

**Individual attitude change and societal dynamics:  
Computational experiments with psychological theories**

Jan Lorenz

Jacobs University Bremen and GESIS Leibniz Centre for Social Science

Martin Neumann, [neumanm@uni-mainz.de](mailto:neumanm@uni-mainz.de)

Johannes-Gutenberg University Mainz

Tobias Schröder, [post@tobiasschroeder.de](mailto:post@tobiasschroeder.de)

Potsdam University of Applied Sciences

**Author Notes**

This research was supported in part by grants from the German Research Council: DFG 265108307 (Jan Lorenz and Martin Neumann) and DFG 396901899 (Jan Lorenz).

Some results were presented at the [“Interdisciplinary Workshop on Opinion Dynamics and Collective Decision 2017”](#) in Bremen (Martin Neumann) and at the DPG Spring Conference (Section Physics of Socio-Economic Systems) 2019, Regensburg (Jan Lorenz).

Correspondence should be addressed to Jan Lorenz [j.lorenz@jacobs-university.de](mailto:j.lorenz@jacobs-university.de).

**© 2021, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/rev0000291**

**Individual attitude change and societal dynamics:  
Computational experiments with psychological theories**

**Abstract**

We present an agent-based model for studying the societal implications of attitude change theories. Various psychological theories of persuasive communication at the individual level are implemented as simulation experiments. The model allows us to investigate the effects of contagion and assimilation, motivated cognition, polarity, source credibility, and idiosyncratic attitude formation. Simulations show that different theories produce different characteristic macro-level patterns. Contagion and assimilation are central mechanisms for generating consensus, however, contagion generates a radicalised consensus. Motivated cognition causes societal polarisation or the fragmentation of attitudes. Polarity and source credibility have comparatively little effect on the societal distribution of attitudes. We discuss how the simulations provide a bridge between micro-level psychological theories and the aggregated macro-level studied by sociology. This approach enables new types of evidence for evaluating psychological theory to complement experimental approaches, thus answering calls to enhance the role of coherent and formalised theory in psychological science.

**Keywords:** Attitudes, Social Influence, Theoretical Integration, Computational Model

### **Individual attitude change and societal dynamics:**

#### **Computational experiments with psychological theories**

We live in a time of political disagreement and attitude polarisation. It is hard to pin down the reasons for conflict and discontent, given that by many measures, this age is more prosperous and less violent than any time in the past (Harari, 2016; Pinker, 2018; <https://ourworldindata.org>). Pessimists have suggested (following Plato's political philosophy) that democracies have an innate tendency to self-destruct through discursive polarisation and fragmentation (Sullivan, 2016), however, a closer look at polling data reveals that not all political attitudes diverge in all democracies. Figure 1 shows recent distributions of political beliefs across four different issues in two different countries, taken from the European Social Survey (ESS, 2018). The distributions are not consistent across countries, and they range from stark polarisation to considerable consensus, with unimodal, bimodal, or multimodal patterns. How do these vastly different distributions come about?

Attitude distributions are macro-level societal phenomena, but they arise from micro-level psychological processes. Social psychological research on mechanisms such as conformity, social comparison, stereotyping, groupthink, and other dynamics has highlighted how psychological biases may contribute to attitude patterns at the societal level.

Psychology's theoretical focus on individual information-processing, however, along with the methodological focus on laboratory experiment, has limited progress on this issue. The relationship between the level of individual cognition and an aggregated social macro-level often remains vague and unspecified. If psychology wants to unleash its potential for better understanding social developments, we need better tools with which to assess the macro-level societal implications of psychological dynamics.

Recent advances in computer science and technology have led to the development of the "computational social sciences", which seek to explain societies as networked patterns of

individual behaviours (e.g., Conte et al., 2012; Lazer et al., 2009; Schweitzer, 2018; Squazzoni, 2012). There is usually little theoretical influence from psychology in these developments, however, as the new field is dominated by scholars from physics, computer science, and sociology. Conversely, the predominant focus on experimental standards of psychology within the discipline has eclipsed the extraordinary potential of novel mathematical tools for theoretical integration and the large-scale empirical study of attitude dynamics (see Vallacher, Read, & Nowak, 2017).

We bridge experimental psychology and computational social science to study the implications of attitude theories for societal dynamics. Capitalising on the pioneering work of Hunter, Danes, and Cohen (1984), we propose a mathematical formalisation and integration of theories of attitude change across several important psychological paradigms. By embedding this model in an agent-based model (Smith & Conrey, 2007; Helbing & Balietti, 2012; Jackson, Rand, Lewis, Norton, & Gray, 2017), we then show how we can use computer simulations in order to scale psychological findings at the individual level to the level of thousands as a proxy for societies. This approach allows us to provide explanations for societal dynamics grounded in psychological knowledge, and to assess the social significance of psychological effects in ways that experiments with individuals and small groups cannot. Finally, we discuss how models such as ours could guide empirical research in a psychology that is both more theoretically coherent and more relevant for understanding societies as a whole.

### **Mathematical and Computational Approaches to the Psychology of Attitudes**

In this section, we briefly review previous research programs in psychology related to the mathematical formalisation of attitude mechanisms. We refer readers interested in more exhaustive reviews to Hunter et al. (1984), on whose earlier attempt to mathematically

formalise theories of attitudes we build, and to the more recent computational approaches of Read and Simon (2012) and Vallacher et al. (2017).

*Cognitive consistency* was a core concept in early theories of attitudes, a tendency of the mind to strive for the mutual fit of all mental representations that sustain an attitude. Heider's (1946) classical "balance" theory states that imbalanced relationships between cognitive objects (e.g., you love a person but they hate the way you dress) will cause attitude changes because of their aversive nature (e.g., either you will lose interest in that person or start dressing differently). Cartwright and Harary (1956) developed a generalisation of Heider's approach based on graph theory, preceding modern network science. Cognitive dissonance theory (Festinger, 1957), perhaps one of the most influential psychological theories of all times, is a special case of cognitive consistency when applied to the justification of one's own actions (e.g., I chose to work as a badly-paid postdoc, therefore I must apparently love science). Osgood and Tannenbaum (1955) developed a mathematical model of attitude change known as congruity theory. The model overcomes the lack of precision in balance-theoretical predictions of who changes their attitude to what extent as a result of communication. Attitudes are assumed to change in inverse relation to their intensity. Strongly felt attitudes are therefore much harder to change than slight preferences.

Gollob (1968) and Heise (1969) built on Osgood's work to develop formal models of how attitudes towards actors, behaviours, and object persons combine into coherent social impressions. This approach was developed by mathematically-minded sociologists into affect control theory, whose main idea is that a motive to maintain and verify mutually compatible affective attitudes about social situations is the driver of social-interaction dynamics (Heise, 1979; 2007; MacKinnon, 1994; Schröder, Hoey, & Rogers, 2016). The mathematical model of affect control theory thus links individual-level attitudes with societal dynamics, but its influence on psychology has remained limited.

Computational social psychology (Vallacher et al., 2017) has taken a different path to mathematical sociology, but is no less traceable to the old ideas about attitude consistency. Connectionist constraint-satisfaction models conceptualise holistic attitudes as a result of an interactive process of mutual excitation and inhibition between specific cognitions or evaluations, which produces the attitude-balance effects discovered by the early *Gestalt* psychologists (e.g., Read & Simon, 2012; Schröder & Wolf, 2017; Simon, Stenstrom, & Read, 2015). More biologically realistic neural network models serve as tools for bridging the neural and cognitive levels of explanation (e.g., Ehret, Monroe, & Read, 2015), similar to the strategy we employ in the present paper for bridging the cognitive and social levels through computational modelling.

Social impact theory, which conceptualises the influence that other people have on individuals, is another example of mathematically formalised attitude research (Latané, 1981; Nowak et al., 1990). According to the model, influence increases by the square root of the number of sources exerting an influence (i.e., assuming a decreasing impact of additional sources), their strength (e.g., their source credibility), and their immediacy (i.e., closeness in time and space). Empirical studies based on social impact theory in various content domains of social psychology have been taken to parameterise the model, resulting in different functional forms for different contexts, while preserving the fundamental mathematical characteristics. As explained below, we will pursue a similar strategy in our present attempt to mathematically integrate theories of attitude change. But before we turn to our own model, we need to review computational techniques aimed at bridging the individual vs. society levels of explanation.

### **Agent-based Models and Social Simulation**

Agent-based modelling allows us to resolve the challenge of going from individual information processing to the aggregated level of society (Hegselmann 2017; Helbing &

Balietti, 2012; Jackson et al., 2017; Smaldino, 2020; Smith & Conrey, 2007; Squazzoni 2012). An agent-based model involves a virtual society, where computer-simulated individuals (*agents*) repeatedly communicate with each other. The resulting flows of information generate macro-level patterns of attitudes and behaviour. Agents in such models can have psychological properties when they are theoretically based on psychological theories (e.g., see Schröder & Wolf, 2017; Smith & Conrey, 2007). However, agent-based modelling is not a widely used method in psychology (Jackson et al., 2017; Smaldino, 2020), and so most agent-based models are not well informed by psychology. Many such models implement simplistic characteristics based on the ad-hoc theories of the modellers, who are often physicists, computer scientists, or sociologists (see Chattoe, 2014; Jager, 2017). These ad-hoc psychological theories are often based on analogies from other disciplines such as modelling humans as particles. While there is debate as to whether psychological realism at the micro-level is necessary to describe social patterns at the macro-level (similar to the question in thermodynamics about whether molecular movements matter for describing the state of a gas; see Schweitzer, 2018), many authors in computational social science have criticised the disconnect between agent-based modelling and psychology (see Jager, 2017, for a review of that debate).

Some social psychologists have used agent-based models to understand the micro-macro link. For example, dynamic social impact theory (Nowak et al., 1990) was developed as an extension of social impact theory (reviewed in the preceding section) with the explicit purpose of simulating population-level attitude change as the result of individuals influencing each other via the social impact function. Schröder and Wolf (2017) used a model with connectionist agents to study the change of attitudes toward sustainability innovations. Muthukrishna and Schaller (2020) modelled social change more generally from a cross-cultural perspective. Undoubtedly, other models exist, however, it is fair to say that

they have not seeded a cumulative, broad research program aimed at formalising psychological theories in similar ways to other scientific disciplines (Smaldino, 2020) and assessing their significance for understanding societal challenges, not just individual processes.

Perhaps the best-known class of micro-macro attitude models are opinion dynamics models, but unfortunately, they are intellectually mostly disconnected from psychology. These models typically rely on a one-dimensional scale, which is treated as abstract in the literature but can easily be connected to evaluative scales from attitude research. The innovation of these models was the introduction of the concept of “bounded confidence” (e.g., Deffuant, Neau, Amblard, & Weisbuch, 2000; Hegselmann & Krause, 2002): Agents ignore other opinions when they are outside a certain confidence interval around their own opinions. We will show in our model below that bounded confidence can be approximated as an extreme case of a more general model of attitude change that is more tightly integrated with psychological attitude theory and research.

The macro-social patterns of consensus, or two or more groups with internal consensus, emerge according to the narrowness of the confidence interval. Assumptions about micro-level information processing thus matter for the macro-level outcome. Since the early opinion-dynamics models, this general finding has been replicated in simulations with a vast number of models built on different assumptions, some of which are informed by specific paradigms of social psychology. For instance, Brousmiche, Kant, Sabouret, and Prenot-Guinard (2016) built a model based on a multicomponent approach to attitudes (Rosenberg & Hovland, 1960). Jager and Amblard (2005) developed a model inspired by social judgment theory (Sherif & Hovland 1961), Salzarulo (2006) referred to self-categorisation theory (Turner, Hogg, Oakes, Reicher, & Wetherell, 1987), and Kurahashi-Nakamura, Mäs, and Lorenz (2016) introduced a facilitation parameter derived from



Fishbein and Ajzen's (1975) theory of reasoned action. Flache et al. (2017) provide a comprehensive review of the research field calling for more integrative work, which we provide in the following from the perspective of psychological theory.

### **Research Goals**

Our first objective in this paper is to move towards a theoretical integration of attitude theories for use in an agent-based model. We will use this model to systematically examine the consequences of different parametric implementations of the theories for the interplay between micro-level attitude change and macro-level societal dynamics. We use agent-based simulation as a virtual laboratory for “experimenting with theory” (Dowling, 1999; Troitzsch, 2017). The simulation approach allows us to systematically compare different psychological theories in an experimental fashion. We systematically manipulate theoretically meaningful parameters and examine the data generated by the simulation, in order to study the effect of the relevant psychological mechanisms on the aggregated level of society. The manipulation of parameter values reflects the uncertainty and lack of knowledge inherent in theoretical assumptions: unlike precise parameters in other scientific disciplines (e.g., the speed of light, earth's gravitational force), there is often more vagueness in psychological theories owing to their predominantly verbal nature (Smaldino, 2020). Using simulation as an epistemic tool (Knuuttila, 2011) for experimentally varying parameter values allows us to gain knowledge even under conditions of uncertainty. In fact, we will see that different parameter values can strongly affect the results. This is positive knowledge, as we can connect emergent simulation results to macro-data patterns, such as those displayed in Figure 1, and thus gain insight as to which parameter ranges correspond to reality. The approach also contributes to a more reliable psychological science, as it can guide empirical research more systematically than the widespread practice of grounding empirical research in intuition and folk theories (Muthukrishna & Henrich, 2019; Van Rooij, 2019).

Various macrodata patterns in attitude distributions are possible in theory and observable in reality. Figure 1 shows example distributions of attitudes toward four different topics in Norway versus Serbia. The differences cannot easily be explained only by characteristics of the countries or topics. In the following, we distinguish essentially unimodal (1) and multimodal (2) patterns, and then further distinguish them with respect to their character of being polarised (A), diversified (B), or condensed (C). “Polarised” means that differences are accentuated. In politics, this is commonly understood as large attitude differences between two groups. We also use this notion when one group is extreme, as conceptualised by “group polarisation” in psychology. In that sense, a unimodal polarised distribution (1-A) resembles an extreme consensus. For example, all people agree that friendship is good and that child abuse is bad. Diversified unimodal attitude distributions (1-B) are all cases where we may reasonably assume a broad normal distribution of attitudes, for example, how much people like or dislike certain flavours of ice cream. A distribution characterised by a neutral consensus (1-C) is probably often unnoticed in our opinions about uninteresting things. The typical case of polarisation in politics involves two opposing camps with extreme attitudes. We call this “bipolarisation” (2-A). Political polarisation currently draws a great deal of attention, and there are typically many intermediate attitudes between extreme political camps and political attitude distributions are multimodal and fragmented (2-B). Attitude distributions can also appear fragmented without sizable groups of extremists (2-C). The latter two cases in particular are common in the attitude landscape according to the European Social Survey (ESS, 2018) as seen in Figure 1. Characteristics of patterns 1-A, 1-B, 1-C, and 2-A also appear in empirical data.

We will link the characteristics of the six distributions of attitudes in stylised form (see Table 2 in the section about the agent-based model) to the interplay of different psychological theories of individual attitude change and social interaction.

## A Coherent Mathematical Model of Individual Attitude Change

### *Formal Modelling Framework for Attitude Change*

Before implementing the simulation model, we need a coherent operationalisation of attitude-change theories in a mathematical model. To this end, we selected mathematical formalisations mostly from Hunter et al. (1984), updated them from a contemporary perspective, and integrated the selected concepts into one attitude change function. This section focuses on the static situation of one individual.<sup>1</sup>

Static analysis of the communicative situation reveals five basic elements: (1) a source who delivers (2) a message to (3) a receiver about (4) some object through (5) some channel (Shannon & Weaver, 1949). In the following, we do not treat the particularities with respect to the object or channel of communication but concentrate on the source, the message, and the receiver, assuming a generic object and channel. Following Hunter et al. (1984, pp. 5-6), we assume that at any moment in time one can use an evaluative ratio scale to measure the attitude of the receiver toward the object  $a$  and the affective or evaluative content of the message about the object  $m$ . We will later treat a simple form of source credibility which affects attitude change but where we do not consider source change. We also refrained from modelling attitude strength as another dynamic variable in order to focus on the dynamics of one variable on the macro-level. The only dynamic variable in our model is thus attitude. Following Hunter et al. (1984, pp. 5-6) and in line with most experimental work, we use a passive communication paradigm with only one direction of influence: the receiver changes their attitude with respect to a message from the source. We do not consider reciprocal

---

<sup>1</sup> Interestingly, Hunter et al. (1984, p. 6) mention in their Introduction: “Volume 2 extends selected attitude change models to group processes and develops mathematical models of group communication and group dynamics.” To the best of our knowledge, Volume 2 never appeared. In some sense, we continue their program here, shifting the focus even further from group processes to societal dynamics.

influence from the receiver to the source. In the society-level simulation model described later, however, sending and receiving agents can switch roles in subsequent time steps.

When  $t$  is the time prior to the message presentation and  $t+1$  is the time after message presentation, then

- $a(t)$  is the *attitude of the receiver towards the object at time  $t$* ,
- $s(t)$  is the *attitude of the receiver towards the source at time  $t$* , and
- $m(t)$  is the *affective or evaluative content of the message at time  $t$* , and

the attitude at time  $t+1$  is

$$(1) \quad a(t+1) = a(t) + f(a(t), s(t), m(t)),$$

where  $f$  is the *attitude change function*, which we specify mathematically in the following.

Consequently, the attitude change is what can be measured post hoc, for example, in an experiment, as the difference between the attitude after processing the message and the attitude before the perception of this message

$$(2) \quad f(a(t), s(t), m(t)) = \Delta a(t) = a(t+1) - a(t).$$

In the following subsections, we enhance and parameterise this model to include different levels of depth of psychological theories of attitude change and persuasion, starting with simple contagion and adding more factors accounting for the agent's previous attitude, such as assimilation, motivated cognition, and polarity, into an integrated framework. We do not include repulsive, contrastive, or boomerang effects (Flache et al., 2017; Mason, Conrey, & Smith, 2007; Hunter et al., 1984) in attitude change in this model because the mathematical formulation with few parameters was not found to be straightforward and the empirical evidence for this effect is mixed (Takács, Flache, & Mäs, 2016). In principle, such effects could easily be included in the framework that follows, in future work.

***Modelling Contagion: Receiving Attitudinal Doses of Positivity or Negativity***

Messages can be seen as simple contagious stimuli driving an attitude in a certain direction with a certain strength. We model attitude change through contagion as

$$(3) \quad \Delta a = \alpha \cdot m,$$

where  $0 \leq \alpha \leq 1$  is a general *strength* parameter. A small value of  $\alpha$  implies that the receiver is more cautious about changing attitudes (i.e. the social influence is only small) whereas in the case of a high value of  $\alpha$  the receiver is more susceptible to social influence and changes more. Contagion in this mathematical representation means that a receiver will always change their attitude in the direction of the sender's message. After a positive message, a receiver with an already positive attitude will be more positive, a receiver with a negative attitude will be less negative, neutral, or even slightly positive, depending on how negative her attitude was and how positive the message was. The attitude of the receiver does not play a role in a contagious change. Attitude change in Equation (3) appears as reinforcement derived from behaviouristic learning models in Hunter et al. (1984, p. 11, Eq. 2.1). It conceptualises the message  $m$  as an attitudinal 'dose' similar to Dodds and Watts' (2005) generalised model of social and biological contagion. There are conceptual similarities to models of biased assimilation (Dandekar, Goel, & Lee, 2013), the homophilous exchange of argument (Mäs & Flache, 2013) and learning from social feedback (Banisch & Olbrich, 2019). In these models, the attitude variable is the likelihood an individual assigns to a certain proposition being true. A message from another individual is then, for example, an argument in favour or against the proposition and pushes the individual's assessment of likelihood in that direction.

***Modelling Assimilation: Accounting for Prior Attitudes***

When contagion occurs, attitude change is independent of the receiver's current attitude.

According to principles of information integration and assimilation (Anderson, 1971; Asch, 1948; Hovland, Harvey, & Sherif, 1957), a message that is identical to the current attitude of the receiver does not trigger an attitude change. If a different attitude is expressed in a message, the individual compares the expressed attitude with their own pre-existing one and assimilates it towards the message, proportionally to the discrepancy. Following Hunter et al. (1984, p. 36, Eq. 3.1), this can be expressed as

$$(4) \quad \Delta a = \alpha(m - a)$$

where  $\alpha$  is the same strength parameter as for contagion in Equation (3). For  $\alpha = 1$  this means the receiver's attitude will become equal to the message<sup>2</sup>. In terms of Anderson's (1971) integration theory, the individual integrates the stimulus message  $m(t)$  with weight  $\alpha$  and the initial attitude with weight  $(1 - \alpha)$ .

A received message can also be processed with a mix of contagion and assimilation. Taking this into account we introduce the novel parameter *degree of assimilation*  $0 \leq \rho \leq 1$  and integrate contagious and assimilative attitude change as

$$(5) \quad \Delta a = \alpha(m - \rho a).$$

For  $\rho = 0$  we obtain contagion (3), for  $\rho = 1$  assimilation (4). Figure 2 shows the difference in attitude change with respect to different degrees of assimilation. Contagion and assimilation differ when the receiver has a positive attitude and receives a message which is also positive, but less so. In this case, contagion would make the receiver's attitude more positive, while assimilation would make it less positive.

Equation (5) brought to the form  $\Delta a = \alpha m - \alpha \rho a$  can also be seen as a combination of a stimulus term  $\alpha m$  and an inhibition term  $-\alpha \rho a$ , where the inhibition term drives the attitude

---

<sup>2</sup> This can be seen after replacement of  $\Delta a$  in Equations (2) and (1) brought to the form  $a(t+1) = \alpha m(t) + (1 - \alpha)a(t)$ .

towards the neutral state of zero, (see Hunter et al., 1984, pp. 24-27, Eq. 2.39), where the inhibition term is  $-\varepsilon \cdot a$ .<sup>3</sup> The degree of assimilation  $\rho$  can blend contagion and assimilation, which to our knowledge has not been studied in models of collective attitude formation.

***Modelling Boundaries and Polarity: More Certainty Close to Maximally Extreme Attitudes***

Repeated attitude change through contagion can drive an individual's attitude to arbitrarily large values. Most evaluative scales, however, are bounded; therefore, we will assume the existence of a most positive and most negative attitude. We assume them to be symmetric with equal distances around the neutral attitude  $a = 0$ . We define every attitude change function  $f(a, s, m)$  such that the new attitude is always capped to stay within the boundaries  $-M < a(t+1) < +M$  where  $M$  is the *maximal absolute attitude*.<sup>4</sup>

Furthermore, there is evidence that people with more extreme attitudes are more resilient to attitude change than people with moderate or neutral attitudes (Cantril, 1944; Hutchinson, 1949; Osgood & Tannenbaum, 1955). To account for such polarity effects on the strength of the change, we introduce a *polarity factor*  $(M^2 - a^2) / M^2$  in the attitude change function:

$$(6) \quad \Delta a = \frac{M^2 - a^2}{M^2} \alpha (m - \rho a).$$

The polarity factor is by definition between zero and one and thus scales down the attitude change of agents with prior attitudes of large absolute values, approaching zero attitude change when the maximal absolute attitude  $M$  is reached. The functional form is taken from Hunter et al. (1984, p. 21, Eq. 2.21) and is shown in Figure 4. We studied models with and without the polarity factor. When we do not use polarity, we still use boundaries to confine

<sup>3</sup> Thus, the degree of assimilation can be interpreted as the ratio of an inhibition parameter and the stimulus strength parameter  $\rho = \varepsilon/\alpha$ . With an inhibition term  $\varepsilon$  that is larger than the strength  $\alpha$ , we would have a model with a strong tendency of relaxation to the neutral state (see Shin & Lorenz, 2010) which we will not treat here.

<sup>4</sup> For any unbounded attitude change function  $\Delta a = f(a, s, m)$  we can always define a bounded version as  $\max(-M - a, \min(M - a, f(a, s, m)))$ , which ensures that  $-M - a < \Delta a < M - a$ , which implies that the new attitude  $a(t+1) = a(t) + \Delta a(t)$  stays within the boundaries. As an equivalent less nested form one could also use the wedge (pairwise minimum) and vee (pairwise maximum) notation  $\Delta a = f(a, s, m) \wedge (M - a) \vee (-M - a)$  which puts the cap at the minimal or maximal attitude at the end of the term.

attitudes to the bounded attitude scale. Polarity has not been studied as such in social simulation models but the way that Deffuant, Amblard, Weisbuch, and Faure (2002) model extremists as individuals, which barely take the attitudes of others into account, is related.

### ***Modelling Motivated Cognition: Evaluating Discrepancy***

So far, the model does not consider the effect of the discrepancy on the judgment of a message: any message is considered equally credible, independent of the degree of discrepancy with prior attitudes. Much social psychological research has shown that this is not realistic, because humans prefer information that is consistent with prior beliefs and they are less convinced by messages that strongly contradict their prior attitudes (e.g., Kunda, 1990; Nickerson, 1998; Sherif & Hovland, 1961). The basic idea for modelling this motivated cognition is that the evaluation of a message by the recipient also depends on the *discrepancy* with the receiver's prior conviction. The discrepancy is the absolute value of the difference between the message and the prior attitude  $|m - a|$ . To that end, we extend our model by scaling attitude change with a *motivated cognition factor*  $\lambda^k / (\lambda^k + |m - a|^k)$ , where we define the parameter  $\lambda > 0$  as the *latitude of acceptance*, and the exponent  $k > 0$  as the *latitude sharpness*.

Attitude change with motivated cognition (but without polarity effects) is thus

$$(7) \quad \Delta a = \frac{\lambda^k}{\lambda^k + |m - a|^k} \alpha(m - \rho a).$$

The motivated cognition factor is by definition between zero and one and diminishes attitude change when discrepancy increases. For a discrepancy equal to the latitude of acceptance  $\lambda$  the factor cuts down attitude change by half. Larger latitude sharpness exponents  $k$  make the function steeper around the latitude of acceptance. Figure 4 illustrates the meaning of the two parameters of the motivated cognition factor as a function of attitude discrepancy for five values of the latitude of acceptance and two different values for latitude sharpness.



The term “latitude of acceptance” stems from the social judgment theory of Sherif and Hovland (1961), and our operationalisation to Hunter et al. (1984) who present the attitude change in Equation (7) with  $k = 2$  (Hunter et al., 1984, p. 56, Eq. 4.1). We extended the function with large values of  $k$ . The functional form approaches a step function with full (zero) acceptance for discrepancies below (above) the latitude of acceptance, exactly matching the bounded confidence model discussed above (Hegselmann & Krause, 2002; Deffuant et al., 2000). It will turn out that the sharpness parameter, which we introduce as a novel generalisation of the equation of Hunter et al. (1984), can be crucial for qualitative differences with respect to fragmentation of the attitude landscape at the societal level.

***Modelling Source Credibility: Taking the source of the message into account***

Modelling the credibility of the source of the message is a major topic in persuasion research. The basic effect is that a message from a source with low credibility triggers less attitude change. Attitude change including source credibility is thus

$$(8) \quad \Delta a = s(i, j) \alpha (m - \rho a)$$

where  $0 < s(i, j) < 1$  is the credibility the receiver of the message  $i$  gives to the sender of the message  $j$ . Source credibility enables us to take into account the effect of different types of individuals. A typical assumption would be that an individual assigns full source credibility ( $s(i, j) = 1$ ) to messages from individuals from their ingroup but a lower value to individuals from outgroups, regardless of the content of the message. For example, individuals may assign a lower source credibility to individuals with a different gender, occupation, or ethnicity.

**Agent-Based Modelling: Societal Effects of Repeated Attitude-Change**

In this section, we describe the agent-based model we developed to study the implications of the micro-mechanisms of attitude change for the aggregated macro-level of society. We assume that all individuals hold attitudes about the same object, influence each other through

social interaction, or change their attitude idiosyncratically. Individuals are repeatedly exposed to the attitudes of others, which serve as messages triggering attitude change according to Equation (1) with a certain specification of the attitude change function based on a combination of Equations (5) - (8). The repeated exchange creates the possibility of systemic effects beyond the situation of one individual presented with a message. The full attitude change function is

$$(9) \quad \Delta a = s(i,j) \lambda^k / (\lambda^k + |m - a|^k) (M^2 - a^2) / M^2 \alpha (m - \rho a) \wedge (M - a) \vee (-M - a)$$

where  $s(i,j)$ ,  $\lambda^k / (\lambda^k + |m - a|^k)$ ,  $(M^2 - a^2) / M^2$ , and  $\alpha$  are factors with values between zero and one modelling source credibility, motivated cognition, polarity and general change strength. Very large  $\lambda$  and  $k$  would essentially be identical to ignoring the motivated cognition factor. The polarity term can be dropped if polarity effects are not considered. The essential part of the equation is the change term  $(m - \rho a)$ , which models contagion and assimilation. The final part with the minimum ( $\wedge$ ) and maximum ( $\vee$ ) ensure that the attitude change cannot drive the attitude  $a$  above or below the boundaries  $M$  and  $-M$ .

### ***Modelling Initial Attitudes***

We assume a group of  $N = 500$  individuals<sup>5</sup> with *initial attitudes* chosen idiosyncratically as random draws from a standard normal distribution  $a_i(0) \sim N(0,1)$ . The initial attitudes of all agents are thus centred around the neutral attitude, with half positive and the other half negative. Furthermore, 68% of initial attitudes are between  $-1$  and  $+1$ , 95% between  $-2$  and  $+2$ , 99.7% between  $-3$  and  $+3$ , and a tiny fraction more extreme. Random draws below or above the maximal absolute attitude are adjusted to  $-M$  or  $M$ , respectively. In the following,

---

<sup>5</sup> Our simulation also runs with 2,000 agents but much slower. We extensively explored whether the system behavior is also sufficiently captured with 500 agents.

we use  $M = 3.5$  as our parameter baseline. Thus, 99.95% of all initial attitudes are within the boundaries. The central panel in Figure 1 shows the probability density function of normal distribution with mean 5 and standard deviation  $5/3.5 \approx 1.43$  corresponding to the same setup within the eleven-point scale used by the European Social Survey.

### ***Modelling idiosyncratic attitude change: New random attitudes***

Individuals do not exclusively form attitudes through social influence but also through idiosyncratic reasoning (Pineda, Toral, & Hernandez-Garcia, 2009). An agent-based model of a whole society should take that into account. As we do not model idiosyncratic reasons for attitude formation explicitly, it makes most sense to model it analogously to initial attitudes as random draws from a normal distribution. The parameter controlling the influx of idiosyncratic attitudes is the *idiosyncrasy probability*  $0 \leq \theta \leq 1$ .

### ***Implementation of the Simulation***

The flowchart in Figure 6 depicts how the simulation is implemented. After initialisation, all agents update their attitudes in random order. When agent  $i$  is selected it chooses a new random attitude with idiosyncrasy probability  $\theta$ . Otherwise, the agent selects another random agent  $j$  for social influence. Agent  $i$  computes the attitude change  $\Delta a_i(t)$  taking the attitude of agent  $j$  as the received message  $m(t) = a_j(t)$ . This process repeats every time step.

The model is implemented in NetLogo 6.2, an agent-based modelling environment (Wilensky, 1999). The code is provided as supplemental material<sup>6</sup> (Lorenz, 2021). It enables easy replication of all simulations presented in the following.

---

<sup>6</sup> The NetLogo model can also be used for additional simulations to the selection described in the following section. Further parameters allow alternative modes of idiosyncratic attitude change, the option to restrict the selection of communication partners to a fixed social network of followers and friends (built upon initialization), and the option to introduce heterogeneous individual parameters in agents for  $\alpha$ ,  $\rho$ ,  $\theta$ , and  $\lambda$ . These can be used for robustness tests which we briefly discuss later. The details of their implementation go beyond the scope of the paper but can be tracked in the code provided.

### **Simulation Experiments: Exploration and Systematic Analysis**

We explored the model with simulation experiments. In every experiment, the goal was to characterise the distribution of attitudes at the societal level, contingent on the attitude-change model and parameters used at the individual level. The goal is to gain insights into the significance of attitude change mechanisms.

The general strategy of analysis followed three steps. First, we explored simulation runs for various parameter configurations to draw qualitative insights into the evolving distributions. The typical outcome under one parameter configuration is either a stable distribution or a typical dynamic pattern. Every simulation run in our model includes random events, in particular, when initial or idiosyncratic attitudes are chosen, and when partners of interaction are selected. This randomness does not prevent stochastically stable distributions or dynamic patterns from evolving at the macrolevel despite potentially large fluctuation at the individual level at the same time. Macroscopic regularities are thus clearly visible and measurable.

The second step was to define macroscopic output measures to characterise the different stable distributions and dynamic patterns properly. In our case, these are measures for bias, diversity, and fragmentation, explained below in detail. In a third step, we set up massive simulations by systematically sweeping through the parameter space and recording the average bias, diversity, and fragmentation for the characterisation of different outcomes.

Based on the results, we briefly discuss the impact of additional parameters not included in the systematic parameter sweep and some robustness tests with respect to two other ways to model idiosyncrasy, namely social networks restricting the selection of communication partners, and parameter heterogeneity between agents.

#### ***Exploration***

We show the main results of the exploration with the help of six examples which we use as an illustration to define macroscopic output measures. The first three examples displayed in Figure 6 show the evolution of extreme consensus, diversity, and the central consensus of society. These examples correspond to the classification examples 1-A, 1-B, and 1-C in Table 2 and in stylised form to the empirical examples in Figure 1. All three panels show the direct output of the NetLogo model (Lorenz, 2021). The large subpanel shows the trajectories of 500 agents evolving over 200 time steps (horizontal axis) and the attitude space (vertical axis) ranging from  $M = -3.5$  to  $3.5$ . The small numerical monitors above the main panel show the input parameters to the left and the output parameters (to be explained later) to the right. The small panels to the right of the main panel show the evolution of the output parameters over time, and the histogram of attitudes at the last time step.

Panel 1-A in Figure 6 shows how almost all agents converge on the same extreme attitude. The simulation had parameters  $\alpha = 0.2$ ,  $\rho = 0$ , and  $\theta = 0.01$ ; that is, only contagion with intermediate strength, rare idiosyncratic attitude formation, and no motivated cognition, no polarity effects, and no source credibility. The evolution of extremity took place as follows. For a short time period, the diversity of the initial condition prevailed, because agents received positive and negative messages with equal likelihood. Many messages were also close to zero and thus not very contagious. Further on, the possibility that a positive agent became more extreme through a positive message was balanced by the possibility that another positive agent became less positive through a negative message. However, this balance was unstable. A small imbalance towards one side emerged by chance and this imbalance quickly started to reinforce in a positive feedback loop because it became more likely to receive messages from one side. This created an unstoppable movement towards one extreme. Thus, almost all agents ended up maximally positive. It was equally likely that they could have ended up almost all maximally negative. In the following, we will call this

configuration ( $\alpha = 0.2$ ,  $\rho = 0$ ,  $\theta = 0.01$ , no motivated cognition, no polarity, equal source credibility) the “baseline case” and focus on deviations from it.

Panel 1-B in Figure 6 shows prevailing diversity without bias. In this simulation run, the only difference to the baseline case is the higher idiosyncrasy-probability parameter set at  $\theta = 0.17$ . In this configuration, there was no drift to one extreme, and thus the initial diversity of attitudes was maintained. A society driven by contagion ( $\rho = 0$ ) with strength  $\alpha = 0.2$  does not collectively drift to one extreme when at least 17% of all instances of attitude change are idiosyncratic. Panel 1-C in Figure 6 shows the evolution of a central rather than extreme consensus. The only difference from the baseline parameters was that these agents assimilated towards the messages they received ( $\rho = 1$ ). Agents adjusting their attitudes towards the attitudes of randomly chosen others naturally tend to decrease diversity.

These three examples show that the three fundamental outcomes of extreme consensus, central consensus, and diversity already provide simple variations of the degree of assimilation  $\rho$  and the idiosyncrasy probability  $\theta$ . Exploratory simulation showed the strength parameter  $\alpha$  to mainly alter the speed of changes, not introducing fundamentally different patterns. We will discuss its impact later.

Figure 7 shows three more examples. They show how bipolarisation and fragmentation emerge through motivated cognition. The simulation in Panel 2-A ran under the parameters of the baseline condition ( $\alpha = 0.2$ ,  $\rho = 0$ ,  $\theta = 0.01$ ), but included the motivated-cognition factor. The latitude of acceptance was  $\lambda = 0.5$  and the latitude sharpness  $k = 2$ . The messages whose evaluative content differed sharply from the receiver’s current attitude therefore had a lower proportional influence, compared to the simulations without motivated cognition displayed in Figure 6. The simulation shows how agents with a positive attitude were more affected by others with positive attitudes. As the mode of processing was contagion, these agents drifted to maximally positive attitudes. As influence decreased with

discrepancy the same could happen on the negative side of the attitude spectrum, driving society to bipolarisation. It is important to note that in the stable bipolarised state, agents were still minimally influenced by messages from the other side, but influence from their own side was always quickly driving them back to the extreme on their side.

Panel 2-B in Figure 7 shows a simulation where the parameters of Panel 2-A were altered in two ways. First, agents mostly assimilated with  $\rho = 0.9$ . Second, the latitude of acceptance was substantially sharper with  $k = 10$ . Similarly, in Panel 2-C the simulation ran with the same sharp motivated cognition but with full assimilation ( $\rho = 1$ ). In both simulations, agent attitudes fragmented into several clusters. These clusters are visible as peaks in the distribution of attitudes. The simulation in Panel 2-B leads to two large and pronounced peaks and three smaller and less pronounced peaks in between and surrounding them. In Panel 2-C a large central cluster and four successively smaller off-central clusters emerged. Such peaks emerged through assimilative forces similar to the central consensus in Panel 1-C of Figure 7. The sharp drop in influence that messages beyond the latitude of acceptance exerted on agents made it possible for several peaks to emerge and prevail. Typically, peaks have a distance of at least  $2\lambda$ , and thus a distance of about one in these cases. The mild degree of contagion in Panel 2-B made the peaks drift slowly outwards.

The visualisation with trajectories is limited in that it does not clearly show the density of agents in the attitude space. When many agents have the same attitude, there is no visual signature. Further, it is notable that the distribution of attitudes in panels 2-B and 2-C has not stabilised after the 200 time steps. To cope with both limitations, we show heatmaps in Figure 8, which display the agent density in the attitude space instead of trajectories. The figure additionally shows simulation runs for the parameter configurations 2-B and 2-C with 4,000 time steps; 20 times more than in Figure 7. The longer runs show that, indeed, the distributions of attitudes do not stabilise but show a persistent dynamic pattern.

With the small level of contagion ( $\rho = 0.9$ ) in 2-B, peaks close to the centre continued to emerge and then drift towards the extremes. Peaks drifted to the left and the right at almost regular intervals. Close to the maximal attitudes, peaks were usually already dissolved through idiosyncratic attitude change. Without contagion ( $\rho = 1$ ) in 2-C, two peaks occasionally merged, while new small peaks emerged off the centre to later merge with the more central peak. The pattern of peaks merging is less frequent and less regular than the pattern in 2-B.

We omitted the microscopic mechanisms of polarity and source credibility in the examples. This was because we did not find substantially different dynamics in the exploration. Indeed, running all six examples with added polarity effects would deliver essentially the same patterns<sup>7</sup>. Differences in source credibility between groups also needed to be quite drastic to change the evolving patterns.<sup>8</sup>

### ***Output Measures: Bias, Diversity, Fragmentation***

With the six handpicked examples of the previous subsection in mind, we developed three output measures to capture characteristics of attitude distributions in quantitative terms: *Bias*, *Diversity* and *Fragmentation*. We measure *Bias* as the deviation of the mean attitude from the neutral attitude and *Diversity* as the standard deviation of attitudes. Diversity can be used to distinguish consensual, diversified, and bipolarised distributions of attitudes. Our measure of *Fragmentation* is a bit more complex and builds on the standard Kolmogorov-Smirnov

---

<sup>7</sup> We found a mild effect of polarity in slowing down the drift to extremes in configurations like 2-B. Further on, when the maximal attitude  $M$  is much less (e.g.  $M = 2$ ) then a larger fraction of attitudes is idiosyncratically extreme. When one of the two extreme peaks is a bit larger by chance, the polarity factor together with greater strength ( $\alpha = 0.2$ ) can bring even assimilating agents to exhibit an emerging substantial, though not extreme, bias. A deeper analysis of these more specific effects is beyond the scope of the paper. One example is provided in the Supplemental Material in Figure S3.

<sup>8</sup> For example, the evolving patterns were essentially the same, when we labeled agents in two groups, defined their initial and idiosyncratic attitudes such that groups averages differed substantially by one, and set source credibility between groups to only half the credibility within the group. In the NetLogo model, this source credibility configuration is called `intergroup_credibility = 0.5` and `initial_groupspread = 1`. Figure S4 in the Supplemental Material gives two examples where source credibility begins to matter under more drastic assumptions.



probability density function<sup>9</sup> which can be numerically computed for each set of many attitude values. To compute *Fragmentation*, we integrate the absolute value of the first derivative of the probability density function numerically. This measures how much the density function goes up and down. Thus, a distribution with several peaks shows higher fragmentation as a normal distribution with the same standard deviation.

All three measures are by definition nonnegative. Further on, we normalise them such that the maximal possible value equals one. This is possible because all attitude distributions are confined between  $-M$  and  $M$ . Thus, *Bias*, *Diversity*, and *Fragmentation* assume only values between zero and one. The mathematical definitions are given in Table 1. In Figures 6 and 8 the three output measures are tracked over time in the upper right subpanel of each panel. Under some configurations with characteristic dynamic patterns instead of a stable distribution, the output measures may fluctuate heavily over time. We therefore measure *Bias*, *Diversity* and *Fragmentation* at every time step (omitting the first time steps as a spin-off phase) and average over time to characterise the emerging attitude landscapes resulting from a parameter configuration.

---

<sup>9</sup> We implemented density with Gaussian kernels and bandwidth of 0.1. The support of the distribution function is discretised by steps of 0.01.

Table 1. Output measures for a set of  $N$  attitudes  $\{a_1, \dots, a_N\}$ .

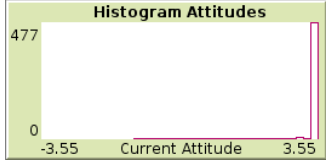
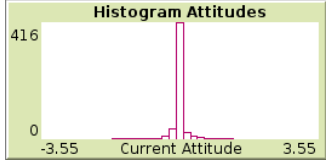
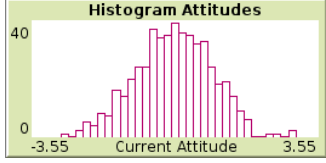
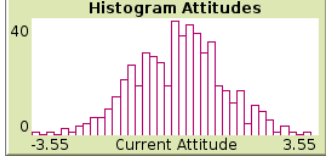
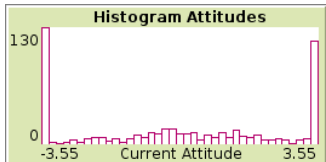
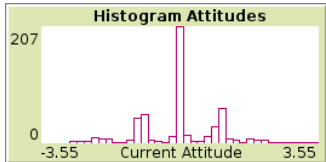
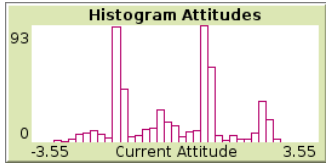
Measure	Equation and Computation	Description	Maximum Achieved for
<b><i>Bias</i></b>	$1/N \cdot  \sum a_i  / M$	Deviation of the mean attitude from neutral	$a_i = M$ for all individuals, or $a_i = -M$ for all individuals
<b><i>Diversity</i></b>	$(\frac{1}{N} \sum (a_i - \underline{a})^2)^{1/2} / M$	Dispersion of attitudes around the mean attitude (std. deviation divided by $M$ )	$a_i = M$ for half of the individuals and $a_j = -M$ for the other half
<b><i>Fragmentation</i></b>	$(\int  p'(a)  da) \cdot \sigma / M / 4$ with $p'$ the first derivative of the probability density function of attitudes, $\sigma$ the standard deviation; computed numerically on the ks-density (bandwidth 0.1, support discretised with $da=0.01$ )	Measure how much the attitude density function goes up and down.	$a_i = M$ for half of the individuals and $a_j = -M$ for the other half

**Note:** *Bias* and *Diversity* output measures range by definition from 0 to 1 whereby 1 represents the maximal possible value which can be achieved for distributions living on the interval  $[-M, M]$ . The same has been achieved by numerical testing for *Fragmentation*.

### ***Classification of Macroscopic Phenomena***

We use the *Bias*, *Diversity*, and *Fragmentation* measures to distinguish extreme, centrally consensual, condensed, diversified, and bipolarised attitude distributions. Moreover, for condensed and diversified distributions we distinguish fragmented from non-fragmented distributions (see Table 2 and Fig. 9 for precise definitions. We consider distributions with a *Bias* larger than  $1/M = 1/3.5 \approx 0.257$  (i.e., more than 1 standard deviation different from the mean of the initial distribution) as extreme. We found that it is of limited interest to further distinguish extreme distributions with respect to *Diversity* or *Fragmentation*. We classify the non-extreme distributions further by increasing *Diversity* as “centrally consensual”, “condensed”, “diversified”, and “bipolarised”. We set the threshold value between “condensed” and “diversified” at the *Diversity* of initial estimates ( $1/M$ ) and the consensual-condensed transition at three quarters of it ( $0.75/M$ ). The maximally possible *Diversity* represents a fully bipolarised distribution with one half of the population at each extreme attitude  $M$  and  $-M$ . We set the diversified-bipolarised threshold at the diversity measure a uniform distribution would have which is  $1/\sqrt{3} \approx 0.577$ . Finally, we distinguish a “fragmented condensed” and “fragmented diversified” distribution from their non-fragmented counterparts when *Fragmentation* is larger than *Diversity*. We found that it is of limited interest to further distinguish consensual and bipolarised distributions with respect to high or low *Fragmentation*. With these seven types of macroscopic outcomes we systematically explored the parameter space with two simulation experiments, one focusing on the transitions between diversity, extreme and central consensus, and the other focusing on the transition between extreme consensus, bipolarisation, and fragmented distributions.

Table 2. Classification of attitude distributions based on *Bias*, *Diversity*, and *Fragmentation* (see Table 1 for their definition).

Type	Definition and Description	Example from Simulation	Figs.
<b>Extreme</b>	$Bias > 1/M$ Bias larger than one standard deviation of the initial attitudes.		6, 8, 10 (1-A)
<b>Central Consensus</b>	$Diversity < 0.75/M$ , $Bias \leq 1/M$ Low <i>Bias</i> . <i>Diversity</i> lower than 75% of <i>Diversity</i> of initial attitudes.		6, 8, 10 (1-C)
<b>Condensed</b>	$0.75/M < Diversity < 1/M$ , $Bias \leq 1/M$ , $Fragmentation \leq Diversity$ Low <i>Bias</i> . Distribution as initial attitudes or slightly more condensed.		Normal
<b>Diversified</b>	$1/M < Diversity < 1/\sqrt{3}$ , $Bias \leq 1/M$ , $Fragmentation \leq Diversity$ Low <i>Bias</i> . <i>Diversity</i> larger than <i>Diversity</i> of initial attitudes but lower than uniform distribution.		6, 8, 10 (1-B)
<b>Bipolarised</b>	$Diversity > 1/\sqrt{3} \approx 0.577$ , $Bias \leq 1/M$ Low <i>Bias</i> . <i>Diversity</i> larger than <i>Diversity</i> of a uniform distribution.		7, 8, 11 (2-A)
<b>Fragmented Condensed</b>	$1/M < Diversity < 1/\sqrt{3}$ , $Bias \leq 1/M$ , $Fragmentation > Diversity$ As <b>Condensed</b> , but with <i>Fragmentation</i> above <i>Diversity</i> .		7, 8, 12 (2-C)
<b>Fragmented Diversified</b>	$0.75/M < Diversity < 1/\sqrt{3}$ , $Bias \leq 1/M$ , $Fragmentation > Diversity$ As <b>Diversified</b> , but with <i>Fragmentation</i> above <i>Diversity</i> .		7, 8, 12 (2-B)

Notes: Concrete values for  $M = 3.5$  are  $0.75/M \approx 0.214$ ,  $1/M \approx 0.257$ , and  $1/\sqrt{3} \approx 0.577$ .

***Simulation Experiment 1: Extremity, Diversity, and Central Consensus***

In a first experiment we explored the basic model behaviour with respect to the degree of assimilation in 51 steps  $\rho = 0, 0.02, \dots, 1$ ; the idiosyncrasy probability in 31 steps  $\theta = 0, 0.01, \dots, 0.3$ ; and with lower precision of four steps the strength of attitude change  $\alpha = 0.1, 0.2, 0.3, 0.4$ . As in examples 1-A, 1-B, and 1-C we did not implement motivated cognition, polarity effects, and source credibility. For each of the  $51 \times 31 \times 4 = 6,324$  parameter combinations we ran a simulation with  $N = 500$  agents for 10,000 time steps and omitted the first 2,500 time steps (spin-off phase) in the computation of the average *Bias*, *Diversity*, and *Fragmentation*. Figure 10 shows the types of macroscopic outcomes in the  $\theta$ - $\rho$ -plane for strength  $\alpha = 0.2$ . This includes configurations 1-A, 1-B, and 1-C of Figure 6.

This large-scale simulation experiment shows that in a society where attitude change is driven by contagion, moving the idiosyncrasy probability from 0.14 to 0.17 drastically changes the outcome from an extreme to an on average neutral and slightly diversified distribution. A similar change happens when the idiosyncrasy probability remains at 0.14, but the degree of assimilation increases from 0 to 0.25. The society then changes to a neutral but condensed distribution. A deeper analysis of *Bias* and *Diversity* (not visible in Figure 10) shows that the transition to extremity is rather abrupt, while the transitions between consensual, condensed and diversified distributions are gradual.

Figure S1 in the Supplemental Material shows the impact of the strength parameter  $\alpha$ . With increasing  $\alpha$ , the region of extremity grows. For example, the transition to neutral diversity under contagion ( $\rho = 0$ ) appears only for idiosyncrasy probability  $\theta = 0.25$ , when the strength is  $\alpha = 0.4$ . Moreover, among the non-extreme regions the central consensus region also increases with  $\alpha$ . A particularly interesting insight from these results is that a call for less idiosyncrasy might be intended to make a society more consensual around a neutral

(more balanced and thus more rational) attitude, however, this can tip a society to become more extreme.

***Simulation Experiment 2: Motivated Cognition Triggers Bipolarisation and Fragmentation***

Our second experiment explored the model behaviour under motivated cognition with respect to the degree of assimilation in 51 steps  $\rho = 0, 0.02, \dots, 1$ ; the latitude of acceptance in 40 steps  $\lambda = 0.05, 0.1, \dots, 2$ ; and (with lower precision) the latitude sharpness with five values  $k = 2, 3, 4, 5$ , and 10. The idiosyncrasy probability is held constant at  $\theta = 0.01$  and the strength at  $\alpha = 0.2$ . As in examples 2-A, 2-B, and 2-C we ignore polarity effects and source credibility. For each of the  $51 \times 40 \times 5 = 10,200$  parameter combinations we ran a simulation with  $N = 500$  agents for 10,000 time steps and omitted the first 2,500 time steps (spin-off phase) in the computation of the average *Bias*, *Diversity*, and *Fragmentation*, as in Simulation Experiment 1. Figure 11 shows the macroscopic outcomes in the  $\lambda$ - $\rho$ -plane for the low level of the latitude sharpness of  $k = 2$ . This includes configuration 2-A of Figure 7.

Figure 11 shows that motivated cognition introduces the new phenomenon of bipolarisation, which appears for latitudes of acceptance up to one and low to moderate degrees of assimilation. The sharp transition between central consensus and extremity with respect to the degree of assimilation at  $\rho \approx 0.95$  remains largely unaffected by decreasing latitudes of acceptance. The persistence of one-sided extremity evolving even with fairly high degrees of assimilation is particularly interesting. For example, for  $\rho = 0.8$  and latitudes of acceptance as low as  $\lambda = 0.5$ , the society drifts collectively to one side. In the bipolarised situation of configuration 2-A, for example, a call for more assimilation might tip the society to one-sided extremity, while central consensus would only be achieved for even higher degrees of assimilation with almost no contagion any more ( $\rho > 0.95$ ).

Figure 12 shows the same  $\lambda$ - $\rho$ -plane but for simulation runs with higher latitude sharpness of  $k = 10$ . This includes the configurations of examples 2-B and 2-C from Figure 6. Figure 12 shows the phenomena of diversified and condensed fragmented attitude distributions. Further on, the whole region of extremity disappeared for the benefit of bipolarisation. Fragmented distributions appear when agents assimilate ( $\rho = 1$ ) for latitudes of acceptance  $\lambda < 0.75$ . When agents mostly assimilate but are also exposed to some contagion, the diversified fragmented distribution also appears for much larger latitudes of acceptance. Figure S2 in the Supplemental Material shows the impact of latitude sharpness also for  $k = 3, 4$ , and  $5$ .

### ***Impact of Polarity Effects and Source Credibility***

Our explorations showed that the effects of polarity and source credibility are not particularly dominant. In particular, all the macroscopic results above remain largely the same with polarity effects as well as with a relatively strong effect from source credibility between two groups with differing average initial attitudes, as pointed out in Footnotes 7 and 8. This can be tested using the NetLogo model we provide (Lorenz, 2021). In particular, all examples for Figures 5 and 6 look similar when additional polarity effects are added or when a non-extreme form of source credibility is added. Only when two groups have strongly different initial and idiosyncratic attitudes and assign each other much lower source credibility can this change the macroscopic dynamics. In such situations of extremely different groups almost not trusting each other, it is also possible to see bipolarisation through contagion without motivated cognition. When agents assimilate, differences between groups must be extreme and intergroup source credibility must be extremely low, so there is no convergence to consensus but stable bimodal distributions. Examples are provided in the Supplemental Material.

Of course, polarity and source credibility are very relevant to the attitude change of an individual, and initial group differences might also trigger differences between the groups in the stable outcomes. For example, when two different groups bipolarise, of course, the group with more negative initial attitudes is over-represented in the final negative extreme peak. However, there are no significant differences at the macroscopic level, implying that source credibility and polarity are probably not the core drivers of emerging fragmentation and bipolarisation in societies.

### *Additional Analyses with the Model*

Here, we briefly report exploratory findings on the robustness of the results regarding alternative modes of idiosyncratic attitude change, network restrictions for the selection of communication partners, and individual parameter heterogeneity of agents. We found that all the effects presented above essentially remain.

Idiosyncratic attitude change can also be modelled as switching back to the initial attitude instead of drawing a new random one; or idiosyncrasy may be interpreted as a weight with which the initial attitude influences attitude change (see Friedkin & Johnsen, 1990). The first leaves all our six examples (1-A to 2-C) essentially unchanged. The latter does the same with two small exceptions. It may lead to a higher bias in example 1-B. In 2-B it induces a less dynamic pattern that is closer to bipolarisation.

The choice of agents' communication partners can be restricted to neighbours from a fixed social network. Our model provides such a network where each agent has a specific number of unidirectional "follower" links and bidirectional "friendship" links. The directed follower links are constructed successively using preferential attachment, and the friendship links are either a random graph, a ring graph, or in cliques. Using a combination of these we



can model social networks with many realistic features such as scale-free distributions of the number of followers, small-world properties, and clustering (see Newman, 2003). The existence of agents with a large number of followers can be seen as a rudimentary implementation of mass media. We mainly explored simulations on networks where each agent follows five others and has five friends. All examples show the same patterns as the simulations without network restrictions. There are some deviations for networks with unrealistically low degrees. Naturally, higher degrees (and thus even more realistic setting) should bring the model even closer to our simulated setting where everyone could potentially interact with every other person. We therefore also expect no change in our results.

Finally, individual heterogeneity of  $\alpha$ ,  $\rho$ ,  $\theta$ , and  $\lambda$  is implemented by assuming a Beta distribution instead of the same value for all. The Beta distribution is used because all four parameters are theoretically bounded between zero and one. We parameterised them by their mean, which coincides with the former homogeneous value and their standard deviation. With moderate dispersion around the average, we find all emerging patterns in the six examples similar to the homogeneous case. Stronger heterogeneity seems to trigger new systemic effects, but their exploration goes beyond the scope of this paper. In conclusion, the individual heterogeneity of parameters seems to be the extension with the strongest impact on macroscopic outcomes, much more than the type of idiosyncrasy and an underlying static social network structure.

### **First Attempt at Comparative Model Evaluation with Macrodata**

After understanding the impact of model parameters on simulation dynamics, the next step for validation is the assessment of the model's capacity to replicate empirical survey data at the macro-level, preferably in comparison with competing models (see Flache et al., 2017).

The macro-level validation of an attitude-dynamics ABM is a daunting task, however, which to the best of our knowledge has never been achieved due to the sheer complexity of communication flows in society. Doing this in a strictly quantitative manner would require common macroscopic measures for simulation output and empirical data, and reliable routines for tuning model parameters to best fit empirical outcomes. In principle, the measures of Bias, Diversity, and Fragmentation could also be computed for the empirical attitude landscapes, however, they do not quantify all aspects of empirical landscapes appropriately. Nevertheless, the macro-data in Figure 1 allows a qualitative comparative assessment to at least some degree if we focus on exploratory data analysis with the aim of finding and quantifying “stylised facts” (see Meyer, 2019). The labels 1-A to 2-C serve as guidelines regarding which empirical observation in Figure 1 can be best compared to which parameter constellations in our model exploration as outlined in Table 2.

In order to compare our ABM based in psychological attitude theory with established models more detached from psychology, we use the bounded confidence model (Deffuant et al., 2000; Hegselmann & Krause, 2002), the model by Jager and Amblard (2005), which extends the bounded confidence model with repulsive effects when opinions of others fall into a latitude of rejection, and the model by Mäs and Flache (2013) with arguments exchanged under homophilous choice of communication partners. The possible outcomes of the bounded confidence model are one, two, three, or more peaks with distances of roughly twice the bound of confidence between adjacent peaks, no peaks at the extremes and no agents with other attitudes.<sup>10</sup> Peaks are internally maximally condensed. The model by Jager and Amblard (2005) additionally produces bipolarisation with sizable peaks at the extremes that have larger distances to the first neighbouring central peak. Mäs’ and Flache’s (2013)

---

<sup>10</sup> In the outcomes for the model by Deffuant et al. (2000) there are peaks at the extremes and between peaks but are typically a thousand times smaller in magnitude than the large peaks. This cannot be represented in the empirical data anyway.

model also has perfect bipolarisation as potential outcome as well as a perfect consensus which can lie anywhere in attitude space.

Our first observation, comparing model outcomes to the eight empirical attitude distributions in Figure 1, is that these never show distributions with totally empty attitude categories, as would be implied by the three models. In reality, there are no fully condensed peaks with gaps without any agents holding attitudes in between. There are three types of possible explanation. In addition to the model misspecifying something, it is also possible that the real world-distribution is still in transition and not yet converged, or that the underlying process of the attitudes expressed in the survey is not solely a result of social influence dynamics. The transition hypothesis is unlikely, because at least some fully converged attitude distributions should be observable somewhere. The model component of idiosyncratic attitude change in our model endogenises the idea that repeated attitude formation through social influence is not the only source of attitude formation. Our model is therefore better able to produce distributions with “blurry” peaks as in Figure 1 Panel 2-C, or a major central (1-C) or extreme consensus (1-A) co-existing with several minor other attitudes. Higher degrees of idiosyncratic attitude formation could reproduce unimodal diversified attitude distributions as in Panel 1-B. The location of the modes in the two Norwegian examples in Panels 1-A and 1-B show either larger or smaller bias. Further analysis will show whether our model can reproduce these. The clear bipolarisation that our model, the model of Jager and Amblard (2005), and the model of Mäs and Flache (2013), are all able to produce, does not show up in empirical data. The closest in the European Social Survey is the example in Panel 2-A. In addition to small amounts of all possible attitudes it also shows a sizable central peak along with the two dominating extreme peaks. The model by Jager and Amblard can produce this situation without the intermediate peaks as a stable

outcome. In our model it appears as a transient situation in parameter constellations which remain changing over time, as shown in Figure 8 (2-B and 2-C).

Overall it appears that all the models, including ours, fail by not reproducing at least some empirical macro data, but ours seems slightly more realistic in generating the noise in observed attitude distributions. As we explain in the subsequent discussion, future validation research will require a complex strategy of calibrating model parameters with micro-experiments systematically aligned with corresponding survey data at the macro-level.

### **Discussion**

Our simulation experiments substantively show that the bipolarisation and fragmentation of attitude landscapes are both driven by motivated cognition, meaning the diminishing influence of discrepant attitudes. The degree of assimilation determines whether there is fragmentation or bipolarisation. Even a small degree of contagion can tip societal dynamics from fragmenting to bipolarising. Fragmentation through assimilation can also only occur with sharp enough latitudes of acceptance. A certain degree of idiosyncratic attitude change may prevent radicalisation or bipolarisation. Central consensus is achieved with a very high degree of assimilation and if fragmentation is prevented by wide or non-sharp latitudes of acceptance. A central consensus is always at risk of becoming an extreme consensus with a decrease in the degree of assimilation in favour of contagion. The drift to the extreme can be prevented to some extent by greater idiosyncrasy.

The particular modelling approach used here has potential for psychological science in general. The simulations enabled “experimenting on theories” (Dowling, 1999; Troitzsch, 2017) by implementing different theoretical assumptions within the framework of one simulation model. Theory informs model construction, and the simulation in turn provides information for theory development. This holds especially true for testing the macro-level

implications of micro-level theories. As the simulations reported above show, the different parameter constellations of the model can yield widely different outcomes at the macro-level, perhaps even to the point where the connection between micro mechanisms and macro-level empirical distributions of attitudes may appear arbitrary. Neither verbal theory nor social simulation research alone can transform such apparent arbitrariness into reliable and testable predictions of the macro-effects of social psychological attitude-change mechanisms. If the magnitudes of parameters matter greatly, as our simulations imply, then experiments could be of help in determining realistic, empirically grounded bounds for the parameters, which would in return enhance the realism of simulations in a recursive interplay between theory, data, and simulation. For a detailed discussion of the epistemology of simulation see, for example, Edmonds (2017), Morgan and Morrison (1999), Humphreys (2004), Knuuttila (2011), and DeLanda (2015).

A computational model such as ours could be taken as theoretical guidance for experimental research in psychology, overcoming the questionable practice of building empirical research programs on incoherent folk theories and intuitions (see Muthukrishna & Henrich, 2019). The simulation experiments provide access to new types of evidence for testing social psychological theories beyond the experimental approach at the level of individual cognition. Simulation experiments allow the effects of the different psychological mechanisms to be studied at the aggregated social macro level, overcoming the vagueness of the information that experimental social psychology provides at the aggregated social macro level. The simulation method thus provides a mediator (Morrison, 1999) between psychology, which makes assumptions at the micro level of the individual, and the aggregated macro level studied by sociology (Squazzoni, 2012). In fact, the results of our experiments show that different micro-level mechanisms do produce different macro-level dynamics. They also show that the macro-level implications are sensitive to some

parameters, but not to others, implying that the simulation method provides more solid constraints on understanding the micro-macro link than verbal theorising alone.

The information about the implications of theoretical assumptions provides new sources for evaluating theory, complementing behavioural experiments as the dominant method in psychology (see Muthukrishna & Henrich, 2019). Here we focus on integrating and evaluating at a theoretical level. Experimenting with theories might become useful for empirically assessing what mechanisms are at work in society, because the simulation experiments provide a means for cross-validation by integrating empirical findings at the micro and the macro level. On the one hand the psychological theories at the micro-level are substantiated by experimental tests, and on the other hand sociological survey data cannot reveal the micro mechanisms at work, but they can show what part of the theoretically possible distributions are empirically realised. As Epstein (2006) remarked, generating a macro-level phenomenon from certain micro-level assumptions provides only a candidate explanation, and there may be other ways of generating the very same macro-phenomenon. It is difficult - if not impossible - to judge which of two candidate explanations is better just from the macroscopic match. For example, bipolarised extreme peaks can be produced by contagion and motivated cognition (as in our model) or repulsive influence (see Jager & Amblard, 2005) as discussed above. However, the psychologically informed mechanisms of our model are more reliable than assumptions which are not based on theoretical influence from psychology, as the sharp bound of confidence, even if they generate similar phenomena, because the model assumptions are validated independently by psychological

experiments, stimulating an iterative research process from theory to simulation and vice versa (Conte, 2009).<sup>11</sup>

One conclusion from our study is that the bipolarisation of political attitudes in Western societies (e.g., DiMaggio, Evans, & Bryson, 1996; Fiorina & Abrams, 2008; Bramson et al., 2016) might indicate a combination of contagion and motivated cognition, since the combination of these mechanisms generates similar macro level patterns even when contagion and motivated cognition are relatively mild. In contrast, polarity effects as well as groups assigning each other low credibility do not generate bipolarisation even under fairly strong assumptions. These mechanisms are thus more likely to be outcomes and intensifiers instead of the origins of bipolarisation. This example provides hints of how the integration of micro-level experimental evidence and macro-level survey data is made possible by capitalising on simulation models. Choosing among competing candidate explanations enables an inference of the best (or at least most credible) explanation (Lipton, 2000; Beisbart, 2019). Bearing in mind Epstein's generativist epistemology, this finding enables further cycles of iterative research to check for other potential factors of influence, such as social norms or additional dynamics of attitude strength.

We think, model-driven approaches to experimenting might help alleviate some of the current replicability problems in experimental psychology, which are not only attributable to bad statistical practices but, arguably, also to a lack of coherent and precisely formalised theory (see also Fried, 2020; Devezzer, Nardin, Baumgaertner, & Buzbas, 2019; Muthukrishna & Henrich, 2019).

### ***Limitations of Our Approach and Outlook for Future Work***

---

<sup>11</sup> A particular challenge in the process of excluding micro-assumptions as explanation when they do not generate a phenomenon is that the same micro-assumption may become relevant in future extensions of the model when combined with other newly implemented mechanisms.

Of course, the ambitious goal of integrating a rich landscape of theories of attitude change across the micro and macro levels is subject to many limitations in practice. One price we pay for the generality of the model is a relative lack of detail concerning information-processing mechanisms, especially when compared with contemporary connectionist or even neurocomputational simulations (e.g., Ehret et al., 2015; Schröder & Wolf, 2017), which model the biological properties of the human brain more closely than the abstract approach in our present model of attitude change. Computational modellers always struggle with the trade-off between principles of EROS (enhancing the realism of simulations) versus KISS (keep it simple, stupid!), as expressed in the famous story of a map that was as detailed as the empire itself, rendering it utterly useless for precisely its detail (Borges, 1946; see Jager, 2017 for discussion). We have sought a middle ground between EROS and KISS by implementing much greater social psychological realism than most social-physics models, while shying away from the basic connectionist mechanisms that become harder to interpret in terms of established verbal theories. As we have shown above, both abstract and connectionist implementations of attitude theories go back intellectually to the same origins in cognitive consistency theories (Read & Simon, 2012); and we thus do consider our approach compatible with connectionism and computational neuroscience, although the level of description is less detailed and more oriented towards verbal interpretability. Our modelling approach is abductive in explaining the real world following the principle of Occam's razor: we search for sufficient but the most parsimonious solutions. A concrete benefit of model parsimony is that a theoretical understanding of the system's dynamics is easier.

A second limitation in terms of detail of the model (i.e., more KISS than EROS) is the way we modelled communication between agents, mostly disregarding realistic properties of social networks such as homophily, the well-known situation where people tend



to communicate more with people similar to them in terms of attitudes (McPherson, Smith-Lovin, & Cook, 2001). Homophily can also be considered a driver of attitude polarisation, as people avoid being exposed to evaluative messages discrepant to their own attitudes. It could thus be argued that motivated cognition at the individual level (as implemented in our model) and homophily at the level of social structure mutually reinforce each other; an additional twist on complexity, which we chose to ignore in our model in service of parsimony.

A third limitation stems from the fact that our model considers one attitude in isolation. More recent research on the malleability and stability of attitudes has pointed to the importance of bundles or networks of attitudes; for example, when vaccine sceptics recently tended to align with organic food consumption (Goldberg & Stein, 2018) or when liberals drink lattes (DellaPosta, Shi, & Macy, 2015). At the individual level, the more embedded an attitude is in an entire network of attitudes, the more resilient it is to change (Dalege et al., 2016) - a fact not reflected in our unidimensional attitude model. At the societal level (DellaPosta, 2020; Goldberg & Stein, 2018), the bundling of attitudes seems to be related to increasing opinion polarisation, which further complicates matters.

### ***Conclusion***

We developed a mathematical model of attitude change that integrates diverse theories in social psychology in a coherent manner, and thus makes them applicable in agent-based models aimed at elucidating societal dynamics of attitude change and opinion polarisation. Simulations with our model showed the complexity involved in scaling individual-level mechanisms of attitudes to society level, and can be taken to guide future research about micro-macro connections in attitude change. With this work, we hope to contribute to closing the gap between social psychology and the interdisciplinary computational social sciences,

opening up new ways of understanding societal phenomena based on established social psychological mechanisms in a networked society.

### References

- Anderson, N. H. (1971). Integration theory and attitude change. *Psychological Review*, 78(3), 171-206.
- Asch, S. E. (1948). The doctrine of suggestion, prestige, and imitation in social psychology. *Psychological Review* 55, 250-276.
- Banisch, S., & Olbrich, E. (2019). Opinion polarization by learning from social feedback. *The Journal of Mathematical Sociology*, 43(2), 76-103.
- Beisbart, C. (2019). What is validation of computer simulations? Towards a clarification of the concept of validation and related notions. In C. Beisbart, N. Saam (eds.), *Computer simulation validation. Fundamental concepts, methodological frameworks, and philosophical perspectives* (pp. 35-67). Cham: Springer.
- Borges, J. L. (1946). Del rigor en la ciencia. *Los Anales de Buenos Aires*, 1(3), 53.
- Brousmiche, K. L., Kant, J. D., Sabouret, N., & Prenot-Guinard, F. (2016). From beliefs to attitudes: Polias, a model of attitude dynamics based on cognitive modeling and field data. *Journal of Artificial Societies and Social Simulation*, 19(4).
- Bramson, A., Grim, P., Singer, D. J., Fisher, S., Berger, W., Sack, G., & Flocken, C. (2016). Disambiguation of social polarization concepts and measures. *The Journal of Mathematical Sociology*, 40(2), 80-111.
- Cantril, H. (1944). *Gauging public opinion*. Princeton: Princeton University Press.
- Cartwright, D., & Harary, F. (1956). Structural balance: A generalization of Heider's theory. *Psychological Review*, 63(5), 277-293.
- Cartwright, N. (1999). Models and the limits of theory: Quantum Hamiltonians and the BCS models of superconductivity. In M. Morgan, M. Morrison (eds.), *Models as mediators. Perspectives on Natural and Social Science* (pp. 241-281). Cambridge: Cambridge University Press.

- Chattoe, E. (2014). Using agent based modelling to integrate data on attitude change. *Sociological Research Online* 19(1).
- Conte, R. (2009). From simulation to theory (and backwards). In F. Squazzoni (ed.), *Epistemological aspects of computer simulation in the social sciences* (pp. 29-47). Berlin: Springer.
- Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., ... & Nowak, A. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1), 325-346.
- Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. (2016). Toward a formalized account of attitudes: The Causal Attitude Network (CAN) model. *Psychological Review*, 123(1), 2-22.
- Dandekar, P., Goel, A., & Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15), 5791-5796.
- Deffuant, G., Amblard, F., Weisbuch, G., & Faure, T. (2002). How can extremism prevail? A study based on the relative agreement interaction model. *Journal of Artificial Societies and Social Simulation*, 5(4).
- Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04), 87-98.
- DeLanda, M. (2015). *Philosophy and simulation. The emergence of synthetic reason* (2<sup>nd</sup>ed). London: Bloomsbury.
- DellaPosta, D. (2020). Pluralistic collapse: The “Oil Spill” model of mass opinion polarization. *American Sociological Review*, 85(3), 507-536.
- DellaPosta, D., Shi, Y., & Macy, M. (2015). Why do liberals drink lattes? *American Journal of Sociology*, 120(5), 1473-1511.

- Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PloS One*, 14(5), e0216125.
- DiMaggio, P., Evans, J., & Bryson, B. (1996). Have Americans' social attitudes become more polarized? *American Journal of Sociology*, 102(3), 690-755.
- Dodds, P. S., & Watts, D. J. (2005). A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 232(4), 587-604.
- Dowling, D. (1999). Experimenting on theories. *Science in Context*, 12(2), 261-273.
- Edmonds, B. (2017). Different modelling purposes. In B. Edmonds, R. Meyer (eds.), *Simulating social complexity* (pp. 39- 58). Cham: Springer.
- Ehret, P. J., Monroe, B. M., & Read, S. J. (2015). Modeling the dynamics of evaluation: A multilevel neural network implementation of the iterative reprocessing model. *Personality and Social Psychology Review*, 19(2), 148-176.
- Epstein, J. (2006). *Generative social science. Studies in agent-based computational modelling*. Princeton: Princeton University Press.
- ESS Round 9: European Social Survey Round 9 Data (ESS) (2018). *Data file edition 3.0*. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC. doi:10.21338/NSD-ESS9-2018.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Palo Alto: Stanford University Press.
- Fiorina, M., & Abrams, S. (2008). Political polarization in the American public. *Annual Review of Political Science*, 11, 563-588.
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading: Addison-Wesley.

- Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of social influence: Towards the next frontiers. *Journal of Artificial Societies & Social Simulation*, 20(4).
- Fried, E. I. (2020). *Lack of theory building and testing impedes progress in the factor and network literature*. <https://doi.org/10.31234/osf.io/zg84s>
- Friedkin, N. E., & Johnsen, E. C. (1990). Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4), 193-206.
- Goldberg, A., & Stein, S. K. (2018). Beyond social contagion: Associative diffusion and the emergence of cultural variation. *American Sociological Review*, 83(5), 897-932.
- Gollob, H. F. (1968). Impression formation and word combination in sentences. *Journal of Personality and Social Psychology*, 10, 341-353.
- Harari, Y. N. (2016). *Homo deus: A brief history of tomorrow*. London: Penguin Random House.
- Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis and simulation. *Journal of Artificial Societies and Social Simulation* 5(3).
- Hegselmann, R. (2017). Thomas C. Schelling and James M. Sakoda: The intellectual, technical, and social history of a model. *Journal of Artificial Societies and Social Simulation* 20(3).
- Heider, F. (1946). Attitudes and cognitive organization. *Journal of Psychology*, 21, 107-112.
- Heise, D. R. (1969). Affectual dynamics in simple sentences. *Journal of Personality and Social Psychology*, 11(3), 204-213.
- Heise, D. R. (1979). *Understanding events: Affect and the construction of social action*. New York: Cambridge University Press.
- Heise, D. R. (2007). *Expressive order. Confirming sentiments in social action*. New York: Springer.

- Helbing, D., & Balietti, S. (2012). Agent-based modeling. In D. Helbing (ed.), *Social self-organization* (pp. 25-70). New York, NY: Springer.
- Hovland, C. I., Janis, I. L., and Kelley, H. H. (1953). *Communication and persuasion*. New Haven: Yale University Press.
- Hovland, C. I., Harvey, O. J., and Sherif, M. (1957). Assimilation and contrast effects in communication and attitude change. *Journal of Abnormal and Social Psychology* 55, 242-252.
- Humphreys, P. (2004). *Extending ourselves. Computational science, empiricism, and scientific method*. Oxford. Oxford University Press.
- Hunter, J., Danes, J., & Cohen, S. (1984). *Mathematical models of attitude change Vol. 1*. Orlando: Academic Press.
- Hutchinson, B. (1949). Some problems of measuring the intensiveness of opinion and attitude. *International Journal of Opinion and Attitude Research* 3, 123-131.
- Jackson, J. C., Rand, D., Lewis, K., Norton, M. I., & Gray, K. (2017). Agent-based modeling: A guide for social psychologists. *Social Psychological and Personality Science*, 8(4), 387-395.
- Jager, W. (2017). Enhancing the realism of simulation (EROS): On implementing and developing psychological theory in social simulation. *Journal of Artificial Societies and Social Simulation*, 20(3).
- Jager, W., & Amblard, F. (2005). Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational and Mathematical Organization Theory*, 10(4), 295-303.
- Knuuttila, T. (2011). Modelling and representing: an artefactual approach to model-based representation. *Studies in History and Philosophy of Science* 42(2), 262-271.

- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480-498.
- Kurahashi-Nakamura, T., Mäs, M., & Lorenz, J. (2016). Robust clustering in generalized bounded confidence models. *Journal of Artificial Societies and Social Simulation*, 19(4).
- Latané, B. (1981). The psychology of social impact. *American Psychologist*, 36, 343-365.
- Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Life in the network: the coming age of computational social science. *Science*, 323(5915), 721-723.
- Lipton, P. (2000). *Inference to the best explanation (2nd ed.)*. London: Routledge.
- Lorenz, J. (2021). Supplemental materials for preprint: Individual attitude change and societal dynamics: Computational experiments with psychological theories. Open Science Framework. <https://doi.org/10.17605/OSF.IO/YQ5B4>
- MacKinnon, N. J. (1994). *Symbolic interactionism as affect control*. Albany: State University of New York Press.
- Mason, W. A., Conrey, F. R., & Smith, E. R. (2007). Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review*, 11(3), 279-300.
- Mäs, M., & Flache, A. (2013). Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. *PLoS ONE* 8(11), e74516.
- McGuire, W. (1985). Attitudes and attitude change. In G. Lindzey, E. Aronson (eds.), *Handbook of Social Psychology 3<sup>rd</sup> ed.* (pp 233 – 346). New York: Random House.
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1), 415-444.



- Meyer, M. (2019). How to use and derive stylized facts for validating simulation models. In Beisbart C. & Saam N. (eds.) *Computer simulation validation* (pp. 383-403). Springer, Cham.
- Morgan, M. & Morrison, M. (1999) (eds.). *Models as mediators. Perspectives on natural and social science*. Cambridge: Cambridge University Press.
- Morrison, M. (1999). Models as autonomous agents. In M. Morgan, M. Morrison (eds.), *Models as mediators. Perspectives on natural and social science* (pp. 38-65). Cambridge: Cambridge University Press.
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behavior*, 3, 221-229.
- Muthukrishna, M. & Schaller, M. (2020). Are collectivistic cultures more prone to rapid transformation? Computational models of cross-cultural differences, social network structure, dynamic social influence, and cultural change. *Personality and Social Psychology Review*, 24(3), 103-120.
- Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, 83(4), 602-629.
- Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review*, 45(2), 167-256.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Annual Review of Psychology*, 2(2), 175-220.
- Nowak, A., Szamrej, J., & Latané, B. (1990). From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review*, 97(3), 362-376.
- Osgood, C. E. & Tannenbaum, P. H. (1955). The principle of congruity in the prediction of attitude change. *Psychological Review*, 62, 42-55.

Pineda, M., Toral, R., & Hernandez-Garcia, E. (2009). Noisy continuous-opinion dynamics.

*Journal of Statistical Mechanics: Theory and Experiment*, 2009(08), P08001.

Pinker, S. (2018). *Enlightenment now: The case for reason, science, humanism, and progress*. London: Penguin.

Read, S. J., & Simon, D. (2012). Parallel constraint satisfaction as a mechanism for cognitive consistency. In B. Gawronski & F. Strack (eds.), *Cognitive consistency: A fundamental principle in social cognition* (pp. 66-88). New York, NY: Guilford.

Rosenberg, M. J., & Hovland, C. I. (1960). Cognitive, affective, and behavioral components of attitudes. In J. Milton & M. Rosenberg (eds.), *Attitude organization and change: An analysis of consistency among attitude components* (pp. 1-14). New Haven: Yale University Press.

Salzarulo, L. (2006). A continuous opinion dynamics model based on the principle of meta-contrast. *Journal of Artificial Societies and Social Simulation* 9(1).

Schröder, T., Hoey, J., & Rogers, K. B. (2016). Modeling dynamic identities and uncertainty in social interactions: Bayesian affect control theory. *American Sociological Review*, 81(4), 828-855.

Schröder, T., & Wolf, I. (2017). Modeling multi-level mechanisms of environmental attitudes and behaviours: The example of carsharing in Berlin. *Journal of Environmental Psychology*, 52, 136-148.

Schweitzer, F. (2018). Sociophysics. *Physics Today*, 71(2), 40-46.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

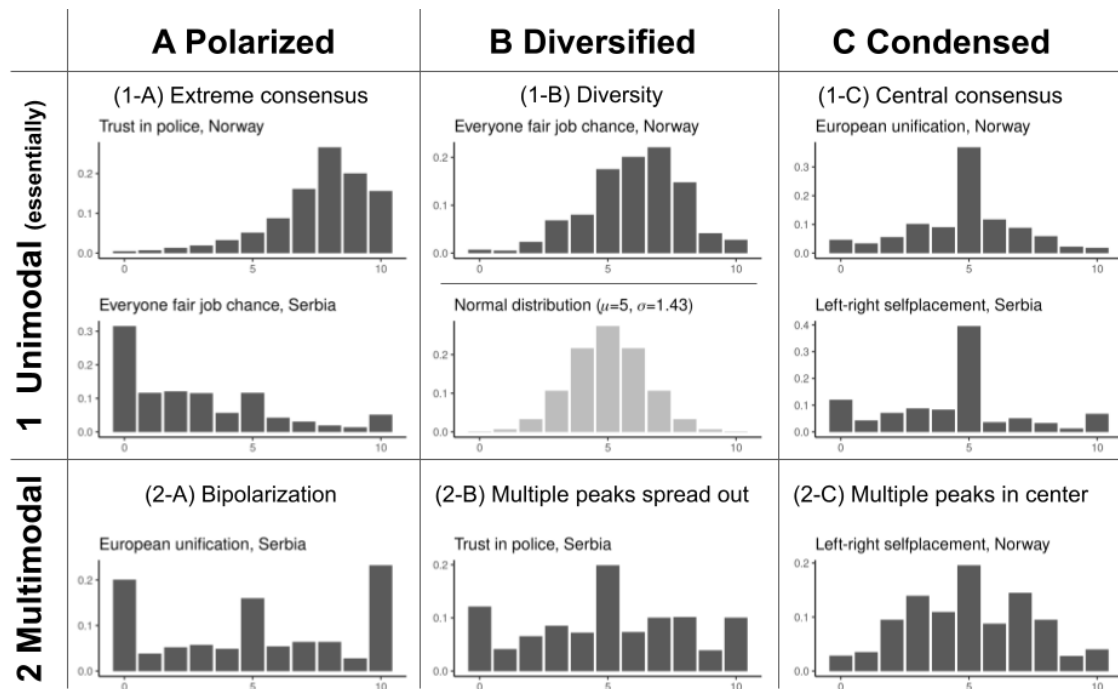
Sherif, M., & Hovland, C. I. (1961). *Social judgement*. New Haven: Yale University Press.

- Shin, J. K., & Lorenz, J. (2010). Tipping diffusivity in information accumulation systems: More links, less consensus. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(06), P06005.
- Simon, D., Stenstrom, D., & Read, S. J. (2015). The coherence effect: Blending cold and hot cognitions. *Journal of Personality and Social Psychology*, 109(3), 369-394.
- Smaldino, P. E. (2020). How to translate a verbal theory into a formal model. *Social Psychology*, 51, 207-218.
- Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, 11(1), 87-104.
- Squazzoni, F. (2012). *Agent-based computational sociology*. Chichester: Wiley.
- Sullivan, A. (2016). Democracies end when they are too democratic. *New York Magazine*, May 1st, 2016. Retrieved January 19th, 2018 from <http://nymag.com/daily/intelligencer/2016/04/america-tyranny-donald-trump.html>.
- Takács, K., Flache, A., & Mäs, M. (2016). Discrepancy and disliking do not induce negative opinion shifts. *PloS One*, 11(6), e0157948.
- Troitzsch, K. G. (2017). Axiomatic theory and simulation. A philosophy of science perspective on Schelling's segregation model. *Journal of Artificial Societies and Social Simulation*, 20(1).
- Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (eds.) (1987). *Rediscovering the social group: A self-categorization theory*. Oxford: Blackwell.
- Vallacher, R. R., Read, S. J., & Nowak, A. (eds.). (2017). *Computational Social Psychology*. New York: Routledge.
- Van Rooij, I. (2019). *Psychological science needs theory development before preregistration*. Retrieved November 29th, 2019 from

<https://featuredcontent.psychonomic.org/psychological-science-needs-theory-development-before-preregistration/>

Wilensky, U. (1999). *NetLogo*. <http://ccl.northwestern.edu/netlogo/>. Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL.

## Figures



*Figure 1.* Eight attitude distributions in Norway and Serbia for four topics from the European Social Survey (ESS, 2018). All attitudes are measured on 11-point rating scales. Frequencies use design weights to maximise representativity. Trust in the police is high in Norway in a comparably consensual unimodal way (1-A). In contrast, it is fragmented and diversified in Serbia (2-B). Similarly, in Serbia, fair job chances for everyone are seen negatively in a comparably consensual way (1-A), while views are distributed in a diversified nonfragmented way (1-B) in Norway. Left-right selfplacement in Serbia and the question of European unification (0 = “gone too far”, 10 = “go further”) in Norway shows characteristics of a central consensus (1-C). Conversely, left-right selfplacement in Norway is fragmented into three central three peaks without strong extremes (2-C). In Serbia, European unification is polarised with the two largest peaks being maximally against and maximally in favour (2-A).

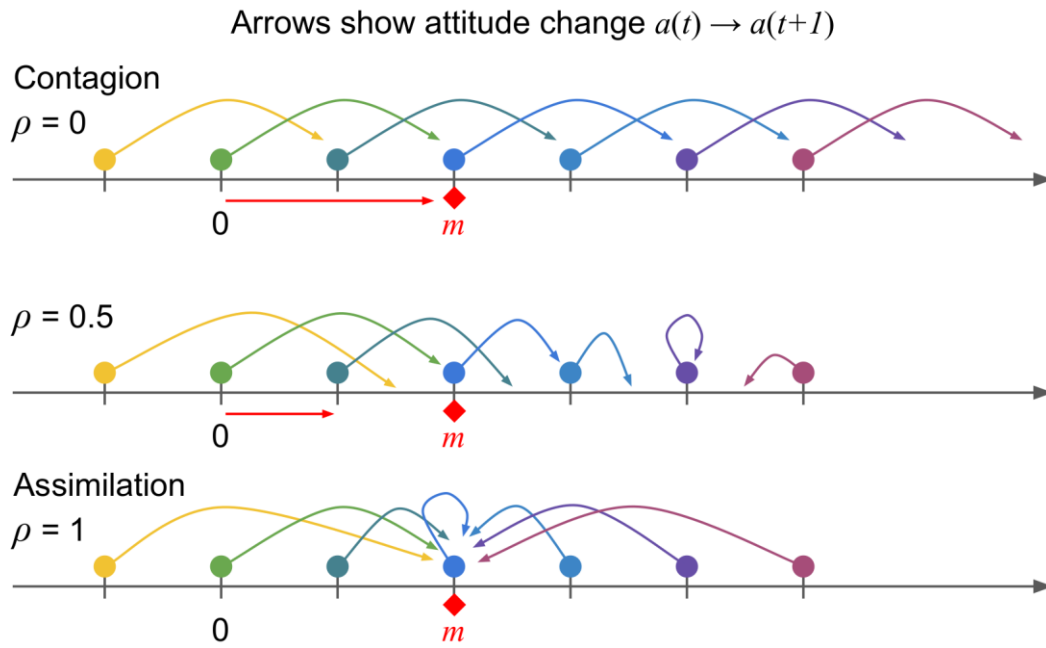


Figure 2. Example of attitude change (Eq. (5)) for different degrees of assimilation  $\rho = 0$ ,  $0.5$ ,  $1$  with respect to a positive message ( $m = 1$ , red diamond), different attitudes of the receiver (coloured dots,  $a(t) = -0.5, 0, 0.5, 1, 1.5, 2, 2.5$ ), and strength  $\alpha = 1$ . In the following, we will use lower  $\alpha$ , which reduces the length of the arrows proportionally.

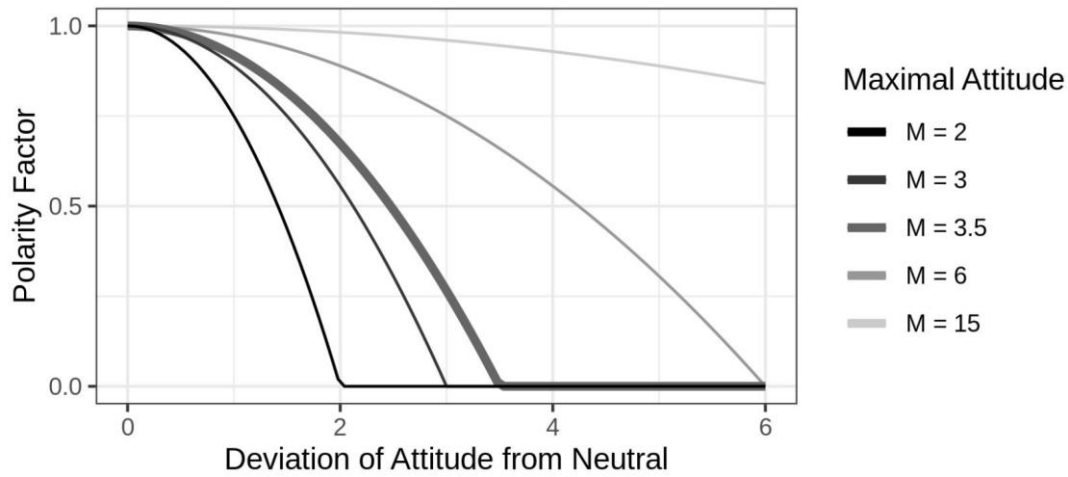


Figure 3. Functional form of the polarity factor  $(M^2 - a^2) / M^2$ .

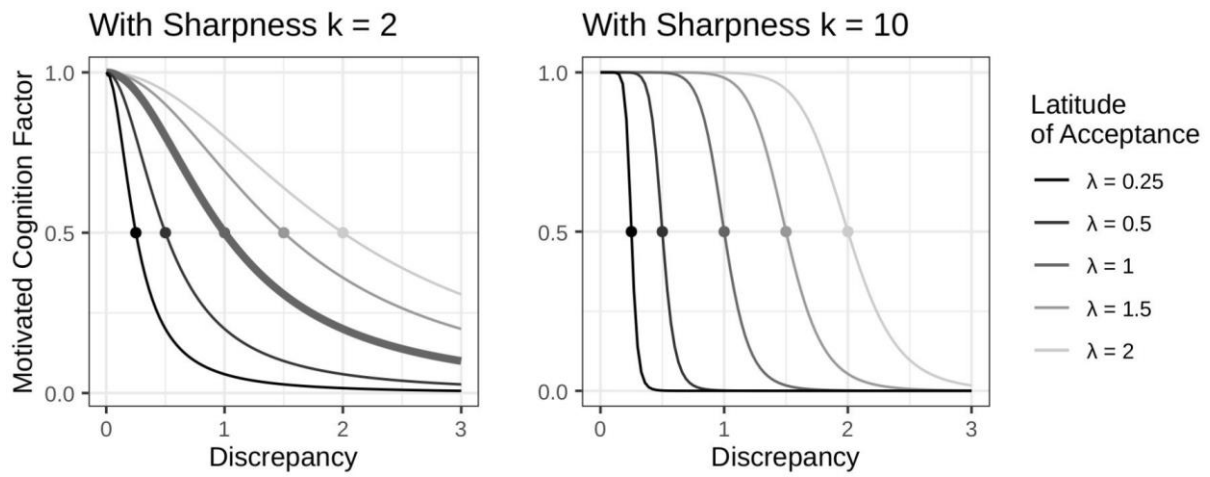


Figure 4. Functional form of the motivated cognition factor  $\lambda^k / (\lambda^k + |m - a|^k)$  with respect to different values of the latitude of acceptance  $\lambda$  and latitude sharpness  $k$ . The discrepancy between attitude and message is  $|m - a|$ .

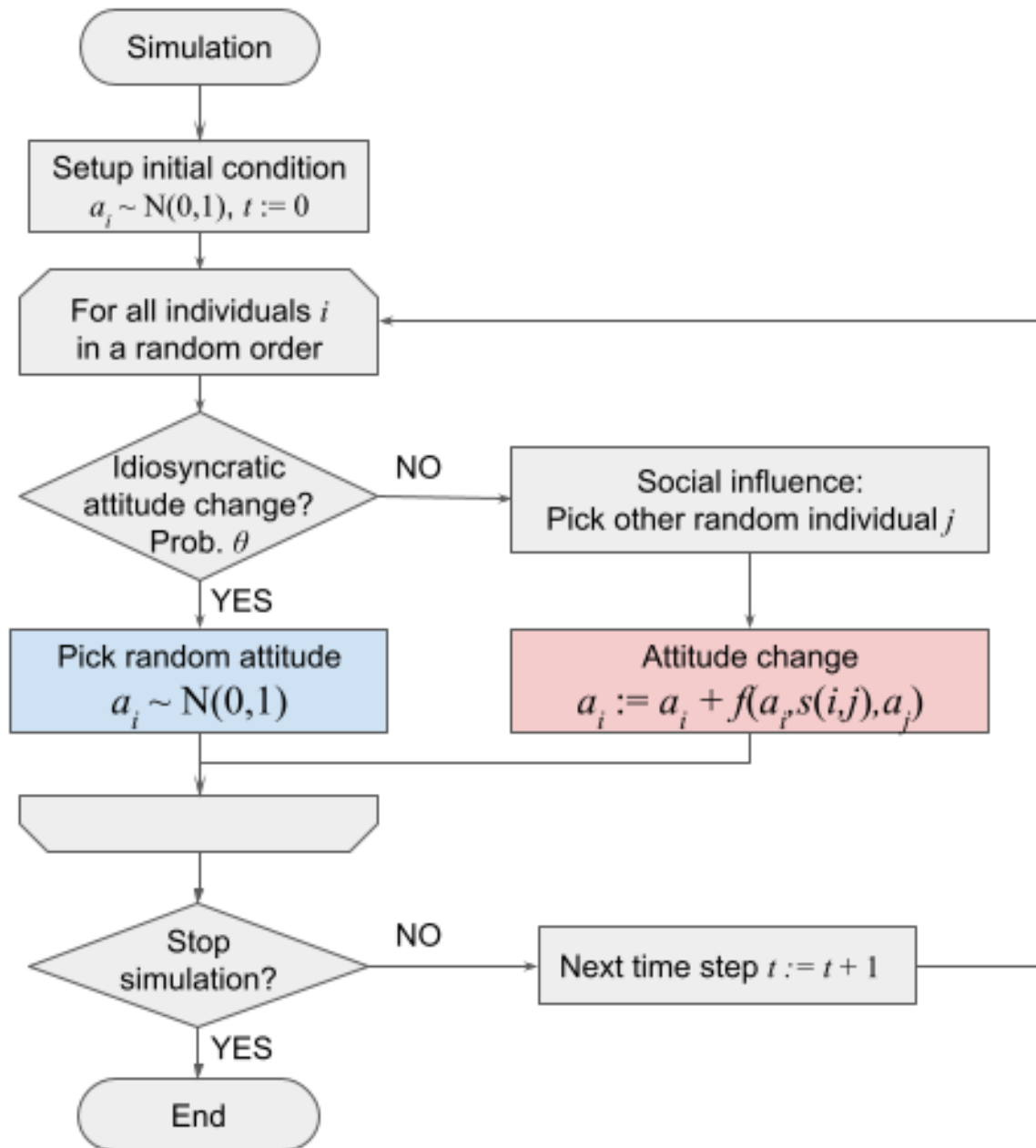
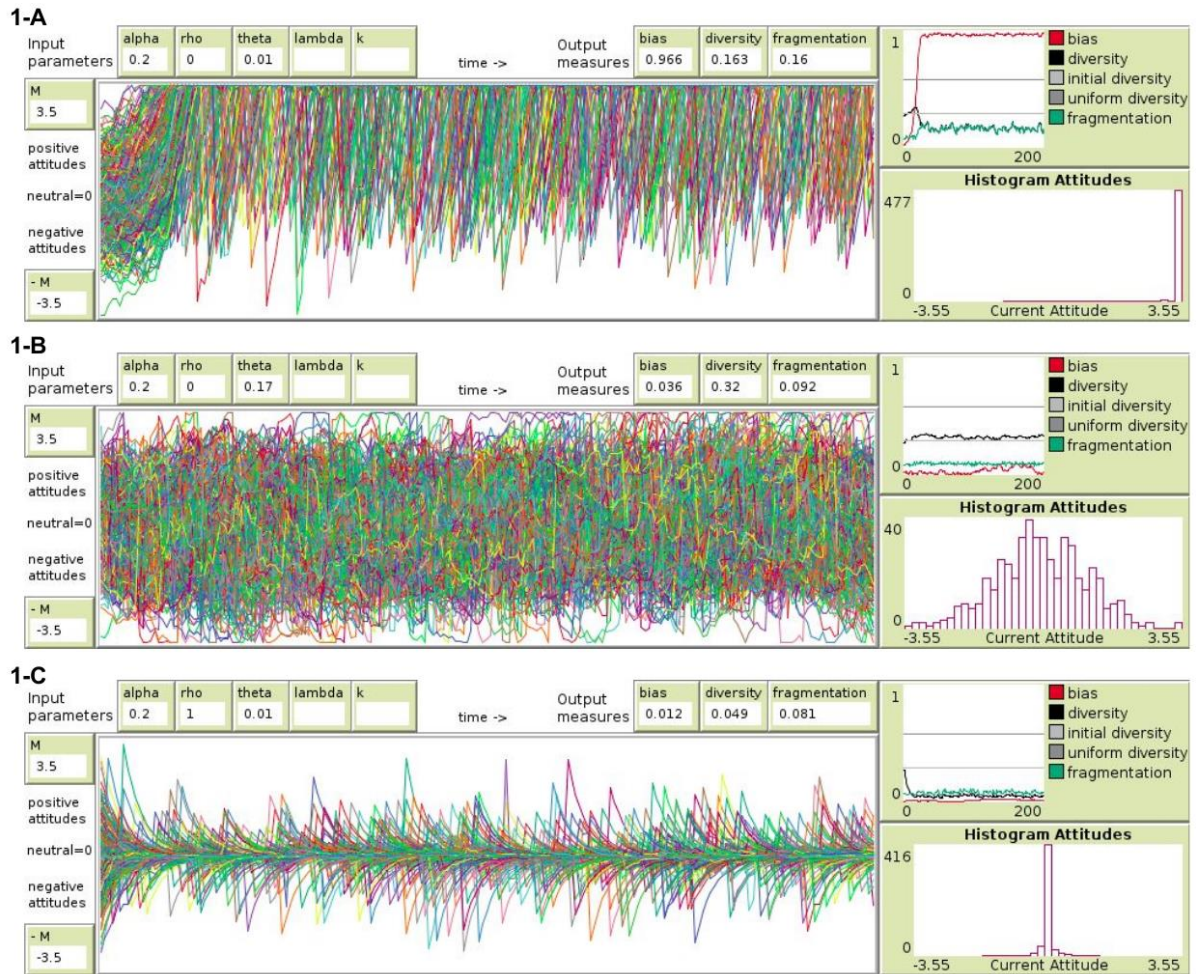
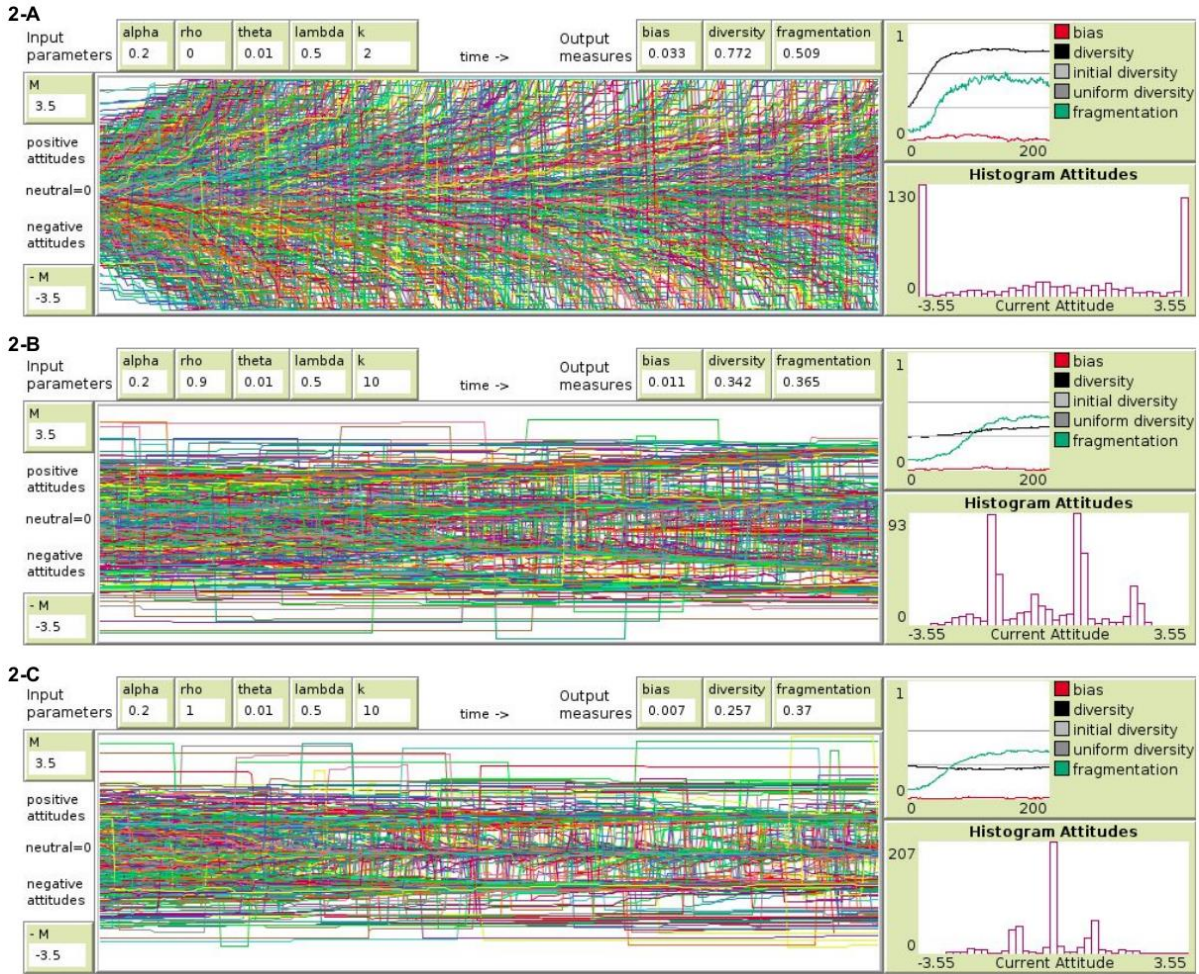


Figure 5. Flowchart of the simulation.

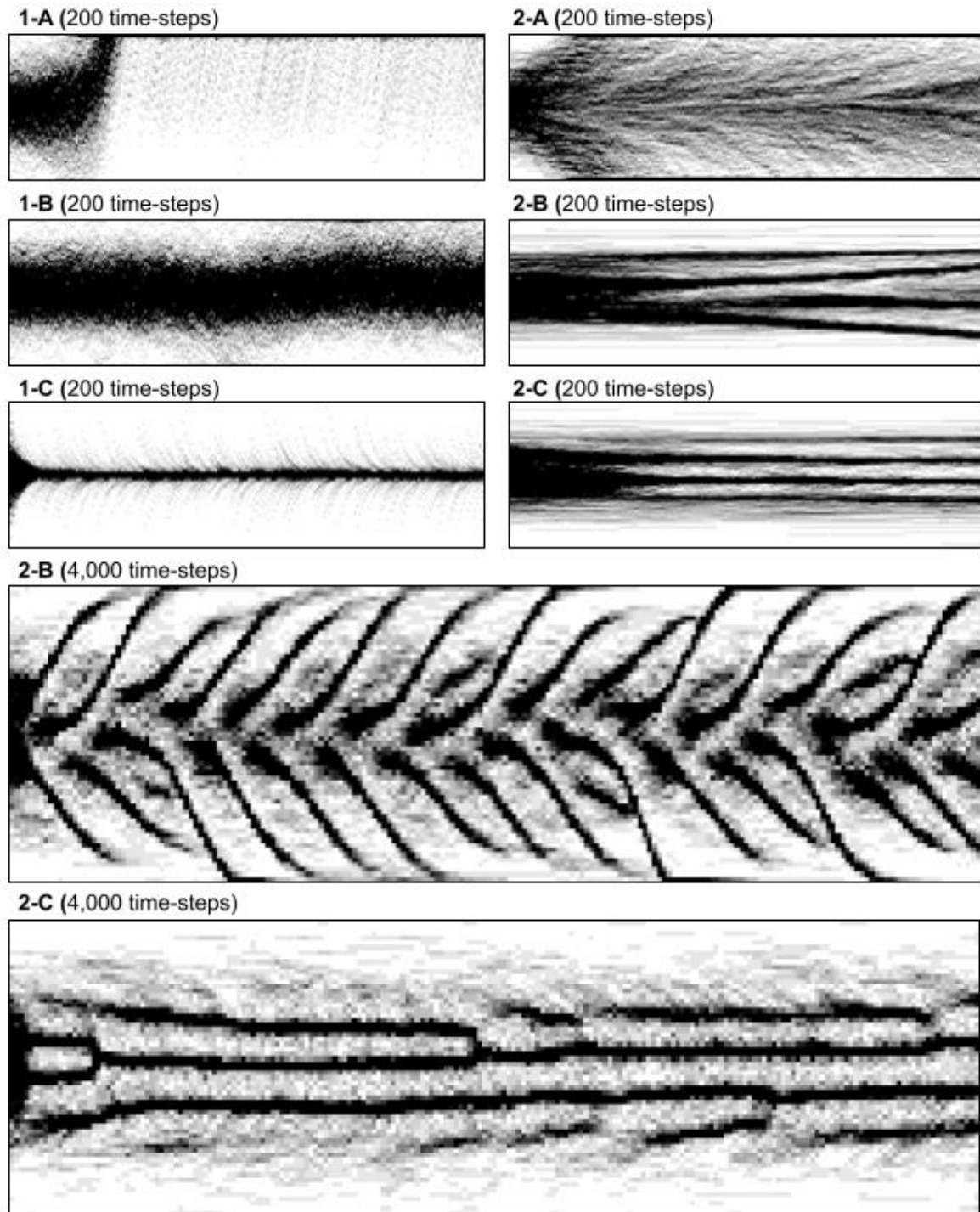




*Figure 6.* Simulation runs each with 500 agents running 200 time-steps. 1-A shows how extreme consensus evolves through contagion. 1-B shows how diversity prevails with more idiosyncrasy. 1-C shows central consensus evolving when individuals assimilate instead of being contagious. The small panels on the right-hand side show the time evolution of the output measures and the histogram of attitudes at time-step 200.



*Figure 7.* Simulation runs each with 500 agents running 200 time-steps. 2-A shows evolving bipolarisation through contagion and motivated cognition. 2-B shows diversified fragmentation into several peaks triggered by assimilation with little contagion and sharp motivated cognition. 2-C shows condensed fragmentation through assimilation under sharp motivated cognition. The outline of panels and subpanels is analogous to Figure 5.



*Figure 8.* Heatmaps of agent densities in simulations 1-A to 2-C (see Figures 5 and 6), and twenty times longer simulations for 2-B and 2-C. Colour code from purple (0 agents) to red (15% of agents or more), space is divided into 61-by-201 patches.

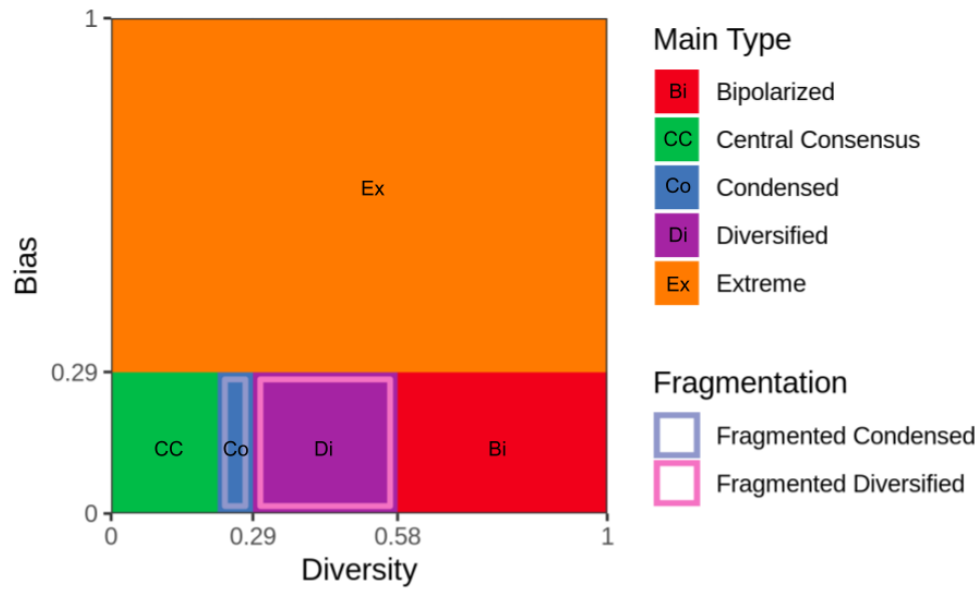


Figure 9. Types of attitude distributions in the *Diversity-Bias* space (see Table 1 for definitions and Table 2 for examples).

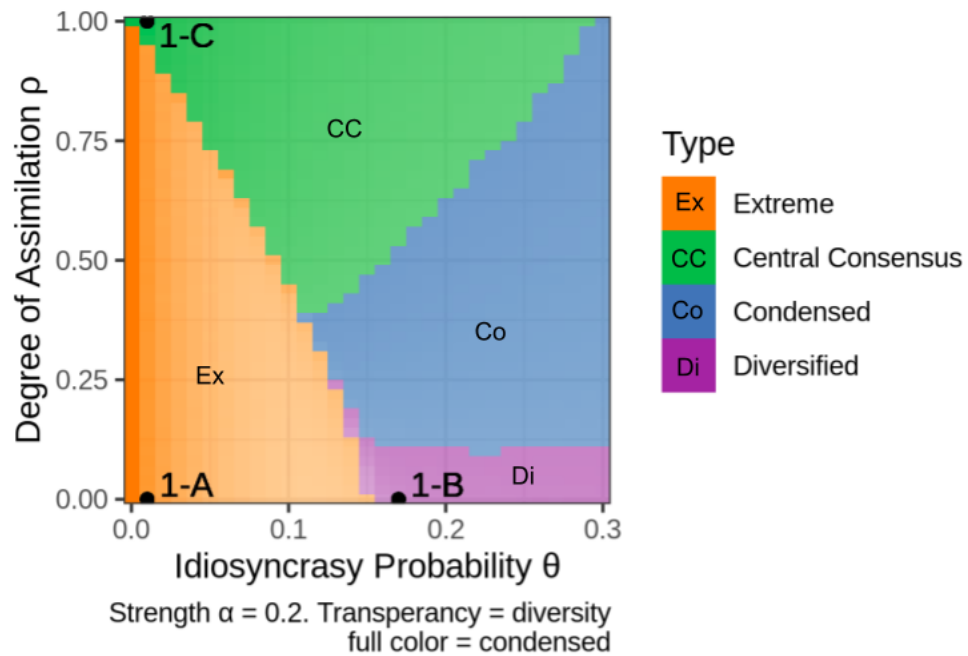
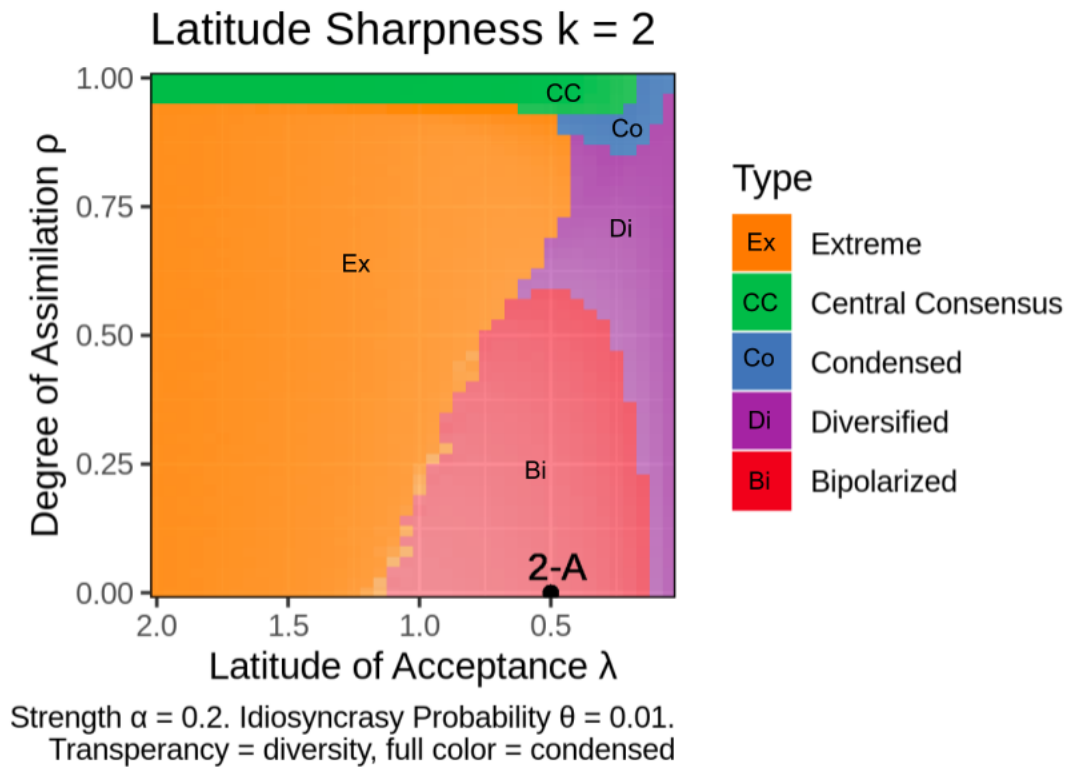
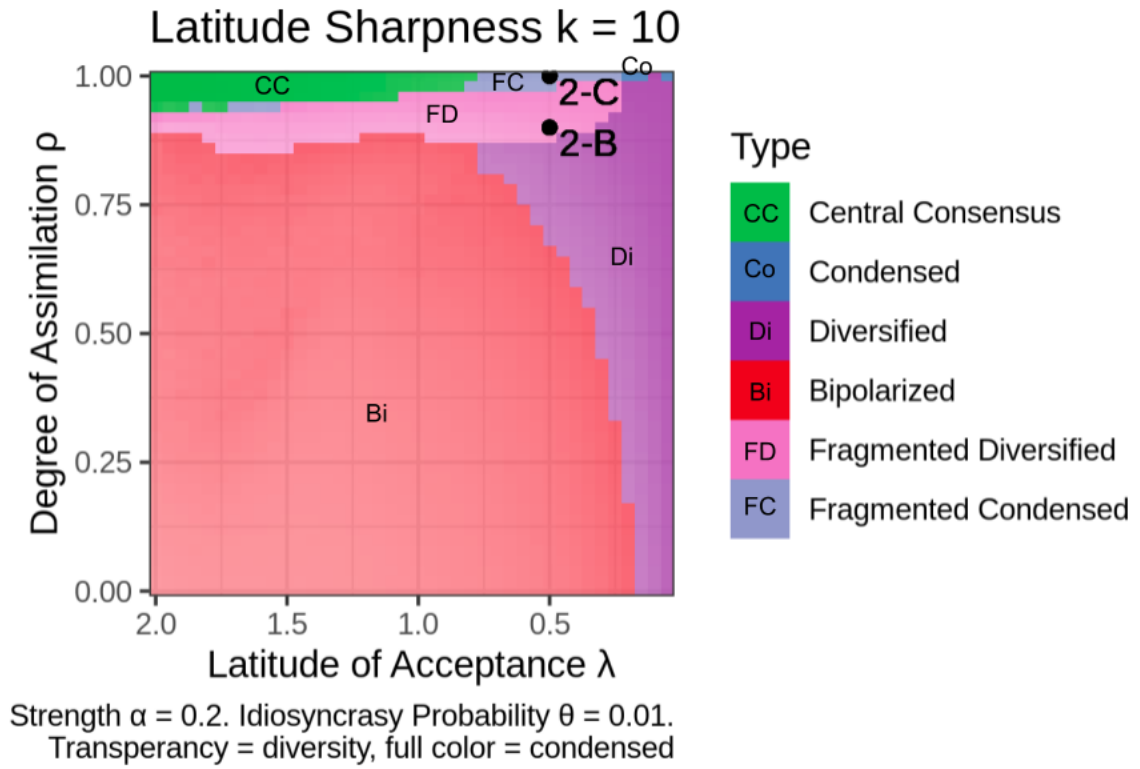


Figure 10. Results of Simulation Experiment 1. The  $\theta$ - $\rho$ -parameter sweep.





*Figure 11.* Results of Simulation Experiment 2. The  $\lambda$ - $\rho$ -parameter sweep with latitude sharpness  $k = 2$ .



*Figure 12.* Results of Simulation Experiment 2. The  $\lambda$ - $\rho$ -parameter sweep with latitude sharpness  $k = 10$ .

## Supplemental Material

### “Individual attitude change and societal dynamics: Computational experiments with psychological theories”

Jan Lorenz, Martin Neumann, Tobias Schröder

Here we provide Figure S1 as an extension of Figure 10 and Figure A.2 as an extension of Figures 11 and 12. Both figures are extended by facets for variations of a third parameter (the strength and the latitude sharpness, respectively).

Figures S3 and S4 show simulations for parameter constellations where polarity and source credibility make a difference, however, the following descriptions will also point out that the effects rely on strong assumptions, or that the emerging phenomena may be seen more as preserving initially phenomena instead of generating emergent ones.

Figure S3 shows two examples showing how polarity is part of a mechanism producing some attitude bias in the population. The setting is that of a neutral consensus through assimilation, as in Figure 6 1-C. The main difference in Figure A.3 is that the maximal attitude  $M$  is not 3.5 but only 2. This has the consequence that 4.55% of all normally distributed initial and idiosyncratic attitudes are maximally extreme instead of only 0.047%. Panel “Polarity 1” shows only a tiny interim deviation of the average attitude from neutral. The simulation in the “Polarity 2” panel only differs from “Polarity 1” with a higher change strength of  $\alpha = 0.67$  (instead of  $\alpha = 0.2$ ). A sizable bias emerges under this additional condition because extremists attract more neutral others while they themselves are not attracted because the polarity factor sets their attitude change to zero. Small random fluctuations in the number of extremists on both sides decide the direction of the bias. Extremists emerge through idiosyncratic attitude change, and thus, their fractions on both sides fluctuate. Nevertheless, once the mass of non-extremists has reached a certain magnitude of bias, they gain a certain determination to stay on that side because their magnitude of change also decreases due to the polarity factor (see Figure 2). This is a scenario where bias also emerges under assimilation through the polarity effect, however this appears only when there is initially a sizable proportion of extremists and when the general propensity to change is strong, meaning that agents assimilate, for example, more than halfway towards the attitude of the other.

Figure S4 provides two examples showing the effects of source credibility. In both examples there are two equally sized groups of agents. The two examples show how source credibility can produce bipolarisation under contagion and bimodality under assimilation. Both mechanisms work without motivated cognition but rely on strong initial and idiosyncratic differences between groups, and extremely low intergroup credibility compared to intragroup credibility. When an agent  $i$  receives information from  $j$  from the other group the credibility factor is  $s(i,j) = 0.25$ . The credibility is  $s(i,j) = 1$  when the other agent is from the same group. Thus, the influence of the other group is four times lower than that of the same group. The “Source Credibility 1” panel shows that bipolarisation can occur due to low intergroup credibility and without motivated cognition under contagion (settings of Figure 6

1-A) when the groups' initial attitudes differ on average by one. As we noted in the main text, an intergroup credibility of 0.5 is not low enough and would still lead to an extreme consensus. The "Source Credibility 2" panel shows that a bimodal attitude landscape can emerge under assimilation (see Figure 6 1-C) without motivated cognition but only with an extremely strong initial average difference of attitudes of 2.5 between the two groups. Additionally, this only works when the inflow of idiosyncratic attitudes with this difference is strong enough. In the simulation, the idiosyncrasy probability is  $\theta = 0.17$  (Figure 6 1-B).

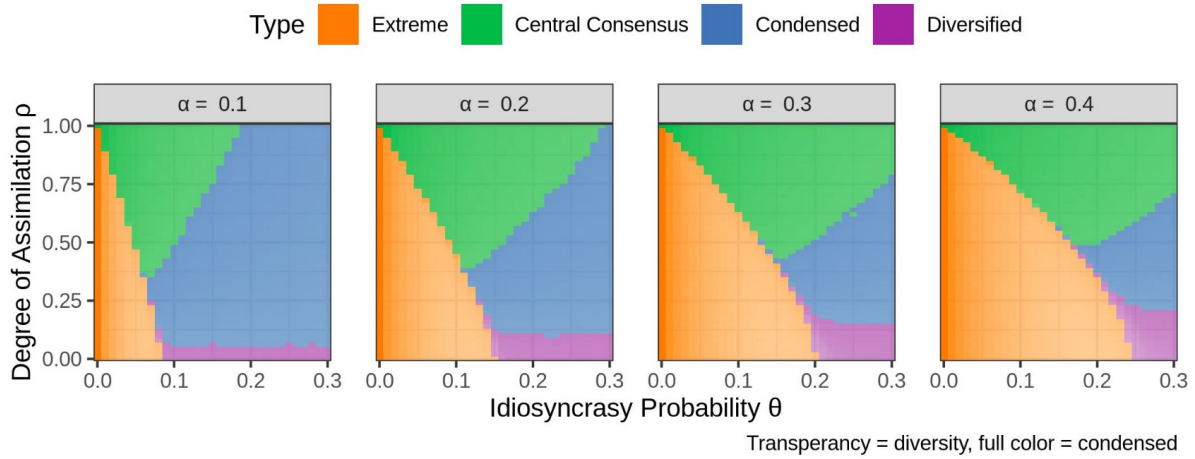


Figure S1. Results Simulation Experiment 1. The  $\theta$ - $\rho$ -parameter sweep for different values of the strength parameter  $\alpha$  (see Figure 9).

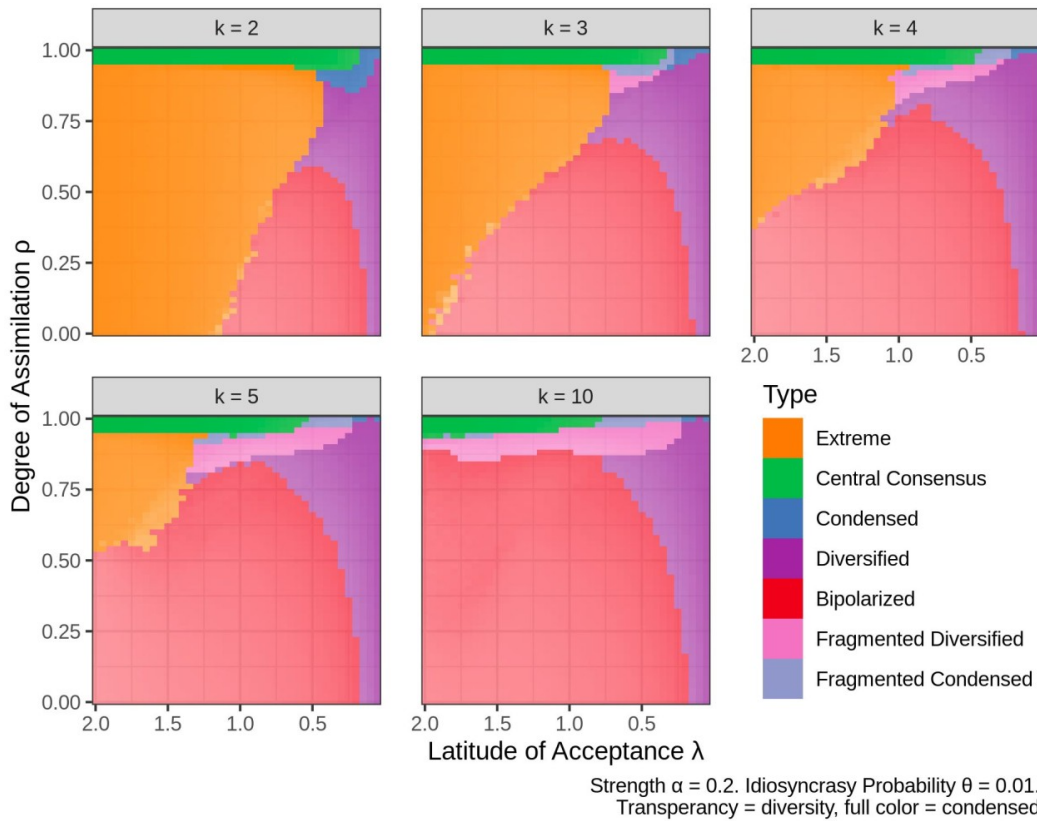


Figure S2. Results of Simulation Experiment 2. The  $\lambda$ - $\rho$ -parameter sweep for  $k = 2, 3, 4, 5, 10$  (see Figures 10 and 11).



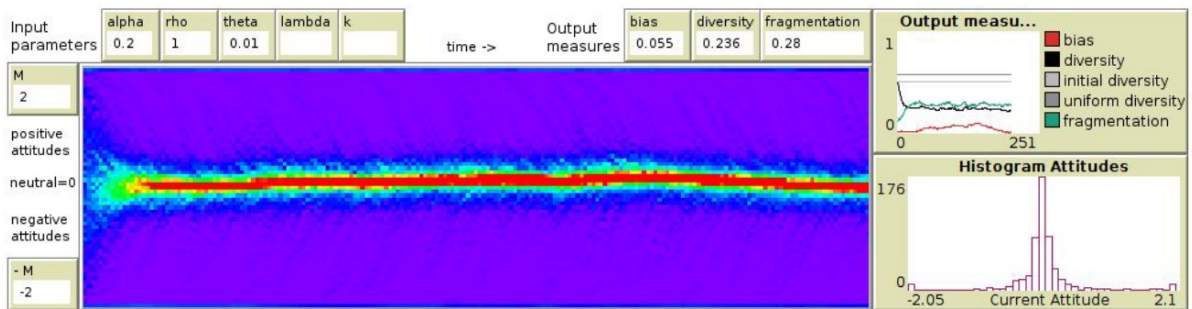
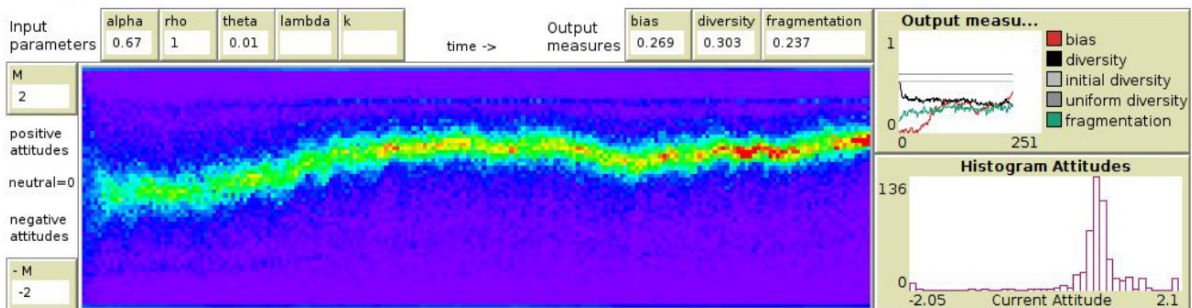
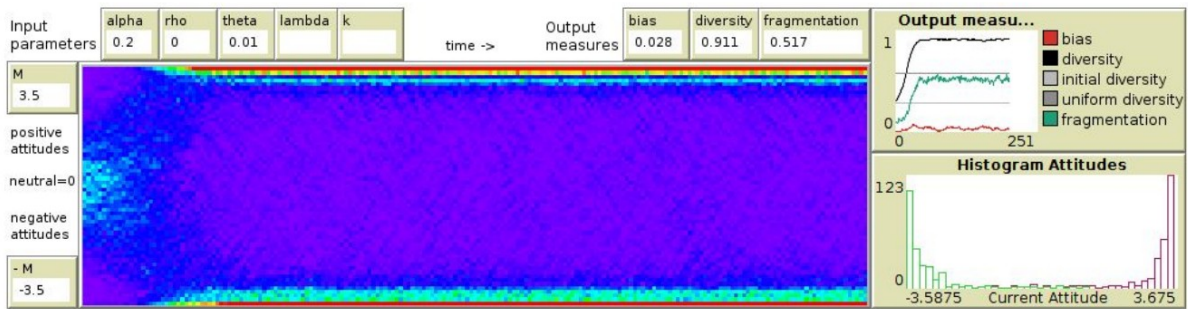
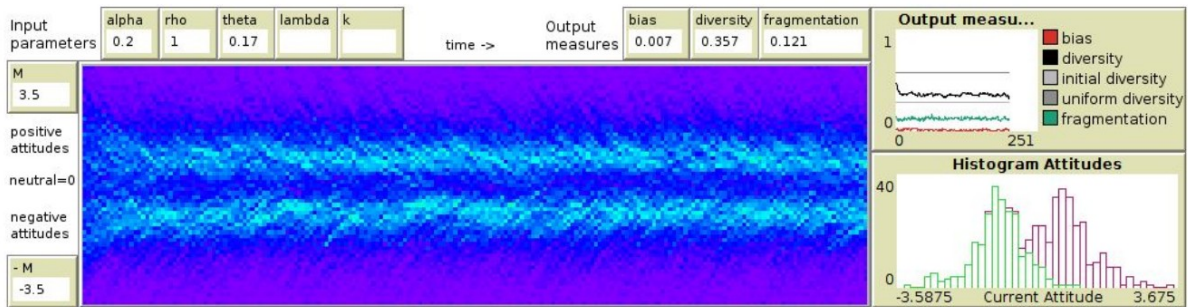
**Polarity 1****Polarity 2**

Figure S3. Simulation runs with parameters as the neutral consensus setting in Figure 6 1-C but with a polarity factor in “Polarity 1” panel and a lower maximal attitude  $M = 2$ . The “Polarity 2” panel additionally has a higher change strength of  $\alpha = 0.67$ . Lower maximal attitudes and higher change strength can cause the emergence of societal bias through assimilation.

### Source Credibility 1



### Source Credibility 2



*Figure S4.* Simulation runs with 500 agents from two equally sized groups with different initial group spread and an intergroup credibility of 0.25 instead of one. In the “Source Credibility 1” panel, the initial and idiosyncratic attitudes of groups differ by one on average. The rest of the parameters are those of the baseline setting (extreme consensus in Figure 6 1-A). The panel shows the emergence of bipolarisation without motivated cognition but two different groups which are initially leaning to opposite sides and which intergroup credibility is only 25% of their within group credibility. In the “Source Credibility 2” panel, the initial and idiosyncratic attitudes of groups differ even more by 2.5, but agents assimilate (see Figure 6 1-C), and with a higher idiosyncrasy probability of  $\theta = 0.17$  (see Figure 6 1-B). The panel shows that bimodal distribution can be preserved even under assimilation without motivated cognition, but only if groups strongly differ and have extremely low intergroup credibility.