Psychological Review

Networks of Beliefs: An Integrative Theory of Individual- and Social-Level Belief Dynamics

Jonas Dalege, Mirta Galesic, and Henrik Olsson

Online First Publication, September 19, 2024. https://dx.doi.org/10.1037/rev0000494

CITATION

Dalege, J., Galesic, M., & Olsson, H. (2024). Networks of beliefs: An integrative theory of individual- and social-level belief dynamics.. *Psychological Review*. Advance online publication. https://dx.doi.org/10.1037/rev0000494

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

ISSN: 0033-295X

https://doi.org/10.1037/rev0000494

Networks of Beliefs: An Integrative Theory of Individual- and Social-Level Belief Dynamics

Jonas Dalege[1, 2], Mirta Galesic[1, 3, 4], and Henrik Olsson[1, 3]

[1] Santa Fe Institute, Santa Fe, New Mexico, United States
[2] Department of Psychology, University of Amsterdam
[3] Complexity Science Hub, Vienna, Austria
[4] Vermont Complex Systems Center, University of Vermont

We present a theory of belief dynamics that explains the interplay between internal beliefs in people's minds and beliefs of others in their external social environments. The networks of belief theory goes beyond existing theories of belief dynamics in three ways. First, it provides an explicit connection between belief networks in individual minds and belief dynamics on social networks. The connection, absent from most previous theories, is established through people's social beliefs or perceived beliefs of others. Second, the theory recognizes that the correspondence between social beliefs and others' actual beliefs can be imperfect, because social beliefs are affected by personal beliefs as well as by the actual beliefs of others. Past theories of belief dynamics on social networks do not distinguish between perceived and actual beliefs of others. Third, the theory explains diverse belief dynamics phenomena parsimoniously through the differences in attention and the resulting felt dissonances in personal, social, and external parts of belief networks. We implement our theoretical assumptions in a computational model within a statistical physics framework and derive model predictions. We find support for our theoretical assumptions and model predictions in two large survey studies ($N_1 = 973$, $N_2 = 669$). We then derive insights about diverse phenomena related to belief dynamics, including group consensus and polarization, group radicalization, minority influence, and different empirically observed belief distributions. We discuss how the theory goes beyond different existing models of belief dynamics and outline promising directions for future research.

Keywords: beliefs, networks, attention, dissonance

Supplemental materials: https://doi.org/10.1037/rev0000494.supp

Why do people's beliefs sometimes become more extreme over time and in other instances stay quite moderate? When are these beliefs shaped by other related issues, and when are they shaped by one's social environment? Why do people sometimes project their personal beliefs onto others, and at other times perceive others' beliefs with high accuracy? How can a single theory of belief change answer these questions and explain a plethora of phenomena including polarization, radicalization, minority influence, and real-world belief patterns? These and other questions about belief formation and change have been investigated in a wide range of disciplines, from psychology (Ajzen, 1991; Dalege et al., 2018; Latané & Wolf, 1981; Vallacher et al., 2017), sociology (Friedkin & Johnsen, 1990; Proskurnikov & Tempo, 2017), cultural evolution (Boyd & Richerson, 1985; Hoppitt & Laland, 2013), and economics (Acemoglu & Ozdaglar, 2011; Golub & Sadler, 2016), to statistical physics (Castellano et al., 2009; Pentland, 2014) and applied mathematics (Hickok et al., 2022). This

Natalie Sebanz served as action editor.

Jonas Dalege https://orcid.org/0000-0002-1844-0528

Mirta Galesic https://orcid.org/0000-0002-1712-9455

Henrik Olsson https://orcid.org/0000-0002-0099-7862

Codes for all simulations in Dalege et al. (2024) are publicly available on the Open Science Framework at https://osf.io/n58h6/?view_only=0da2fb3267574e4fa53978bc4a36ba32. Jonas Dalege, Mirta Galesic, and Henrik Olsson were supported by a grant from the National Science Foundation (Grant BCS-1918490; principal investigator: Henrik Olsson; coprincipal investigator: Mirta Galesic), Mirta Galesic and Henrik Olsson were supported by a grant from the National Science Foundation (Grant DRMS-1949432), and Jonas Dalege was supported by a European Union Horizon 2020 Marie Curie Global Fellowship (Grant 889682).

The authors thank Tamara van der Does for comments and discussion and for designing empirical studies and collecting data; Seungwoong Ha, Jens Lange, Han L. J. van der Maas, Simon Schweighofer, and Peter Steiglechner for comments and discussion; and Maria Calderon Leon, Sara Elisha Lepine, and Zofia Washington for assistance in study design.

Jonas Dalege played a lead role in conceptualization, investigation, methodology, and writing. Mirta Galesic played a lead role in conceptualization, investigation, methodology, and writing. Henrik Olsson played a lead role in funding acquisition, project administration, supervision, conceptualization, investigation, methodology, and writing.

Correspondence concerning this article should be addressed to Jonas Dalege, Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, Netherlands. Email: j.dalege@uva.nl

large body of work has provided a plethora of empirical findings and theories about the structure, formation, and change of beliefs.

However, there is a disconnect between lines of work that study belief change at the level of a single individual mind (the internal level) and at the level of social networks (the external level). At the internal level, modeling individual belief networks has been a promising framework for explaining empirical findings on how and why people change their beliefs (Dalege et al., 2016; Monroe & Read, 2008; Shultz & Lepper, 1996; Van Overwalle & Siebler, 2005). At the external level, belief change has been fruitfully studied within the context of social networks, whereby interactions with others' actual or perceived beliefs can influence one's own beliefs (Christakis & Fowler, 2010; DeGroot, 1974; Festinger, 1954; French, 1956; Friedkin & Johnsen, 1990; Harary, 1959).

The internal and external levels of beliefs have been studied largely in isolation from each other. Models of internal belief networks generally disregard external social networks, although a number of classic social-cognitive theories recognize the importance of social environments in changing individual beliefs and behaviors (Ajzen, 1991; Cialdini & Trost, 1998; Festinger, 1954; Fishbein & Ajzen, 1975; Petty & Cacioppo, 1986). On the other side, models of belief dynamics on external social networks, typically developed outside psychology in fields from sociology, economics, computational social science, to statistical physics (Albert & Barabási, 2002; Easley & Kleinberg, 2010; Friedkin & Johnsen, 1990; Jackson, 2008; Newman, 2003; Watts, 2004), largely disregard the richness of individual belief structures and individual differences in perceptions of others' beliefs (Galesic, Bruine de Bruin, et al., 2021).

To acquire a comprehensive understanding of belief dynamics, we need to integrate findings and models at the internal and external levels, ideally using theoretically driven quantitative models that are simple but empirically testable (Galesic, Olsson, et al., 2021). In this article, we develop such an integrative theory of belief dynamics, the networks of belief (NB) theory. We build on the attitudinal entropy (AE; Dalege et al., 2018) framework and the hierarchical Ising opinion model (HIOM; van der Maas et al., 2020). Like these theories, we conceptualize beliefs as nodes in attitude networks within a statistical physics framework which enables rigorous investigations and predictions of belief dynamics. We adopt an inclusive definition of beliefs (Galesic, Olsson, et al., 2021) that encompasses beliefs as assumptions about the state of the world, views on moral and political issues, evaluation of attitudes, or own preferences.

The NB theory goes beyond past theories in three ways. First, it makes a direct connection between internal beliefs and external social worlds, through people's social beliefs or perceived beliefs of others. This is a crucial step that enables connecting models and findings on internal networks, largely developed in psychology, and belief dynamics models operating solely on the level of external networks, of which multitudes exist in fields from computational social science to statistical physics.

Our second contribution is explaining how social beliefs can become more or less accurate representations of the actual beliefs of others and how that affects the resulting belief dynamics. In this way, we bridge the gap between "objectively" measured beliefs in social networks that are the major type of data used to study belief dynamics and "subjective" perception of these beliefs that are what eventually matters for social influence (Thomas & Swaine Thomas, 1928).

Our third contribution is explicit modeling of attention to different sources of dissonance between these different beliefs in internal and external networks. We postulate and show empirically that differences in attention to different parts of the internal and external networks, and the resulting differences in felt dissonance regarding those parts, lead to changes in consistency of these network parts. This in turn leads to predictable patterns in people's beliefs and to well-established effects such as polarization, radicalization, and minority influence.

In what follows, we first outline the main premises of the NB theory, describe the relationships between the central psychosocial constructs in the NB theory and statistical physics constructs, present our formal implementation of the theory together with illustrative simulations and predictions, and report empirical tests of our theoretical predictions. We then demonstrate that the NB theory can account for pervasive phenomena in the belief dynamics literature such as group polarization, group radicalization, and minority influence, as well as for empirical trends in real-world data such as the patterns of political self-identification at a given moment and its increasing polarization over time. We end with a thorough discussion of how the NB theory contributes to the existing models of belief dynamics in different disciplines.

Main Premises

The NB theory rests on three premises about the underlying psychosocial constructs. First, we assume that beliefs and their relations can be represented as network structures. Second, we assume that individuals aim to reduce dissonance within their belief networks and increase correspondence between perceived and actual beliefs of others around them. Third, dissonance between beliefs only affects belief change when it is *felt*, and this depends on whether individuals pay attention to their beliefs and their relationships. In their general forms, these premises have been implemented and tested in a range of models and empirical studies.

Belief Networks

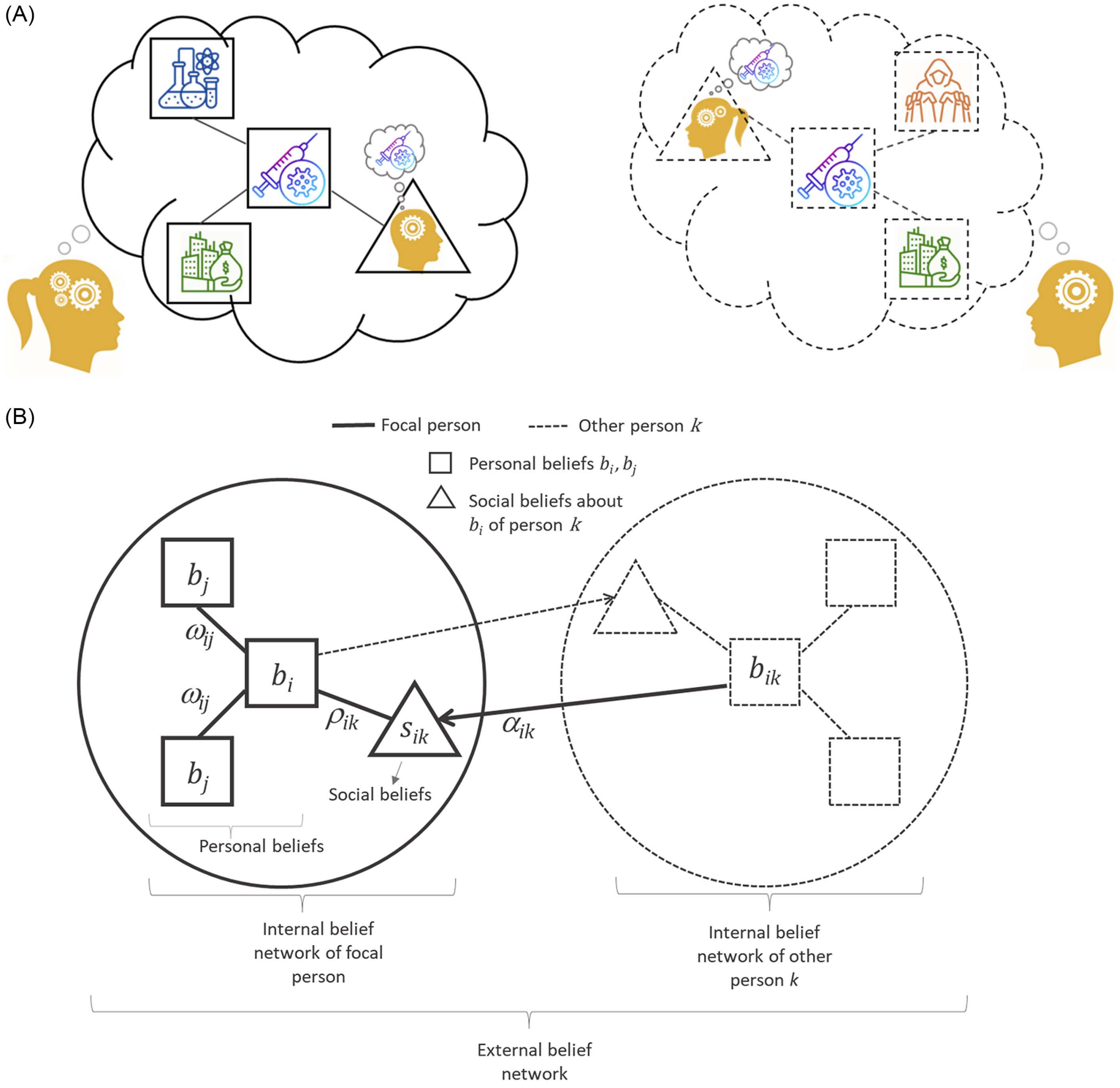
The first main premise of the NB theory is that the dynamics of a single belief depends on related beliefs one has (e.g., Brandt & Sleegers, 2021; Dalege et al., 2016, 2018). For example, in Figure 1A, beliefs about vaccination of the person on the left are related to beliefs she has about science, economic situation, and perceived beliefs of her friend on the right. For her friend, beliefs about vaccination are related to his beliefs about a secret conspiracy behind vaccination, economic situation, and perceived beliefs of his friend on the left.

As shown in Figure 1B, these beliefs can be represented as nodes in a network, with edges representing mutual influence of beliefs that different individuals have about various issues and about each other. Going beyond previous theories, the NB theory assumes that beliefs of interconnected individuals form two distinct classes of networks that interact with each other (Figure 1B).

One network represents each individual's *internal network*, which consists of *personal beliefs* and *social beliefs*. *Personal beliefs* represent beliefs related to an issue, such as various facts, preferences, and more general moral and other values that one considers relevant (e.g., whether one believes that one should get vaccinated might depend on one's beliefs about the efficacy and

Figure 1

Networks of Belief Theory Premises Illustrated on a Minimal External Network of Two People, Seen From the Perspective of the Focal Person on the Left



Note. Panel A shows an example of beliefs about vaccination. Panel B shows a more formal visual representation of the model. The focal person has an internal belief network that consists of personal (squares) and social (triangles) beliefs. The personal beliefs include a belief about a certain issue (focal belief b_i, e.g., about safety of vaccines) and other related beliefs (b_j, e.g., beliefs about science, economy, or various conspiracies). The personal beliefs are connected by edges ω_{ij} that represent the mutual influence between the personal beliefs. The focal person also has a social belief, s_{ik}, which is their perceived belief of person k about the issue b_i (e.g., whether a friend believes vaccines are safe). The mutual influence between the social belief and the personal belief about issue b_i is represented by the edge ρ_{ik}. The influence of the actual belief b_{ik} of person k on the social belief of the focal person is represented by the directed edge α_{ik}. Subscripts for the focal person and the edges between the other beliefs b_j are omitted for simplicity. The other person k has the same belief structure as the focal person. See the online article for the color version of this figure.

safety of the vaccine and whether getting vaccinated aligns with one's religious and political beliefs). *Social beliefs* represent perceived beliefs of other individuals one is connected to.

A second, *external network*, describes the connections between individuals. Each connection between two individuals consists of the relation between Individual A's *social beliefs* about Individual B's *actual beliefs* and of the relation between Individual B's *social beliefs* about Individual A's *actual beliefs*.

In sum, the two networks, internal and external, are connected through individuals' social beliefs, which are in turn affected by both their own personal beliefs and by the actual beliefs of others in their social circle. This assumption is based on findings that individuals perceive their social environment relatively accurately (Galesic et al., 2018; Nisbett & Kunda, 1985), while also showing ego projection in some instances (Gagné & Lydon, 2004; Goel et al., 2010).

Dissonance Reduction

The second main premise of the NB theory is that people have a need to *reduce dissonance* they might feel because of inconsistency between their different personal and social beliefs (Festinger, 1957; Gawronski, 2012; Gawronski & Strack, 2012; Heider, 1946, 1958), as well as because of a lack of correspondence between their social beliefs (perceptions of others) and what the others actually believe (Dhami & Olsson, 2008; Hammond, 1965). We assume three types of dissonance. Inconsistency between personal beliefs (e.g., believing that vaccines are effective at preventing diseases but also dangerous is inconsistent, because they have opposing implications for one's decision to get vaccinated) can cause *personal dissonance*. Inconsistency between personal and social beliefs (e.g., believing that vaccines are effective but perceiving some or all friends as believing otherwise) can cause *social dissonance*. Finally, lack of correspondence between social beliefs (perceived beliefs of others) and others' actual beliefs can give rise to *external dissonance*. Note that in contrast to personal and social dissonance, external dissonance requires that one is able to observe lack of correspondence between one's social beliefs and others' actual beliefs (such as when a friend behaves in stark contrast to one's social beliefs). The study of the accuracy of the perception of the characteristics of other individuals and groups has a long history in psychology (for reviews, see Funder, 1995; Gagné & Lydon, 2004; Galesic et al., 2018), but these studies have not investigated the processes of belief change or dissonance reduction. Of more relevance to belief change is literature on metaperception correction that investigates ways of changing inaccurately held beliefs. Here too, the focus is not on the processes of belief change or dissonance reduction. In addition, corrections are predominantly done on the group level (Mernyk et al., 2022), while in NB theory, the focus is on connected individuals.

The NB theory is the first to explicitly differentiate between personal, social, and external dissonances. Other models implement belief change processes through dissonance reduction, but they do not explicitly differentiate dissonances due to the lack of consistency of personal and social beliefs (social dissonance) and the lack of correspondence of social and actual others' beliefs (external dissonance; for related models, see Ellinas et al., 2017; Goldberg & Stein, 2018; Rodriguez et al., 2016; Schweighofer et al., 2020; van der Maas et al., 2020; more in the Discussion section). To fully understand when and why individuals change their beliefs, we need to understand how these dissonances together lead to different social phenomena (e.g., see the Minority Influence section).

In line with social psychological research on dissonance, ambivalence, and related phenomena (Festinger, 1957; Newby-Clark et al., 2002), we distinguish between *potential and felt dissonance*. Potential dissonance refers to a lack of consistency and correspondence in one's beliefs, while felt dissonance refers to the psychological discomfort arising when attending to potential dissonance. The concept of felt dissonance is similar to the concept of felt ambivalence (Dalege et al., 2018; Priester & Petty, 1996; van Harreveld et al., 2009). As pointed out by several authors (e.g., Newby-Clark et al., 2002; Priester & Petty, 1996; van Harreveld et al., 2009), potential dissonance due to incongruent beliefs does not have to result in felt dissonance.

Attention to Dissonance

The third main premise of the NB theory is that the extent to which potential dissonance translates into felt dissonance is moderated by the amount of *attention* to different inconsistencies between beliefs. This attention will depend on the overall importance of the issue, as well as on individual and cultural differences in sensitivity to potential dissonance and its sources. One example of what could moderate attention is the need for closure. People might differ in how much they desire closure in their internal and external networks, as reflected in their propensity to seek different opinions, to like or dislike questions that can be answered in different ways, or to be irritated when one person disagrees with everyone else in a group (Webster & Kruglanski, 1994; but see Stalder, 2010). Another example could be cultural differences. For example, Japanese but not European American participants are showing more personal dissonance reduction in the presence of relevant others (Kitayama et al., 2004).

Individuals can direct different amounts of attention to personal, social, or external dissonance. For example, when considering who to vote for, some individuals might seek consistency between their beliefs about a certain candidate and their other personal beliefs such as beliefs about moral and economic issues. Other individuals might be more interested in achieving consistency between their beliefs about the candidate and what they perceive others around them think about the candidate. Still, others might feel that it is important to know accurately what others think, that is, achieve correspondence between their social beliefs and others' actual beliefs. The attention to dissonance in different parts of one's internal and external network can also depend on the issue at hand. For example, individuals might care more about their personal dissonance when evaluating a political candidate, about their social dissonance when choosing what to wear, and about the correspondence dissonance when deciding whether to get vaccinated. Directing more attention to different parts of one's belief network results in a higher impact of those parts of the network, which will be more likely to motivate a change in one's beliefs according to the felt dissonance. Importantly, these different dissonances can be measured and modeled, providing a nuanced picture of when and why different people change (or do not change) beliefs about different issues.

Relating Psychosocial and Statistical Physics Constructs

The premises of the NB theory need to be translated into a computational model that allows to investigate the theory's implied dynamics and to derive testable predictions. There are many frameworks that can be used to develop such computational models (Borsboom, van der Maas, et al., 2021; Page, 2018). These include general constraint satisfaction frameworks (e.g., Shultz & Lepper, 1996) and their neural network implementations (e.g., Monroe & Read, 2008). We choose a statistical physics framework, in which beliefs are represented as spins, potential dissonance is represented as energy, and attention is represented as temperature. We find the statistical physics framework useful for two main reasons. First, statistical physics models are well-suited to describe higher level cognitive and social dynamics emerging from lower level dynamics in a parsimonious way. This has been recognized in previous models and frameworks (Castellano et al., 2009; Galesic, Olsson, et al., 2021; Rodriguez et al., 2016; van der Maas et al., 2006). In the present case, statistical physics models enable us to investigate how belief dynamics emerge at the level of internal networks (from interactions between personal and social beliefs within individuals) and at the level of external networks (from interactions between individuals). Second, we can build on already existing statistical physics models of beliefs at the internal level (Dalege et al., 2018; Schweighofer et al., 2020; van der Maas et al., 2020) and at the external level (Castellano et al., 2009; Pham et al., 2022; Redner, 2019). Of course, as every analogy in science (Gigerenzer, 1991), the statistical physics analogy has limitations, and it is important to not force statistical physics assumptions on human psychology and sociality when they are not applicable (see the Discussion section). Similarly, we acknowledge that other factors than dissonance reduction play a role in the formation and change of beliefs such as information integration (Anderson, 1971; see the Discussion section).

In Table 1, we describe the relationships between the psychosocial constructs in our main premises, the corresponding statistical physics constructs, and the empirical measures of these constructs. The formal implementation of the NB theory rests on three core assumptions about the relationship between psychosocial and statistical physics constructs, described next. These three assumptions map onto and enable modeling of the three premises described in the previous section (the Main Premises section).

Beliefs as Spins

First, we assume that belief nodes can be modeled analogous to what is traditionally referred to as *spins* in statistical physics systems. This is a standard assumption in belief dynamic models based on a statistical physics framework (for a review, see Castellano et al., 2009). In physics, the term spin refers to a variable that can take two or more states. Beliefs are assumed to range from very positive to very negative; a person can, for example, have a highly positive opinion about Politician A but a weak negative opinion about Politician B. Conceptualizing beliefs as spins represents these magnitudes of beliefs as values ranging from −1 to 1, with higher positive or negative values indicating stronger endorsement of the given belief in either direction.

Furthermore, like in physical systems where interacting particles are considered to be coupled, we can assume that beliefs within a single individual as well as beliefs of different individuals can be coupled, that is, they can influence each other. Couplings, or edges, can range in strength, representing different amounts of influence between one's own and between own and others' beliefs. Some beliefs are fairly independent, while other beliefs have a strong influence on each other. For example, one's beliefs about the effectiveness and safety of COVID-19 vaccines are probably more strongly related than one's beliefs about the effectiveness of COVID-19 vaccines and the safety of flu vaccines. The sign of a given coupling represents whether two beliefs have an inhibitory or excitatory influence on each other. For example, the beliefs that vaccines are effective and safe might have an excitatory influence on each other, while the beliefs that vaccines are effective and unnatural might have an inhibitory influence on each other.

Besides influencing each other, belief nodes can also be affected by an outside influence such as new information (e.g., scientific findings about an issue), a media source (e.g., arguing for or against an issue), a transformative event (e.g., surprising elections or a war), fitness benefits or drawbacks of different beliefs (e.g., antivaccination views), and others. More general beliefs such as those about moral values and cultural norms can also affect beliefs about specific issues, without being affected in turn, at least not in the short run. Theoretically, all these influences can be represented in a network model, but practically, it is not feasible to measure or include all of them. Therefore, to model those more durable exogenous influences, we use the statistical physics analogy of *local fields* whose positive (negative) values lead to a higher probability to (not) endorse a given belief. These local fields influence belief nodes in the network but do not get influenced in turn.

Potential Dissonance as Energy

Second, we formalize potential dissonance as *energy*. In statistical physics systems, energy is high when systems include particles that are strongly connected but misaligned. Similarly, potential dissonance is a function of both misalignment of beliefs and the strength of the influence between beliefs. Formally, if two beliefs have a strong excitatory influence between each other but are misaligned, energy will be higher than if two beliefs, which have weak excitatory influence between each other, are misaligned. The concept that energy represents potential dissonance, consistency, or harmony has a long history in psychology and related fields. For example, in Hopfield's 1980s work (Hopfield, 1982), he linked energy to bidirectionally connected neural networks, and researchers conceptualized energy as lack of harmony (Smolensky, 1986) or coherence (e.g., Thagard & Verbeurgt, 1998), and the energy concept has been related to dissonance in belief dynamic models (Shultz & Lepper, 1996).

To implement our assumption that there are three types of dissonances (personal, social, and external, see the Main Premises section), we assume three different contributions to energy—one for the misalignment between different personal beliefs, one for the misalignment between personal and social beliefs, and one for the misalignment between social beliefs and actual beliefs of others. That potential dissonance can be formalized as energy and that people can differentiate between the dissonances in different parts of their belief networks are core assumptions of the theory that have not yet been investigated empirically.

Empirical Tests of the Assumption of Three Separate Dissonances

To investigate the validity of our assumption that there are three separate dissonances, we conducted two empirical studies (see Appendix A for details and Table 1 for examples of main questions). Study 1 was a survey with 973 U.S. participants from Mechanical Turk which assessed their beliefs on genetically modified (GM) food safety and related moral and political beliefs, consistency and dissonance in personal and social beliefs, and reactions to an informational intervention. Study 2 was a two-wave survey which involved 669 U.S. participants from the University of Southern California's Understanding America Panel, examining beliefs on GM food, flu vaccination, and climate change, plus nominated friends' actual beliefs. The focus was on personal/social beliefs and

Table 1

Main Psychosocial Constructs, Their Associated Statistical Physics Constructs, and Their Empirical Measures/Questions

Psychosocial construct	Statistical physics construct	Empirical measures
Beliefs (personal or social)	Spins (b_i, s_{ik}, for personal and social beliefs), taking any value between −1 and 1.	Example questions: • Thinking about childhood diseases, such as measles, mumps, rubella, and polio, what comes closer to your view? 1 = Parents should be allowed to choose to NOT vaccinate their children. 7 = All children should be required to vaccinate. • When answering the question (above) what do you think (Contact) would answer? All answers were recoded to range from −1 to 1.
Influence between beliefs	Couplings, or edges, between personal beliefs (ω_{ij}), personal and social beliefs (ρ_{ik}), and social and actual others' beliefs (α_{ik}), taking any negative value (inhibitory influence) or positive value (excitatory influence), with 0 representing absence of influence between beliefs.	Correlation between different beliefs or individual-level estimates of the relationship between different beliefs.
Exogenous influences	Local field (τ_i), taking any value, with 0 representing absence of exogenous influence, and positive (negative) values representing positive (negative) exogenous influence.	External information relevant for beliefs, for example, from scientific studies, media, current events, and fitness consequences.
Potential dissonances between different personal beliefs, between personal and social beliefs, and between social beliefs and actual beliefs of others	Energy (H), given by the misalignment of coupled spins with each other and the local field. Misalignment between different personal beliefs is represented by personal energy (H_{pers}), misalignment between personal and social beliefs is represented by social energy (H_{soc}), and misalignment between social beliefs and actual beliefs of others is represented by external energy (H_{ext}).	Difference between respective beliefs (Equations 3–5).
Attention directed at personal, social, and external dissonance	Inverse temperature (β), taking any value from 0 to ∞, determines to what extent the belief updating process is more or less probabilistic. Low β (high temperature) results in more probabilistic updates of the beliefs, while high β (low temperature) results in a more deterministic belief updating process. In the NB theory, attention to the three different sources of dissonance is represented by three inverse temperatures, one for dissonance between personal beliefs (β_{pers}), one for dissonance between personal and social beliefs (β_{soc}), and one for dissonance between social beliefs and actual beliefs of others (β_{ext}). The relative size of β for the different networks determines the relative importance of the dissonances in these networks for the belief updating process.	Example questions: • It is important to me that my beliefs toward childhood vaccination are not in conflict with each other. • It is important to me that my personal beliefs and the beliefs of (Contact 1) toward childhood vaccination are not in conflict with each other. • It is important to me that I accurately perceive the belief of (Contact) about childhood vaccination. All scales ranged from 1 = *strongly disagree* to 7 = *strongly agree.*
Felt dissonance of internal (personal and social) and external beliefs	Interactive effect between energy (H) and inverse temperature (β), represented by D, with high energy and low temperature producing the highest felt dissonance.	Example question: • I have completely mixed reactions toward the issue of genetically modified food.

Note. NB = networks of belief.

felt dissonance, influenced by seeing friends' responses and scientists' views. All questions in both studies except for the demographics were answered on 7-point scales with labeled extremes. Potential dissonances were measured by the incongruence of beliefs multiplied by the estimated connection between them.

First, the results from confirmatory factor analyses from Study 1 showed that personal and social felt dissonances formed two separate factors, which were moderately related. This finding is in line with the assumptions of our theory. Second, regression analyses showed that felt personal dissonance was reliably best predicted by personal potential dissonance and felt social dissonance was reliably best predicted by social potential dissonance. Regression analyses of the data in Study 3 showed that felt external dissonance was best predicted, albeit not reliably, by potential external dissonance (see Appendix B for details).

Attention as Temperature

Third, the moderating factor of attention is implemented by using the statistical physics concept of (inverse) *temperature*. Temperature represents to what extent the belief updating process is more or less probabilistic. High temperature results in more randomness in the belief updating process so that beliefs change in a way that does not necessarily reduce potential dissonance, while low temperature results in a more a deterministic belief change in a way that reduces dissonance. The NB theory therefore implies that belief change will be related to dissonance only when a person pays attention to the dissonance, while lack of attention results in beliefs changing in an almost random fashion. The formal implementation in the model is inverse temperature in the form of a parameter β. This parameter is multiplied with the dissonance of the beliefs, and therefore, it amplifies the differences in dissonances between different belief states (see Equations 2 and 6). This amplification in turn leads to a more pronounced pressure to reduce dissonance and leads to more deterministic, predictable network dynamics. In contrast, low inverse temperature leads to a belief network behaving in a random, unpredictable fashion.

Note that the concept of temperature in statistical physics is different than the intuitive concept of temperature in everyday language. In everyday experience, when people are excited about something they might experience higher heart rate, blood pressure, and related physiological changes that make one feel warmer. Hence, the relationship between excitement and temperature translates to "higher temperature" in everyday language meaning "stronger opinions." However, in statistical physics, the concept of temperature is very different. Here, higher temperature (typically induced exogenously) causes particles to behave more erratically, not necessarily in a way that reduces the overall energy. Hence, the relationship between temperature and the probabilistic nature of particle systems can be mapped on the relationship between the lack of attention and the probabilistic nature of belief systems. In the latter, lower attention causes beliefs to update more probabilistically, not necessarily in a way that reduces the overall dissonance.

To implement our assumption that individuals direct different amounts of attention to inconsistencies among their personal and social beliefs, as well as to the lack of correspondence between their social beliefs and actual beliefs of people around them, the NB theory features three temperatures—one for directing attention to the misalignment between different personal beliefs, one for directing attention to the misalignment between personal and social beliefs, and one for directing attention to the misalignment between social beliefs and actual beliefs of others. High temperature (or low attention) of one part in the network results in more probabilistic updating of beliefs for this network, implying low impact of this network on belief change. Low temperature (or high attention) of one part in the network results in a more deterministic belief updating process, implying high impact of this network on belief change. Implementing the NB theory in this way leads to the consequence that attention is necessary for dissonance reduction. With enough attention, potential dissonance translates into felt dissonance, which in turn typically leads to lower (potential) dissonance. In some cases, it can be difficult to lower the dissonance, for example, when one's social environment is against vaccines but one's personal beliefs are highly positive toward vaccines. Such situations are probably the ones in which individuals feel the most dissonance for a prolonged time.

Empirical Test of the Separability of Different Attention Parameters

To test the separability of the different attention parameters, we again used the empirical studies described in the previous section and in Appendix A. As described in detail in Appendix C, we first tested whether questions about the importance of reducing different dissonances (see also example questions in Table 1) formed three separated factors. Confirmatory factor analyses of the data in Study 2 indeed showed factors for the importance of reducing personal dissonance, for the importance of reducing social dissonance, and for the importance of reducing external dissonance.

We then proceeded to test whether these different factors uniquely predict their associated dissonances. NB theory predicts that higher attention to dissonance predicts a reduced potential dissonance (energy): When people pay attention to their potential dissonance (i.e., when they experience felt dissonance), their beliefs will tend to become more consistent, reducing potential dissonance. We find that attention to dissonance in the personal, social, and external parts of the belief network relates negatively to the potential dissonance (energy) in those parts of the network. Study 1 lacked data on participants' social contacts' actual beliefs, limiting the measurement of energy in external networks. However, accurate social beliefs are likely to raise internal network energy due to potential inconsistencies between social and personal beliefs. Findings weakly support this, potentially due to homogeneous social networks or the greater influence of belief expression frequency on external network temperature.

Formal Implementation

In this section, we formalize the NB theory to study the interactions of internal and external belief network dynamics and derive quantitative empirical predictions. We first present the general implementation of the NB theory connecting different parts of the theory and then introduce each part separately. We illustrate each part using simulations, discuss the main predictions following from the formalization, and present empirical tests of our assumptions and predictions. In Appendix D, we discuss several ways to estimate and measure the core constructs of the NB theory in empirical data.

We define the probability that an individual updates their personal belief b_i from its current state to a new state $b_i^{'}$ as:

$$P(b_i \rightarrow b_i^{'}) = \frac{1}{1 + e^{\Delta(D_{\text{pers}})}} \Big/ \sum_m \frac{1}{1 + e^{\Delta(D_{\text{pers}})}}, \qquad (1)$$

where $b_i^{'}$ is one of m possible states ranging from -1 to 1 and $\Delta(D_{\text{pers}})$ represents the change in felt dissonance that would occur if b_i flips to $b_i^{'}$ (the felt dissonance of b_i is subtracted from the felt dissonance of $b_i^{'}$). Thus, if the felt dissonance of $b_i^{'}$ is lower than the felt dissonance of b_i, the probability of flipping to $b_i^{'}$ is higher than keeping b_i. Conversely, if the felt dissonance of $b_i^{'}$ is higher than the felt dissonance of b_i, the probability of flipping to $b_i^{'}$ is lower than keeping b_i. Equation 1 is an adapted form of Glauber's (1963) dynamics for ordinal belief states. In contrast to Glauber's dynamics, where only energies of two states (i.e., −1 and 1) are compared to each other, the current updating compares the current state's energy to the energies of each of the m possible states. To ensure that the probabilities of remaining in the current state and the probabilities of flipping to any other state together sum to 1, we divide them by the sum of the energy comparisons for each state. Additionally, while the standard Glauber dynamics formulation includes a division of change in energy by the temperature of the system, we account for the (inverse) temperature by including it as attention parameters in the equation for felt dissonance. Specifically, we define the felt dissonance D_{pers} as:

$$D_{\text{pers}} = \beta_{\text{pers}} H_{\text{pers}} + \beta_{\text{soc}} H_{\text{soc}}, \tag{2}$$

where H_{pers} is the potential dissonance between personal beliefs (defined by Equation 3), H_{soc} is the potential dissonances between personal and social beliefs (defined by Equation 4), and β_{pers} and β_{soc} represent the attention paid to each of the dissonances. The values of β ranges from zero to infinity. We assume that the correlation between these attention parameters is positive but lower than 1 (for empirical validation, see Appendix C). The ratio between these parameters determines the amount of weight given to different sources of dissonance. If, for example, one values the consistency of their different personal beliefs more than the consistency of their personal and social beliefs, then β_{pers} would be higher than β_{soc}.

Because the attention parameters are multiplied with the dissonances, high attention parameters amplify differences in dissonance. A minor difference in dissonance will lead to a minor difference in the probability to flip when the attention parameter is also low, but to a pronounced difference in the probability to flip if the attention parameter is high. Note that the way we implemented belief updating is a local process—each belief can only reduce its dissonance with connected nodes and the global dissonance of the network is reduced only indirectly. A belief network can therefore get stuck in a local minimum where parts of the network have reduced their local dissonance, but the global dissonance of the network is not at its potential minimum. This effect usually occurs when attention is very high, so that beliefs have a high pressure to reduce their dissonance locally. The network then can only "escape" from such local minima if attention is somewhat lowered.

For the potential dissonance between personal beliefs, we extend the formulation from the AE framework (Dalege et al., 2016, 2018; see also van der Maas et al., 2020). In the AE framework, the dynamics of belief networks is modeled using the Ising (1925) model and beliefs are modeled as binary. Here, we extend this formulation to a version of Potts model (Wu, 1982) that allows for an approximation of continuous beliefs (measured on a scale from −1, representing *complete disagreement*, to 1, representing *complete agreement*). Similar representations have been used in a variety of different models (e.g., Hopfield, 1982, 1984). The potential dissonance between personal beliefs is given by:

$$H_{\text{pers}} = -\sum_{i} \tau_i b_i - \sum_{ij} \omega_{ij} b_i b_j, \tag{3}$$

where b_i represents the current state of given personal belief i on a continuum between −1 and 1, ω_i represents the exogenous influence on b_i, and the edge weights ω_{ij} represent the influence between b_i and a different personal belief b_j. If the influence between two beliefs is excitatory, their dissonance is high when they are in different states and low when they are in similar states. Conversely, if the influence between two beliefs is inhibitory, their dissonance is high when they are in similar states and low when they are in different states. Note that both the values of ω_{ij} and τ_i are always relative to β_{pers}. For example, a value of 1 for both ω_{ij} and τ_i and a value of 0.1 for β_{pers} would result in the exact same dynamics as a value of 10 for both ω_{ij} and τ_i and a value of 1 for β_{pers}. The same holds for the edges and attention parameters in the other energy functions. Similarly, ω_{ij} and τ_i need to be in a similar range to both be able to affect the dynamics of the personal beliefs.

In addition to the potential personal dissonance, personal beliefs are also affected by the potential social dissonance, given by:

$$H_{\text{soc}} = -\sum_{i} \rho_{ik} b_i s_{ik}, \tag{4}$$

where the ρ_{ik} represents the influence between personal belief b_i and an associated social belief s_i one has about a person k (in other words, one's perceived belief of the person k's belief). Note that the social beliefs have no local field (such as τ_i in Equation 3), because their dispositions are determined implicitly by the local fields acting on personal beliefs of others in one's social environment. We thus assume that social beliefs are *only* affected by one's own beliefs and actual beliefs in one's external network, given by the following equation:

$$P(s_i \rightarrow s_i^{'}) = \frac{1}{1 + e^{\Delta(D_{\text{soc}})}} \Big/ \sum_{m} \frac{1}{1 + e^{\Delta(D_{\text{soc}})}}, \tag{5}$$

where $\Delta(D_{\text{soc}})$ represents the difference in felt dissonance between the current state s_i and a new state $s_i^{'}$, stemming from potential dissonances between personal and social beliefs (H_{soc}; defined by Equation 4) and potential dissonances between social and actual beliefs of connected individuals (H_{ext}; defined by Equation 7), each weighted by an attention parameter β as defined by Equation 6:

$$D_{\text{soc}} = \beta_{\text{soc}} H_{\text{soc}} + \beta_{\text{ext}} H_{\text{ext}}, \tag{6}$$

where β_{soc} and β_{ext} represent the attention paid to each of the dissonances. We again assume that the correlation between these attention parameters is positive but lower than 1 (for empirical validation, see Appendix C). The ratio between these parameters determines the amount of weight given to different sources of dissonance. If, for example, one is motivated to hold accurate social beliefs and discusses beliefs with others a lot, it is likely that their β_{ext} will be higher than β_{soc}.

We define the external potential dissonance as:

$$H_{\text{ext}} = -\sum_{i} \sum_{k} \alpha_k s_{ik} b_{ik}, \tag{7}$$

where s_{ik} is the individual's social belief about belief i of person k, b_{ik} is the actual belief i of the connected person k, and α_k represents the influence of the actual belief b_{ik} of person k on the social belief s_{ik}. In contrast to the edges in the internal network ω_{ij} and ρ_{ik}, α_k is directed as it is assumed that only actual beliefs can influence social beliefs directly, but that social beliefs can influence actual beliefs only indirectly. The dynamics of each person k's beliefs b_{ik} are modeled by Equations 1 and 2.

Because the dissonance between two beliefs in our implementation is expressed as multiplication of belief states ($b_i b_j$ in Equation 3, $b_i s_{ik}$ in Equation 4, and $s_{ik} b_{ik}$ in Equation 5), the lowest dissonance is achieved when two beliefs have the same (respectively different) and extreme states when they are positively (respectively negatively) connected. We think that it is reasonable to assume that more extreme values result in lower energies, because extreme values represent a more unambiguous belief, which probably serves the need for consistency more than a more ambiguous belief. Additionally, in Appendix E, we investigate several other distance measures and show that multiplication is empirically the best suited distance measure for our theory. As can be seen in Figure E1, only the multiplication implementation can reproduce the increasing bimodality that we see in empirical data (Figure 2). Furthermore, this multiplication implementation has been used in several other constraint satisfaction models of belief change (e.g., Shultz & Lepper, 1996).

Note that different potential dissonances will have different sizes, reflecting the sum over different numbers of pairs of personal beliefs (Equation 3), personal and social beliefs (Equation 4), and social and actual beliefs (Equation 7). This formulation as sums allows that a person who has relatively more social contacts than beliefs related to an issue might feel higher social or external dissonance than a person who has relatively more personal beliefs about an issue.

Illustrations of the Networks of Beliefs Theory's Dynamics

To illustrate the belief dynamics implied by the different attention parameters, we ran several illustrative simulations varying one attention parameter while keeping the other two attention parameters constant. We turn to simulations of the interactions between the different attention parameters in the Explaining Established Belief Dynamics Phenomena section. Codes for all simulations can be found at https://osf.io/n58h6/?view_only=0da2fb3267574e4fa5397 8bc4a36ba32.

General Simulation Setup

In all simulations reported in this article, each individual has a fully connected personal belief network of 10 nodes, as illustrated by the square nodes in the leftmost panel in Figure 2A. One of the personal beliefs of each individual represents a focal belief that is communicated in the external network. We focus on the focal beliefs about an issue and its related personal and social beliefs, where this part of the belief network is represented as a fully connected network. The whole belief system will have a more nuanced pattern of relationships. The simulations mimic interactions about one important (focal) topic at a time, but can be extended to other network structures and more beliefs. Consequently, each individual has social beliefs about the focal belief of each of the other individuals in the network, represented by the triangle nodes in Panel (a) of Appendix G. For the illustrative simulations shown in Appendix G, we used a fully connected 10-node external network, shown in Panel (j) of Appendix G. Note that each circle in this panel corresponds to one of the internal networks shown in Panel (a).

We set the edge weights in all simulations to $\omega = .4$ for the influence between personal beliefs, $\rho = 1$ for the influence between personal and social beliefs, and $\alpha = 1.4$. These values lead to moderate correlations between beliefs in all parts of the belief network ($r \cong .3$) at moderate levels of attention ($\beta = 1$; see Appendix F for details). Each belief (both personal and social) can have seven different states: -1, $-.66$, $-.33$, 0, .33, .66, and 1. Simulations with other numbers of belief states produce equivalent patterns of results. We calculate probabilities of different belief states $b_i^{'}$ and $s_i^{'}$ according to Equation 1 and Equation 5 and assign a new belief state according to these probabilities. In each iteration, all beliefs are updated in turn and in random order. Every simulation run consists of 100 iterations unless specified otherwise. Unless otherwise noted, beliefs were initialized randomly and all τ were set to 0. For the simulations in Appendix G, we compared the belief dynamics with the attention parameter of interest set to either 0.5 or 2, keeping the other attention parameters constant. The reason we used these values for the attention parameter is that we wanted to compare low and high values for the attention parameter that meaningfully differ, but which are still in a realistic range for the networks we investigate in this article. Note that interpreting absolute values of the attention parameter is not instructive, as the effects of it are contingent on other factors such as the size and connectivity of the network. The illustrations provided in this section show that the values for the attention parameter function in this way—we see pronounced differences in the dynamics for the networks between low and high values for the attention parameter and these dynamics also match roughly what we observe in empirical data. For each illustrative simulation, we ran each attention condition 1,000 times.

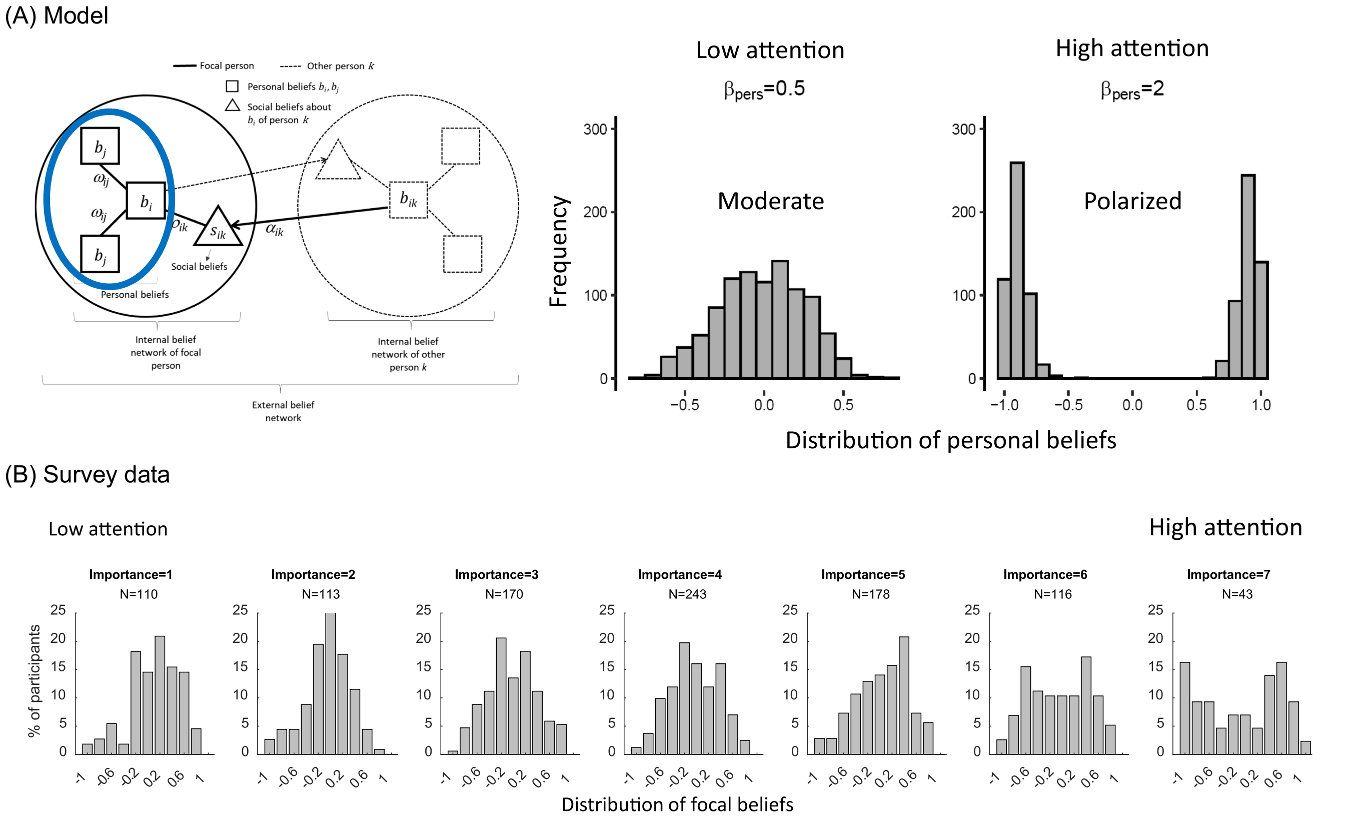
Implications of Attention Directed at Personal Dissonance

For the implications of varying attention directed at personal dissonance, we investigate the mean distribution of personal beliefs within each individual, as well as the dynamics of the mean beliefs over time. The two leftmost panels in Figure 3A show the distributions of mean belief states at the end of each of the 100 runs for each individual, for each level of attention. These results suggest that low attention leads to a normal distribution of mean end beliefs, with most networks ending up in a neutral state (equivalent to how they started). In contrast, high attention leads to a bimodal distribution with most networks ending up in an extreme state. To better understand the dynamics over time, Panels (d) and (e) in Appendix G show the results of five randomly selected simulation runs over time, assuming either low or high attention. Low attention leads to networks fluctuating around the neutral middle point, while high attention leads to more stable networks that settle at more extreme belief states.

These simulation results lead to two basic implications of the NB theory. First, if individuals pay much attention to their potential personal dissonance, the theory predicts a bimodal distribution of belief averages, while it predicts a normal distribution of belief averages when individuals pay little attention to this dissonance. Second, paying attention to potential personal dissonance should lead to a higher stability and resistance of beliefs.

Figure 2

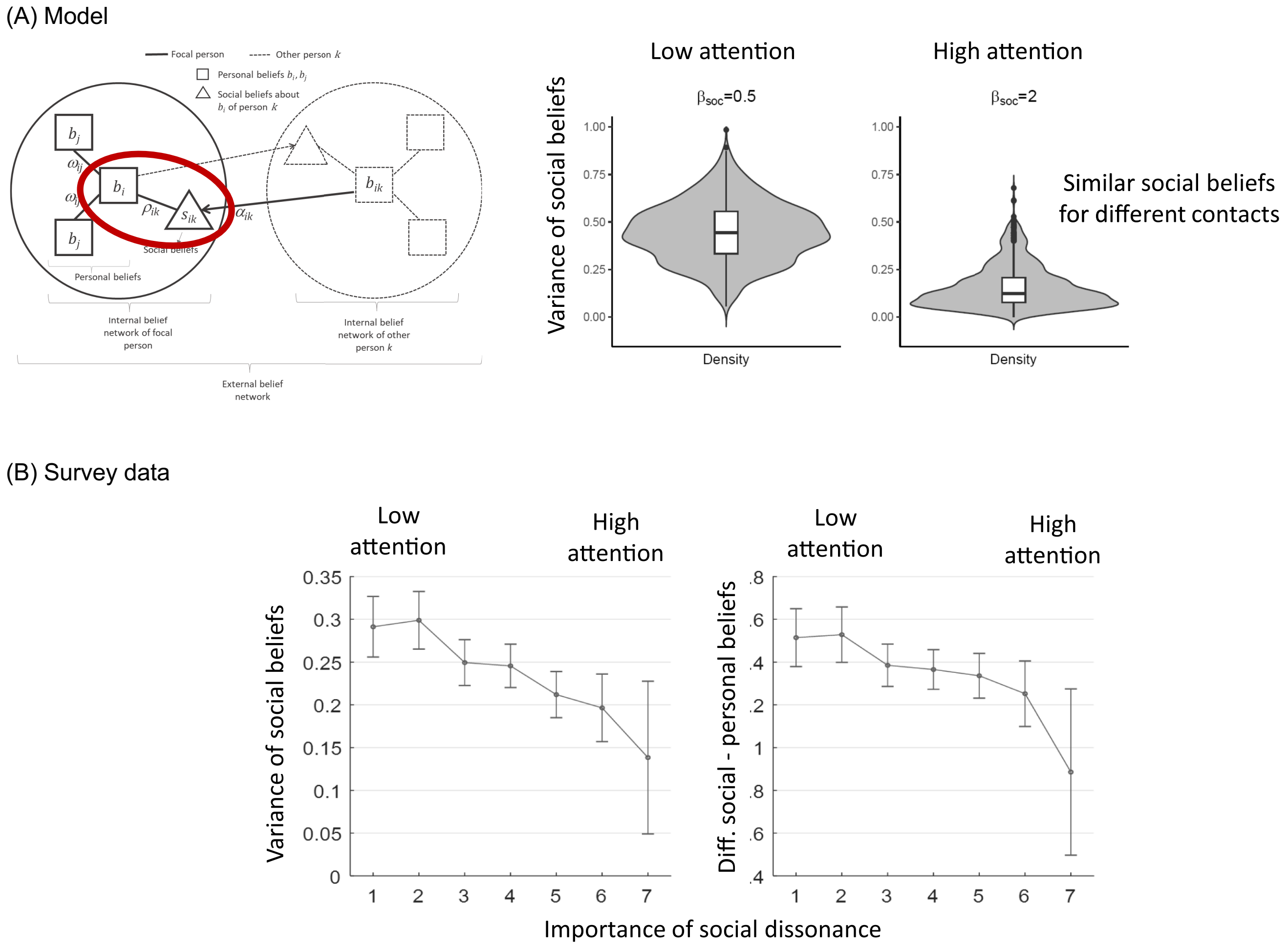
Simulated and Empirical Results Related to the Dynamics of Personal Belief Networks



Note. The predicted distributions of personal beliefs under low and high attention to personal dissonance (Panel A) are echoed in the distributions of average beliefs for participants who assigned low versus high levels of importance to their potential personal dissonance (Panel B, results from Study 1, *N* = 973). See the online article for the color version of this figure.

Figure 3

Simulated and Empirical Results Related to the Dynamics of Social Belief Networks



Note. The predicted differences in variance of social beliefs under low and high attention to social dissonance (Panel A) are echoed in the empirically measured variance of social beliefs about five social contacts, for participants assigning different levels of importance to their potential social dissonance (Panel B, left). As predicted by the theory, this reduction in variance is related to reduced difference between social and personal beliefs of these same participants (Panel B, right, results from Study 1, N = 973). See the online article for the color version of this figure.

Empirical Tests of Implications of Attention Directed at Personal Dissonance

We tested these implications and found support for them. First, directing attention to personal dissonance should result in a bimodal belief distribution among individuals valuing belief consistency, compared to a normal distribution for others. Study 1 (see Appendix A) tested this by examining GM food belief averages among participants differing in consistency importance. Figure 2 shows that high-importance individuals displayed more bimodal belief distributions, confirming the implication.

Second, focusing on personal dissonance should enhance belief stability and resistance, resulting in extreme, consistent beliefs that require strong intervention to alter. Study 1's educational intervention on GM food safety demonstrated that individuals who attribute higher importance to their belief consistency, and thus pay higher attention to their personal dissonance, showed less belief change postintervention. Structural equation modeling confirmed this relationship: Higher attention to dissonance was related to less belief change after the intervention.

Implications of Attention Directed at Social Dissonance

For the implications of varying attention directed at social dissonance, we focus on the correlation between social and personal beliefs and the variance in social beliefs within each individual. As can be seen in the two rightmost panels in Figure 3A, higher attention leads to lower variance of the social beliefs at the end of a simulation run. This is the result of the social beliefs becoming more interdependent with the focal personal belief.

These simulation results lead to two further basic implications of the NB theory. First, the theory predicts that those individuals will have social beliefs that are more consistent with their personal beliefs (Panels f and g in Appendix G). This effect can occur either because these individuals project their beliefs more to others, because they change their personal beliefs to be aligned with their social beliefs, or due to a mix of both these processes. Second, if individuals pay much attention to the potential dissonance between their personal beliefs and their social beliefs, the theory predicts that variance of their social beliefs will decrease. When one's social network is heterogeneous (contacts have different beliefs), this effect will resemble ego projection.

Empirical Tests of Implications of Attention Directed at Social Dissonance

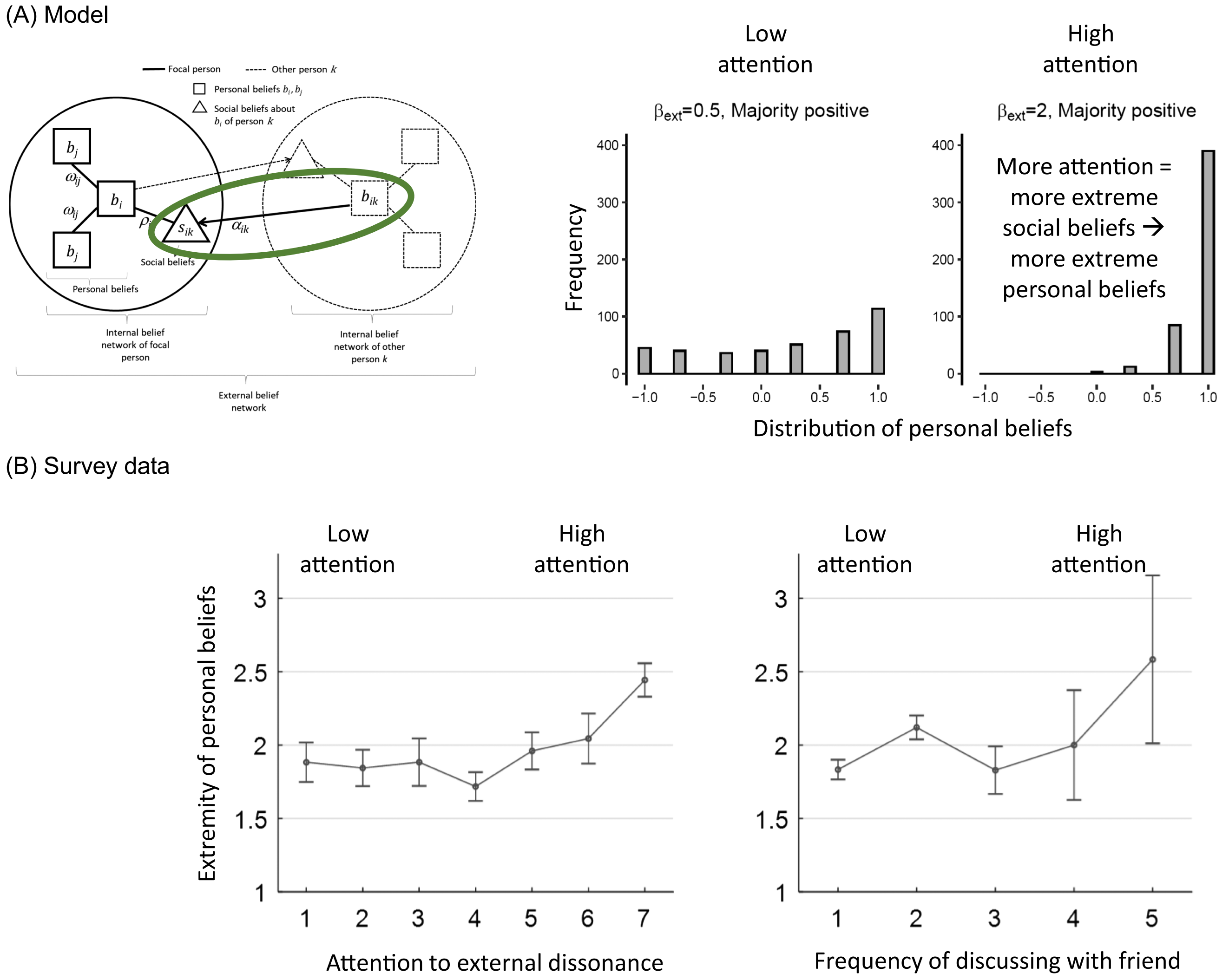
Again, we tested these implications and found support for them. The implications of directing attention to social dissonance are that (a) people who pay more attention to social dissonance will have lower variance of social beliefs and (b) their personal and social beliefs will be more consistent. To test these predictions, we calculate the variance of social beliefs (across the five social contacts) that participants reported in Study 1, as well as the average difference of each of the social beliefs and personal beliefs about the same issues. As Figure 3b suggests, both the variance (left panel) and the difference of personal and social beliefs (right panel) tend to be smaller for those participants who reported a higher importance of social consistency (more attention to the social dissonance).

Implications of Attention Directed at External Dissonance

For the implications of varying attention directed at external dissonance, we focus on how likely it is that group radicalization occurs and how closely social beliefs are aligned with actual beliefs of other individuals. First, as can be seen in the two rightmost

Figure 4

Simulated and Empirical Results Related to the Dynamics of External Belief Networks



Note. The predicted differences in the distribution of personal beliefs under low and high attention to external dissonance (Panel A) are echoed in the empirically measured extremity of personal beliefs, for participants assigning different levels of importance to their potential external dissonance (Panel B, left) and discussing their beliefs with friends more often (Panel B, right, results from Study 2, $N = 669$). See the online article for the color version of this figure.

panels of Figure 4A, higher attention to external dissonance leads to group radicalization in the focal beliefs. The distribution of end focal beliefs becomes skewed toward the extreme end of the initial average view (here, we show only the groups where the majority has positive beliefs) and more so when attention to external dissonance is high. Second, as can be seen in Panels (m) and (n) of Appendix G, social beliefs (perceptions of actual others' focal beliefs) are closer to the actual beliefs when attention to external dissonance is high.

These simulation results lead to the basic implications of the NB theory. If people pay a lot of attention to the dissonance between their social beliefs (what they believe others think) and what others actually believe, group radicalization is more likely to occur and individuals are more likely to agree with each other.

Empirical Tests of Implications Attention Directed at External Dissonance

As before, we tested this implication and found support for it. The implication of directing attention to external dissonance is that groups of people who pay a lot of attention to their external dissonance (i.e., to the correspondence between their social beliefs and others' actual beliefs) will be more likely to show higher levels of radicalization. To investigate this, we analyze data about the extremity of participants' beliefs in the first wave of Study 2 and regress them on our two proxies of attention to external dissonance: the importance of accurately perceiving friend's beliefs and frequency of discussing a topic with the friend. In a mixed model adjusting for clustering of participants within, we find as expected that both of these proxies for attention were positively related to more radical beliefs, as illustrated in Figure 4B (for importance, $\beta = .08$, $p < .001$; for frequency, $\beta = .08$, $p = .02$). These results remain robust after including interactions with topics and controlling for felt social dissonance.

Explaining Established Belief Dynamics Phenomena

In this section, we discuss how the NB theory can explain established phenomena in belief dynamics, including group polarization, group radicalization, and minority influence. First is a note on the terminology. Some authors use the term group polarization to describe extreme differences in beliefs of different segments of the population (e.g., Axelrod et al., 2021; Iyengar & Westwood, 2015; Yardi & Boyd, 2010) or strong dislike among different segments (Mason, 2016). Here, we will use the term "group polarization" in this sense. Other authors (e.g., Friedkin, 1999; Isenberg, 1986; Sunstein, 2002) use the term "polarization" to denote radicalization of a whole group toward the same more extreme position, usually occurring after a group discussion or a mere exposure to arguments in line with an existing view. We will therefore call this type of polarization "group radicalization." Finally, we will discuss separately the phenomena of minority influence on group polarization and radicalization.

Consensus and Polarization

A number of different models have been developed to explain consensus and polarization in societies (Flache et al., 2017; Levin et al., 2021). Long-run consensus is associated mostly with unconditional social influence models. These models assume that if individuals are connected by an edge, they will always influence each other toward reducing their belief differences. For nominal beliefs, examples include voter models (Clifford & Sudbury, 1973; Holley & Liggett, 1975; Redner, 2019), while for continuous beliefs, the most prominent ones are averaging models (DeGroot, 1974; Friedkin & Johnsen, 1990, 2011).

Group polarization has been extensively studied with conditional influence models that assume a threshold or a tolerance level that determines if an individual will be influenced by a neighbor (Axelrod, 1997; Deffuant et al., 2000; Hegselmann & Krause, 2002; van der Maas et al., 2020).

Several models go beyond this threshold dynamics of influence weights and include other psychosocial mechanisms that can produce similar patterns. For example, in Nowak et al.'s (1990) dynamic version of Latané's (1981) theory of social impact, group polarization is explained through an interplay of two individual parameters: persuasiveness—the likelihood that one will influence those who have a different point of view, and supportiveness—the likelihood that one will help those with the same point of view to resist change. In Leonard et al. (2021; see also Franci et al., 2021), it is explained through a dynamic interaction of group-level self-reinforcement and reflective partisanship, which determine the extent of within- and between-group influences. Finally, some models explain polarization through the process of balancing the relationships between beliefs and between individuals holding different beliefs (Pham et al., 2020; 2022; Rodriguez et al., 2016; Schweighofer et al., 2020; see the Discussion section).

In the NB theory, we aim to explain consensus and polarization that occur in structured networks where people are already grouped in clusters due to reasons other than their beliefs, for example, geography, workplaces, and family groups. These structures are ubiquitous in most human collectives. We ask, if such a clustered society starts with equally distributed, largely moderate beliefs, in what circumstances does it become polarized? We posit that consensus and polarization can emerge simply because of different levels of attention people pay to their personal, social, and external dissonances. For example, people's attention to these dissonances can change after prominent events (e.g., surprising elections or court rulings about issues such as abortion and gay marriage). As we show next, when people pay a lot of attention to their personal dissonance (how their own beliefs align to each other), the society remains rather diverse (heterophilous). However, when they pay more attention to their social and external dissonances (how their social beliefs align with their other beliefs and others' beliefs), either complete (radical) consensus or polarization can occur.

To investigate this, we run a simulation with 100 individuals, who each have an internal fully connected network ($\omega = 1$) of 10 personal belief nodes, initialized at random belief states. One of these internal nodes functions as the focal belief in this simulation. This belief is connected, via social beliefs, to focal beliefs of other individuals (see Figure 1). We generate a social network using stochastic block model with two clusters of 50 individuals. We set the probability that any individual is connected to another individual in the same cluster to .4 for both clusters, and the probability that an individual is connected to an individual from the other cluster to .01. We set each type of attention to dissonance (β_{pers}, β_{soc}, and β_{ext}) to either 0.5 (low attention) or 2 (high attention). For each of the resulting eight combination of attention parameters, we simulated 100 runs. In each simulation run, nodes were first randomly initialized and then each updated over 100 iterations. In this section, we summarize the results in Figure 5 (detailed results can be found in Appendix H).

Figure 5 summarizes how NB theory can explain patterns of both consensus and polarization due to attention to personal, social, and external dissonance. Moderate consensus—clustered around a neutral opinion—is likely to occur when individuals do not pay too much attention to personal and social dissonances. On the other hand, either extreme consensus or extreme polarization can occur when individuals do not pay too much attention to personal dissonance and attend strongly to their social and external dissonances. Finally, if individuals pay a lot of attention to personal dissonance, individuals tend to hold extreme beliefs that are independent of the social network, resulting in extreme heterophily (or polarization within groups).

These simulations also point to ways in which radical polarization might be reduced. One is reducing attention to social dissonance, or more generally the need to align personal beliefs with the perceived beliefs of our social contacts. As shown, for example, in Panels (f) and (g) of Figure H1, reducing β_{soc} from 2 to 0.5 leads to occurrence of moderate or extreme heterophily instead of radical consensus/polarization. Note that reducing β_{soc} while keeping β_{ext} high (attention to accuracy of social beliefs) will generally lead to accurate social beliefs. Taken together, our results suggest that promoting originality and nonconformism could be sufficient to reduce unfavorable societal outcomes such as complete polarization or (seemingly less damaging but potentially dangerous in the long run) radical consensus.

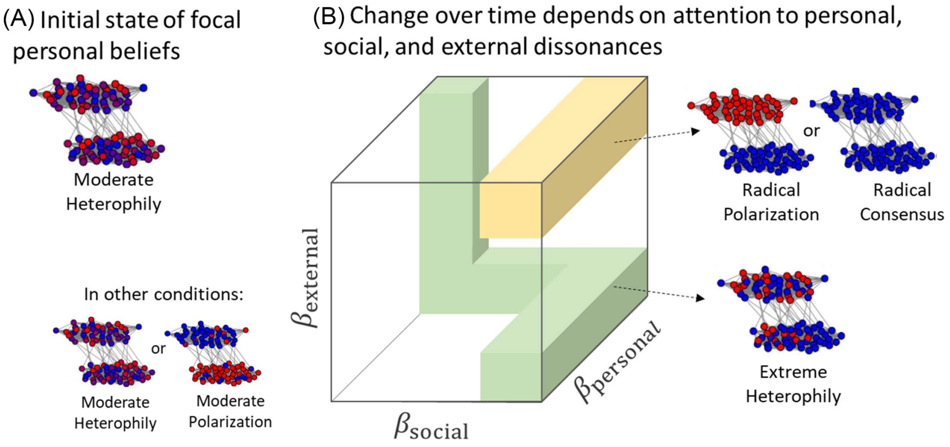
Group Radicalization

Group radicalization refers to the phenomenon that group discussion can lead to a more extreme average group opinion than it was before the discussion. For instance, in Sunstein et al. (2007), liberal groups that discussed issues such as affirmative action or civil unions adopted more extreme beliefs about those issues than they started with, although such radicalization does not always occur (Isenberg, 1986). A variety of accounts have been proposed to explain this phenomenon. Some involve motivational processes such as a desire to be “better” than the group average (Isenberg, 1986) and to be distinct from an outgroup (Turner, 1985). Other accounts postulate sampling processes that lead to higher likelihood of sharing arguments that are in line with majority the view (Burnstein & Vinokur, 1973). Yet, other accounts propose that the core mechanism is realignment of internal belief networks, so that beliefs about an issue align with one’s preexisting and more stable internal preferences or core values such as political or religious orientation. Group discussion, or even just thinking about an issue (Dalege et al., 2018; Monroe & Read, 2008), can make people more aware of the alignment (or lack of thereof) of their belief about a particular issue and their core values. The threshold models described in the previous section, and their extensions, can also produce group radicalization. For example, one mechanism is repulsive social influence. Here, influence weights can be negative or positive depending on the difference in beliefs between individuals (Flache & Macy, 2011; Huet & Deffuant, 2010). This leads these models to naturally settle in bipolar belief states (Flache & Macy, 2011; Huet & Deffuant, 2010; Jager & Amblard, 2005).

The NB theory predicts group radicalization through a novel process where attention to different dissonances produces the phenomenon. To show this, we simulate a group of 10 individuals who are all equally strongly connected to each other (all $\alpha = 1$). All these individuals have an internal network of 10 beliefs. Each of these networks is fully connected (all $\omega = 1$). Each simulation run

Figure 5

Conditions for Different Patterns of Consensus and Polarization



Note. From the initial state of moderate heterophily of focal personal beliefs in the two communities (Panel A), beliefs can change depending on the attention people pay to their personal, social, and external dissonances (Panel B). If they do not pay attention to any of these dissonances, the model predicts no or little change. If they pay attention to both their social and external dissonance, the model predicts either radically different beliefs in the two communities, or complete consensus across communities—depending on small differences in initial conditions (yellow shadings). In all other conditions, the model predicts that beliefs will become more extreme within groups (extreme heterophily; green shadings). Detailed results are shown in Appendix H. See the online article for the color version of this figure.

consisted of two phases of 100 iterations each, with the first phase modeling the period before the group discussion and the second phase the period after the group discussion. In the first phase, both social dissonance β_{soc} and external dissonance β_{ext} were set to 0, representing that individuals are not communicating with each other yet. In the second phase, both β_{soc} and β_{ext} were set to either 0.5 or 2, representing less and more engaged group discussion. In addition, attention to personal dissonance β_{pers} was also set to either 0.5 or 2, representing less or more pressure to align one's personal beliefs with each other. Belief states at the beginning of each simulation were initialized at random. A summary of the results of these simulations is shown in Figure 6. Detailed results can be found in Appendix I.

Figure 6 shows that when people pay a lot of attention to how well their personal beliefs align with their social beliefs, group radicalization is likely to occur. This is even more likely when people also pay attention to how their social beliefs align with the actual beliefs of others. More intense group discussions and in particular those that focus on (dis)agreement between individuals might raise the attention to both of these sources of dissonance (social and external), making group radicalization more likely. In contrast, paying attention to personal dissonance can reduce group radicalization. These results also suggest ways to reduce group radicalization: by reducing attention to social dissonance, for example, by encouraging nonconformity and originality during group discussions, and by increasing attention to personal dissonance, for example, by reminding people of their core values that might be in contrast with beliefs of others.

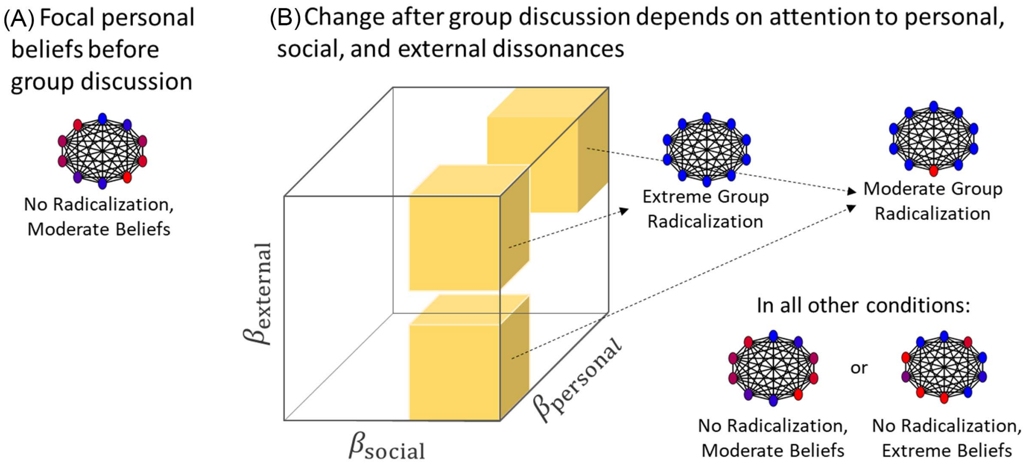
Minority Influence

Minority influence refers to the process by which a small group of individuals with dissenting beliefs and behaviors can gradually influence and change the beliefs and behaviors of the majority (Moscovici et al., 1969; for reviews, see, e.g., Martin & Hewstone, 2009). A plethora of factors can affect how successful a minority is in influencing a majority (Crano, 2010; David & Turner, 2001; Quiamzade et al., 2010; C. M. Smith & Tindale, 2010). Minorities can directly influence focal beliefs of the majority or indirectly influence related beliefs even when focal beliefs of the majority remain unchanged.

Here, we focus on modeling the conditions for direct minority influence. We ran a simulation with 30 individuals embedded in a network with two clusters generated by a stochastic block model. The majority cluster consisted of 20 individuals whose belief states were initialized randomly in the range from −1 to 0. The minority cluster consisted of 10 individuals whose belief states were initialized randomly in the range from 0 to 1. The probability that any individual in the majority cluster was connected to any other individual in the majority cluster was set to .33, while the probability that any individual in the minority cluster was connected to any other individual in the minority cluster was set to .67. These settings resulted in an average degree of 6 within both clusters taking into account only the connections within each cluster. The probability of a link between the minority and majority clusters was set to .1. We also investigated stochastic block model networks with more sparsely connected clusters, as well as random networks, but the patterns of results were similar.

Figure 6

Conditions for Group Radicalization Before and After Group Discussion About Personal Beliefs


(A) Focal personal beliefs before group discussion
No Radicalization, Moderate Beliefs
(B) Change after group discussion depends on attention to personal, social, and external dissonances
Extreme Group Radicalization
Moderate Group Radicalization
$\beta_{external}$
β_{social}
$\beta_{personal}$
In all other conditions:
or
No Radicalization, Moderate Beliefs
No Radicalization, Extreme Beliefs

Note. In network representations, circles represent individuals with their personal and social beliefs (equivalent to the large circles in Figure 1), and edges represent the directed influence of individuals' actual beliefs to the social beliefs of their contacts (each equivalent to the two edges connecting the individual circles in Figure 1). Blue nodes indicate negative beliefs and red nodes indicate positive beliefs, with higher color saturation corresponding to higher extremity of beliefs. We track four types of networks that occur at the end of simulations assuming different combinations of attention parameters, starting from a network where people have moderate beliefs and there is no group radicalization (Panel A). After group discussion (Panel B), the model predicts that when people pay attention to their social dissonance, moderate or extreme radicalization will occur, depending on the level of attention to personal and external dissonances. If people do not pay attention to their social dissonance, beliefs can become more extreme but there will be no whole group radicalization (detailed results are shown in Appendix I). See the online article for the color version of this figure.

We varied the attention parameters independently for the majority and minority clusters. For ease of interpretation, we varied the attention to social and external dissonances together, setting both β_{soc} and β_{ext} to either 0.5 or 2. The attention to personal dissonance β_{pers} was set independently to either 0.5 or 2. For each combination of attention parameters, in both majority and minority clusters, we simulated 100 runs. Belief states at the beginning of each simulation were initialized at random. A summary of the simulations is shown in Figure 7 (detailed results can be found in Appendix J).

Taken together, as the summary of the simulations in Figure 7 suggest, the minority will have the most influence on the rest of the network if they strive to keep their personal beliefs consistent, and if the majority is distracted and does not pay attention to their personal, social, or external dissonance. These predictions fit well with empirical findings showing that minorities must hold consistent beliefs to influence majorities (Moscovici et al., 1969; Wood et al., 1994). At the same time, the majority will have the most influence on the minority if they pay attention to those different sources of dissonance and strive for consistency of beliefs, independently of what the minority does.

Comparing Model Predictions With Real-World Trends

In this section, we discuss whether the NB theory can reproduce belief dynamics found in real-world data. We first model increasing polarization in the United States over time and then turn to cross-sectional distributions of beliefs in different European countries.

Empirical Trends in Group Polarization

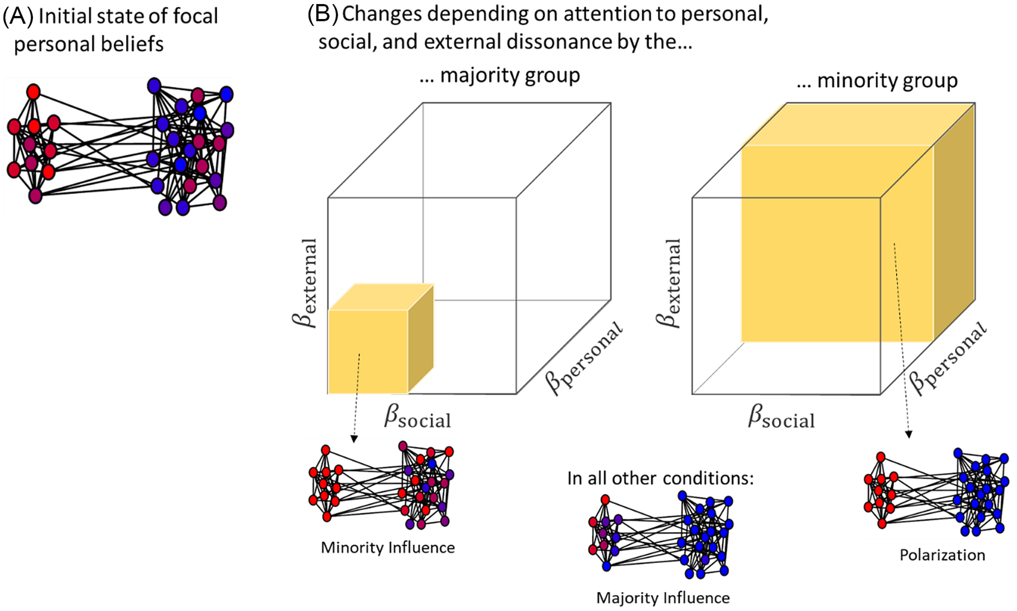
A trend observed in many real-world settings is, as mentioned before, an increased polarization of societies into groups holding extreme and opposing beliefs. For example, as Figure 8 shows, evaluations of Democrat and Republican political candidates, as recorded in American National Election Studies, have become more and more extreme over the last several decades. Notably, interest in political campaigns has also increased in the same period.

The NB theory explains group polarization and related phenomena through the interplay of attentions to personal, social, and internal dissonances (see the Consensus and Polarization section). However, these patterns and their trends over time can also be explained by increased attention to all dissonances together with lower connectivity. In particular, when people pay a lot of attention to the consistency of their personal and social beliefs, as well as to correspondence of others' perceived and actual beliefs, either radical consensus or polarization can occur, depending on small biases in initial conditions. If the two groups are even a little bit biased toward opposing positions to begin with, an increased attention to these dissonances will likely lead to radical polarization.

It is reasonable to assume that the increased interest in political campaigns relates to increased attention to political beliefs of others one knows and to the dissonance between own and others' perceived and actual beliefs. According to the NB theory, this would then naturally lead to increased radicalization of political beliefs within initially moderate groups of Democrats and Republicans, and consequently to group polarization.

Figure 7

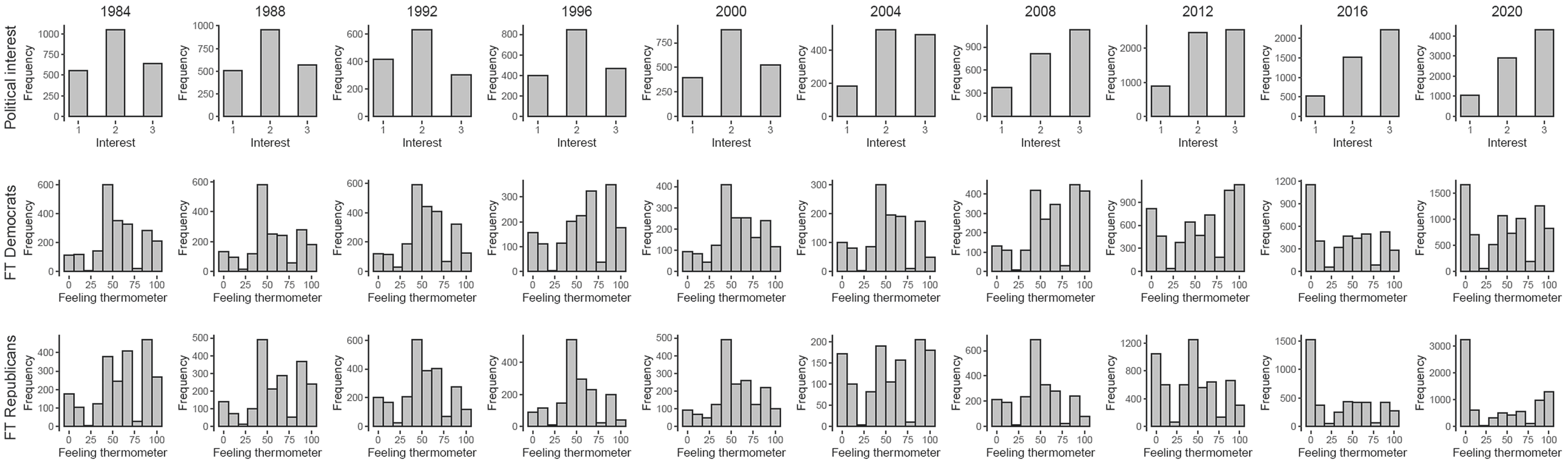
Conditions for Minority and Majority Influence



Note. Starting from an initial state where minority and majority have substantially different beliefs (Panel A), minority can prevail only if the majority group does not pay attention to any of their potential dissonances (Panel B; averaged over all levels of attention for the minority group). If the minority group pays attention to their personal dissonance, they can avoid majority influence at the cost of strong polarization between the two groups. For all other combinations of attention parameters, majority will influence the minority group. Detailed results are shown in Appendix J. See the online article for the color version of this figure.

Figure 8

Trends in Answers to Questions About Political Interest



Note. Trends in answers to questions about political interest (top row, question "Would you say that you have been [not much interested-somewhat interested-very much interested] in the political campaigns so far this year?") and evaluations of Democrat and Republican political candidates that year ("How would you rate candidate [on the feeling thermometer]?"). Data and full text of questions are available at https://electionstudies.org. FT = feeling thermometer.

To investigate this prediction in more detail, we ran a simulation where we varied network structure (from two initially disconnected clusters to two strongly interrelated clusters, top to bottom rows of Figure 9) and increased all attention parameters gradually over time (from $\beta_{\text{pers}} = \beta_{\text{soc}} = \beta_{\text{ext}} = 0.25$ to 1.5 over five time steps). Each network included 30 people, with 15 people in each cluster. The clusters differed in their exogenous influence; the first cluster had negative exogenous influence, $\tau = -.5$, and the second cluster had positive exogenous influence, $\tau = .5$. These differing exogenous influences could, for example, represent one cluster consisting of people living in a rural area, making them more predisposed to identify as conservatives, and the other cluster consisting of people living in an urban area, making them more predisposed to identify as liberals. Within-cluster probability of connection was .35, and between-cluster probability was either 0, .01, .1, or .35. Initial beliefs in each cluster were set randomly. We ran 100 runs with eight iterations in each run.

Figure 9 shows the results for each time step, suggesting that group polarization can occur from random initial beliefs purely because of the increase in people's attention to dissonance. This is likely to happen whenever there are sparse or no connections between the two clusters (the first two rows in Figure 9). When the two clusters are moderately related (connectivity 0.1, third row in Figure 9), both polarization or consensus are about equally likely. Finally, when the two clusters are tightly related, consensus will occur in the direction of whatever the initial bias was.

Importantly, as shown in Figure 5, extreme differences in beliefs (extreme heterophily) can occur even in tightly connected communities, in particular when people pay a lot of attention to their personal dissonances rather than external and social dissonances. Thus, the NB theory provides explanations for both polarization between groups (Figure 9) and within groups (extreme heterophily in Figure 5).

These results have implications for interventions that could be used to reduce the likelihood of group polarization. Paradoxically, reducing people's attention to political campaigns would help, but so would an increased level of intergroup contact between different groups. Note that such an increased contact could lead to adoption of whatever option was in slight majority initially. Note, however, that the effects of intergroup contact are multifaceted and that the nature of the contact, the specifics of group characteristics, and the beliefs in question most certainly also determine to what extent polarization or consensus will be achieved, or if the behavior of the individuals and groups will change. For example, surface-level contact as Anglo-Whites interacting with Spanish-speaking confederates on rail platforms can increase exclusionary beliefs (Enos, 2014), tolerant behaviors against other groups of people might not generalize beyond the specifics of the intervention (e.g., Mousa, 2020), and exposure to opposing views can increase political polarization (Bail et al., 2018).

Cross-Sectional Distribution of Beliefs

We now look at how well the NB theory can explain specific distribution of beliefs observed at a single time point in different countries. Figure 10 shows data from four countries about respondents' own political placement and opinion about the European Union. Flache et al. (2017) argued that these data showed evidence of several belief dynamic processes. The central peak suggests assimilation to the central or moderate belief in the population, the nonextreme clusters just next to the center suggest similarity bias, and the (small) extreme peaks might suggest polarization produced by a mix of unconditional and repulsive processes.

The NB theory can reproduce this general pattern conditional on two empirically testable assumptions. First, most individuals pay little attention to the consistency of their personal beliefs about these topics, meaning that their attention parameters β_{pers} is low. Second, political placement represents an average of several more specific beliefs. To show that these assumptions can reproduce the empirically observed distributions, we ran a simulation with 2,000 individuals (this number was chosen in order to have a similar sample size as in the empirical data) who each have a fully connected network consisting of 50 belief nodes, representing the many specific beliefs that give rise to the answer on the political placement and the European Union questions. In this simulation, we focus only on dissonance arising from internal beliefs and omit the external network by setting the attention parameters for both social and external dissonances to 0. We draw the values of the attention toward personal dissonances from exponential distributions with rates of either 2 (representing an exponential distribution with many extreme values), 5 (representing an exponential distribution with moderately many extreme values), or 10 (representing an exponential distribution with few extreme values). These distributions represent populations in which most individuals have low levels of attention to the consistency of their beliefs, with more or fewer individuals paying some attention. Exponential distribution is appropriate for such cases with a cutoff at 0 and a high upper bound.

As shown in Figure 11, our simulations can reproduce some of the main patterns in the empirical answer distributions, depending on the attention distribution in the population (insets in Figure 11). The assumption that most people are not paying much attention to the consistency of their beliefs (the middle and right panels in Figure 11) produces predicted answer distributions that are closest to the empirical data. Attention distributions alone cannot explain biases toward one or the other end of the answer distributions noticeable in Figure 10. Those biases could be explained by assuming exogenous influences on people's beliefs, such as political leaders and specific country-level events. Although NB theory can reproduce these patterns in the answer distributions, it is important to note that assumptions of the amount of attention to the consistency of personal beliefs about these topics and that the political placement represents an average of several more specific beliefs need empirical validation.

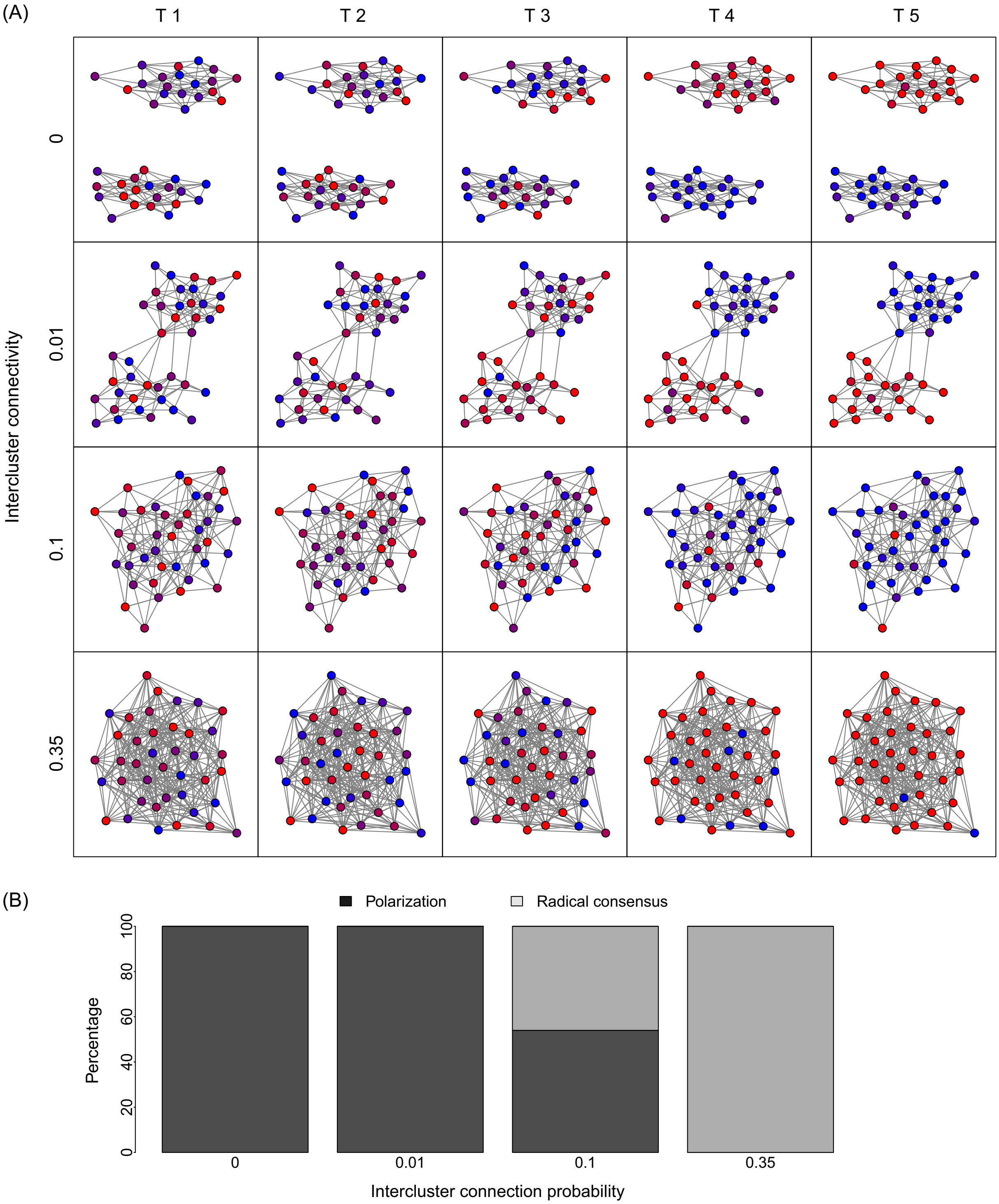
Discussion

In this article, we develop an integrative theory of belief dynamics, the NB theory. It goes beyond past theories in three ways. First, it makes a direct connection between internal beliefs and external social worlds, through people's social beliefs or perceived beliefs of others. Second, it can explain how social beliefs can become more or less accurate representations of the actual beliefs of others, and how that affects the resulting belief dynamics. Third, the theory explains diverse belief dynamics phenomena parsimoniously through the differences in attention and the resulting felt dissonances in personal, social, and external parts of belief networks.

We implement the NB theory as a computational model in a statistical physics framework in which beliefs are represented as spins, potential dissonance is represented as energy, and attention is

Figure 9

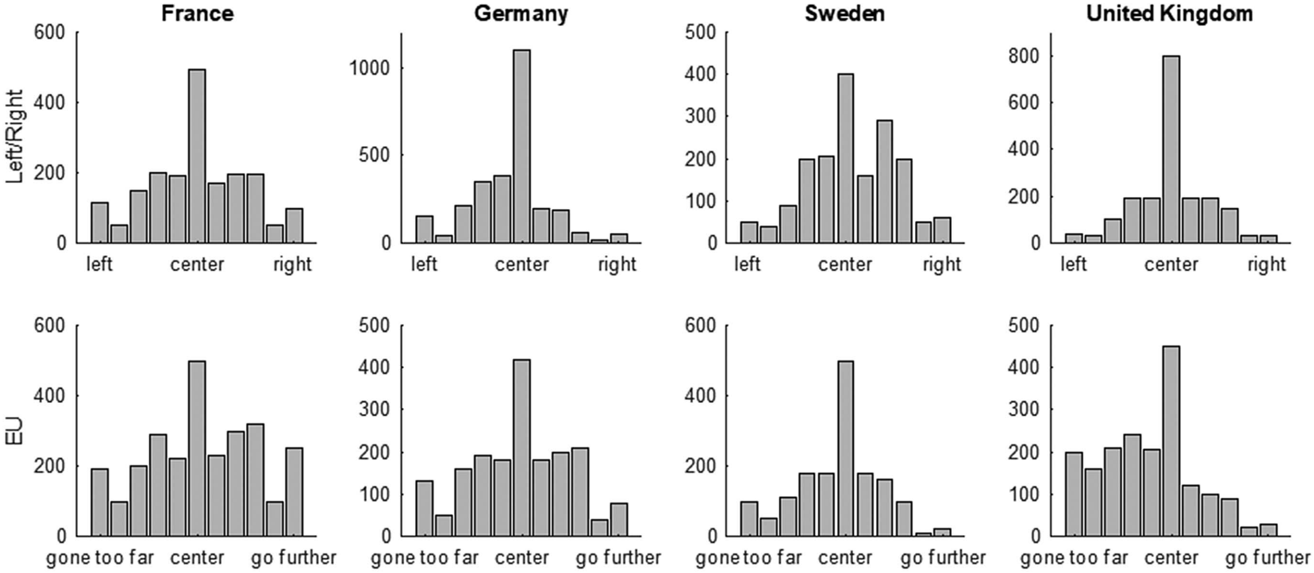
Simulation of Belief Dynamics Occurring on Networks With Different Levels of Intercluster Connectivity



Note. Simulation of belief dynamics occurring on networks with different levels of intercluster connectivity (Panel (A), top to bottom row: from no to high connectivity). At each time step, attention to dissonance increases, resulting in different end patterns of beliefs. When intercluster connectivity is zero or low, the dynamics always leads to radical polarization (the first two bars in Panel (B)). When it is medium (0.1), the dynamics ends up in radical consensus half of the time, and in radical polarization the other half, depending on small differences in initial belief distribution in the two clusters (the third bar in Panel (B)). When the connectivity is high (0.35), the end result is always radical consensus (the fourth bar in Panel (B)). See the online article for the color version of this figure.

Figure 10

Empirical Distributions of Answers to Questions About Political Position



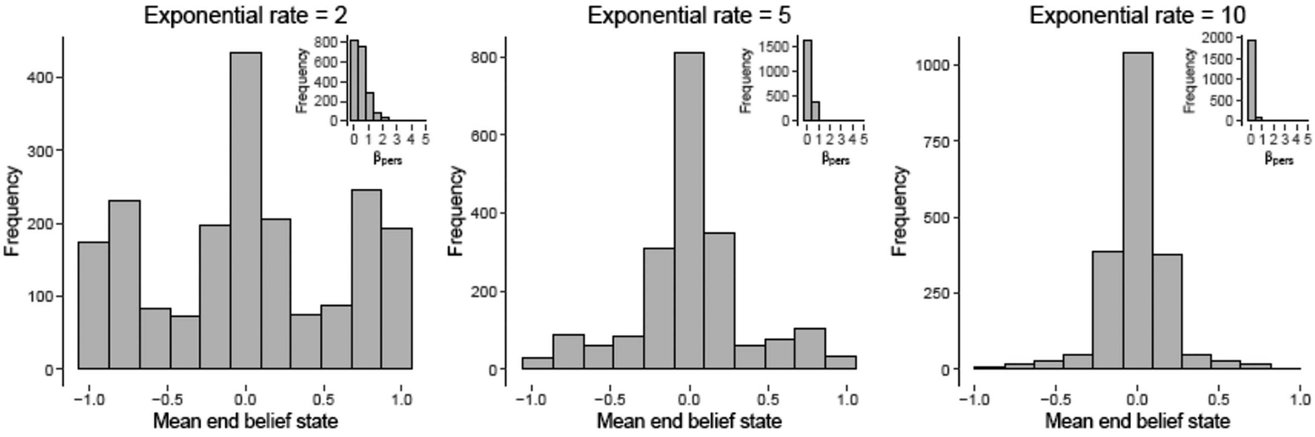
Note. Empirical distributions of answers to questions about political position (top row, question "in politics people sometimes talk of 'left' and 'right.' Where would you place yourself on this scale?") and the European Union (bottom row, question "Now thinking about the European Union, some say European unification should go further. Others say it has already gone too far. What best describes your position?") in four European countries, from the 2012 European Social Survey (adapted from Figure 3 in Flache et al., 2017).

represented as temperature. We test our theoretical assumptions in two large survey studies and show that they are justified (see the Relating Psychosocial and Statistical Physics Constructs section and Appendices A–C).

We then pose different predictions based on simulations of the internal and external networks and find empirical support for them using survey data and survey experiments. Specifically, we predict the following: (a) If individuals pay a lot of attention to their potential personal dissonance, their average beliefs will show a bimodal distribution, but if they pay little attention, their average beliefs will be distributed normally; (b) paying attention to potential personal dissonance leads to more stable beliefs; (c) if individuals pay attention to the potential dissonance between their personal beliefs and their social beliefs, variance of their social beliefs will decrease; (d) those individuals will also have social beliefs that are more consistent with their personal beliefs; (e) when people pay a lot of attention to the dissonance between their social beliefs (what they believe others think) and what others actually believe, group radicalization is more likely to occur and individuals are more likely to agree with each other; and (f) individuals are predicted to have more accurate social beliefs if a belief is discussed and/or expressed often in a group.

Figure 11

Predicted Distribution of Answers to Questions Shown in Figure 10



Note. Insets show hypothesized distributions of attention parameters.

Finally, we show that the computational implementation of NB theory can account for established belief dynamics phenomena including consensus and polarization, group radicalization, and minority influence, providing new parsimonious explanations for these diverse phenomena. We also show that the model can reproduce real-world trends in group polarization over time and in cross-sectional distribution of beliefs.

Relationship to Other Models

How does our model compare to the numerous previous accounts? Models of belief dynamics can be compared on many different dimensions. The sheer number of models that have been proposed in the literature, and the lack of tests against empirical data, makes it difficult to compare the predictive accuracy and the empirical validity of the models' assumptions. Many of these models can in principle account for a wide range of phenomena, but what mechanisms are necessary and/or sufficient to explain these phenomena are unknown. The scope of such an investigation of belief dynamics models is beyond the scope of this article, therefore focus on structural and process features of models: (1) Does the model represent one or more personal beliefs; (2) if it has more than one belief represented, does it represent them as independent beliefs, in a network, or does it assume a summary representation; (3) does the model include social influence, and if so, does it include social influence from only one agent at the time or the whole network; (4) are other's beliefs based on the factual beliefs, or are they represented subjectively; (5) what are the belief updating mechanisms for belief nodes; (6) what are the mechanisms for updating edge weights; (7) what are the mechanisms for updating the social network; and (8) to what extent does the model been tested on empirical data. In this section, we will focus on comparing our model with models that are most similar to ours according to these features.

Our model represents several personal beliefs in a network (Features 1 and 2 above) together with a subjective representation of others' beliefs (Feature 4) and an external factual belief network that can influence the subjective representation of others' beliefs (Feature 3). The updating mechanism for nodes follows the minimization of the weighted sum of the personal, social, and external energies (Feature 5), which in turn relies on the weighted average of all beliefs (personal beliefs have unique weights between beliefs, while social and external beliefs rely on one weighting parameter each, Feature 7). The edge weights in our model can be derived by the relations in empirical data (6; i.e., partial correlations). In terms of tests on empirical data, we provide several tests of core assumptions (Feature 8, see Appendices B and C). Given this feature list, this means that many prominent belief dynamic models will be left out in this comparison. These include models that only represent one belief or vectors of noninteractive beliefs (Features 1 and 2), such as French's (1956) formal model of social power, Harary's (1959) generalization of French's model, and DeGroot's (1974) consensus formation model and others that followed in this and other traditions such as the bounded confidence models (Deffuant et al., 2000; Hegselmann & Krause, 2002; Weisbuch et al., 2002), vector models based on attraction or assimilation and rejection or repulsion mechanisms (Flache & Macy, 2011; Huet & Deffuant, 2010; Jager & Amblard, 2005), social influence network theory (e.g., Friedkin & Johnsen, 1990, 2011), computational implementations of social impact theory (Nowak et al., 1990), models of the dissemination of culture (Axelrod, 1997), vector models that combine demographic and belief representations (Flache & Mäs, 2008), and several models inspired by statistical physics (e.g., Galesic & Stein, 2019; Pham et al., 2022; for reviews of models inspired by statistical physics, see Castellano et al., 2009, and for a general overview of social influence models, see Flache et al., 2017).

There are models that represent beliefs as networks (Feature 2 above), but only include social influence implicitly (Feature 3). For example, in the AE framework (Dalege et al., 2016, 2018), which we use as the basis for the representation of the internal belief network in our model, exogenous or social influence can only be represented with the external field parameter (τ in Equation 3 above). A similar model with only nonspecific exogenous influences is Brandt and Sleegers (2021). An influential class of models that also represent beliefs as networks is constraint satisfaction models. Shultz and Lepper (1996) formulated a constraint satisfaction model that aims at reducing dissonance in a network of cognitions. In this model, there is no explicit representation of social influence, but the general formulation of this and other constraint satisfaction models allows for nodes that represent various external or social influences as they are represented in the network. A more elaborate neural network model of constraint satisfaction is Monroe and Read's (2008) attitudes as constraint satisfaction model. Their model has separate banks of nodes for cognitive representations and for persuasive communications, and they investigate the impact of external messages represented on the persuasion units. There are also other neural network models that are not based on constraint satisfaction. For example, Van Overwalle and Siebler's (2005) model is based on an autoassociative recurrent network that focuses on assimilating new information instead of the settling of the network in a steady state.

There are not many models that both have a network representation of internal beliefs and a representation of an external network (Features 2 and 3). A model that explicitly represents belief networks and social networks is Goldberg and Stein's (2018) *associative diffusion model*. The aim of this model is to explain cultural differentiation without the need for segregated networks. It is also based on the perception of behaviors of others (Feature 4), but in contrast to NB theory, it does not include the effect of actual beliefs on the perception of actual belief. In G&S, observing others behavior directly influences their own beliefs, while in our model that updates the social beliefs that might update the personal beliefs. The belief updating mechanism (Feature 5) is similar to that of NB theory in that it is based on a constraint satisfaction mechanism. In contrast to NB theory, the edge weights (Feature 6) in the associative diffusion model are determined by a reinforcement scheme with decay. The assumptions and predictions of the associative diffusion model have not been tested against data (Feature 6).

A model that shares the same basic assumption about the internal network is the HIOM (van der Maas et al., 2020), only without the social beliefs. In the HIOM, the internal network of personal beliefs is described by a mean-field approach resulting in modeling these networks as cusp catastrophes: Strongly involved individuals have resistant beliefs but change them drastically if they do change, and weakly involved individuals have less resistant beliefs and change them in a more gradual manner. These dynamics are also implied by the NB theory as it is an extension of the AE framework (for an illustration of these dynamics implied by the AE framework, see Dalege et al., 2022). The HIOM additionally assumes that involvement increases when individuals interact. A crucial difference between the HIOM and the NB theory lies in the representation of social beliefs—the HIOM does not differentiate between actual beliefs and perceived beliefs of others (Feature 4). The HIOM therefore cannot explain that individuals differ in how much importance they attach to their personal versus social beliefs. This difference also leads to a different coupling between individuals in the HIOM than our NB theory. While the HIOM assumes that averages of more specific beliefs are communicated between individuals, the NB theory assumes that beliefs are coupled between individuals through perceived beliefs. Additionally, the HIOM treats beliefs as binary while the NB theory treats beliefs as continuous. A potentially fruitful avenue for future research would be to integrate the dynamics of increasing attention due to their interaction into the NB theory.

Another model that shares features with ours is the social cognitive–social belief model Rodriguez et al. (2016). The aim of this model is to integrate insights from network models of social influence and models that represent personal beliefs as networks (Features 2 and 3 above), and as such, this is very much in line with the aims of our model. Another similarity to our model is the reliance on a statistical physics formalism in the sense that the model relies on minimizing an energy function that combines energies in the internal network with weights attached to each of the personal and social components in the energy function. The main difference between our model and the Rodriguez et al. model is that their nodes represent concepts and the signed edges between them are the associative beliefs, while in our model the nodes are the beliefs, and the edges determine the relation between the beliefs. In contrast to our energy function for the personal beliefs which takes a weighted average over all pairs of connected nodes, in Rodriguez et al.'s model, the average is taken over the edge weights of all triads. Rodriguez et al. allowed for the possibility that the beliefs are not factual (Feature 4), although they do not separate personal beliefs from social beliefs. The external network in Rodriguez et al.'s model represents the individuals' and the factual beliefs of these individuals. A comparison of mechanisms of edge weight updating is not straightforward as the edge weights in Rodriguez et al.'s belief network are the beliefs, and as such, they can be updated by accepting a social contact's belief if it decreases their individual energy (Feature 6). In terms of empirical tests (Feature 8), the simulations of Rodriguez et al.'s model are compared in broad terms to different patterns that have been observed in empirical data, such as how the existence of "zealots" shapes belief dynamics and the entrenchment of fringe groups.

Another model that integrates insights from network models of social influence and models that represent personal beliefs as networks (Features 2 and 3) is the one presented in Ellinas et al. (2017). In contrast to our model, Ellinas et al.'s model is not based on the statistical physics notion of minimizing energy and is focused on the probability of accepting associations from neighboring individuals (Feature 7). The probability that an individual accepts an incoming association from a randomly selected other individual is a function of both the portion of the individual's neighbors that agree with that belief and the social rank difference between the receiver and its neighbors, with different weights allowed for each of these components. In Ellinas et al.'s model, the belief networks and the social rank of individuals are estimated from survey data. In contrast to our model, this model has a mechanism to change the value of edges between personal beliefs by accepting proposed link values by other neighboring individuals (Feature 6). However, the node values are static and do not change (Feature 5). Ellinas et al. conducted theoretical explorations of different parameter values, but there are no empirical tests of assumptions or predictions (Feature 8).

In contrast to Ellinas et al.'s (2017) model that does not have a mechanism for updating of beliefs, the weighted network balance model described in Schweighofer et al. (2020) also has a mechanism for updating node values (Feature 5 above). The model is a computational extension of the ideas in Heider's (1946) balance theory and Cartwright and Harary's (1956) structural balance theory and is particularly aimed at explaining hyperpolarization, which Schweighofer et al. (2020) defines as the simultaneous occurrence of belief extremeness and issue constraint for multidimensional beliefs. In contrast to our model, where we distinguish between internal and external belief networks, the weighted network balance model considers relations between two Individuals A and B and a policy issue, D. The model is a combination of network representations and vector representations (Features 1 and 2). The beliefs of any individual are represented with a vector that is all beliefs of an individual toward the policy issue. These interpersonal beliefs in the model do not need to reflect reality but might be purely subjective (Feature 4). These vector representations of beliefs are then combined with signed geometrical means. These are then used to determine edge weights between the individuals the policy issue under consideration (Feature 6). Beliefs are updated by moving them closer to a perfectly balanced opinion vector. In Schweighofer et al. (2020), the weighted network balance model is tested with data from American National Election Study (Feature 8).

In sum, our comparison of NB theory and other models suggests three main contributions of the NB theory. It includes both internal beliefs and external social worlds, connected through people's social beliefs or perceived beliefs of others. It explains how these social beliefs can differ from the actual beliefs of others, which is often disregarded in models of belief dynamics. Third, while many of these models implement belief change processes through some form of dissonance, incoherence, or imbalance reduction in internal belief networks or in the relation between beliefs in social networks, they do not explicitly differentiate dissonances due to the consistency of personal and social beliefs and the correspondence of social and actual others' beliefs.

We also make a contribution by providing extensive empirical tests of the core assumptions of our theory as well as of specific empirical predictions. This is very rare, with most models being presented on the conceptual level without empirical tests (Castellano et al., 2009; Redner, 2019). Going forward, given the plethora of models that can account for phenomena such as consensus, polarization, and radicalization, an estimate of model mimicry (Wagenmakers et al., 2004) in the area of belief dynamic models would be very helpful to tease out what representations and mechanisms are really needed to explain pertinent phenomena.

Limitations and Possible Extensions

The statistical physics framework that we used here is just one the many analogies that have been used to explore and understand human belief dynamics. Many other analogies have been used for this purpose, including epidemiological models, where transmission of belief is like a transmission of disease (Newman, 2003); percolation, where beliefs seep through a society like liquid through a substance (Duffie et al., 2010); balance, where beliefs and individuals align in a way that leads to most consistent relationships on the level of pairs and triads (Heider, 1958; Pham et al., 2020); expected utility, where beliefs change to maximize the product of value and likelihood of different cognitions (Ajzen, 1991; Fishbein & Ajzen, 1975); evolution, where beliefs evolve in the process of cultural learning (Richerson & Boyd, 2008); Bayesian networks, where networks of beliefs change in line with their conditional dependencies (Cook & Lewandowsky, 2016); forces, where belief change under combined influence of several distinct social forces (Latané, 1981); and networks, where systems of beliefs are conceptualized as networks that aim to minimize the overall energy (Dalege et al., 2016).

While analogies can be very useful to cope with novelty and uncertainty in science and life in general (Dunbar, 1997; Holyoak & Thagard, 1996), it is important to keep in mind that they were "borrowed" from another domain that does not necessarily correspond to all of the intricacies of the domain we wish to explain (Gigerenzer, 1991; Olson et al., 2019), in this case belief dynamics. We need to recognize the unnecessary "baggage" of worldview, assumptions, and methodologies that are transferred to the domain of belief dynamics along with the analogy.

In the case of statistical physics analogy, such potential baggage includes its assumptions that beliefs are like spins and that there is no strategic behavior, no individual differences, no emotions, and no institutions. Another important baggage is the assumption that people always want to minimize dissonance among their beliefs, while it is clearly possible that people sometimes can consciously maintain some dissonance between their beliefs. For example, scientists are trained to forego changing their beliefs about scientific issue just because they do not fit with their moral values.

More specifically, in the NB theory, we assume that the main mechanism for belief change is dissonance reduction, and we implement it within a statistical physics framework that assumes that people find belief states that minimize energy. We want to make clear that we do not believe that all belief change is caused by an optimization process of one criterion. For a full understanding of belief dynamics, a comprehensive theory of belief dynamics should include mechanisms that go beyond the goal of dissonance or energy reduction, although several mechanisms of belief change can be incorporated in a framework that is assumed to minimize energy. For example, another form of consistency criterion that has been applied to several belief dynamic models is triangle balance (e.g., Pham et al., 2022; Rodriguez et al., 2016; Schweighofer et al., 2020). That is, a triangle of nodes is balanced if it includes zero or two negative links; otherwise, it is unbalanced. This form of consistency mechanism could be incorporated as a separate term in the energy function (see, e.g., Pham et al., 2022). Going beyond dissonance, balance, and minimization of energy, other mechanisms have been proposed as fundamental for belief change. For example, information integration in the form of summation and averaging can produce belief change (Anderson, 1971). For social and external beliefs, in this view, people are not trying to minimize dissonance or achieve balance, but might just be influenced by people around them without ever considering dissonance or balance.

Another limitation of our statistical physics implementation of NB theory is that it assumes primacy to edge weights in the sense that these influence node values but nodes do not influence edge weights. There is no mechanism in the model that can differentially change the edge weight between two nodes. The only mechanism that can change the values of all edge weight is the attention parameter β. That is, increasing or decreasing the value of β has the same effect as increasing or decreasing all edge weights. Possible extensions could include edge weight updating mechanisms from neural network models of belief dynamics (e.g., Monroe & Read, 2008).

Finally, in the current formulation, NB theory and its implementation is only concerned with beliefs and their dynamics and not the relation between beliefs, preferences, and behavior The beliefs are not at present connected to a preference to choose a specific option or act in certain way. In addition, we have not modeled the strategic considerations related to beliefs. For example, people might choose to signal some beliefs, but not others, and some beliefs might be signaled covertly (van Der Does, Galesic, et al., 2022).

Conclusion

NB theory connects internal and external belief networks. Unlike many other models of belief dynamics, it does not assume that people perceive others' beliefs accurately, but allows that these perceptions can be shaped by both personal beliefs and others' actual beliefs. The theory can explain different phenomena in belief dynamics largely by one basic mechanism—attending to different parts of belief networks. We find empirical support for the different assumptions and predictions of the theory in two large studies. Some of the confirmed predictions reproduce previous findings, and others are unique to the NB theory, for example, that paying attention to the dissonance between personal and social beliefs decreases the variance of social beliefs and increases their consistency with personal beliefs, and that group radicalization is more likely when people pay attention to the accuracy of their social beliefs. The theory also offers a parsimonious explanation for group consensus, polarization, radicalization, and minority influence and reproduces empirically observed patterns of beliefs in different studies. We hope that in addition to providing answers to crucial questions on the dynamics of beliefs, the NB theory will inspire further work on the integration of different research areas on belief dynamics and addressing some of the most pressing issues we currently face as a society.

References

Acemoglu, D., & Ozdaglar, A. (2011). Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, *1*(1), 3–49. https://doi.org/10.1007/s13235-010-0004-1

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Akerlof, G. A. (1997). Social distance and social decisions. *Econometrica*, *65*(5), 1005–1027. https://doi.org/10.2307/2171877

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97. https://doi.org/10.1103/RevModPhys.74.47

Anderson, N. H. (1971). Integration theory and attitude change. *Psychological Review*, *78*(3), 171–206. https://doi.org/10.1037/h0030834

Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. *Journal of Conflict Resolution*, *41*(2), 203–226. https://doi.org/10.1177/0022002797041002001

Axelrod, R., Daymude, J. J., & Forrest, S. (2021). Preventing extreme polarization of political attitudes. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(50), Article e2102139118. https://doi.org/10.1073/pnas.2102139118

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

Bhatia, N., & Bhatia, S. (2021). Changes in gender stereotypes over time: A computational analysis. *Psychology of Women Quarterly*, *45*(1), 106–125. https://doi.org/10.1177/0361684320977178

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.-M., Wysocki, A. C., van Borkulo, C. D., van Bork, R., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*(1), Article 58. https://doi.org/10.1038/s43586-021-00055-w

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. University of Chicago Press.

Brandt, M. J. (2022). Measuring the belief system of a person. *Journal of Personality and Social Psychology*, *123*(4), 830–853. https://doi.org/10.1037/pspp0000416

Brandt, M. J., & Morgan, G. S. (2022). Between-person methods provide limited insight about within-person belief systems. *Journal of Personality and Social Psychology*, *123*(3), 621–635. https://doi.org/10.1037/pspp0000404

Brandt, M. J., & Sleegers, W. W. A. (2021). Evaluating belief system networks as a theory of political belief system dynamics. *Personality and Social Psychology Review*, *25*(2), 159–185. https://doi.org/10.1177/1088868321993751

Burnstein, E., & Vinokur, A. (1973). Testing two classes of theories about group induced shifts in individual choice. *Journal of Experimental Social Psychology*, *9*(2), 123–137. https://doi.org/10.1016/0022-1031(73)90004-8

Card, D., Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., Abramitzky, R., & Jurafsky, D. (2022). Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(31), Article e2120510119. https://doi.org/10.1073/pnas.2120510119

Cartwright, D., & Harary, F. (1956). Structural balance: A generalization of Heider's theory. *Psychological Review*, *63*(5), 277–293. https://doi.org/10.1037/h0046049

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics*, *81*(2), 591–646. https://doi.org/10.1103/RevModPhys.81.591

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science*, *30*(2), 174–192. https://doi.org/10.1177/0956797618813087

Christakis, N. A., & Fowler, J. H. (2010). Social network sensors for early detection of contagious outbreaks. *PLOS ONE*, *5*(9), Article e12948. https://doi.org/10.1371/journal.pone.0012948

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 151–192). McGraw-Hill.

Clifford, P., & Sudbury, A. (1973). A model for spatial conflict. *Biometrika*, *60*(3), 581–588. https://doi.org/10.1093/biomet/60.3.581

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science*, *8*(1), 160–179. https://doi.org/10.1111/tops.12186

Crano, W. D. (2010). Majority and minority influence in attitude formation and attitude change: Context/categorization—Leniency contract theory. In R. Martin & M. Hewstone (Eds.), *Minority influence and innovation: Antecedents, processes and consequences* (pp. 53–78). Psychology Press.

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. https://doi.org/10.1037/a0039802

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. L. (2018). The attitudinal entropy (AE) framework as a general theory of individual attitudes. *Psychological Inquiry*, *29*(4), 175–193. https://doi.org/10.1080/1047840X.2018.1537246

Dalege, J., Galesic, M., & Olsson, H. (2024). *Networks of beliefs (NB) theory* [Simulation code]. OSF. https://osf.io/n58h6/?view_only=0da2fb3267574e4fa53978bc4a36ba32

Dalege, J., Haslbeck, J. M. B., & Marsman, M. (2022). Idealized modeling of psychological dynamics. In A. M. Isvoranu, S. Epskamp, L. Waldorp, & D. Borsboom (Eds.), *Network psychometrics with R: A guide for behavioral and social scientists* (pp. 233–246). Routledge, Taylor & Francis Group. https://doi.org/10.4324/9781003111238-17

Dalege, J., & van der Does, T. (2022). Using a cognitive network model of moral and social beliefs to explain belief change. *Science Advances*, *8*(33), Article eabm0137. https://doi.org/10.1126/sciadv.abm0137

David, B., & Turner, J. C. (2001). Majority and minority influence: A single process self-categorization analysis. In C. K. W. De Dreu & N. K. De Vries (Eds.), *Group consensus and minority influence: Implications for innovation* (pp. 91–121). Blackwell Publishing.

Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, *3*(01n04), 87–98. https://doi.org/10.1142/S0219525900000078

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, *69*(345), 118–121. https://doi.org/10.1080/01621459.1974.10480137

Dhami, M. K., & Olsson, H. (2008). Evolution of the interpersonal conflict paradigm. *Judgment and Decision Making*, *3*(7), 547–569. https://doi.org/10.1017/S1930297500000802

Duffie, D., Giroux, G., & Manso, G. (2010). Information percolation. *American Economic Journal. Microeconomics*, *2*(1), 100–111. https://doi.org/10.1257/mic.2.1.100

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 461–493). American Psychological Association. https://doi.org/10.1037/10227-017

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press. https://doi.org/10.1017/CBO9780511761942

Ellinas, C., Allan, N., & Johansson, A. (2017). Dynamics of organizational culture: Individual beliefs vs. social conformity. *PLOS ONE*, *12*(6), Article e0180193. https://doi.org/10.1371/journal.pone.0180193

Enos, R. D. (2014). Causal effect of intergroup contact on exclusionary attitudes. *Proceedings of the National Academy of Sciences*, *111*, 3699–3704. https://doi.org/10.1073/pnas.1317670111

Epskamp, S. (2020). Psychometric network models from time-series and panel data. *Psychometrika*, *85*(1), 206–231. https://doi.org/10.1007/s11336-020-09697-3

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1–18. https://doi.org/10.18637/jss.v048.i04

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*(4), 617–634. https://doi.org/10.1037/met0000167

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press. https://doi.org/10.1515/9781503620766

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Addison-Wesley.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Flache, A., & Macy, M. W. (2011). Local convergence and global diversity: From interpersonal to social influence. *Journal of Conflict Resolution*, *55*, 970–995. https://doi.org/10.1177/0022002711414371

Flache, A., & Mäs, M. (2008). How to get the timing right. A computational model of the effects of the timing of contacts on team cohesion in demographically diverse teams. *Computational & Mathematical Organization Theory*, *14*(1), 23–51. https://doi.org/10.1007/s10588-008-9019-1

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 and Social Simulation*, *20*(4), Article 2. https://doi.org/10.18564/jasss.3521

Franci, A., Bizyaeva, A., Park, S., & Leonard, N. E. (2021). Analysis and control of agreement and disagreement opinion cascades. *Swarm Intelligence*, *15*(1–2), 47–82. https://doi.org/10.1007/s11721-021-00190-w

French, J. R., Jr. (1956). A formal theory of social power. *Psychological Review*, *63*(3), 181–194. https://doi.org/10.1037/h0046123

Friedkin, N. E. (1999). Choice shift and group polarization. *American Sociological Review*, *64*(6), 856–875. https://doi.org/10.1177/000312249906400606

Friedkin, N. E., & Johnsen, E. C. (1990). Social influence and opinions. *The Journal of Mathematical Sociology*, *15*(3–4), 193–206. https://doi.org/10.1080/0022250X.1990.9990069

Friedkin, N. E., & Johnsen, E. C. (2011). *Social influence network theory: A sociological examination of small group dynamics* (Vol. 33). Cambridge University Press. https://doi.org/10.1017/CBO9780511976735

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Gagné, F. M., & Lydon, J. E. (2004). Bias and accuracy in close relationships: An integrative review. *Personality and Social Psychology Review*, *8*(4), 322–338. https://doi.org/10.1207/s15327957pspr0804_1

Galesic, M., Bruine de Bruin, W., Dalege, J., Feld, S. L., Kreuter, F., Olsson, H., Prelec, D., Stein, D. L., & van der Does, T. (2021). Human social sensing is an untapped resource for computational social science. *Nature*, *595*(7866), 214–222. https://doi.org/10.1038/s41586-021-03649-2

Galesic, M., Olsson, H., Dalege, J., van der Does, T., & Stein, D. L. (2021). Integrating social and cognitive aspects of belief dynamics: Towards a unifying framework. *Journal of the Royal Society Interface*, *18*(176), Article 20200857. https://doi.org/10.1098/rsif.2020.0857

Galesic, M., Olsson, H., & Rieskamp, J. (2018). A sampling model of social judgment. *Psychological Review*, *125*(3), 363–390. https://doi.org/10.1037/rev0000096

Galesic, M., & Stein, D. L. (2019). Statistical physics models of belief dynamics: Theory and empirical tests. *Physica A: Statistical Mechanics and its Applications*, *519*, 275–294. https://doi.org/10.1016/j.physa.2018.12.011

Gawronski, B. (2012). Back to the future of dissonance theory: Cognitive consistency as a core motive. *Social Cognition*, *30*(6), 652–668. https://doi.org/10.1521/soco.2012.30.6.652

Gawronski, B., & Strack, F. (2012). Cognitive consistency as a basic principle of social information processing. In B. Gawronski & F. Strack (Eds.), *Cognitive consistency: A fundamental principle in social cognition* (pp. 1–16). Guilford Press.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*(2), 254–267. https://doi.org/10.1037/0033-295X.98.2.254

Glauber, R. J. (1963). Time-dependent statistics of the Ising model. *Journal of Mathematical Physics*, *4*(2), 294–307. https://doi.org/10.1063/1.1703954

Goel, S., Mason, W., & Watts, D. J. (2010). Real and perceived attitude agreement in social networks. *Journal of Personality and Social Psychology*, *99*(4), 611–621. https://doi.org/10.1037/a0020697

Goldberg, A., & Stein, S. K. (2018). Beyond social contagion: Associative diffusion and the emergence of cultural variation. *American Sociological Review*, *83*(5), 897–932. https://doi.org/10.1177/0003122418797576

Golub, B., & Sadler, E. (2016). Learning in social networks. In Y. Bramoulle, A. Galeotti, & B. Rogers (Eds.), *The Oxford handbook of the economics of networks* (pp. 504–542). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199948277.013.12

Hammond, K. R. (1965). New directions in research on conflict resolution. *Journal of Social Issues*, *21*, 44–66. https://doi.org/10.1111/j.1540-4560.1965.tb00505.x

Harary, F. (1959). On the measurement of structural balance. *Behavioral Science*, *4*(4), 316–323. https://doi.org/10.1002/bs.3830040405

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, *5*(3).

Heider, F. (1946). Attitudes and cognitive organization. *The Journal of Psychology*, *21*(1), 107–112. https://doi.org/10.1080/00223980.1946.9917275

Heider, F. (1958). *The psychology of interpersonal relations*. Wiley. https://doi.org/10.1037/10628-000

Hickok, A., Kureh, Y., Brooks, H. Z., Feng, M., & Porter, M. A. (2022). A bounded-confidence model of opinion dynamics on hypergraphs. *SIAM Journal on Applied Dynamical Systems*, *21*(1), 1–32. https://doi.org/10.1137/21M1399427

Holley, R. A., & Liggett, T. M. (1975). Ergodic theorems for weakly interacting infinite systems and the voter model. *Annals of Probability*, *3*(4), 643–663. https://doi.org/10.1214/aop/1176996306

Holyoak, K. J., & Thagard, P. (1996). *Mental leaps: Analogy in creative thought*. MIT Press.

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

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(10), 3088–3092. https://doi.org/10.1073/pnas.81.10.3088

Hoppitt, W., & Laland, K. N. (2013). *Social learning: An introduction to mechanisms, methods, and models*. Princeton University Press. https://doi.org/10.1515/9781400846504

Huet, S., & Deffuant, G. (2010). Openness leads to opinion stability and narrowness to volatility. *Advances in Complex Systems*, *13*(3), 405–423. https://doi.org/10.1142/S0219525910002633

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, *50*(6), 1141–1151. https://doi.org/10.1037/0022-3514.50.6.1141

Ising, E. (1925). Contribution to the theory of ferromagnetism. *Zeitschrift für Physik*, *31*(1), 253–258. https://doi.org/10.1007/BF02980577

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, *59*(3), 690–707. https://doi.org/10.1111/ajps.12152

Jackson, M. O. (2008). *Social and economic networks*. Princeton University Press. https://doi.org/10.1515/9781400833993

Jager, W., & Amblard, F. (2005). Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory*, *10*(4), 295–303. https://doi.org/10.1007/s10588-005-6282-2

Kitayama, S., Snibbe, A. C., Markus, H. R., & Suzuki, T. (2004). Is there any "free" choice? Self and dissonance in two cultures. *Psychological Science*, *15*(8), 527–533. https://doi.org/10.1111/j.0956-7976.2004.00714.x

Latané, B. (1981). The psychology of social impact. *American Psychologist*, *36*(4), 343–356. https://doi.org/10.1037/0003-066X.36.4.343

Latané, B., & Wolf, S. (1981). The social impact of majorities and minorities. *Psychological Review*, *88*(5), 438–453. https://doi.org/10.1037/0033-295X.88.5.438

Leonard, N. E., Lipsitz, K., Bizyaeva, A., Franci, A., & Lelkes, Y. (2021). The nonlinear feedback dynamics of asymmetric political polarization. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(50), Article e2102149118. https://doi.org/10.1073/pnas.2102149118

Levin, S. A., Milner, H. V., & Perrings, C. (2021). The dynamics of political polarization. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(50), Article e2116950118. https://doi.org/10.1073/pnas.2116950118

Martin, R., & Hewstone, M. (Eds.). (2009). *Minority influence and innovation: Antecedents, processes and consequences*. Psychology Press. https://doi.org/10.4324/9780203865552

Mason, L. (2016). A cross-cutting calm: How social sorting drives affective polarization. *Public Opinion Quarterly*, *80*(1), 351–377. https://doi.org/10.1093/poq/nfw001

Mernyk, J. S., Pink, S. L., Druckman, J. N., & Willer, R. (2022). Correcting inaccurate metaperceptions reduces Americans' support for partisan violence. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(16), Article e2116851119. https://doi.org/10.1073/pnas.2116851119

Monroe, B. M., & Read, S. J. (2008). A general connectionist model of attitude structure and change: The ACS (Attitudes as Constraint Satisfaction) model. *Psychological Review*, *115*(3), 733–759. https://doi.org/10.1037/0033-295X.115.3.733

Moscovici, S., Lage, E., & Naffrechoux, M. (1969). Influence of a consistent minority on the responses of a majority in a color perception task. *Sociometry*, *32*(4), 365–380. https://doi.org/10.2307/2786541

Mousa, S. (2020). Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq. *Science*, *369*, 866–870. https://doi.org/10.1126/science.abb3153

Newby-Clark, I. R., McGregor, I., & Zanna, M. P. (2002). Thinking and caring about cognitive inconsistency: When and for whom does attitudinal ambivalence feel uncomfortable? *Journal of Personality and Social Psychology*, *82*(2), 157–166. https://doi.org/10.1037/0022-3514.82.2.157

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*(2), 167–256. https://doi.org/10.1137/S003614450342480

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

Nisbett, R. E., & Kunda, Z. (1985). Perception of social distributions. *Journal of Personality and Social Psychology*, *48*(2), 297–311. https://doi.org/10.1037/0022-3514.48.2.297

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. https://doi.org/10.1037/0033-295X.97.3.362

Olson, M. E., Arroyo-Santos, A., & Vergara-Silva, F. (2019). A user's guide to metaphors in ecology and evolution. *Trends in Ecology & Evolution*, *34*(7), 605–615. https://doi.org/10.1016/j.tree.2019.03.001

Page, S. E. (2018). *The model thinker: What you need to know to make data work for you*. Basic Books.

Pentland, A. (2014). *Social physics: How good ideas spread-the lessons from a new science*. Penguin.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–205). Academic Press. https://doi.org/10.1016/S0065-2601(08)60214-2

Pham, T. M., Kondor, I., Hanel, R., & Thurner, S. (2020). The effect of social balance on social fragmentation. *Journal of the Royal Society Interface*, *17*(172), Article 20200752. https://doi.org/10.1098/rsif.2020.0752

Pham, T. M., Korbel, J., Hanel, R., & Thurner, S. (2022). Empirical social triad statistics can be explained with dyadic homophylic interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(6), Article e2121103119. https://doi.org/10.1073/pnas.2121103119

Priester, J. R., & Petty, R. E. (1996). The gradual threshold model of ambivalence: Relating the positive and negative bases of attitudes to subjective ambivalence. *Journal of Personality and Social Psychology*, *71*(3), 431–449. https://doi.org/10.1037/0022-3514.71.3.431

Proskurnikov, A. V., & Tempo, R. (2017). A tutorial on modeling and analysis of dynamic social networks. Part I. *Annual Reviews in Control*, *43*, 65–79. https://doi.org/10.1016/j.arcontrol.2017.03.002

Quiamzade, A., Mugny, G., Falomir-Pichastor, J. M., Butera, F., Martin, R., & Hewstone, M. (2010). The complexity of majority and minority influence processes. In R. Martin & M. Hewstone (Eds.), *Minority influence and innovation: Antecedents, processes and consequences* (pp. 21–52). Psychology Press.

Redner, S. (2019). Reality-inspired voter models: A mini-review. *Comptes Rendus Physique*, *20*(4), 275–292. https://doi.org/10.1016/j.crhy.2019.05.004

Richerson, P. J., & Boyd, R. (2008). *Not by genes alone: How culture transformed human evolution*. University of Chicago Press.

Rodriguez, N., Bollen, J., & Ahn, Y. Y. (2016). Collective dynamics of belief evolution under cognitive coherence and social conformity. *PLOS ONE*, *11*(11), Article e0165910. https://doi.org/10.1371/journal.pone.0165910

Schweighofer, S., Schweitzer, F., & Garcia, D. (2020). A weighted balance model of opinion hyperpolarization. *Journal of Artificial Societies and Social Simulation*, *23*(3), Article 5. https://doi.org/10.18564/jasss.4306

Shultz, T. R., & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review*, *103*(2), 219–240. https://doi.org/10.1037/0033-295X.103.2.219

Sîrbu, A., Loreto, V., Servedio, V. D., & Tria, F. (2013). Opinion dynamics with disagreement and modulated information. *Journal of Statistical Physics*, *151*, 218–237. https://doi.org/10.1007/s10955-013-0724-x

Smith, C. M., & Tindale, R. S. (2010). Direct and indirect minority influence in groups. In R. Martin & M. Hewstone (Eds.), *Minority influence and innovation: Antecedents, processes and consequences* (pp. 263–284). Psychology Press.

Smith, E. R., & Zarate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, *99*(1), 3–21. https://doi.org/10.1037/0033-295X.99.1.3

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group. (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 194–281). MIT Press.

Stalder, D. R. (2010). Competing roles for the subfactors of need for closure in moderating dissonance-produced attitude change. *Personality and Individual Differences*, *48*(6), 775–778. https://doi.org/10.1016/j.paid.2010.01.028

Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature Human Behaviour*, *4*(4), 361–371. https://doi.org/10.1038/s41562-019-0800-6

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(37), 9210–9215. https://doi.org/10.1073/pnas.1807222115

Sunstein, C. R. (2002). The law of group polarization. *Journal of Political Philosophy*, *10*(2), 175–195. https://doi.org/10.1111/1467-9760.00148

Sunstein, C. R., Schkade, D., Ellman, L. M., & Sawicki, A. (2007). *Are judges political? An empirical analysis of the federal judiciary*. Brookings Institution Press.

Thagard, P., & Verbeurgt, K. (1998). Coherence as constraint satisfaction. *Cognitive Science*, *22*(1), 1–24. https://doi.org/10.1207/s15516709cog2201_1

Thomas, W. I., & Swaine Thomas, D. (1928). *The child in America: Behaviour problems and programs*. Knopf.

Turner, J. C. (1985). Social categorization and the self-concept: A social cognitive theory of group behavior. In E. Lawler (Ed.), *Advances in group processes* (pp. 77–121). JAI Press.

Vallacher, R. R., Read, S. J., & Nowak, A. (Eds.). (2017). *Computational social psychology*. Routledge. https://doi.org/10.4324/9781315173726

van Borkulo, C. D., Borsboom, D., & Schoevers, R. A. (2016). Group-level symptom networks in depression—Reply. *JAMA Psychiatry*, *73*(4), 411–412. https://doi.org/10.1001/jamapsychiatry.2015.3157

van der Does, T., Stein, D. L., Fedoroff, N., & Galesic, M. (2022). *Science communication in light of moral and social concerns: Testing a statistical physics model of belief change*. OSF. https://doi.org/10.31219/osf.io/zs7dq

van Der Does, T., Galesic, M., Dunivin, Z. O., & Smaldino, P. E. (2022). Strategic identity signaling in heterogeneous networks. *Proceedings of the National Academy of Sciences*, *119*(10), Article e2117898119. https://doi.org/10.1073/pnas.2117898119

van der Maas, H. L., Dalege, J., & Waldorp, L. (2020). The polarization within and across individuals: The hierarchical Ising opinion model. *Journal of Complex Networks*, *8*(2), Article cnaa010. https://doi.org/10.1093/comnet/cnaa010

van der Maas, H. L., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842

van Harreveld, F., van der Pligt, J., & de Liver, Y. N. (2009). The agony of ambivalence and ways to resolve it: Introducing the MAID model. *Personality and Social Psychology Review*, *13*(1), 45–61. https://doi.org/10.1177/1088868308324518

Van Overwalle, F., & Siebler, F. (2005). A connectionist model of attitude formation and change. *Personality and Social Psychology Review*, *9*(3), 231–274. https://doi.org/10.1207/s15327957pspr0903_3

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*(1), 28–50. https://doi.org/10.1016/j.jmp.2003.11.004

Watts, D. J. (2004). *Six degrees: The science of a connected age*. W.W. Norton.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, *67*(6), 1049–1062. https://doi.org/10.1037/0022-3514.67.6.1049

Weisbuch, G., Deffuant, G., Amblard, F., & Nadal, J. P. (2002). Meet, discuss, and segregate! *Complexity*, *7*(3), 55–63. https://doi.org/10.1002/cplx.10031

Wood, W., Lundgren, S., Ouellette, J. A., Busceme, S., & Blackstone, T. (1994). Minority influence: A meta-analytic review of social influence processes. *Psychological Bulletin*, *115*(3), 323–345. https://doi.org/10.1037/0033-2909.115.3.323

Wu, F. Y. (1982). The Potts model. *Reviews of Modern Physics*, *54*(1), 235–268. https://doi.org/10.1103/RevModPhys.54.235

Xie, S. Y., Flake, J. K., Stolier, R. M., Freeman, J. B., & Hehman, E. (2021). Facial impressions are predicted by the structure of group stereotypes. *Psychological Science*, *32*(12), 1979–1993. https://doi.org/10.1177/09567976211024259

Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society*, *30*(5), 316–327. https://doi.org/10.1177/0270467610380011

(*Appendices follow*)

Appendix A

Study Descriptions

Study 1 was a survey of 973 U.S. participants from Mechanical Turk, conducted from May 11–13, 2022. Among the participants, 55% were male, and 44.6% female; mean/minimum/maximum age was 40.6/10/78 years; 9.6% had high school or less, 30.8% some college, and 59.6% college degree. They answered questions (see the full text in Supplemental Materials) about the following: (a) their personal beliefs about safety of GM food and related moral and political beliefs (personal belief nodes); (b) importance of consistency of their personal beliefs (attention to potential personal dissonance); (c) experience of dissonance in their personal beliefs (felt personal dissonance); (d) their social beliefs about three beliefs about GM safety held by each of five of their social contacts (social belief nodes); (e) importance of consistency of their personal and social beliefs (attention to potential social dissonance); (f) experience of dissonance between their personal and social beliefs (felt social dissonance); (g) importance of accuracy of social beliefs (attention to potential external dissonance); (h) frequency of contact with social contacts; (i) overall importance of the topic of GM food; (j) personal beliefs about safety of GM food after receiving a brief informational intervention (a quote from a National Academies of Sciences, Engineering, and Medicine report saying that there is no evidence GM food currently on the market is harmful when consumed); and (k) basic demographics. All questions except for the demographics were answered on 7-point scales with labeled extremes. The study was approved by University of New Mexico Institutional Review Board No. 10819. This study was not preregistered.

Study 2 was a two-wave survey conducted on a probabilistic national sample of U.S. population from October 13 to December 15, 2022, by the Center for Economic and Social Research at the University of Southern California. Participants, who were members of the University of Southern California's Understanding America Panel, answered questions about their personal and social beliefs related to GM food, flu vaccination, and climate change. The social beliefs were about a friend they nominated themselves from a different household. This friend was invited to complete a short survey about their actual personal beliefs. Participants also answered questions about the attention to and feelings of social and external dissonance as well as questions about their interaction with their social contacts and their experience with scientists. They received an informational intervention which consisted of seeing their friend's actual answers as well as scientists' views on GM food, flu vaccination, and climate change. In total, 669 participants completed both survey waves, and each had a friend completing the short survey as well. Most of the results will be presented elsewhere (Olsson et al., in prep), and here, we focus on the participants' answers about their (a) personal and (b) social beliefs, their (c) felt social and (d) felt external dissonance, and (e) their attention to potential external dissonance (see those questions in the Supplemental Materials). The study was approved by University of Southern California. BRANY Social, Behavioral, and Educational Research Institutional Review Board No. 22-065-1044. This study was not preregistered.

Appendix B

Energy as Potential Dissonance, and Separability of Energies: Validation Tests

A central assumption of the NB theory is that the statistical physics concept of energy provides a formalization of potential dissonance. If this is the case, then potential dissonance should be related to subjectively measured felt dissonance, provided that participants pay some attention to the potential dissonance. Furthermore, we assume that people can differentiate between potential dissonances in different parts of their belief networks. If this is correct, then felt personal dissonance should be more related to potential personal dissonance than to potential social dissonance, and vice versa for the felt social dissonance. Similarly, felt social dissonance should be more related to potential social dissonance than to potential external dissonance, and vice versa for the felt external dissonance.

To test the assumption about the relationships of potential and felt personal and social dissonances, we analyze questions about Study 1 participants' personal beliefs about GM food and about perceived beliefs of five of their social contacts. These questions enabled us to calculate potential personal and social dissonances for each participant, as described below. The survey also included measures of felt dissonances among personal and social beliefs, adapted from the felt ambivalence scale developed by Priester and Petty (1996; see Supplemental Materials for more details). For example, one of three items measuring personal dissonance was "I experience no conflict at all towards the issue of GM food" and for social dissonance "I experience no conflict at all between my beliefs and the beliefs of [*social contact*] towards the issue of GM food," on the scale from 1 to 7.

We checked whether the personal and social felt dissonances are indeed two different psychological constructs, by investigating whether the three items measuring each construct loaded on two separate factors. To test whether the personal and social felt dissonance items loaded on two separate factors, we fitted a confirmatory factor model with two correlated factors. Items of the different scales were only allowed to load on their respective factor. This model did not fit the data well, root-mean-square error of approximation (RMSEA) = .15, comparative fit index (CFI) = .93. We therefore investigated the reasons for the bad fit of the model and found that this was generally caused by higher correlations between the negatively worded items of the personal and social felt dissonance scales than would be expected by the model. We therefore allowed for correlated errors between these items. This resulted in a model with good fit to the data, RMSEA = .08, CFI = .98. In this model, personal and social felt dissonances were moderately correlated, $r = .32$, $p < .001$. We then tested whether a

(*Appendices continue*)

two-factor model fits the data better than a one-factor model (we allowed for correlated errors between the negatively worded items in both these models). The one-factor model showed poor fit to the data, RMSEA = .29, CFI = .73, and also showed worse fit than the two-factor model, Akaike information criterion (AIC) = 21,980.53 versus AIC = 21,368.35 for the two-factor model, Bayesian information criterion (BIC) = 22,043.98 versus BIC = 21,436.68 for the two-factor model. The results thus indicate that personal and social felt dissonances form two separate factors, which are moderately related. Both these findings are in line with the assumptions of our theory.

We then proceeded to test whether personal and social felt dissonances relate to personal and social potential dissonances. We expected that personal felt dissonances predict personal energies and that they do so better than social dissonances. Conversely, we expected that social dissonances predict social energies and that they do so better than personal dissonances. To calculate potential personal and social dissonance, we first estimated edges between personal beliefs (ω_{ij}) and between personal and social beliefs (ρ_{ik}), consisting of the average of one's own beliefs and the average of the social beliefs for each of five social contacts. Edges represent regularized partial correlations, which were estimated using the *EBICglasso* function in the R-package *qgraph* (Epskamp et al., 2012; Epskamp & Fried, 2018). Using these networks, we calculated potential dissonance for each participant using Equations 3 and 4, and setting $\tau_i = 0$. We then added these energies to the factor model described above and regressed each of them on personal and social felt dissonances. Personal energies were significantly and moderately to strongly predicted by personal felt dissonances, $\beta = .42, p < .001$, but not by social felt dissonances, $\beta = .05, p = .14$. Social energies were significantly and strongly predicted by social dissonances, $\beta = .55, p < .001$, but only weakly by personal dissonances, $\beta = .12$, $p < .001$. To summarize, we found that personal and social felt dissonances load on two separate factors, these dissonances are moderately related to each other, and they are more related to their associated potential dissonances (energies) than to the energies stemming from the other part of the internal belief network. Taken together, these results indicate that indeed felt dissonances arise from lack of consistency between beliefs, and are separate for personal and social beliefs.

To test the assumption about the relationships of potential and felt social and external dissonances, we analyze questions about Study 2 participants' personal and social beliefs about climate change, GM food, and vaccination. Here, because of time constraints in the survey, we measured only one personal belief for each topic (e.g., "What comes closer to your view on climate change? 1 = There is solid evidence that the climate is NOT changing because of human activity, 7 = There is solid evidence that the climate is changing because of human activity"). Similarly, we used just one question per topic to measure social felt dissonance (e.g., "I experience a lot of conflict between my beliefs about climate change and the beliefs of [friend].") and external felt dissonance (e.g., "I feel uneasy about the discrepancy between what I thought [friend] believes about climate change and what [friend] actually believes."). We measure potential social dissonance as the absolute difference between personal beliefs and perceived friend's beliefs, and external social dissonance as the absolute difference between social beliefs and actual friend's beliefs.

We use a mixed-effects regression models to investigate how social and external felt dissonance relates to social and external potential dissonances. In the model for each of the two felt dissonances, we also control for the other felt dissonance, as well as for the clustering of the three topics within participants. As expected, social potential dissonance was reliably but weakly positively related to social felt dissonance ($\beta = .07, p < .001$) but not to external felt dissonance ($\beta = -.01$, $p = .43$). And, external potential dissonance was positively, although not reliably related to external felt dissonance ($\beta = .03, p = .12$) and negatively and not reliably to social felt dissonance ($\beta = -.01$, $p = .59$).

Appendix C

Attention to Dissonance as Temperature, and Separability of Temperatures: Validation Tests

Another central assumption of the NB theory is that there are three separate types of attention (or temperature in statistical physics terms), one each for the dissonance in personal, social, and external belief networks. To test this assumption, in Study 1 (see Appendix A), we also included questions tapping participants' subjective importance of consistent personal beliefs (a proxy for temperature of the personal part of the internal network, e.g., "It is important to me that my beliefs toward GM food are not in conflict with each other"), importance of consistent social beliefs (a proxy for temperature of the social part of the internal network, e.g., "It is important to me that my personal beliefs and the beliefs of *social contact* toward GM food are not in conflict with each other."), and importance of accurate social beliefs (a proxy for temperature of the external network, e.g., "It is important to me that I know what *social contact* thinks about GM food.").

We first test whether the importance items load on three separate factors corresponding to personal, social, and external belief networks. We first tested whether the importance items from Study 1 load on three separate factors. To do so, we fitted a confirmatory factor model with three correlated factors. Items of the different scales were only allowed to load on their respective factor. This model did not fit the data well, RMSEA = .20, CFI = .89. We therefore investigated the reasons for the bad fit of the model and found that this was generally caused by higher correlations between similarly worded items of the importance of consistent social beliefs scale and the importance of accurate social beliefs scale than would be expected by the model. We therefore allowed for correlated errors between these items. This resulted in a model with good fit to the data, RMSEA = .07, CFI = .99. In this model, the importance of consistent personal beliefs was strongly correlated with the importance of consistent social beliefs, $r = .53$, $p < .001$, and with the importance of accurate social beliefs, $r = .51, p < .001$. The importance of consistent social beliefs was very strongly correlated with the importance of accurate social beliefs, $r = .77$, $p < .001$.

To test whether indeed three factors are needed to explain the different importance, we first compared the three-factor model with a two-factor in which the importance of consistent personal beliefs items loaded on the first factor and the importance of consistent social beliefs and the importance of accurate social beliefs items loaded on the second factor (both models included correlated errors between similarly worded items). The two-factor did not fit the data well, RMSEA = .27, CFI = .81, and also showed worse fit than the three-factor model, AIC = 27,348.60 versus AIC = 25,800.25 for the three-factor model, BIC = 27455.97 versus BIC = 25,917.37 for

(*Appendices continue*)

the three-factor model. We then proceeded to test whether a one-factor model fitted the data. This model did not fit the data well, RMSEA = .35, CFI = .66, and also showed worse fit than the three-factor model, AIC = 28,627.50, BIC = 28,729.99. We therefore conclude that three factors are needed to explain the correlational patterns in the data. This finding provides first support that there are indeed three separate forms of temperatures in belief networks.

Furthermore, we investigate whether higher attention to dissonance predicts a reduced potential dissonance (energy). This is what would be expected from the NB theory: When people pay attention to their potential dissonance (i.e., when they experience felt dissonance), their beliefs will tend to become more consistent, reducing the potential dissonance. We find that attention to dissonance in a particular part of belief network relates negatively to the potential dissonance (energy) in that part of the network. For personal and social beliefs, we use measures of importance of personal and social potential dissonances collected in Study 1 as proxies for attention, and find the expected negative relationship of attention and potential dissonances ($\beta = -.21$, $p < .001$ for the personal beliefs, and $\beta = -.17$, $p = .002$ for the social beliefs). For external beliefs, as proxies for attention, we use measures of importance of external dissonance as well as of frequency of discussing different topics with a friend, collected in Study 2. We again find the expected negative relationship between external potential dissonance (the difference between perceived and actual beliefs of friends) and both proxies for attention ($\beta = -.06$, $p = .002$ for importance, and $\beta = -.24$, $p < .001$ for frequency of discussion).

While in Study 1 we could not measure the energy in external networks (as we did not have data about the actual beliefs of participants' social contacts), we can predict that the importance of accurate social beliefs might *increase* energies of the social part of the internal network. The reason is that, assuming at least moderately heterogeneous social networks, higher accuracy of one's social beliefs would lead to higher inconsistency of one's social and personal beliefs. We find weak support for this hypothesis ($\beta = .09$, $p = .10$), possibly because social networks of our participants were homogeneous. Another reason might be that external temperature is more strongly affected by how often a belief is expressed and/or discussed in a social network.

Appendix D

Measurement of Theoretical Constructs

There are several different ways to measure or estimate the core constructs and predictions specified in the NB theory. First, the state of nodes in belief networks can be measured through survey questionnaires (Dalege et al., 2016; van der Does, Stein, et al., 2022), by inferences from behavior (e.g., from voting records), or by inferring networks of related beliefs from textual corpora such as books, newspapers, congressional speeches, and social media (Bhatia & Bhatia, 2021; Card et al., 2022; Charlesworth & Banaji, 2019). Edges between belief nodes can be estimated using partial correlations or regression weights based on co-occurrences of beliefs within or between participants at a single time point or over time or by other methods for fitting parameters of network models (Borsboom, Deserno, et al., 2021). There are obvious challenges of using cross-section data to represent within-person associations (e.g., Fisher et al., 2018). The relationship between individual belief networks and group-level networks are still being debated in the literature, but it seems that when there is some nonzero correlation on the individual level, it is typically positively related to group-level correlations (Brandt & Morgan, 2022), which was also noted by van Borkulo et al. (2016).

To overcome problems with cross-section data, edges could also be set by asking participants how strongly their beliefs are related (Brandt, 2022; see also Stolier et al., 2018, 2020; Xie et al., 2021). Second, energy can be estimated by combining the state of nodes with the estimated network structure and using the equations in the Formal Implementation section to calculate potential dissonance between beliefs in different parts of the belief network (cf. Dalege & van der Does, 2022). Third, temperature or attention to potential dissonance can be estimated by comparing belief networks of different groups (Epskamp, 2020) or by directly asking individuals about the importance they attach to their beliefs and the dissonance between them (as we do here).

Appendix E

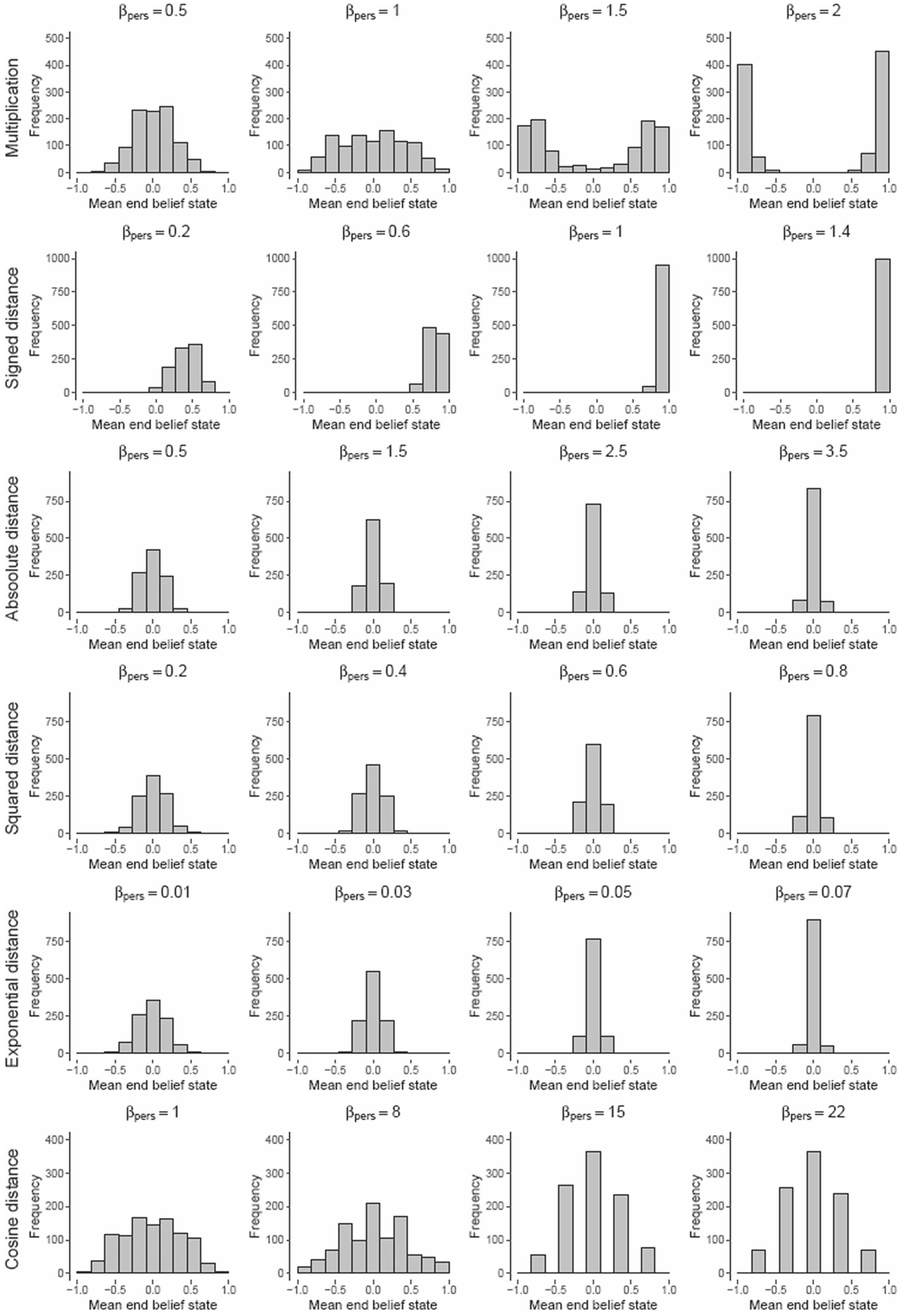
The Choice of Distance Measures

In Figure E1, we present results comparing different distance measures.

(*Appendices continue*)

Figure E1

Steady State Distributions of Different Implementations of Energy Minimization (Distance Measures) Under Different Inverse Temperatures



Note. The simulations were run on a network of 10 belief nodes, using the settings described in the Formal Implementation section. The first row represents the implementation we use in this article, relying on multiplication. The second row represents calculating energies using signed distance (Akerlof, 1997). The third row represents calculating energies using absolute distance (Akerlof, 1997). The fourth row represents calculating energy using squared distance (Deffuant et al., 2000). The fifth row represents calculating energy using exponential distance (E. R. Smith & Zarate, 1992). The sixth row represents calculating energy using cosine distance (Sîrbu et al., 2013). As can be seen, only the multiplication implementation can reproduce the increasing bimodality that we see in empirical data (see Figure 2).

(*Appendices continue*)

Appendix F

Determining Edge Weights

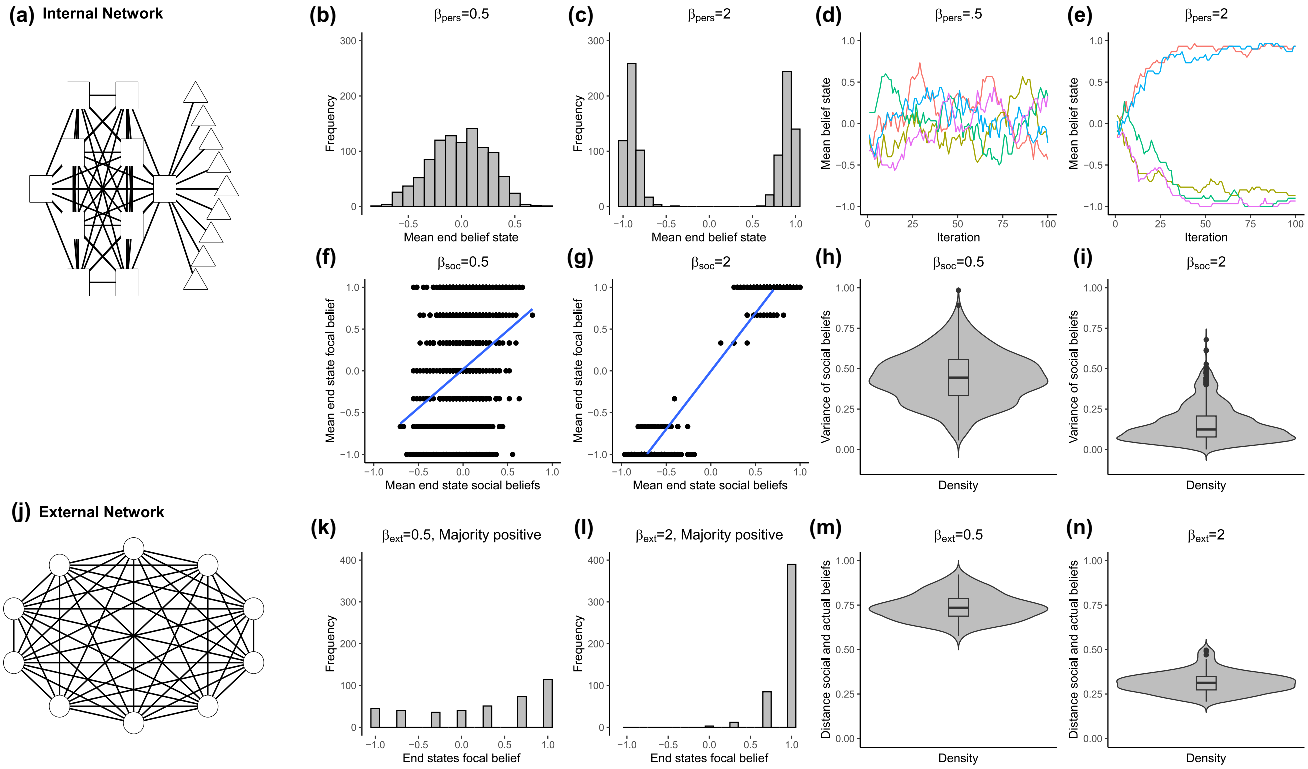
To determine the edge weights used in all simulations reported in this article, we initialized beliefs randomly and then allowed them to settle in the steady state achievable for $\beta = 1$. We then calculated correlations over runs between one of the personal beliefs (the focal belief) and all the other personal beliefs, between the focal belief and the social beliefs, and between social and actual beliefs. We searched through different values of edge weights until we found those that lead to a moderate size of all those three types of correlations. In this way, we aimed to establish an intermediate case where all parts of the belief network have similar correlations given a moderate attention to potential dissonance, enabling us to compare effects of lower or higher attention under similar conditions for all parts of the belief network. In other words, attention to potential dissonance of $\beta = 1$ represents the baseline for each part of the belief network and results in beliefs belonging to that part of the network being moderately correlated. A β lower than 1 will result in less correlated beliefs, and a β higher than 1 will result in more strongly correlated beliefs. The resulting edge weights, used in all simulations, are $\omega = .4$ for the influence between personal beliefs, $\rho = 1$ for the influence between personal and social beliefs, and $\alpha = 1.4$ for the correspondence between social and actual beliefs.

Appendix G

Illustrations of the Networks of Beliefs Theory's Dynamics

Figure G1

Results of Illustrative Simulations on the Dynamics of the Networks of Beliefs Theory



Note. Panel (a) shows the internal network of every individual (squares represent personal beliefs and triangles represent social beliefs) used in the simulations. Panels (b)–(e) show the dynamics for varying attention to personal dissonance, either low ($\beta_{pers} = .5$, in (b) and (d)) or high ($\beta_{pers} = 2$, in (c) and (e)). Panels (b) and (c) show the distributions of mean belief states at the end of simulation runs, and (d) and (e) the whole dynamics of five randomly selected runs. Panels (f)–(i) show the dynamics for varying attention to social dissonance, either low ($\beta_{soc} = .5$, in (f) and (h)) or high ($\beta_{soc} = 2$, in (g) and (i)). Panels (h) and (i) show the density plots of the variance of the social beliefs at the end of simulations runs. Panel (j) shows the external network used in the simulations, with each circle representing one individual equivalent to that shown in Panel (a). Panels (k)–(n) show the dynamics for varying attention to external dissonance, either low ($\beta_{ext} = .5$, in (k) and (m)) or high attention ($\beta_{ext} = 2$, in (l) and (n)). Panels (k) and (l) show the histograms of the end states of the focal beliefs of groups which ended with a majority of positive belief states. Bars indicate frequency of different end states of focal beliefs across 10 individuals. Panels (m) and (n) show the density plots of the distance between social beliefs and actual others' beliefs.

(*Appendices continue*)

Appendix H

Consensus and Polarization

We use both qualitative and quantitative measures to understand how different levels of attention to dissonance affect the likelihood of consensus and polarization. On the qualitative level, we visualize the networks representative of those occurring at the end of simulation runs in different conditions. Panels (a)–(e) of Figure H1 show five types of networks that we observe in our simulations, showing radical polarization (RP) where the two social network clusters end up having completely opposing beliefs, radical consensus (RC) where the two clusters end up having the same beliefs, moderate polarization (MP) where individuals in the two clusters have different but moderate beliefs, moderate heterophily (MH) where the two clusters have similar distribution of mostly moderate beliefs, and extreme heterophily (EH) where the two clusters have similar distribution of mostly extreme beliefs. Abbreviations of different networks in Panels (f)–(i) of 4 show in which conditions they appear.

On the quantitative level, we use two measures of consensus and polarization. One is variance of focal beliefs at the end of simulation runs in different conditions, shown in Panels (f) and (g) of Figure H1. High values of variance show that network members hold very different opinions (as in the network EH described above) while low values of variance indicate that the opinions are not very different from each other (as in RC networks). Note that in some conditions, RP (extreme variance) and RC (no variance) networks are equally likely outcomes, producing a moderate average variance. For this reason, we also use another quantitative measure, modularity, showing how likely is that connected individuals hold similar focal beliefs (Panels (h) and (i); Newman, 2006). High values of modularity indicate that connected individuals hold similar beliefs (corresponding to RP, RC, and MP networks above), while low values of modularity indicate that connected individuals are not more likely to hold similar beliefs than unconnected individuals (corresponding to MH and EH networks). To calculate modularity, we binarized the focal beliefs into positive and negative.

Taken together, the patterns in Figure H1 suggest four main insights. First, when attention to both personal and social dissonance is low ($\beta_{\text{pers}} = \beta_{\text{soc}} = 0.5$), beliefs almost always remain moderate and the final state resembles either moderate polarization (MP, where the two clusters have different, but moderate beliefs) or moderate heterophily (MH, where there is little difference between the two clusters), independently of attention to external dissonance. In other words, when individuals pay little attention to either how much their personal beliefs align to each other or how much their personal beliefs align with the perceived beliefs of their social contacts, they will likely end up holding moderate beliefs. If in addition they pay little attention whether their perceived beliefs about social contacts are accurate ($\beta_{\text{ext}} = 0.5$), there will be a range of moderate beliefs throughout the network (MH). If they do pay attention to accuracy ($\beta_{\text{ext}} = 0.5$), the result will be a moderately polarized network (MP).

Second, when attention to personal dissonance is high and attention to external dissonance is low ($\beta_{\text{pers}} = 2$ and $\beta_{\text{ext}} = 0.5$), extreme heterophily (EH) is likely to emerge almost independently of the level of attention to social dissonance (β_{soc}). In this case, individuals try to align their personal beliefs and are not pressured to align with their actual social environments. This allows their beliefs to radicalize unrestricted by others' beliefs.

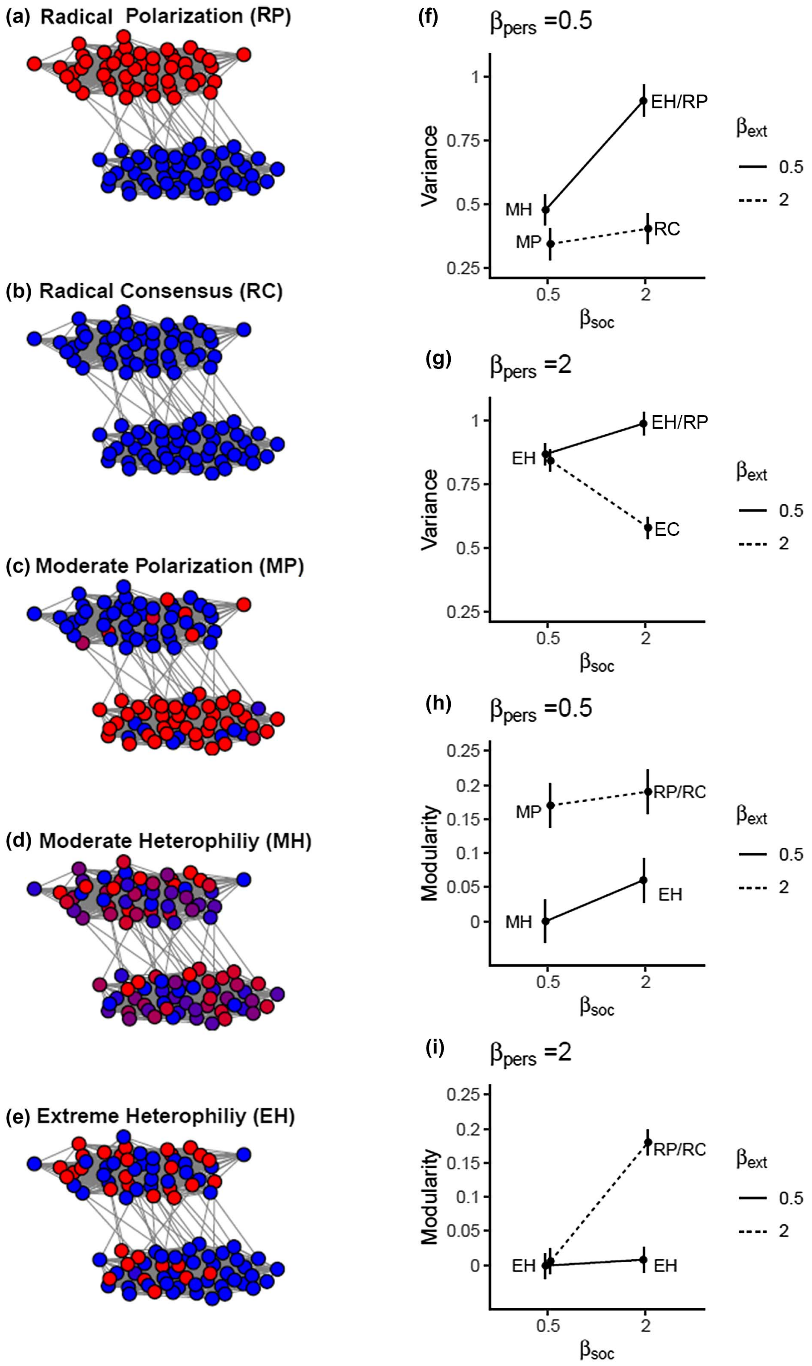
Third, when attention to social dissonance is high and attention to external dissonance is low ($\beta_{\text{soc}} = 2$ and $\beta_{\text{ext}} = 0.5$), extreme heterophily (EH) is again likely to emerge, independently of the level of attention to personal dissonance (β_{pers}). In this case, individuals try to align their personal beliefs and perceived beliefs of their social contacts, unfettered by concerns about what their social contacts actually believe. These dynamics allows for radicalization of personal beliefs independently of actual social environments.

Fourth, when attentions to social dissonance and external dissonance are both high ($\beta_{\text{soc}} = \beta_{\text{ext}} = 2$), either radical polarization or radical consensus will emerge (RP or RC) with equal probability, depending on minor biases in the initial random configuration of beliefs. If both clusters randomly initialize with the same tendency in beliefs, we observe extreme consensus; if the clusters randomly initialize with different tendencies in beliefs, we observe extreme polarization. In this condition, individuals are strongly motivated to conform to their social environments, radicalizing their personal beliefs in the process because the overall dissonance is lowest when all beliefs in one's network are extreme.

(*Appendices continue*)

Figure H1

Conditions for Different Patterns of Consensus and Polarization



Note. Panels (a)–(e) show the representative illustrations of five types of networks that occur at the end of simulations assuming different combinations of high and low attention to personal, social, and external dissonances (β_{pers}, β_{soc}, and β_{ext}, respectively). Panels (f) and (g) show the variance of focal beliefs, and Panels (h) and (i) their modularity at the end of simulation runs in different conditions. Each point in Panels (f)–(i) is marked by the abbreviation of the network type that is most likely to occur in that case. Error bars indicate 95% confidence intervals. In network illustrations, circles represent individuals with their personal and social beliefs (equivalent to the circles in Figure 1), and edges represent the directed influence of individuals' actual beliefs to the social beliefs of their contacts (each equivalent to the two edges connecting the individual circles in Figure 1). See the online article for the color version of this figure.

(*Appendices continue*)

Appendix I

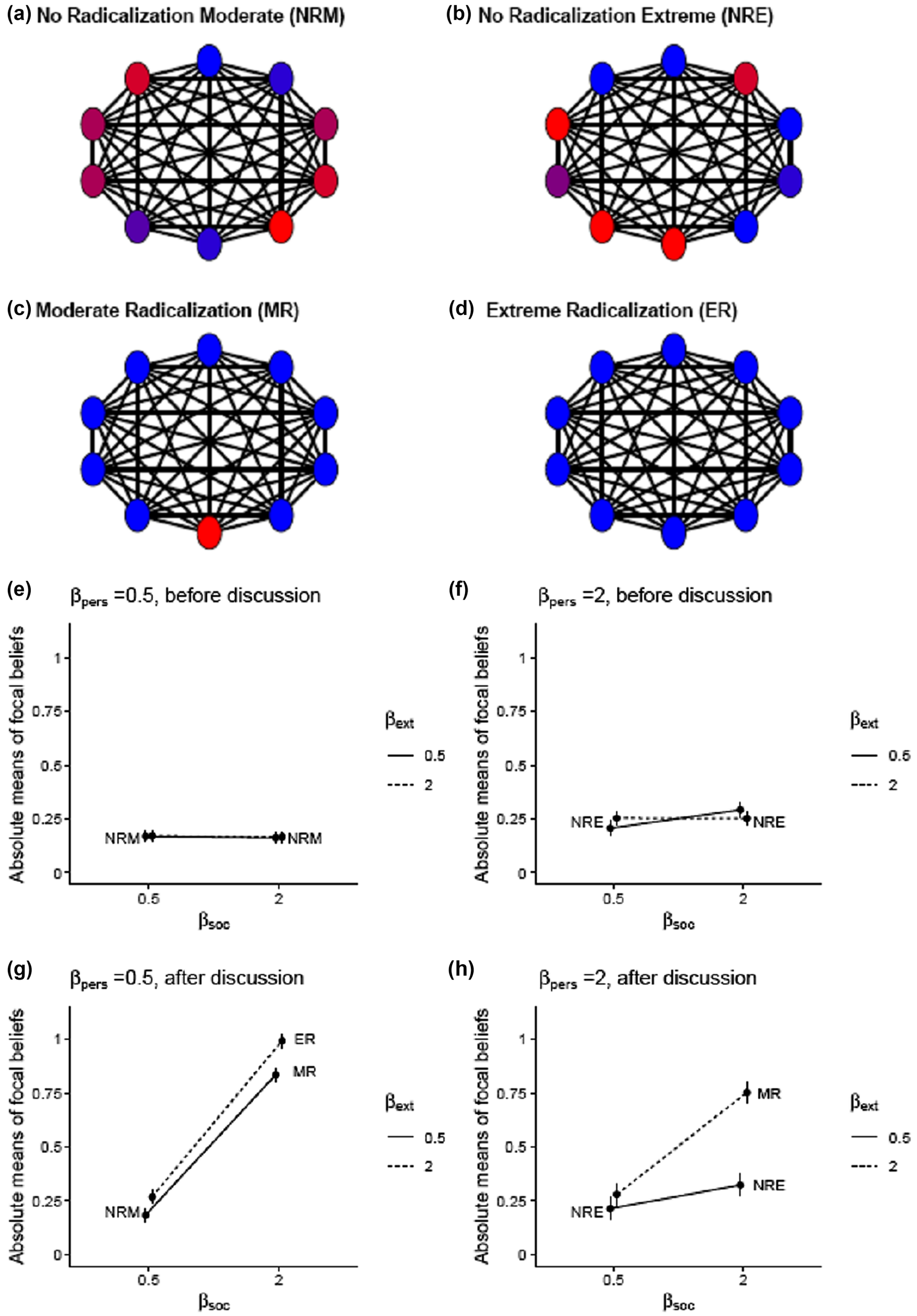
Group Radicalization

We examine the results both through network visualizations and quantitative indicators. Panels (a)–(d) in Figure I1 show four different distributions of beliefs in networks observed at the end of our simulations. In the no radicalization moderate network, members have a range of relatively moderate beliefs. In the no radicalization extreme network, members have mostly extreme, but diverse beliefs. In moderate radicalization and extreme radicalization, almost all or all members, respectively, have the same extreme beliefs. For more nuanced results, Panels (e)–(h) provide absolute means of focal beliefs at the end of the two phases of our simulation, for different conditions. The higher those means, the more radicalized the group. Panels (e) and (f) show that, as expected, without group discussion beliefs in all conditions show no group radicalization, with different individuals still holding different beliefs after 100 iterations. The only difference between conditions is that a stronger attention to the consistency of personal beliefs ($\beta_{pers} = 2$ vs. 0.5) leads to more extreme personal beliefs (Panels (f) vs. (e) in Figure I1). After group discussion, however, beliefs can radicalize but not necessarily. When attention to personal dissonance is low ($\beta_{pers} = 0.5$), high attention to social dissonance ($\beta_{soc} = 2$) is sufficient to foster moderate (MR) or extreme (MR) radicalization independently of external dissonance. In other words, when people try to align their personal beliefs with perceived beliefs of others, while disregarding the consistency between their personal beliefs, this will lead to radicalization even if their perceptions do not correspond to the actual beliefs of others. When attention to personal dissonance is high ($\beta_{pers} = 2$), high attention to social dissonance ($\beta_{soc} = 2$) is not sufficient to overcome the need to align personal beliefs with each other. In that case, high attention to external dissonance ($\beta_{ext} = 2$) is also needed to produce at least moderate radicalization.

(*Appendices continue*)

Figure I1

Conditions for Group Radicalization Before and After Discussion



Note. Panels (a)–(d) show the representative illustrations of four types of networks that occur at the end of simulations assuming different combinations of attention parameters. Panels (e) and (f) show the absolute means of focal beliefs after 100 runs before group discussion, and Panels (g) and (h) show the same at the end of additional 100 runs after group discussion. Error bars indicate 95% confidence intervals. In network illustrations, circles represent individuals with their personal and social beliefs (equivalent to the circles in Figure 1), and edges represent the directed influence of individuals' actual beliefs to the social beliefs of their contacts (each equivalent to the two edges connecting the individual circles in Figure 1). Blue nodes indicate negative beliefs and red nodes indicate positive beliefs, with higher color saturation corresponding to higher extremity of beliefs. See the online article for the color version of this figure.

(*Appendices continue*)

Appendix J

Minority Influence

To investigate under which circumstances we observe majority influence, minority influence, or no influence of either group on the other, we study both visual representations of networks typically observed at the end of our simulations and mean focal beliefs at those end points. As shown in Panels (a)–(c) of Figure J1, we observe three representative end networks: minority influence where minority belief overcomes the whole network, majority influence where majority influence wins, and polarization where both groups become radicalized and extremely different from each other.

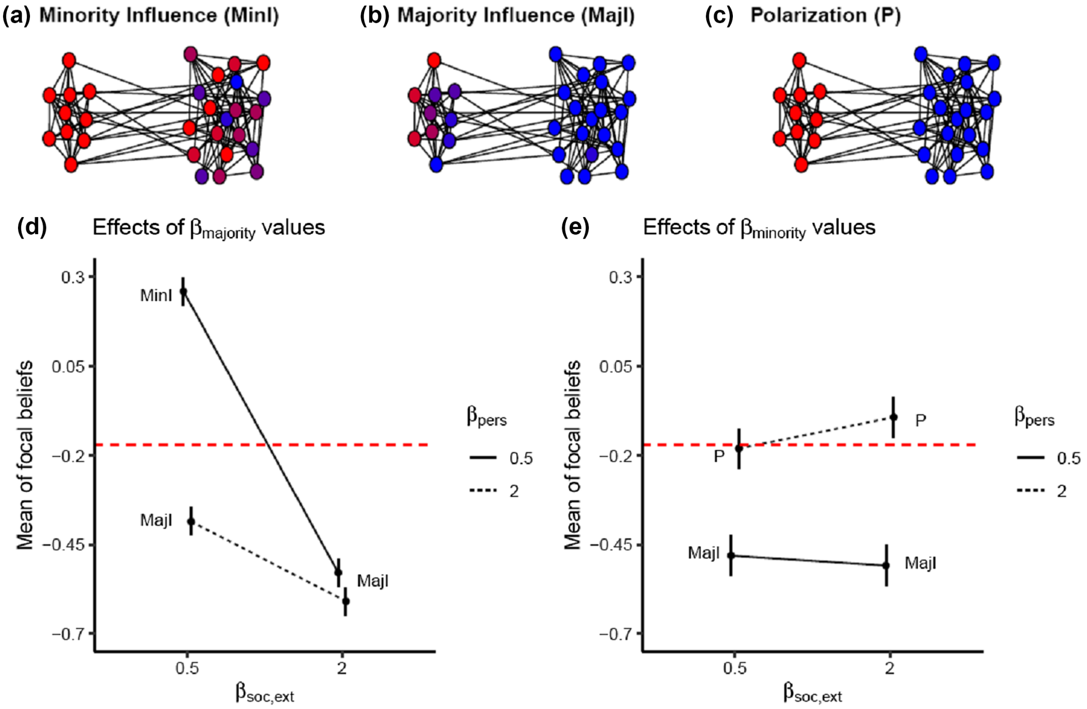
Furthermore, Panel (d) of Figure J1 shows mean end focal beliefs for different combinations of attention parameters of the majority group, averaging over the attention parameters of the minority group. Panel (e) shows equivalent results for the attention parameters of the minority group. Note that in this setup, where as described above the majority comprises 2/3 of the population and has an average initial belief of −.5, and the minority comprises 1/3 of the population and has an average initial belief of .5, the average belief without any influence is expected to be −.17, as indicated by the red dashed line in Panels (d)–(e) of Figure J1. Mean end focal belief that is around −.17 indicates no influence of either group, a mean that is clearly below indicates majority influence, and a mean that is clearly above indicates minority influence.

The results in Figure J1 suggest that a strong minority influence can occur only when majority group pays little attention to the dissonances in personal, social, and external belief networks (whenever $\beta_{\text{pers}} = \beta_{\text{soc}} = \beta_{\text{ext}} = 0.5$; top left point in Panel (d)). When the majority attends to either of these dissonances, majority influence is more likely to occur. Independently of majority's levels of attention, minority group can at least avoid majority influence (if not influence the majority itself) by paying a lot of attention the consistency of own personal beliefs (top two points in Panel (e)). In this case, group polarization occurs with the two groups holding radically different beliefs.

(*Appendices continue*)

Figure J1

Conditions for Minority and Majority Influence



Note. Panels (a)–(c) show the representative illustrations of three types of networks that occur at the end of simulations assuming different combinations of attention parameters. Panel (d) shows the effects of majority β values, averaging over minority β values, and Panel (e) shows the effects of minority β values averaging over majority β. Red dashed lines indicate no majority or minority influence. In network illustrations, circles represent individuals with their personal and social beliefs (equivalent to the circles in Figure 1), and edges represent the directed influence of individuals' actual beliefs to the social beliefs of their contacts (each equivalent to the two edges connecting the individual circles in Figure 1). Blue nodes indicate negative beliefs and red nodes indicate positive beliefs, with higher color saturation corresponding to higher extremity of beliefs. See the online article for the color version of this figure.

Received April 11, 2023
Revision received March 22, 2024
Accepted May 22, 2024 ■