - 2|c_{-1X} - c_{1X}|}{2} \left( \frac{c_{1X}}{c_{0X} + c_{1X}} + \frac{c_{-1X}}{c_{-1X} + c_{0X}} \right). \tag{C.1}$$

The following conditions are then used to determine the phase for a given system state:

*Between-Group Polarization (BGP).*

$$\left[\mu_A < \frac{1}{2}\right] \wedge \left[\mu_B < \frac{1}{2}\right] \wedge \left[\mu > \frac{1}{2}\right] \tag{C.2}$$

*In-Group Polarization (IGP).*

$$\left[\mu_A > \frac{1}{2}\right] \wedge \left[\mu_B > \frac{1}{2}\right] \wedge \left[\mu > \frac{1}{2}\right] \tag{C.3}$$

*Pole Consensus (PC).*

$$\left[\mu_A < \frac{1}{2}\right] \wedge \left[\mu_B < \frac{1}{2}\right] \wedge \left[\mu < \frac{1}{2}\right] \wedge [\max\{c_{-1}, c_0, c_1\} \in \{c_{-1}, c_1\}] \tag{C.4}$$

*Middle-Ground Consensus (MGC).*

$$\left[\mu_A < \frac{1}{2}\right] \wedge \left[\mu_B < \frac{1}{2}\right] \wedge \left[\mu < \frac{1}{2}\right] \wedge [\max\{c_{-1}, c_0, c_1\} = c_0] \tag{C.5}$$

Note that the proposed phase classification is based mostly on the polarization index, and the global distribution of opinion is explicitly used only to differentiate between the **PC** and **MGC** phases. The schematic opinion distributions visualized in Fig. 2 correspond to the solutions that we found to be most common in each phase. In our extensive computational study some systems described by the **dynamic model of anticonformity and conformity with BC** evolved to states that are not recognized by the classification presented above. However, for every group size specification the percentage of unclassified states appeared in less than 0.1% of the parameter space, and mostly in the vicinity of the phase boundaries. Additionally, in the extended study ($Q \in \{2, ..., 10\}$, $q \in \{\lfloor Q/2 \rfloor + 1, ..., Q\}$) we found that for $q \geq 4$ the **static model of conformity with BC and conformity without BC** starts to exhibit an unclassified mixed phase within a significant proportion of the parameter space. In such cases, the parameter region previously corresponding to **PC** (see the bottom right panel in Fig. 3.) shrinks and the opinion distributions in this region evolve to states in which one faction exhibits polarization and the other one – consensus. However, this observation in no way contradicts the general conclusions that we have drawn in the paper – conformity without BC does not allow to depolarize the system if the probability of out-group interactions is not sufficiently large. Plots from the extended numerical study, along with the source codes for obtaining them, are available in the replication package.

*Relation between the polarization index and opinion distribution.* Since the phase classification is based on the polarization index $\mu$, it is worth mentioning how it translates to the distribution of opinions in the agent system. In Fig. C.8, the polarization index is plotted with respect to $c_{-1}$ and $c_1$, with a marked boundary between polarization and consensus. As stated previously, for symmetric distributions ($c_{-1} = c_1$), the distributions characterized by $c_0 < 1/3$ correspond to polarization, which agrees with the intuitive understanding of polarization as a bimodal opinion distribution. Furthermore, the vast majority of the final distributions we have obtained (excluding the ones corresponding to the Pole Consensus phase) were symmetric. For asymmetric distributions, the boundary between polarization and consensus is a convex function passing through points $(1/4, 3/4)$, $(1/3, 1/3)$ and $(3/4, 1/4)$. In this case, deriving exact and intuitive classification is non-trivial, but it is possible to infer simple upper and lower bounds. If $\min\{c_{-1}, c_1\}$ is less than $1/4$, then the distribution corresponds to the consensus state; otherwise, if it is greater than $1/3$, the system is in a polarized phase. Note that although this simplified approach does not deal with all distributions, the ones remaining correspond to relatively moderate polarization index values (close to $1/2$), and are therefore less interesting from the perspective of studying polarization and consensus.

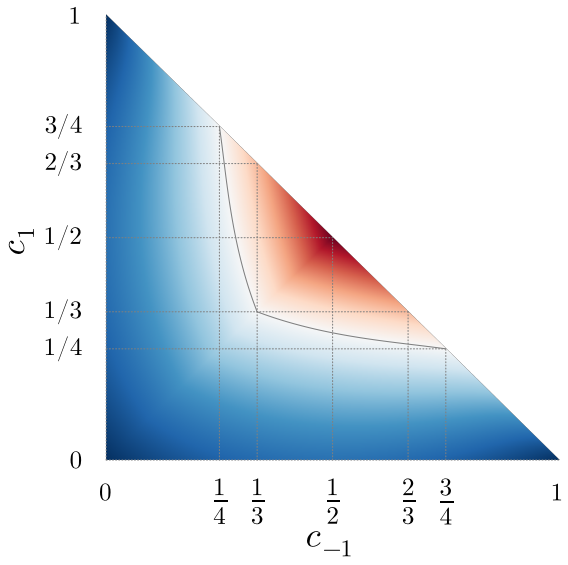


Figure C.8: Polarization index $\mu$ with respect to the concentrations of opinions $-1$ and $1$. Reddish colors indicate $\mu > 1/2$, corresponding to polarized states, while bluish colors indicate $\mu < 1/2$, corresponding to consensus states. The darker the color, the stronger the degree of polarization or consensus. Solid grey line indicates the boundary between polarized and consensus states.

## Appendix D. Algorithms

The results presented in the main body of the paper were obtained by the numerical evolution of dynamic equations. This approach is based on the assumption that the agent system is large enough to treat an elementary change in agent concentrations $1/N$ as infinitesimal. In the following, we present pseudocodes for Monte Carlo simulations which can be used to implement the considered models for finite-size systems. Note that the algorithms provided leave some freedom on which weighted random sampling variant is used to construct the $Q$-panel. For large systems, the differences between some of the sampling methods vanish. Hence, whether we use the sampling with or without replacement, or even allow the target to be selected as a member of the $Q$-panel, the ODE systems derived in Appendix A will remain unchanged.

*Appendix D.1. Dynamic approach*

**Algorithm 1:** One Monte Carlo step of the dynamic model with anticonformity and conformity with BC. $A$ and $B$ are the factions, i.e. subsets of agents, and $\mathcal{S} : (A \cup B) \to \{-1, 0, 1\}$ maps agents to their opinions.

**Input:** $Q > 1$, $q \in (\lfloor Q/2 \rfloor, Q]$, $p \in [0, 1]$, $A, B, \mathcal{S}$
**Output:** updated $\mathcal{S}$

**for** $i \leftarrow 1$ **to** $|A \cup B|$ **do**
  Draw a random target $x$ from $A \cup B$;
  Draw a sample $\mathcal{Q}$ of $Q$ agents from $(A \cup B) \backslash \{x\}$ (or $A \cup B$) using a weighted random sampling (with or without replacement), with weights $w_z = 1 - \beta$ for agents in the same faction as $x$ and $w_z = \beta$ for the rest of the agents;
  Draw $r \in [0, 1)$ from a uniform distribution;
  **if** $r < p$ **then**
    **if** $|\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) = \mathcal{S}(x)\}| \geq q$ **then**
      **if** $\mathcal{S}(x) = 0$ **then**
        $\mathcal{S}(x) \leftarrow \mathcal{S}(x) + (-1)^b$, where $b \sim$ Bernoulli$(1/2)$;
      **else**
        $\mathcal{S}(x) \leftarrow 0$;
  **else**
    **if** $\exists s : |\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) = s\}| \geq q \wedge |s - \mathcal{S}(x)| = 1$ **then**
      $\mathcal{S}(x) \leftarrow s$;
**return** $\mathcal{S}$;

**Algorithm 2:** One Monte Carlo step of the dynamic model with anticonformity and conformity without BC. $A$ and $B$ are the factions, i.e. subsets of agents, and $\mathcal{S} : (A \cup B) \to \{-1, 0, 1\}$ maps agents to their opinions.

**Input:** $Q > 1$, $q \in (\lfloor Q/2 \rfloor, Q]$, $p \in [0, 1]$, $A, B, \mathcal{S}$
**Output:** updated $\mathcal{S}$

**for** $i \leftarrow 1$ **to** $|A \cup B|$ **do**
  Draw a random target $x$ from $A \cup B$;
  Draw a sample $\mathcal{Q}$ of $Q$ agents from $(A \cup B) \backslash \{x\}$ (or $A \cup B$) using a weighted random sampling (with or without replacement), with weights $w_z = 1 - \beta$ for agents in the same faction as $x$ and $w_z = \beta$ for the rest of the agents;
  Draw $r \in [0, 1)$ from a uniform distribution;
  **if** $r < p$ **then**
    **if** $|\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) = \mathcal{S}(x)\}| \geq q$ **then**
      **if** $\mathcal{S}(x) = 0$ **then**
        $\mathcal{S}(x) \leftarrow \mathcal{S}(x) + (-1)^b$, where $b \sim$ Bernoulli$(1/2)$;
      **else**
        $\mathcal{S}(x) \leftarrow 0$;
  **else**
    **if** $|\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) > \mathcal{S}(x)\}| \geq q = 1$ **then**
      $\mathcal{S}(x) \leftarrow \mathcal{S}(x) + 1$;
    **else if** $|\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) < \mathcal{S}(x)\}| \geq q = 1$ **then**
      $\mathcal{S}(x) \leftarrow \mathcal{S}(x) - 1$;
**return** $\mathcal{S}$;

**Algorithm 3:** One Monte Carlo step of the dynamic model with conformity without BC and conformity with BC. $A$ and $B$ are the factions, i.e. subsets of agents, and $\mathcal{S} : (A \cup B) \to \{-1, 0, 1\}$ maps agents to their opinions.

**Input:** $Q > 1$, $q \in (\lfloor Q/2 \rfloor, Q]$, $p \in [0, 1]$, $A, B, \mathcal{S}$
**Output:** updated $\mathcal{S}$

**for** $i \leftarrow 1$ **to** $|A \cup B|$ **do**
  Draw a random target $x$ from $A \cup B$;
  Draw a sample $\mathcal{Q}$ of $Q$ agents from $(A \cup B)\backslash\{x\}$ (or $A \cup B$) using a weighted random sampling (with or without replacement), with weights $w_z = 1 - \beta$ for agents in the same faction as $x$ and $w_z = \beta$ for the rest of the agents;
  Draw $r \in [0, 1)$ from a uniform distribution;
  **if** $r < p$ **then**
    **if** $|\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) > \mathcal{S}(x)\}| \geq q = 1$ **then**
      $\mathcal{S}(x) \leftarrow \mathcal{S}(x) + 1$;
    **else if** $|\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) < \mathcal{S}(x)\}| \geq q = 1$ **then**
      $\mathcal{S}(x) \leftarrow \mathcal{S}(x) - 1$;
  **else**
    **if** $\exists s : |\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) = s\}| \geq q \wedge |s - \mathcal{S}(x)| = 1$ **then**
      $\mathcal{S}(x) \leftarrow s$;
**return** $\mathcal{S}$;

*Appendix D.2. Static approach*

**Algorithm 4:** One Monte Carlo step of the static model with anticonformity and conformity with BC. $A$ and $B$ are the factions, i.e. subsets of agents, $\mathcal{S} : (A \cup B) \to \{-1, 0, 1\}$ maps agents to their opinions, and $\mathcal{B} : (A \cup B) \to \{1, 0\}$ maps agents to their behavior (1 for anticonformity and 0 for conformity). Note that in the case of the static approach $p = |A \cup B|^{-1} \sum_{x \in A \cup B} \mathcal{B}(x)$.

**Input:** $Q > 1$, $q \in (\lfloor Q/2 \rfloor, Q]$, $A, B, \mathcal{S}, \mathcal{B}$
**Output:** updated $\mathcal{S}$

**for** $i \leftarrow 1$ **to** $|A \cup B|$ **do**
  Draw a random target $x$ from $A \cup B$;
  Draw a sample $\mathcal{Q}$ of $Q$ agents from $(A \cup B) \backslash \{x\}$ (or $A \cup B$) using a weighted random sampling (with or without replacement), with weights $w_z = 1 - \beta$ for agents in the same faction as $x$ and $w_z = \beta$ for the rest of the agents;
  **if** $\mathcal{B}(x) = 1$ **then**
    **if** $|\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) = \mathcal{S}(x)\}| \geq q$ **then**
      **if** $\mathcal{S}(x) = 0$ **then**
        $\mathcal{S}(x) \leftarrow \mathcal{S}(x) + (-1)^b$, where $b \sim$ Bernoulli$(1/2)$;
      **else**
        $\mathcal{S}(x) \leftarrow 0$;
  **else**
    **if** $\exists s : |\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) = s\}| \geq q \wedge |s - \mathcal{S}(x)| = 1$ **then**
      $\mathcal{S}(x) \leftarrow s$;
**return** $\mathcal{S}$;

**Algorithm 5:** One Monte Carlo step of the dynamic model with anticonformity and conformity without BC. $A$ and $B$ are the factions, i.e. subsets of agents, $\mathcal{S} : (A \cup B) \to \{-1, 0, 1\}$ maps agents to their opinions, and $\mathcal{B} : (A \cup B) \to \{1, 0\}$ maps agents to their behavior (1 for anticonformity and 0 for conformity). Note that in the case of the static approach $p = |A \cup B|^{-1} \sum_{x \in A \cup B} \mathcal{B}(x)$.

**Input:** $Q > 1$, $q \in (\lfloor Q/2 \rfloor, Q]$, $A, B, \mathcal{S}$
**Output:** updated $\mathcal{S}$

**for** $i \leftarrow 1$ **to** $|A \cup B|$ **do**
  Draw a random target $x$ from $A \cup B$;
  Draw a sample $\mathcal{Q}$ of $Q$ agents from $(A \cup B) \backslash \{x\}$ (or $A \cup B$) using a weighted random sampling (with or without replacement), with weights $w_z = 1 - \beta$ for agents in the same faction as $x$ and $w_z = \beta$ for the rest of the agents;
  **if** $\mathcal{B}(x) = 1$ **then**
    **if** $|\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) = \mathcal{S}(x)\}| \geq q$ **then**
      **if** $\mathcal{S}(x) = 0$ **then**
        $\mathcal{S}(x) \leftarrow \mathcal{S}(x) + (-1)^b$, where $b \sim$ Bernoulli$(1/2)$;
      **else**
        $\mathcal{S}(x) \leftarrow 0$;
  **else**
    **if** $|\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) > \mathcal{S}(x)\}| \geq q = 1$ **then**
      $\mathcal{S}(x) \leftarrow \mathcal{S}(x) + 1$;
    **else if** $|\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) < \mathcal{S}(x)\}| \geq q = 1$ **then**
      $\mathcal{S}(x) \leftarrow \mathcal{S}(x) - 1$;
**return** $\mathcal{S}$;

**Algorithm 6:** One Monte Carlo step of the static model with conformity without BC and conformity with BC. $A$ and $B$ are the factions, i.e. subsets of agents, $\mathcal{S} : (A \cup B) \to \{-1, 0, 1\}$ maps agents to their opinions, and $\mathcal{B} : (A \cup B) \to \{1, 0\}$ maps agents to their behavior (1 for conformity without BC and 0 for conformity with BC). Note that in the case of the static approach $p = |A \cup B|^{-1} \sum_{x \in A \cup B} \mathcal{B}(x)$.

**Input:** $Q > 1$, $q \in (\lfloor Q/2 \rfloor, Q]$, $A, B, \mathcal{S}, \mathcal{B}$
**Output:** updated $\mathcal{S}$

**for** $i \leftarrow 1$ **to** $|A \cup B|$ **do**
  Draw a random target $x$ from $A \cup B$;
  Draw a sample $\mathcal{Q}$ of $Q$ agents from $(A \cup B) \backslash \{x\}$ (or $A \cup B$) using a weighted random sampling (with or without replacement), with weights $w_z = 1 - \beta$ for agents in the same faction as $x$ and $w_z = \beta$ for the rest of the agents;
  **if** $\mathcal{B}(x) = 1$ **then**
    **if** $|\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) > \mathcal{S}(x)\}| \geq q = 1$ **then**
      $\mathcal{S}(x) \leftarrow \mathcal{S}(x) + 1$;
    **else if** $|\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) < \mathcal{S}(x)\}| \geq q = 1$ **then**
      $\mathcal{S}(x) \leftarrow \mathcal{S}(x) - 1$;
  **else**
    **if** $\exists s : |\{z : z \in \mathcal{Q} \wedge \mathcal{S}(z) = s\}| \geq q \wedge |s - \mathcal{S}(x)| = 1$ **then**
      $\mathcal{S}(x) \leftarrow s$;
**return** $\mathcal{S}$;

## References


Axelrod, R., 1997. The dissemination of culture: A model with local convergence and global polarization. Journal of Conflict Resolution 41, 203–226. doi:10.1177/0022002797041002001.

Axelrod, R., Daymude, J.J., Forrest, S., 2021. Preventing extreme polarization of political attitudes. Proceedings of the National Academy of Sciences of the United States of America 118. doi:10.1073/pnas.2102139118.

Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Hunzaker, M.B.F., Lee, J., Mann, M., Merhout, F., Volfovsky, A., 2018. Exposure to opposing views on social media can increase political polarization. Proceedings of the National Academy of Sciences 115, 9216–9221. doi:10.1073/pnas.1804840115.

Baldassarri, D., Gelman, A., 2008. Partisans without constraint: Political polarization and trends in american public opinion. American Journal of Sociology 114, 408–446. doi:10.1086/590649.

Baldassarri, D., Page, S.E., 2021. The emergence and perils of polarization. Proceedings of the National Academy of Sciences 118, e2116863118. doi:10.1073/pnas.2116863118.

Balietti, S., Getoor, L., Goldstein, D.G., Watts, D.J., 2021. Reducing opinion polarization: Effects of exposure to similar people with differing political views. Proceedings of the National Academy of Sciences of the United States of America 118. doi:10.1073/pnas.2112552118.

Borges, H.M., Vasconcelos, V.V., Pinheiro, F.L., 2024. How social rewiring preferences bridge polarized communities. Chaos, Solitons & Fractals 180, 114594. doi:10.1016/j.chaos.2024.114594.

Castellano, C., Muñoz, M.A., Pastor-Satorras, R., 2009. Nonlinear $q$-voter model. Phys. Rev. E 80, 041129. doi:10.1103/PhysRevE.80.041129.

Centola, D., Macy, M., 2007. Complex contagions and the weakness of long ties. American Journal of Sociology 113, 702–734. doi:10.1086/521848.

Chen, T.H.Y., Salloum, A., Gronow, A., Ylä-Anttila, T., Kivelä, M., 2021. Polarization of climate politics results from partisan sorting: Evidence from finnish twittersphere. Global Environmental Change 71, 102348. doi:https://doi.org/10.1016/j.gloenvcha.2021.102348.

Combs, A., Tierney, G., Guay, B., Merhout, F., Bail, C.A., Hillygus, D.S., Volfovsky, A., 2023. Reducing political polarization in the united states with a mobile chat platform. Nature Human Behaviour 7, 1454–1461. doi:10.1038/s41562-023-01655-0.

Cota, W., Ferreira, S.C., Pastor-Satorras, R., Starnini, M., 2019. Quantifying echo chamber effects in information spreading over political communication networks. EPJ Data Science 8, 35. doi:10.1140/epjds/s13688-019-0213-9.

David Caldwell, G.C., Vivyan, N., 2024. Long-run trends in partisan polarization of climate policy-relevant attitudes across countries. Environmental Politics 0, 1–26. doi:10.1080/09644016.2024.2403957.

Debnath, R., van der Linden, S., Alvarez, R.M., Sovacool, B.K., 2023. Facilitating system-level behavioural climate action using computational social science. Nature Human Behaviour 7, 155–156. doi:10.1038/s41562-023-01527-7.

Deffuant, G., Neau, D., Amblard, F., Weisbuch, G., 2000. Mixing beliefs among interacting agents. Advances in Complex Systems 03, 87–98. doi:10.1142/S0219525900000078.

Desportes, J.P., Lemaine, J.M., 1988. The Sizes of Human Groups: An Analysis of their Distributions. Springer Netherlands, Dordrecht. pp. 57–65. doi:10.1007/978-94-009-2802-2_5.

Doornbosch, L.M., van Vuuren, M., de Jong, M.D., 2024. Moving beyond us-versus-them polarization towards constructive conversations. Democratization doi:10.1080/13510347.2024.2401127.

Druckman, J.N., Klar, S., Krupnikov, Y., Levendusky, M., Ryan, J.B., 2021. Affective polarization, local contexts and public opinion in america. Nature Human Behaviour 5, 28–38. doi:10.1038/s41562-020-01012-5.

Dvorak, F., Fischbacher, U., Schmelz, K., 2024. Strategic conformity or anti-conformity to avoid punishment and attract reward. The Economic Journal doi:10.1093/ej/ueae085.

Ebeling, R., Nobre, J., Becker, K., 2023. A multi-dimensional framework to analyze group behavior based on political polarization. Expert Systems with Applications 233, 120768. doi:10.1016/j.eswa.2023.120768.

Falkenberg, M., Galeazzi, A., Torricelli, M., Di Marco, N., Larosa, F., Sas, M., Mekacher, A., Pearce, W., Zollo, F., Quattrociocchi, W., Baronchelli, A., 2022. Growing polarization around climate change on social media. Nature Climate Change 12, 1114–1121. doi:10.1038/s41558-022-01527-x.

Falkenberg, M., Zollo, F., Quattrociocchi, W., Pfeffer, J., Baronchelli, A., 2024. Patterns of partisan toxicity and engagement reveal the common structure of online political communication across countries. Nature Communications 15, 9560. doi:10.1038/s41467-024-53868-0.

Fasching, N., Iyengar, S., Lelkes, Y., Westwood, S.J., 2024. Persistent polarization: The unexpected durability of political animosity around us elections. Science Advances 10, eadm9198. doi:10.1126/sciadv.adm9198.

Fay, N., Garrod, S., Carletta, J., 2000. Group discussion as interactive dialogue or as serial monologue: The influence of group size. Psychological Science 11, 481 – 486. doi:10.1111/1467-9280.00292.

Flores, A., Cole, J.C., Dickert, S., Eom, K., Jiga-Boy, G.M., Kogut, T., Loria, R., Mayorga, M., Pedersen, E.J., Pereira, B., Rubaltelli, E., Sherman, D.K., Slovic, P., Vastfjall, D., Van Boven, L., 2022. Politicians polarize and experts depolarize public support for COVID-19 management policies across countries. Proceedings of the National Academy of Sciences of the United States of America 119. doi:10.1073/pnas.2117543119.

Fraxanet, E., Pellert, M., Schweighofer, S., Gómez, V., Garcia, D., 2024. Unpacking polarization: Antagonism and alignment in signed networks of online interaction. PNAS Nexus , pgae276doi:10.1093/pnasnexus/pgae276.

Glynn, C.J., Herbst, S., Lindeman, M., O'Keefe, G.J., Shapiro, R.Y., 2018. Public Opinion, 3rd Edition. Routledge.

Goldstein, N.J., Cialdini, R.B., Griskevicius, V., 2008. A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. Journal of Consumer Research 35, 472–482. doi:10.1086/586910.

Gugushvili, D., 2024. Political polarization in the willingness to make sacrifices for the environment: a cross-national analysis. Environmental Sociology doi:10.1080/23251042.2024.2400421.

Górski, P.J., Atkisson, C., Hołyst, J.A., 2023. A general model for how attributes can reduce polarization in social groups. Network Science 11, 536 – 559. doi:10.1017/nws.2023.13.

Hegselmann, R., Krause, U., 2002. Opinion dynamics and bounded confidence: models, analysis and simulation. Journal of Artificial Societies and Social Simulation 5.

Hewstone, M., Rubin, M., Willis, H., 2002. Intergroup bias. Annual Review of Psychology 53, 575 – 604. doi:10.1146/annurev.psych.53.100901.135109.

Hohmann, M., Devriendt, K., Coscia, M., 2023. Quantifying ideological polarization on a network using generalized euclidean distance. Science Advances 9, eabq2044. doi:10.1126/sciadv.abq2044.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., Westwood, S.J., 2019. The Origins and Consequences of Affective Polarization in the United States. Annual Review of Political Science 22, 129–146. doi:10.1146/annurev-polisci-051117-073034.

de Jong, J.F., 2024. Cross-partisan discussions reduced political polarization between uk voters, but less so when they disagreed. Communications Psychology 2. doi:10.1038/s44271-023-00051-8.

Judge, M., Kashima, Y., Steg, L., Dietz, T., 2023. Environmental decision-making in times of polarization. Annual Review of Environment and Resources 48. doi:10.1146/annurev-environ-112321-115339.

Kempe, D., Kleinberg, J., Tardos, E., 2015. Maximizing the spread of influence through a social network. Theory of Computing 11, 105–147. doi:10.4086/toc.2015.v011a004.

Kozlowski, A.C., Murphy, J.P., 2021. Issue alignment and partisanship in the american public: Revisiting the 'partisans without constraint' thesis. Social Science Research 94, 102498. doi:10.1016/j.ssresearch.2020.102498.

Lipiecki, A., Sznajd-Weron, K., 2022. Polarization in the three-state q-voter model with anticonformity and bounded confidence. Chaos, Solitons & Fractals 165, 112809. doi:10.1016/j.chaos.2022.112809.

Liu, S., Mäs, M., Xia, H., Flache, A., 2022. When intuition fails: The complex effects of assimilative and repulsive influence on opinion polarization. Advances in Complex Systems 25. doi:10.1142/S0219525922500114.

MacDonald, G., Nail, P.R., Harper, J.R., 2011. Do people use reverse psychology? an exploration of strategic self-anticonformity. Social Influence 6, 1–14. doi:10.1080/15534510.2010.517282.

Macy, M.W., Ma, M., Tabin, D.R., Gao, J., Szymanski, B.K., 2021. Polarization and tipping points. Proceedings of the National Academy of Sciences of the United States of America 118. doi:10.1073/pnas.2102144118.

Magdaci, O., Matalon, Y., Yamin, D., 2022. Modeling the debate dynamics of political communication in social media networks. Expert Systems with Applications 206, 117782. doi:10.1016/j.eswa.2022.117782.

Mason, L., 2018. Ideologues without issues: The polarizing consequences of ideological identities. Public Opinion Quarterly 82, 866–887. doi:10.1093/poq/nfy005.

McCoy, J., Rahman, T., Somer, M., 2018. Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. American Behavioral Scientist 62, 16–42. doi:10.1177/0002764218759576.

Morales, A.J., Borondo, J., Losada, J.C., Benito, R.M., 2015. Measuring political polarization: Twitter shows the two sides of Venezuela. Chaos: An Interdisciplinary Journal of Nonlinear Science 25, 033114. doi:10.1063/1.4913758.

Mäs, M., Flache, A., 2013. Differentiation without distancing. explaining bi-polarization of opinions without negative influence. PLOS ONE 8, 1–17. doi:10.1371/journal.pone.0074516.

Nail, P.R., Di Domenico, S.I., MacDonald, G., 2013. Proposal of a double diamond model of social response. Review of General Psychology 17, 1–19. doi:10.1037/a0030997.

Nyczka, P., Byrka, K., Nail, P., Sznajd-Weron, K., 2018. Conformity in numbers—does criticality in social responses exist? PLoS ONE 13. doi:10.1371/journal.pone.0209620.

Ojer, J., Starnini, M., Pastor-Satorras, R., 2023. Modeling explosive opinion depolarization in interdependent topics. Physical Review Letters 130. doi:10.1103/PhysRevLett.130.207401.

Park, J., Warner, B.R., 2024. Does collectivism make a difference? a comparative test of common ingroup strategies to reduce affective polarization. International Journal of Public Opinion Research 36. doi:10.1093/ijpor/edae033.

Santoro, E., Broockman, D.E., 2022. The promise and pitfalls of cross-partisan conversations for reducing affective polarization: Evidence from randomized experiments. Science Advances 8, eabn5515. doi:10.1126/sciadv.abn5515.

Schweighofer, S., Garcia, D., 2024. Raising the spectrum of polarization: Generating issue alignment with a weighted balance opinion dynamics model. Journal of Artificial Societies and Social Simulation 27, 15. doi:10.18564/jasss.5323.

Siedlecki, P., Szwabiński, J., Weron, T., 2016. The interplay between conformity and anticonformity and its polarizing effect on society. JASSS 19. doi:10.18564/jasss.3203.

Smith, L., Thomas, E., Bliuc, A., et al., 2024. Polarization is the psychological foundation of collective engagement. Communications Psychology 2, 41. doi:10.1038/s44271-024-00089-2.

Van Der Maas, H., Dalege, J., Waldorp, L., 2020. The polarization within and across individuals: The hierarchical Ising opinion model. J. Complex Netw. 8. doi:10.1093/comnet/cnaa010.

Vasconcelos, V.V., Levin, S.A., Pinheiro, F.L., 2019. Consensus and polarization in competing complex contagion processes. Journal of the Royal Society Interface 16. doi:10.1098/rsif.2019.0196.

Vieira, A., Anteneodo, C., 2018. Threshold q -voter model. Physical Review E 97. doi:10.1103/PhysRevE.97.052106.

Vieira, A., Peralta, A., Toral, R., Miguel, M., Anteneodo, C., 2020. Pair approximation for the noisy threshold q-voter model. Physical Review E 101. doi:10.1103/PhysRevE.101.052131.

Wojcieszak, M., Warner, B.R., 2020. Can interparty contact reduce affective polarization? a systematic test of different forms of intergroup contact. Political Communication 37, 789 – 811. doi:10.1080/10584609.2020.1760406.

Ye, D., Lin, H., Jiang, H., Du, L., Li, H., Chen, Q., Wang, Y., Yuan, L., 2025. Simplex bounded confidence model for opinion fusion and evolution in higher-order interaction. Expert Systems with Applications 272, 126551. doi:10.1016/j.eswa.2025.126551.