The diversity puzzle

Mäs, Michael

III. The Polarizing Effects of Argument Exchange[13]

Abstract
In the previous chapter, we analyzed a social-influence model that includes the assumption of negative influence and demonstrated that this model can generate opinion polarization. However, in the following we will show that empirical research provided mixed evidence for negative influence. For this reason, we will develop in this chapter an alternative theory of opinion polarization.
Our new theory explains polarization (denoted bi-polarization here) without the negative-influence assumption. This approach assumes that opinions are based on arguments and that these arguments are exchanged during interaction. When individuals with similar opinions interact, they likely provide each other with new arguments that support their opinions. In this way, opinions become more extreme. We will propose that in combination with homophilious selection of interaction partners, the exchange of arguments can lead to polarization of opinions.
We will proceed in two steps. First, we will develop a formal model of opinion dynamics that is based on argument exchange and demonstrate that this model can indeed generate polarization. A computational experiment will reveal that *strong* homophily is a crucial condition of polarization in this model.
Second, we have put the new theory to the test. In a laboratory experiment, we tested our central proposition that argument exchange can cause polarization. Groups of 8 participants discussed an issue and could either exchange arguments or only opinions. We found polarization, but as predicted only when arguments were transmitted and when there was strong homophily in the selection of interaction partners.

III.1. Introduction

Sociological and socio-psychological theories of intergroup processes, like social differentiation (Bourdieu 1984[1979]; Mark 2003), ingroup favoritism (Tajfel 1981), outgroup discrimination (Mummendey et al. 1999), and intergroup conflict (Sherif 1966; Tajfel and Turner 1986) build on the assumption that individuals seek to accentuate differences between their own group and salient outgroups. Sociological approaches, for

[13] This chapter is co-authored with Andreas Flache and is submitted for publication in a social psychological journal. Note that we adjusted the terminology. In particular, we refer to the development of clusters with increasingly distant opinions as 'bi-polarization' and not 'polarization'. This was necessary because the term 'polarization' describes collective extremization tendencies in the social psychological literature.

instance, argue that people develop elaborated cultural norms in order to distinguish themselves from groups with a lower status (Bourdieu 1984[1979]; Bryson 1996; Elias 1969[1939]; Simmel 1957; Turner 1995). In the same line of reasoning, psychological theories building on the self-categorization paradigm (Brewer 1991; Tajfel and Turner 1986; Turner 1987) hold that humans adjust their opinions and behavior in a way to minimize the heterogeneity within their ingroups and to maximize differences to outgroups (Hogg, Turner and Davidson 1990).

The notion that humans seek to accentuate intergroup differences has recently been incorporated into theories of opinion dynamics (Baldassarri and Bearman 2007; Salzarulo 2006). In particular, it has been assumed that in an intergroup context, individuals may adapt their opinions in order to maximize disagreement with a perceived outgroup opinion. With two groups perceiving each other as outgroup, this suggests *bi-polarization*, the development of a bimodal opinion distribution in the course of social interaction, with gradually increasing distance between the opposite modes. In fact, empirical studies of opinion dynamics have provided some evidence of bi-polarization tendencies with regard to salient opinions, for example among college students (Feldman and Newcomb 1969) or in ethnically mixed work teams (Early and Mosakowski 2000). In a similar vein, observers of the dynamics of political opinions found tendencies towards bi-polarization on controversial issues in the American public during election periods (Abramowitz and Saunders 2008; Evans 2003; Fiorina and Abrams 2008; Levendusky 2009). Existing sociological and socio-psychological theories of bi-polarization explicitly assume what we denote here "negative influence", the tendency of individuals to adjust their opinions in a way to increase opinion differences to dissimilar others (Baldassarri and Bearman 2007; Hogg, Turner and Davidson 1990; Macy et al. 2003; Salzarulo 2006). Yet, as will be elaborated in more detail further below, empirical research on opinion dynamics provided mixed evidence for negative influence and has been criticized on methodological grounds (e.g. Krizan and Baron 2007).

This raises the question how bi-polarization might be explained without the assumption that individuals seek to increase opinion differences to dissimilar others. In this paper, we present and test a theory of bi-polarization that does not rely on negative influence. Moreover, our theory does also not need to assume that bi-polarization results from the interaction between two distinct groups who regard each other as salient outgroup. Instead, we model bi-polarization as consequence of an *intra*-group process.

Our theory has two main ingredients. First, we draw on Persuasive Argument Theory (PAT) (Myers 1982; Vinokur and Burnstein 1978). PAT holds that individuals base their opinions on arguments and influence each others' opinions when they exchange arguments. When individuals with similar opinions exchange arguments, they may provide each other with new arguments that support their opinions. As a consequence, their opinions may be intensified and become more extreme. We combine PAT with the assumption of *homophily* (Lazarsfeld and Merton 1954) (Ibarra 1992; Lazarsfeld and Merton 1954; McPherson, Smith-Lovin and Cook 2001; Moody 2001), or more specifically the notion that individuals tend to interact with others who hold similar opinions. Sociological research has shown that homophily, or the tendency of "birds of a feather to flock together" is a robust empirical regularity with regard to similarity in a wide range of characteristics, including opinion similarity (Byrne 1971; McPherson, Smith-Lovin and Cook 2001).

In the theory section of this paper we present an informal reasoning that describes how the interplay of persuasive argument exchange and homophily may give rise to bi-polarization. However, without precise modeling it is hard to gain solid intuitions about the social outcomes ensuing from simultaneous interactions of multiple individual group members driven by these two principles. Recently, an increasing number of theorists advocate employing agent-based computational modeling (Bonabeau 2002; Macy and Willer 2002; Smith and Conrey 2007) to understand the "large-scale consequences of the theoretical assumptions about individual behavior when the behaviors are carried out in the context of many other agents and iterated dynamically over an extended period of time" (Smith and Conrey 2007: 88). To assure the logical consistency of our reasoning, we therefore elaborate a computational agent-based model that shows how and under what conditions the interplay of persuasive argument exchange and homophily can generate bi-polarization.

Furthermore, we put the theory to a test. We conducted a laboratory experiment (N=96) with a controlled group discussion process in groups of 8 participants. The focus was on testing how bi-polarization is affected by the two key mechanisms, argument exchange and homophily. For this, we manipulated independently the possibility for participants to exchange the arguments on which they build their opinions (rather than, or in addition to being only exposed to each others' opinions) and homophily, operationalized as the extent to which interaction was restricted to pairs of participants with similar opinions. We carefully avoided that participants received information on which to base

social categorizations of others into in- our outgroup members. This precluded bi-polarization driven by the mechanism of negative influence. We hypothesized that bi-polarization would nevertheless occur, but only in the experimental condition in which participants could exchange the arguments underlying their opinions, and in which homophily was imposed in the matching of interaction partners.

III.2. The critical role of negative influence

Recent contributions to the literature on social-influence dynamics (for recent review see Mason, Conrey and Smith 2007) demonstrate the critical role of the negative-influence assumption in theories of bi-polarization. Early formal models failed to generate bi-polarization (Abelson 1964; Berger 1981; French 1956; Harary 1959; Wagner 1982) because only positive influence was assumed. Instead, these models imply that a group inevitably ends up in consensus as long as there are no subgroups that are entirely cut off from outside influences. In search for processes that give rise to bi-polarization, an increasing number of models have therefore been proposed that combine both positive influence from similar and negative influence from dissimilar sources (Baldassarri and Bearman 2007; Flache and Mäs 2008a; Macy et al. 2003; Mark 2003; Salzarulo 2006).

These theoretical accounts of bi-polarization hinge critically upon the assumption of negative influence. However, experimental tests have hitherto not provided unequivocal evidence in support of this assumption. In laboratory experiments, researchers typically informed participants about the opinions of fictitious members of both ingroup and outgroup and then measured pre-test–post-test opinion shifts. These studies have led to very mixed results. Many did not find increasing differences between in- and outgroup opinions at all (Hogg, Turner and Davidson 1990; Krizan and Baron 2007; Lemaine 1975). In addition, research illustrates that individuals may publicly distance themselves from others but their private opinions actually do not shift (Berger and Heath 2008).

Moreover, methodological issues cast doubt on the conclusiveness of those studies that researchers interpreted as support for negative influence (Berscheid 1966; Hogg, Turner and Davidson 1990; Mazen and Leventhal 1972; Sampson and Insko 1964; Schwartz and Ames 1977; van Knippenberg and Wilke 1988; van Knippenberg, De Vries and van Knippenberg 1990). Krizan and Baron (2007) raised a number of issues with regard to experiments in the group polarization tradition. We point here to two major additional problems. First, some experimental designs do not allow to disentangle positive

influence from the ingroup and negative influence from the outgroup in the explanation of opinion shifts (e.g. Hogg, Turner and Davidson 1990; van Knippenberg and Wilke 1988; van Knippenberg, De Vries and van Knippenberg 1990). In these studies, participants have been exposed two sources of social influence, ingroup members and outgroup members. Participants were exposed to ingroup members who held opinions relatively similar to their own. Some of these ingroup members held more extreme opinions than the participant. Outgroup members always had opinions distinct from those of the participants. With such a design, opinion changes away from the outgroup opinion may have been caused by both negative influence from the outgroup or positive influence from more extreme ingroup members (Mackie 1986).

The second problem is that some studies did not control for general opinion drifts during the experiment (Mazen and Leventhal 1972; Sampson and Insko 1964). For example, Mazen and Leventhal (1972) confronted expectant mothers with a favorable description of breast feeding and measured how this affected the mothers' opinions on this issue. They found that mothers developed more positive opinions when they received information from a communicator with a similar skin color (positive influence). But, when the communicator and the mother were dissimilar in skin color the opinions of the mothers turned more negative. This suggests that these mothers were influenced negatively by the communicator. However, we argue that this result may have been caused by a general trend towards more negative opinions. In this study, the second opinion measurement took place one week after the first. In this period, participants might have developed more negative opinions. However, those mothers who where similar to the communicator were positively influenced by them and changed their minds back to more positive opinions. The opinions of the dissimilar mothers, however, might have been unaffected by the communicator's information and remained more negative. Unfortunately, the authors did not control for trend effects in their analyses. It is therefore not clear whether the reported opinion dynamics are the result of negative influence or opinion drifts.

In sum, existing theories of bi-polarization critically hinge on the assumption that individuals tend to adjust their opinions in a way to increase opinion differences to dissimilar others. However, there is hitherto no conclusive empirical evidence for negative influence. In the following section, we elaborate a theory of bi-polarization which does not rely on negative influence.

III.3. Theory

Our theory of bi-polarization builds on earlier theorizing on demographic faultlines (Lau and Murnighan 1998) and group polarization (Myers 1982; Myers and Bishop 1970). These approaches already combined insights from PAT (Isenberg 1986; Vinokur and Burnstein 1978) and research on homophily (Ibarra 1992; Lazarsfeld and Merton 1954; McPherson, Smith-Lovin and Cook 2001; Moody 2001).

PAT has been developed to explain why discussion groups which initially have a tendency towards one side of an issue will become more extreme in their opinions as a result of discussion. This phenomenon, called *polarization*, has been robustly demonstrated by a range of experimental studies (Isenberg 1986; Myers 1982). PAT assumes that individuals base their opinions on pro and con arguments. During discussion, individuals are exposed to the arguments their interaction partners consider relevant. In groups where members tend towards a specific opinion already prior to discussion, mainly those arguments will be brought up that favor the prevailing tendency. Discussion members, thus, provide each other with further arguments that support their initial position. This intensifies opinions and aggregates to a *collective* opinion shift towards more extreme positions.

Building on earlier work (Lau and Murnighan 1998; Myers 1982; Myers and Bishop 1970) we argue that the interplay of the persuasive argument exchange described by PAT with homophily can give rise to bi-polarization. The idea is that small initial opinion differences in a group are gradually amplified when argument exchange occurs more frequent between those individuals who initially have relatively similar opinions than between those whose opinions are relatively dissimilar. Due to homophily, individuals with opinions leaning towards the same pole of the opinion spectrum interact more likely with each other than with those who lean towards the opposite pole. Thus, persuasive argument exchange reinforces existing opinion tendencies, but in opposing directions in the separate subsets of group members who share the same initial tendency. This further reduces the likelihood of interaction between initially dissimilar pairs of individuals, which in turn further strengthens existing tendencies. This process unfolds simultaneously at both sides of the opinion spectrum, such that a self-reinforcing dynamic may arise that entails bi-polarization even in the absence of negative influence.

Bi-polarization requires homophily according to this reasoning. However, it remains unclear how strong homophily needs to be to render bi-polarization a likely outcome of the

dynamic. As long as there is some probability of interaction also between actors with dissimilar opinions, bi-polarization tendencies might be very unlikely. When actors with dissimilar opinions interact, they likely exchange arguments that speak against their current tendency and lead to more moderate opinions. Furthermore, in subsequent interaction actors will transmit these counter arguments to similar others. This will lead to further opinion convergence. In sum, this reasoning suggests that even though actors may tend to interact with similar others, occasional deviations from this rule may suffice to impede the bi-polarization tendencies of argument exchanged and homophily. In a non-deterministic world, there is no guarantee that a self-reinforcing dynamic eventually leads a social system into the state towards which the dynamic moves. This has for example been demonstrated for formal stochastic models of residential segregation (e.g. Stauffer and Solomon 2007), or cultural dissemination (e.g. Klemm et al. 2003a). In these models, the "ordered" outcomes towards which individual decision rules drive the system, such as highly segregated residential distributions, or local clustering of similar cultures, only arise when the level of randomness in individual decision making is relatively small.

To identify how strong homophily needs to be to give rise to bi-polarization, we developed and applied an agent-based computational model of our theory. In the following section, we present the model. Subsequently, we report results from a computer simulation experiment designed to assess the relationship between the strength of homophily and bi-polarization.

III.3.1. The formal model

The agent-based model implements the substantive assumptions of PAT and homophily for each of N interdependent individuals who simultaneously participate in an artificial influence process. Each individual is represented as an agent i, with a numerically valued opinion o_i ($-1 \leq o_i \leq +1$) that represents the agent's stance on a given issue. We assume that there is a limited number of arguments that address the issue. The valence of an argument is expressed numerically. More precisely, P pro arguments ($a_l = 1$) and C con arguments ($a_l = -1$) are available. This is summarized in the argument vector, an array of arguments with $P+C$ elements. Elements with a row number smaller than $P+1$ hold pro arguments, i.e. $a_l = +1$. The remaining elements contain con arguments, i.e. $a_l = -1$.

Empirical research suggests that people have limited capacities to remember and process information (Cowan 2001; Miller 1956). Accordingly, we assume that agents base their opinion only on a subset of S relevant arguments ($S \leq P+C$). The remaining

arguments are not relevant in the opinion formation. Technically, an agent's opinion is the average value of the arguments a_l that the agent considers relevant (see equation 1). For simplicity, we assume that all relevant arguments have the same persuasiveness. Technically, this is expressed by the assumption that all relevant arguments are equally weighted in the calculation of the opinion.

$$o_i = \frac{1}{S}\sum_{l=1}^{S} a_l \qquad (1)$$

For example, an agent i that bases her opinion on 6 pro arguments (S=6) holds an opinion of o_i=1. However, if the agent considers e.g. 3 pro and 3 con arguments relevant, the opinion will take the value zero.

Following research on memory processes (Brown and Chater 2001), we assume that agent's disregard pieces of information if they are not sufficiently recent. Thus, the more recent an argument is at a given point in time, the longer this argument remains relevant for the formation of the agent's opinion. This is implemented for each agent in a recency vector. This vector has P+C elements. Each element indicates how recent the respective argument is for the agent. Elements of the recency vector with a row number smaller than P+1 identify the relevance of pro arguments. The remaining elements determine the relevance of con arguments. Arguments are either relevant or not, but agents rank the S relevant arguments according to their recency. We denote the recency of an argument ($s_{l,i}$) with integer values between 0 and S ($s_{l,i} \in \{0,...S\}$). A value of $s_{l,i}$ = 0 indicates that the argument a_l is *not* sufficiently recent and therefore *not* relevant for actor i. Values above zero indicate that this argument *is* sufficiently recent and therefore affects actor i's opinion. The most recent argument has the value of $s_{l,i} = S$, the second most recent argument has the value S-*1*, and so on. Thus, if an agent considers three arguments (S=3) then one has a recency of 1, one has a recency of 2, and one has a recency of 3. The recency rank of an argument does *not* affect the extent to which an argument shapes the current opinion (see equation 1). However, the recency determines *how long* an argument affects the agent's opinion in the influence process. The exact rules for updating argument-recency will be elaborated further below.

We model the opinion formation process as a sequence of events, each event corresponding to one interaction between two agents. An interaction consists of a partner selection phase and a subsequent social influence phase. In the partner selection phase, two agents from the population are matched for interaction, based on opinion-homophily.

Subsequently, an opinion of one of the interacting agents is updated as a result of the interaction. The updating rule operationalizes the argument exchange mechanism of PAT.

We implement the *partner selection* phase as follows. In each event, the computer first randomly picks an agent i^*. Then an interaction partner j ($j \neq i^*$) is selected. The probability that agent j is chosen as interaction partner depends on the similarity between i^* and j, $sim_{i^*,j}$, that varies between 0 and 1. A similarity of zero expresses maximal dissimilarity, whereas $sim_{i^*,j} = 1$ if both actors hold exactly the same opinion. Formally,

$$sim_{i^*,j} = \frac{1}{2}\left(2 - \left|o_{i^*} - o_j\right|\right) \tag{2}$$

The probability that agent i^* chooses j as interaction partner (p_j) derives from their relative similarity, that is: the degree to which j is more similar to i than other group members are. Technically,

$$p_j = \frac{\left(sim_{i^*,j}\right)^h}{\sum_{j=1, j\neq i^*}^{N}\left(sim_{i^*,j}\right)^h} \tag{3}$$

Equation 3 implements homophily. The more similar j is to i^* the higher is the probability that they will interact. If two actors differ maximally then the probability of interaction equals zero. To vary the *strength of homophily* we include the parameter h into the model. The higher the value of h, the steeper is the increase of the likelihood that j will be chosen by i^* as an interaction partner in the relative similarity of i^* and j. The actual selection of the interaction partner of i^* is implemented by a random draw of one agent from the set of all other group members, based on the probabilities p_j given by (3).

Next, i^* is socially influenced by the selected interaction partner j^* based on the persuasive arguments mechanism. For this, the computer randomly picks one argument, a_{l^*}, out of the S arguments that j^* considers relevant. Each relevant argument has the same probability to be chosen ($1/S$). Arguments that are not relevant for j^* are not chosen. The chosen argument is then adopted by i^* . Technically, its recency for i^* is updated to a value of $S+1$ ($s_{l^*,i^*} = S+1$). Subsequently, the recency of all arguments that have non-zero recency in i^*'s recency vector is reduced by one, if prior to the interaction the corresponding argument was more recent for i^* than the argument adopted from j^*. As a result, the argument that was communicated by j^* becomes relevant for i^* and attains the highest recency of all argument that i^* considers relevant ($s_{l^*,i^*} = S$).

This updating procedure implements the assumptions about agents' limited capacity to memorize information (Cowan 2001; Miller 1956) and their bias towards considering only recent information (Brown and Chater 2001). It implies in particular that agents "forget" one of the arguments previously relevant for them, if they have learned a new argument in the interaction. This assures that the number of arguments that is relevant for an agent is kept constant at S throughout the influence process.

Interaction events are iterated until the system reaches equilibrium. Our model has exactly two equilibria, perfect consensus and maximal bi-polarization. Perfect consensus is reached when all agents hold the same opinion and base it on the same set of arguments. Perfect consensus is a stable situation because agents can transmit only arguments which their interaction partners already consider relevant. This implies that opinions will not be affected by argument exchange. Maximal bi-polarization obtains if there are two maximally distinct subgroups and the members of each subgroup agree on opinions and arguments with each other. That is, the members of the subgroups have coordinated on the opposite poles of the opinion scale and the pairwise similarity ($sim_{i,j}$) between agents of different subgroups is zero. In this situation, the probability is zero that agents interact who belong to different subgroups (see equation 3). Argument exchange between the subgroups is thus precluded. In addition, interaction of agents that belong to the same subgroup can not lead to opinion changes because these agents base their opinion on either exclusively pro-arguments or exclusively con- arguments. Any outcome of the process that is not perfect consensus or perfect bi-polarization can not be an equilibrium. The reason is that any other outcome implies that there are differences in opinions or arguments between agents, and a positive probability of interaction between the agents who hold different opinion or arguments. There is thus a positive probability that the distribution of arguments and opinions in the population will change due to interaction.

III.3.2. Dynamics of Bi-polarization: an illustrative simulation run

We began by testing whether the model can generate bi-polarization. For this, we imposed conditions for which we expected bi-polarization tendencies to be very strong. Accordingly, we imposed relatively strong homophily, assuming h=9. With this value, homophily is so strong that interaction between agents who do not hold perfectly similar opinions is extremely unlikely. Furthermore, we assumed that thirty pro and con arguments are available (P=C=30) and all agents consider 10 relevant arguments at the same time

(S=10). For this condition, we simulated a population of 100 agents and studied the change of agents' opinions and argument vectors over 30,000 simulation events.

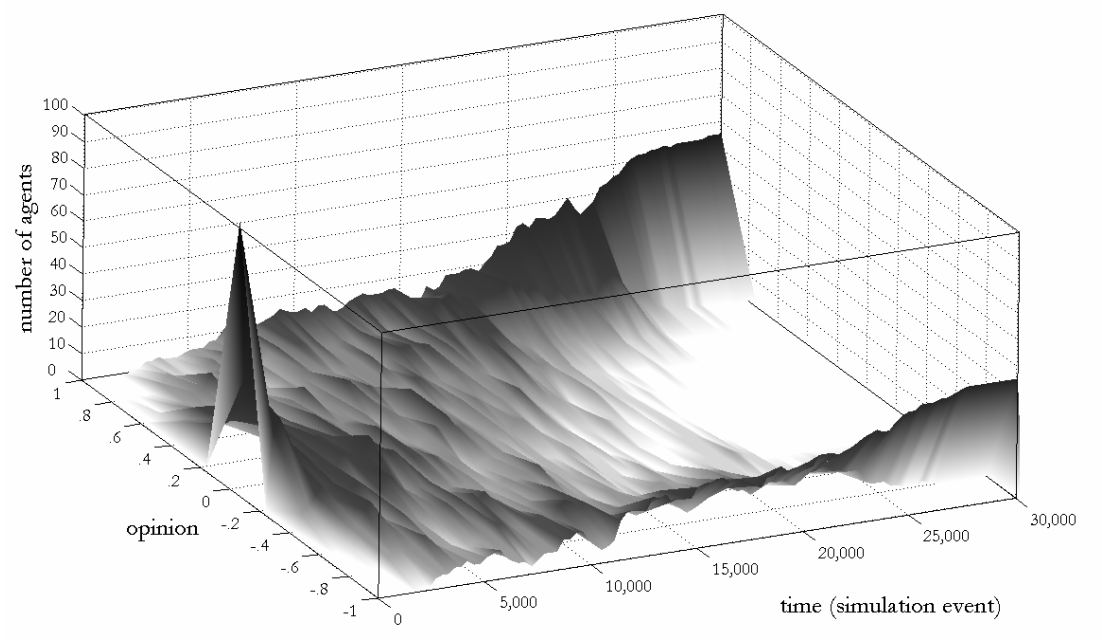
The initial distribution of arguments and opinions was created by assigning to each agent a random set of 5 pro and 5 con arguments. With this, all agents started with the same opinion at the middle of the opinion scale (o_i=0 for all i). Thus, at the outset there are no opinion differences between agents. Notice that this does not rule out opinion changes, as there are still differences between agents in the arguments on which their opinion is based.

Figure III.1 shows a surface which depicts the development of the opinion distribution during the typical simulation run. The shading of the surface's areas and the respective value on the z-axis indicate how many agents hold a certain opinion at a certain simulation event. White areas indicate that no agent holds the respective opinion. The darker the area, the more agents hold this opinion. At the beginning of the simulation (event zero), all 100 agents hold the same opinion. The figure shows how bi-polarization emerges in this simulation run. While opinions are approximately uniformly distributed after about 10,000 simulation events, the distribution becomes bimodal after about 15,000 events. Subsequently, the two modes gradually become more accentuated and shift towards the opposite ends of the opinion spectrum until, after about 30,000 events, the population is almost entirely split into two approximately equally large subsets of agents with opinions of -1 and +1, respectively.

Opinion change is possible despite initial uniformity, because agents base their opinion on different (randomly assigned) sets of five pro and five con arguments. Thus, in some interactions agents' opinion shifts away from the initial consensus, because they learn a new pro (con) argument and forget a con (pro) argument. Their new opinion is then based on more pro (con) than con (pro) arguments and takes a positive (negative) value. Figure III.1 shows that this results in an increase of the variance of the opinion distribution in the first phase of the simulation run. After about 10,000 simulation events, the opinion is uniformly distributed. Due to the strong homophily, agents are matched with interaction partners that have adjusted their opinion in the same direction. These interaction partners will more likely provide each other with arguments that further intensify their opinion tendency rather than to communicate arguments that render their opinions more moderate again. Eventually the opinion trajectories of all agents move to one of the two outer ends of the opinion scale. At this point, the opinion distribution stabilizes, because agents base their opinions on either only pro or only con arguments such that interaction is only

possible between agents who already hold identical opinions. Agents can no longer learn arguments that could change their opinions.

Figure III.1: Bi-polarization generated by argument exchange and homophily (N=100, P=C=30, S=10, h=9)



To summarize, this illustrative simulation run confirms that the interplay of argument exchange and homophily can generate bi-polarization even in the absence of negative influence. What is more, in this run bi-polarization emerged even though we assumed perfect opinion consensus at the outset. In sharp contrast, existing models of continuous influence dynamics (Baldassarri and Bearman 2007; Flache and Mäs 2008a; Macy et al. 2003; Salzarulo 2006) imply that bi-polarization can only arise when there are initially differences between members of a population which can form the basis of group categorization and negative influence.

III.3.3. Effects of Homophily

Next, we wanted to know whether homophily always entails bi-polarization, or whether bi-polarization can only arise when homophily is sufficiently strong. We conducted a simulation experiment in which we varied the model parameter h between 0 (no homophily) and 8 (strong homophily) in steps of 1. Per condition, we ran 500 independent replications of the simulation. In all simulations of this experiment, we studied populations of 20 agents (N=20). This is a plausible group size for school classes and work

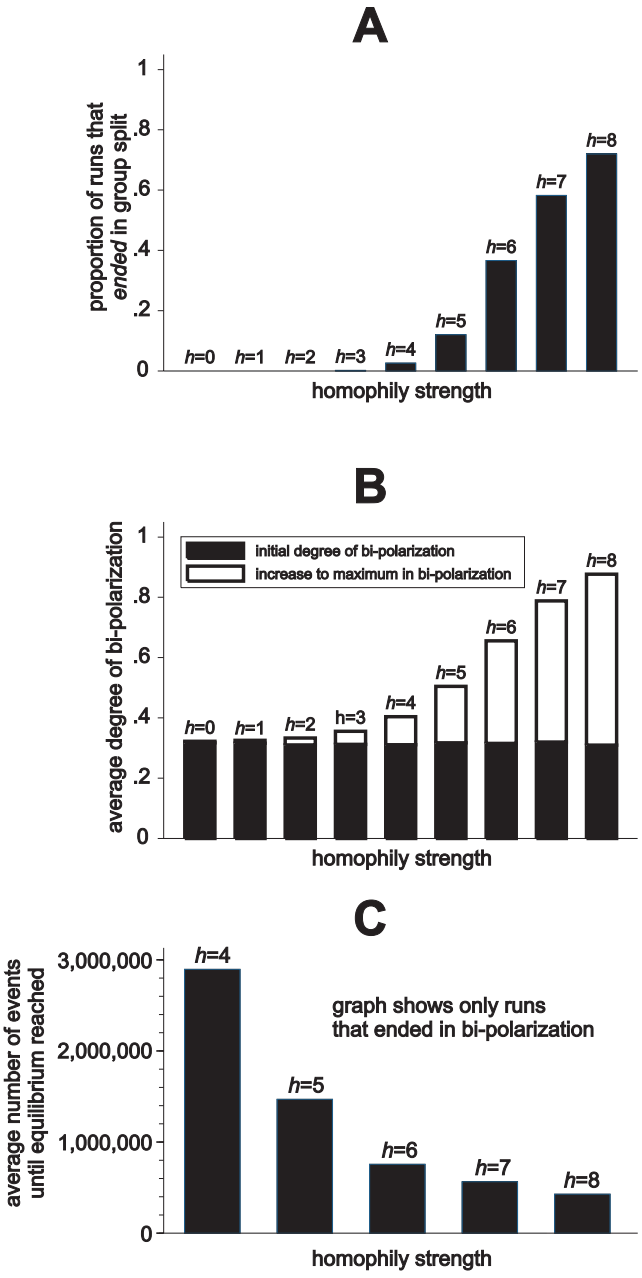
teams (see e.g. Wegge et al. 2008), two of the settings for which theory and empirical accounts of intra-group conflicts suggest the possibility of bi-polarization dynamics (see e.g. Lau and Murnighan 1998). We assume that there are 20 pro and 20 con arguments (P=C=20) available, and that agents can only take 6 arguments into account simultaneously (S=6). Values for P, C and S were selected to create sufficient variation in the initial argument sets also between agents who hold identical opinions. For this, P and C should considerably exceed S. Otherwise agents with similar opinions very likely base their opinions on similar sets of arguments. This would preclude the possibility that argument exchange between agents with similar opinions renders their opinions more extreme because they provide each other with arguments which they already consider relevant. Furthermore, we created the initial condition such that opinions are uniformly distributed. For this, we randomly assigned to each agent one of the S+1 possible opinion values and then randomly picked one of the possible sets of S arguments which correspond to the selected opinion value.

Figure III.2 summarizes the results. Panel A of figure III.2 shows how homophily strength h affected the proportion of runs that ended in bi-polarization. When homophily strength was below h=3 all runs ended in consensus. At h=3, only one out of the 500 replications for this condition ended in a group split with two subgroups at the opposing poles. For higher values of homophily strength h, panel A shows that the stronger homophily was, the more runs ended in a perfect group split.

If a simulation runs ends in perfect consensus, there may nonetheless have been a temporary period of significant bi-polarization in the dynamic (see chapter IV). To test for this possibility, we assessed for each simulation run the degree of bi-polarization at the outset and the maximal degree of bi-polarization that occurred during the simulation. Following Flache and Mäs (2008a; 2008b), the degree of bi-polarization was measured with the standard deviation of the distribution of pairwise opinion distances between all pairs of agents in the population. This measure takes its maximal value (1) when there are two equally large and maximally different subgroups and. The minimal value of the polarization measure (0) is obtained for perfect opinion consensus. In between these two extremes, the polarization measure increases in the extent to which the opinion distribution is bimodal, with equally large modes at opposite extreme ends of the opinion spectrum. Panel B in figure III.2 shows the average degree of bi-polarization at the beginning of the runs and its average increase. Under all conditions, the simulations started with random, uniform opinion distributions. This resulted for all conditions in a low degree of bi-polarization in

the initial situation (indicated by the black areas of the bars). The white areas of the bars show that the average maximal degree of bi-polarization obtained in the simulation runs increased in the strength of homophily, *h*. Furthermore, bi-polarization increased only slightly in the course of a simulation run when homophily was weak (*h*<4). In other words, in these conditions the simulated populations hardly bi-polarized. Only strong homophily could give rise to significant levels of bi-polarization.

Figure III.2: Results from simulation experiment on the effects of homophily on bi-polarization (500 runs per condition, *N*=20, *P*=*C*=20, *S*=6)



Finally, for those runs which ended in bi-polarization, panel C in figure III.2 informs about the average number of simulations events that it took to reach the equilibrium. This measure serves as an indicator of the duration of the bi-polarization process. The conditions with weak homophily (b<3) are neglected in panel C, because only a single run ended in bi-polarization in these conditions. The graph shows that the weaker homophily the more events it took until bi-polarization was reached. These results indicate that bi-polarization is not only possible under strong homophily. The self-reinforcing process that leads to bi-polarization may evolve also if homophily is only moderately strong. However, the graph shows that for the corresponding conditions it may take a considerable amount of time until a group splits up into opposing factions. But the longer it takes before the equilibrium of bi-polarization is reached, the more likely it is that in the process agents interact with dissimilar others and learn counter arguments to their current opinion tendency. If this happens, it is likely that agents further spread these counter arguments within the subset of the population that leans towards the same pole of the opinion spectrum. As a consequence, bi-polarization declines again and the population becomes more likely to move towards the other possible equilibrium, perfect consensus. This explains why under moderate homophily perfect group splits occur only rarely.

To summarize, our computational experiments yield two main findings. First, the informal reasoning that we proposed above is consistent: The interplay of homophily and argument exchange can entail bi-polarization. Second, the social influence dynamics of our model only generate bi-polarization if homophily is sufficiently strong. To assure that these conclusions can be generalized beyond the specific parameter setting that we inspected in the experiments which we report in this paper, we have conducted extensive additional tests, varying the remaining parameters of our model (N, P, C, S). We have not found any combination of these parameters that generated bi-polarization under weak homophily (b<2). This suggests that strong homophily is a *necessary* condition of bi-polarization. Accordingly, the strength of homophily is one of the two major manipulations in the experimental test of our theory of bi-polarization.

III.4. Experimental Test

III.4.1. Overview

The two key mechanisms that according to our theory jointly underlie bi-polarization are argument exchange and homophily. To test this theory, we designed and conducted a

controlled group discussion experiment in which both mechanisms were manipulated independently from each other. To assure that bi-polarization in the group discussions can not be attributed to the negative influence mechanism, it was carefully avoided that participants received information which would allow them to perceive in- or outgroups within the set of their interaction partners. The key hypothesis of the experiment was that bi-polarization would nevertheless occur in the group discussions, but only in those conditions in which participants could exchange the arguments underlying their opinions, and in which selection of interaction partners was driven by homophily based on opinion similarity.

The group discussion was designed such that groups of eight participants interacted in a computer network which we implemented with the software z-tree (Fischbacher 2007). An experimental session consisted of seven interaction periods. In each period, participants were matched in pairs and exchanged information concerning an artificial issue. At the beginning of the experiment and after every interaction, we measured the opinions of the participants and calculated the degree of bi-polarization, the dependent variable of this study.

Computer-based communication was chosen to preclude that participants could be negatively influenced by interaction partners which the subject would categorize as members of some subjectively perceived outgroup. Participants could transmit in the communication process only opinion ratings and/or standardized arguments. In this way, participants were not aware of any social characteristic of their interaction partners.

Homophily was manipulated via the procedure with which participants were matched to interaction partners in the course of the group discussion. We compared in a within-subject design two different matching procedures. The entire group discussion process consisted of seven interaction periods. In the first three interaction periods, only participants with similar opinions interacted with each other (homophilious matching). In the remaining four periods, only participants with dissimilar opinions were brought into contact. We expected that – given the possibility of argument exchange – bi-polarization will increase in the homophilious matching phase, while it will decline in the phase of heterophilious interaction.

In the homophilious matching phase (first three interactions), participants are matched with partners who hold similar opinions. According to existing theories of bi-polarization (Baldassarri and Bearman 2007; Hogg, Turner and Davidson 1990; Macy et al.

2003), there should be no systematic tendency for opinion change either towards or away from interaction partners' opinions in this phase. In contrast, our theory of bi-polarization posits that opinions will change when participants can exchange arguments. More specifically, our theory predicts that if there is argument exchange, participants will shift in the homophilious interaction phase their opinions towards the pole of the opinion spectrum towards which they tended already prior to interaction. We have assured in the experimental design that already at the outset arguments are distributed such that there is some tendency for the group to split into two subsets of participants which lean towards opposite poles of the opinion spectrum. Accordingly, our theory further predicts that bi-polarization will increase in the homophilious interaction phase, if argument exchange is possible. Bi-polarization will, however, not increase in the homophilious interaction phase according to our theory if participants can not learn arguments in the course of interaction.

To test these predictions, we compare opinion dynamics in groups where participants discussed the issue only with arguments (*Only-argument-condition*) with groups where participants could only inform each other about their opinions (*Only-opinion-condition*). During the first three interactions of the *Only-argument-condition*, both preconditions of bi-polarization are satisfied. Hence, we expect a significant increase in the degree of bi-polarization during the first interactions in this condition. We expect *no* bi-polarization in the *Only-opinion-condition*, because negative influence tendencies are precluded in the experiment. For the same reason, we predicted bi-polarization during the first three interaction periods (homophilious matching) to be stronger in the *Only-argument-condition* than in the *Only-opinion-condition*.

As an additional test, we included an experimental condition where interaction partners exchanged both arguments and opinions (*Opinions and arguments-condition*). This serves as robustness test of our theory. Two outcomes are plausible. First, one might expect that the effect of learning a new argument on a participant's opinion is reduced, if participants learn that their interaction partners do not hold more extreme opinions than they themselves do. Following this reasoning, one would expect weaker bi-polarization in the *Opinions and arguments-condition* compared to the *Only-argument-condition*. Contrary to this reasoning, information about the interaction partner's opinion could also intensify the effects of persuasive arguments. In the *Opinions and arguments-condition*, participants were made aware of the fact that their first three interaction partners held similar opinions (due to homophilious matching). As a consequence, participants in this condition might have felt more attracted (Byrne 1971) by the first three interaction partners than participants of

the *Only-argument-condition*. This could, in turn, let the interaction partner appear more credible and thus increased the persuasiveness of the transmitted argument. This reasoning implies that the strongest bi-polarization should be observed in the homophilious interaction phase in the *opinions and arguments-condition*.

To summarize, our experiment is designed to test the following main hypotheses.

Hypothesis 1: In the homophilious matching phase, there will be more bi-polarization in the *only-argument condition* than in the *only-opinion condition*.

Hypothesis 2: In the heterophilious matching phase, bi-polarization will decrease in the *only-argument condition*, in the *only-opinion condition* and in the *opinions and arguments-condition*.

III.4.2. Method

Participants. Members of a general subject pool at the Department of Sociology at the University of Groningen had been invited to participate in this experiment. This subject pool comprises students and alumni of the two universities in Groningen. Interested students could register for a specific session using an online form (Greiner 2004). We assigned the sessions randomly to the three experimental conditions. Participants received monetary compensation for participation. After excluding problematic sessions (see below), we included data of 65 female and 31 male participants in the analyses (N=96). On average, participants were 23 years old.

Procedure. In each experimental session, we invited 8 participants to a computer laboratory where they sat in separate cubicles. We informed them that they would not be deceived during the experiment and that we had designed the experiment to study the formation of individual opinions in a social setting. Participants were asked to imagine that they were member of a discussion group, talking about the best location for building a new leisure center. This new center could be constructed in one of two hypothetical towns (town A and town B) or at any place in between these two towns. We chose this artificial issue because participants had no previous knowledge about it. This made it possible to impose the arguments that were known to each of the participants. In addition, the best spot for the leisure center can be identified on an interval scale, providing the participants an unambiguous means to inform each other about their opinion. After all participants had confirmed that they had understood the instructions, we started the computer program which ran the experiment. From now on, all instructions and communications took place on the computers screens.

In the first phase of the experiment, each participant received a different set of three arguments. Each argument suggests that either town A or town B is the better place for the new leisure center. For example, one of the pro town A arguments reads: "There is a university in town A. The nearer the leisure center will be build to town A, the more students will be attracted". Altogether there were 6 arguments pro town A and 6 pro town B. Half of the participants received 2 arguments pro town A and one pro town B and the other half received one pro town A and two pro town B. In the following, we therefore refer to those participants who received two pro town A arguments as "A-types" and to the others as "B-types". Whether a specific participant was of type A or B was assigned randomly. Furthermore, all participants who received two pro town A (B) arguments received the same pro town B (A) argument. This was done to assure that the degree of opinion homophily between members of the same type remained approximately constant throughout the homophilious interaction phase. We made the participants explicitly aware of the fact that they had received different sets of arguments. However, we did not tell them how we distributed the arguments. Hence, we expect that the participants were not aware of the two types and thus no social categorization was possible on basis of the initial distribution of arguments.

After the instructions and assignment of initial arguments, participants were asked to rate for each of their arguments how relevant it was for them on a 7-point scale. We included this to force participants to read each argument carefully and to allow us to check later the plausibility of participants' opinion ratings (see below). Subsequently, participants expressed the first time their opinion about the best location for the new leisure center. We used a scale ranging from -50 (*town A*) to +50 (*town B*). Participants could choose any value between the two extremes.

In the second phase, each participant interacted once with each of the seven other participants of the session. In the first three rounds (homophilious interaction phase) interactions took only place between participants who had received the same number of pro town A and pro town B arguments. In the remaining 4 interactions, participants were subsequently matched with the 4 participants of their session who had another number of pro-A and pro-B arguments (heterophilious interaction phase). We used this interaction schedule in all three between-subject conditions. Participants were not aware of the schedule. We only told them that they would interact once with each participant of the experiment. All interactions did really take place. Participants were not deceived.

In the *Only-opinion-condition*, each interaction consisted of two steps. First, the computer informed the participants about their partners' opinion on the best location for the leisure center, showing the partner's most recent opinion rating. Second, all participants rated again where they personally thought the best place for the leisure center was. The interactions in the *Only-argument-condition* consisted of three steps. First, both interaction partners were asked to select which of their arguments should be transmitted to their current interaction partner. Second, participants read which argument their respective partner had transmitted. Whenever a participant had received a new argument then this argument was added to this participant's list of arguments and could later be transmitted to interaction partners. Finally, the participants expressed their opinion again. The new opinion rating, however, was not communicated to the current interaction partners. The *Opinions and arguments-condition* was very similar to the *Only-argument-condition* except for the fact that in step 2, participants did not only read the transmitted argument but also learned the opinion of the respective partner about the best location of the leisure center.

At the end of each interaction, participants read on the screen that the interaction was finished and that they would be matched with a new interaction partner. When all 7 interactions were completed, participants answered a short questionnaire and received monetary compensation for participation.

Altogether, we conducted 18 sessions with 8 participants per session. However, we excluded 6 sessions from the analysis because the manipulation of the initial opinions did not work out. Even though participants received at the beginning of the experiment two arguments favoring one of the two towns, it was still possible that participants considered the one argument in favor of the other town as most relevant. In some cases, participants' initial opinion therefore tended towards the town for which less arguments were given. All sessions in which this happened for more than one of the participant were excluded from the analysis. This was necessary to ensure that the interaction schedule imposed homophilious matching during the first three interactions. Altogether, we used data from twelve sessions with eight participants each for the statistical analyses (N=96). For each of the three conditions, four sessions are available (N=32 each).

Six participants misunderstood the answering scale and entered their opinion with the wrong sign at the very first measurement (e.g. 20 instead of -20). At the second measurement, these participants corrected their opinion to the intended value (-20). We could identify these participants because the relevance measures for the initial arguments where not in line with the initial opinion (the counter argument was rated less relevant than

the other two arguments). This misunderstanding and, in particular, the participants' corrections at the second measurement might be problematic for parts of our analyses. The reason is that the initial degree of bi-polarization may be underestimated and thus the change of bi-polarization between the first and the second interaction round be overestimated. To avoid these complications, we reversed the sign of the initial opinion of the six participants. As a result of these changes, the increase in the degree of bi-polarization is unaffected by the participants' misunderstandings. To be sure, these changes are a conservative correction, because they make it more likely that we refute the hypothesis of *increasing* bi-polarization in the homophilious matching phase with argument exchange. As a further control, we present in the Appendix additional analyses that used the absolute value of the participants' opinions as an indicator of polarization. This variable is unaffected by the corrections that we conducted. The additional analyses show that all results could be replicated with this method.

III.4.3. Results

Figure III.3 describes bi-polarization in the three conditions of the experiment. In each graph, the upper (lower) thin solid line depicts the average opinion of the 16 participants of type B (A). The distance between the two lines (highlighted by the gray area) serves as a measure of the opinion distance between the two types. It neglects, however, the degree to which opinions vary between participants of the *same* type. To also take this into account, we calculated for each session and each interaction period the bi-polarization measure that we have also used in the computational experiments. The change of the average value across all sessions of this bi-polarization measure is shown with bold solid lines in figure III.3. For the opinion scale of the experiment the bi-polarization measure can maximally take the value of 50 (maximal degree of bi-polarization). The measure decreases to zero if all participants of a session have exactly the same opinion. The dotted lines in the graphs highlight the changes in average degree of bi-polarization during the first three (homophilious matching) and the last four interactions (heterophilious matching). To also quantify bi-polarization in the three conditions, we estimated for each condition separately a linear regression with the degree of bi-polarization as dependent and two period effects as independent variables. The first period effect models the change in bi-polarization during the first three interactions (Periods are coded: 0 1 2 3 3 3 3 3) and the second quantifies the change during the remaining 4 interactions (Periods are coded: 0 0 0 0 1 2 3 4). The regression coefficients of the two period effects indicate whether there was bi-polarization (positive coefficient); no change (insignificant coefficient); or whether

opinion distance between the two types decreased (negative coefficient). For each condition there were 32 observations available (4 sessions × 8 opinion measurements).

Figure III.3: Distance between subgroup averages and Bi-polarization dynamics

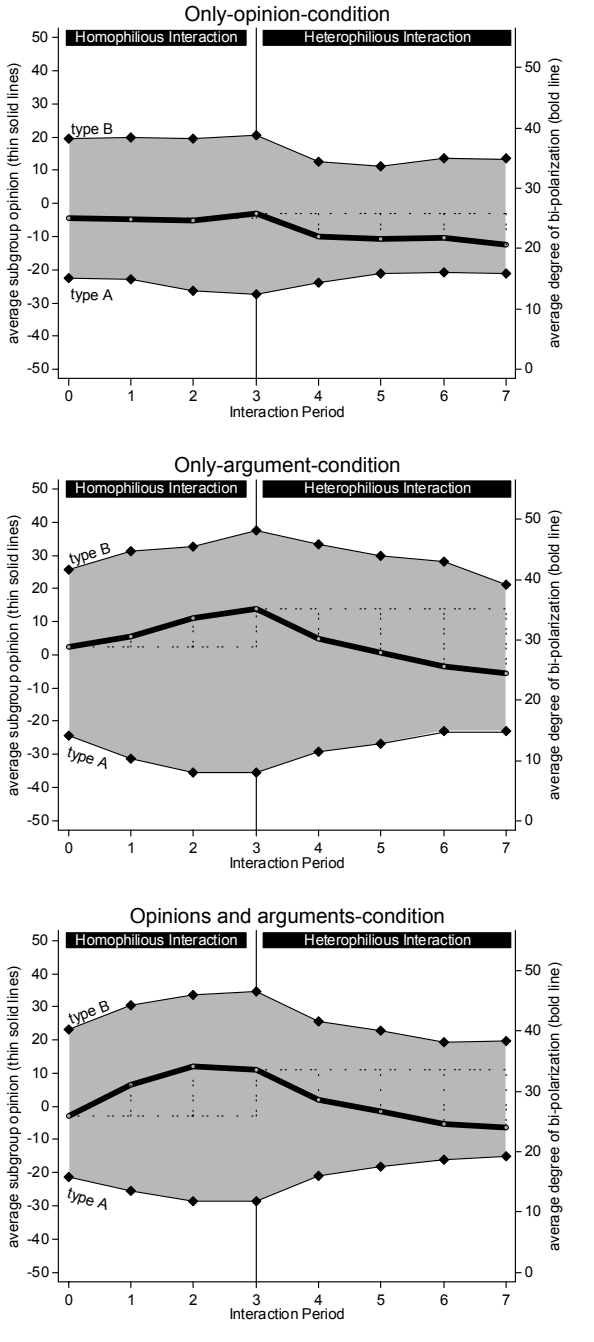


Figure III.3 shows for all three conditions that there have been significant differences between the opinion averages of the two types of participants already before the first interaction (interaction period=0). Also the initial degree of bi-polarization was in all

conditions significantly different from zero (t-values of intercepts in the regressions range from 15.03 to 24.43). This demonstrates that the assignment of arguments led to the desired opinion differences between the two types.

In the *Only-opinion-condition*, the degree of bi-polarization hardly changed during the first three interactions. Actually, it *de*creased on average by 0.21 during the first three interactions. The decrease is not significantly different from zero (t=-0.42). This is different for the *Only-argument-condition* and the *Opinions and arguments-condition*. In both conditions the degree of bi-polarization significantly *in*creased per interaction, by 1.71 (t=2.98) and 1.91 (t=2.15) respectively.

Dynamics changed when interaction partners with different opinions were matched (interaction period>3). Under all three conditions, figure III.3 shows decreasing opinion differences between the two types of participants. This confirms our second hypothesis. In the *Only-opinion-condition*, the degree of bi-polarization decreased on average by 1.01. This effect differs significantly from zero (t=-2.77) but the confidence interval of the effect reveals that it does not differ significantly from the weak decrease during the first three interactions. In the *Only-argument-condition* and the *Opinions and arguments-condition*, the degree of bi-polarization decreased from interaction period 4 on by 2.75 (t=-6.5) and 2.74 (t=-4.17) on average. In both conditions, this decrease during the interactions 4 thru 7 differs significantly from zero and therefore also from the *in*crease during the first three interactions.

We also wanted to know whether the dynamics of bi-polarization differed significantly between the three conditions. To test this, we began by estimating a regression that tested differences between the *Only-opinion-condition* (reference category) on the one hand and the *Only-argument-condition* and the *Opinions and arguments-condition* on the other hand. For this purpose, we used the same regression approach as for the separate models and included main and interaction effects for the experimental conditions. It turned out that the increase in the degree of bi-polarization during the first three interactions was in significantly stronger in both the *Only-argument-condition* (t=2.01) and the *Opinions and arguments-condition* (t=2.22) than it was in the *Only-opinion-condition*. This supports our hypothesis 1. We observed bi-polarization even in the absence of negative influence, but only in the conditions in which both homophily and argument exchange were imposed.

We found furthermore that the *de*crease in the degree of bi-polarization during the final four interaction periods was significantly stronger in the *Only-argument-condition* (t=-

2.47) and the *Opinions and arguments-condition* (t=-2.44) as compared to the *Only-opinion-condition*. A comparison of the differences between the *Only-argument-condition* and the *Opinions and arguments-condition* revealed that both the developments during the first three interactions (t=.19) and the subsequent four interactions (t=.02) did not differ significantly between the two conditions.

In conclusion, the results of the experiment confirm our hypotheses. We found significant bi-polarization in both conditions where participants exchanged arguments under homophilious matching, despite the fact that negative influence was precluded by the experimental design (Hypothesis 1). In the control condition, where participants only exchanged opinions, we did not find significant bi-polarization. Thus, the empirical test supports the theoretical results of our computational experiments. Bi-polarization can arise without negative influence, but only if homophily in the selection of interaction partners is sufficiently strong. In addition, we do not find significant differences in bi-polarization between the *Only-argument-condition* and the *Opinions and arguments-condition*. This suggests that the bi-polarization dynamics of persuasive arguments are robust to the effects of opinion exchange. At the same time, opinion exchange alone does not entail bi-polarization, as we found in the *Only-opinion-condition*. Finally, the results support that homophily is a precondition of bi-polarization. Under all three experimental conditions, we find decreasing differences between the two types of participants under heterophilious matching (Hypothesis 2).

III.5. Summary and Discussion

Previous explanations of bi-polarization hinge on the assumption that individuals seek to adopt opinions that maximize opinion differences to outgroup members (e.g. Hogg, Turner and Davidson 1990; Mason, Conrey and Smith 2007). Yet, empirical research which tested the assumption of negative influence has produced very mixed results and is open to methodological criticism (e.g. Krizan and Baron 2007). This has led us to propose a new theory of bi-polarization that does not rely on negative influence. Our theory has two main ingredients. First, it is assumed that individuals tend to interact with others who hold similar opinions (homophily). Second, during interaction individuals influence each others' opinions by exchanging arguments. Based on PAT, we assume that when individuals with similar opinions interact, they likely provide each other with new arguments that support their opinions. The interplay of homophilious selection of interaction partners and influence with persuasive arguments can create a feedback process

which triggers the formation of subgroups that independently from each other develop increasingly distant opinions.

We studied the theory first from a theoretical angle and then tested it empirically. For the theoretical study, we developed a formal model of the theory. Computer simulations led to three main results. First, we demonstrated that our formal model can generate bi-polarization, although the model does not include the assumption of negative influence. Second, we showed that the interplay of homophily and argument exchange can trigger bi-polarization even in a population where initially all individuals hold perfectly similar opinions. This implication of our theory contradicts self-categorization theories, which predict bi-polarization only when individuals perceive differences between members of the population. Third, a simulation experiment revealed that according to our theory bi-polarization can only arise if homophily is sufficiently strong.

Future research should use the formal model to identify further conditions of bi-polarization. Mäs et al. (see chapter IV) have recently explored the effects of including demographic characteristics into the formal model (Mäs et al. 2008). Interestingly, their simulations revealed that demographic diversity in a group may intensify bi-polarization only if two further conditions are met. First, the demographic attributes need to form a sufficiently strong faultline (Lau and Murnighan 1998) in the sense that several demographic dimensions are sufficiently correlated. Second, demographic attributes need to be correlated with agents' opinions (Homan et al. 2007; Phillips 2003; Phillips et al. 2004). If one of the conditions is not met, then subgroups fail to become sufficiently distinct from each other. The reason is that similarity between subgroups in demographic characteristics or initial opinions provides a basis for argument exchange accross subgroup boundaries which, in turn, prevents bi-polarization. In addition, simulation results suggest that demographic faultline may fire up bi-polarization in the short run, but at the same time they help to overcome opinion differences in the long run as long as there is at least some demographic overlap between subgroups (Mäs et al. 2008).

In the second part of the present paper, we tested our theory with a laboratory experiment. We created a setting for which existing theories of bi-polarization assume no negative influence and therefore predict no bi-polarization. Yet, we predicted and found bi-polarization for specific experimental conditions. In particular, we found significant bi-polarization under the conditions that our formal model identified, strong homophily and the possibility to exchange arguments.

We have deliberately created an artificial setting in our experiment. This allowed us to exclude other explanations of bi-polarization than the mechanisms that we wanted to test. Yet, we believe that it would be fruitful for future experimental research to test our theory of bi-polarization in less constrained settings, like standard group discussion experiments (Johnson and Johnson 1982). At the same time, we also believe that the setting of our experiment may be a realistic model of some parts of the social world. For instance, internet search engines and online social networks make it very easy to connect only to others with similar opinions (Sunstein 2008) and exchange very condensed pieces of information, like in our experiment. We therefore believe that future empirical research may gain a better understanding of opinion dynamics in real life situations by focusing on those settings in which both argument exchange and homophily are possible and could thus, according to our theory, give rise to bi-polarization.

III.6. Appendix to chapter III

In laboratory experiment, it turned out that six participants misunderstood the answering scale and entered their opinion with the wrong sign at the very first measurement (e.g. 20 instead of -20). At the second measurement, they corrected their opinion and entered the intended value (-20). We had to rectify this because otherwise the statistical model would spuriously indicate that the degree of bi-polarization increased. Above, we have argued that these changes make it more likely that our hypothesis is refuted. Still, we provide here additional analyses which demonstrate that our results are not affected by the changes.

In particular, we have focused on the participants level and used the absolute value of each participant's opinions as the dependent variable, a measure which is unaffected by the corrections. According to the hypotheses, we expect that in the conditions where arguments were exchanged the absolute value of the participants' opinions increased during the first three interaction periods. This reflects that the participants developed more extreme opinions. In contrast, we expect that the opinions remained constant in the *Only-opinion-condition*.

To take into account that the opinion measurements of each participant are interdependent, we estimated linear multi-level regressions (Snijders and Bosker 1999) with random intercept effects on the participant level. An empty model revealed that the explained variance on the group level was not significant. Including random period effects (random slopes) did not affect the significance decisions.

We followed the same strategy as with the group-level analyses in the main paper. Models 1 thru 3 (see table III.1) tested the two period effects separately for each condition. In line with the analyses of the group-level data, we found no significant opinion extremization during homophilious matching in the *Only-opinion-condition* (model 1). In contrast, we found significantly more extreme opinions in the *Only-argument-condition* and the *Opinions and arguments-condition* (models 2 and 3). Also in line with the group-level analyses, we found for all conditions that the participants held increasingly moderate opinions during heterophilious matching of interaction partners.

Table III.1: Multi-level regression of participants' absolute opinion value on interaction period

Fixed Effects	Model 1	Model 2	Model 3	Model 4	Model 5
I: only opinions					
Intercept	25.89 **			25.89 **	
	(2.20)			(2.28)	
Homophil. interaction	-.17			-.17	
	(.34)			(.46)	
Heterophil. interaction	-1.41 **			-1.41 **	
	(.25)			(.34)	
II: only arguments					
Intercept[a]		30.48 **		4.59	30.48 **
		(2.33)		(3.22)	(2.31)
Homophil. Interaction[b]		1.89 **		2.07 **	1.89 **
		(.49)		(.65)	(.51)
Heterophil. Interaction[b]		-2.95 **		-1.54 **	-2.95 **
		(.36)		(.48)	(.37)
III: arguments and opinions					
Intercept[a,c]			29.19 **	3.30	-1.29
			(2.29)	(3.22)	(3.27)
Homophil. Interaction[b,d]			2.11 **	2.28 **	.22
			(.52)	(.65)	(.72)
Heterophil. Interaction[b,d]			-3.66 **	-2.25 **	-.70
			(.39)	(.48)	(.53)
Random Effects					
Sd(intercept)	11.80	11.86	11.42	11.70	11.65
	(1.53)	(1.57)	(1.53)	(.89)	(1.10)
Sd(resid)	4.80	6.91	7.43	6.49	7.18
	(.23)	(.33)	(.35)	(.18)	(.24)
Number of participants	32	32	32	96	64
Number of observations	256	256	256	768	512

Note. Table reports unstandardized effects. Values enclosed in parentheses represent standard errors.
** significant on 0.01 level
[a] in model 4 included as condition dummy. [b]in model 4 included as interaction term with condition dummy. [c] in model 5 included as condition dummy. [d]in model 5 included as interaction term with condition dummy.

Model 4 compares the *Only-opinion-condition* with the *Only-argument-condition* and the *Opinions and arguments-condition*. In line with the group-level results, we find that the opinions turned significantly more extreme during homophilious matching in the two conditions where arguments were exchanged. Also the opposite effect during heterophilious matching

is stronger in the two conditions with argument exchange, compared to the *Only-opinion-condition*.

Finally, model 5 compares the *Only-argument-condition* and the *Opinions and arguments-condition*. Supporting the group-level results, table III.1 shows that there are no significant differences between the two conditions.

We replicated all five models and included the respondents' age and gender as control variables. In all five models, neither the main nor interaction effects with the period effects and the between-subject conditions were significant.