Predicting Polarization Using Personality Traits - Calibrating an Agent-Based Model on Survey Data

Rahel Geppert[1,2] and Jürgen Maes[1]

[1]Conflict and Social Psychology, University of the Bundeswehr München

[2]Differential Psychology and Psychological Assessment, HMU Health and Medical University Erfurt

Author Note

This research is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU. The Authors declare that there is no conflict of interest. The data in this study is available per request to the first author, Rahel Geppert, at rahel.geppert@unibw.de.

Abstract

We examine whether classic opinion-dynamics models can be strengthened by grounding their core parameters in psychological domain knowledge. First, we fielded a preregistered survey experiment in Germany (N = 1,804) that mimics a single ABM updating step. Participants reported justice-relevant traits, empathy, ambiguity tolerance, justice sensitivity, and moral indignation, then engaged with counter-attitudinal content. From these data we estimated generalized linear models linking traits to two Hegselmann–Krause parameters: the confidence bound c (openness to differing opinions) and the self-influence weight m (inverse to susceptibility to influence). Empirically, a subset of theorized links was supported: cognitive empathy widened c, moral indignation narrowed c, and victim-oriented justice sensitivity increased m. We then calibrated a bounded-confidence ABM with these mappings and evaluated it against four canonical polarization findings. The empirically informed model substantially improved qualitative replication of group extremization and signed directional change under mixed evidence, matched the backfire contrast as good as the baseline, and, like the baseline, failed to reproduce a superlinear polarization slope. These results show how justice-focused traits, particularly empathy and moral indignation, provide a parsimonious, data-driven bridge from micro-level psychology to macro-level polarization dynamics, while also revealing limits that motivate extensions (e.g., identity cues and framing) in future work.

Keywords: Polarization, Simulation, Personality, Agent-Based Model

Introduction

Classic models of opinion dynamics capture how beliefs change through local interaction, but they often abstract away from the rich mechanisms documented in the social sciences Bramson et al. (2016) and Smaldino (2023). We argue that such models can be strengthened by explicitly grounding their core mechanisms in psychological domain knowledge. Here, we do so in four steps: (1) we develop a theoretical link between four justice-relevant traits, empathy, ambiguity tolerance, justice sensitivity and moral indignation, and the micro-processes that drive opinion change; (2) we test these links in an empirical study designed to mimic a single updating step in an agent-based model (ABM), estimating how the traits shape openness to differing opinions and susceptibility to influence; (3) we feed the resulting empirical associations into a classic ABM by mapping traits to its key parameters; and (4) we evaluate whether this empirically informed ABM better reproduces canonical polarization findings than a baseline version of the same model that omits the social-science layer.

Societal Polarization

Explaining polarization is vital given its societal relevance. The increasing polarization of societies has become a central concern for both scientific research and public discourse, as evidenced by its growing prominence in academic literature and the media (Kreiss & McGregor, 2024).This interest results from the fact that polarization, among other things, threatens democracy and social cohesion (Kish Bar-On et al., 2024), it threatens citizens' willingness to compromise and interact with one another(Kish Bar-On et al., 2024), and it fosters the distribution of fake news within a population (Cole et al., 2025). Scholars and policymakers alike are engaged in ongoing debates about how to measure polarization, uncover its drivers, predict its evolution, evaluate its societal consequences, and design interventions to mitigate its adverse effects (Berntzen et al., 2024; Cakanlar, 2024; Pereira et al., 2025; Wagner, 2024).

Explaining polarization plausibly requires considering interindividual differences that foster it (Devia, Giordano, et al., 2023; Nyhan & Reifler, 2010; Schweighofer et al., 2020; Taber & Lodge, 2006). For instance, high neuroticism and low agreeableness increase polarization (Lou & Xu, 2025), high levels of intellectual humility decrease it (Bowes & Tasimi, 2025). Given the limited attention paid to such effects in the modeling community, we investigated whether personality traits, especially those with a strong interpersonal component, can predict societal polarization. If such traits, including empathy, tolerance of ambiguity, justice sensitivity and moral indignation, are shown to influence opinion dynamics, they could enhance our ability to forecast polarization trajectories across contexts for which reliable psychological data are available.

More specifically, we contribute to the literature on issue polarization, which refers to the extent to which societal opinions are distributed across extremes of an ideological spectrum. Issue polarization is marked by bimodal or multimodal opinion distributions, with peaks drifting towards the more extreme points (Bramson et al., 2016; DiMaggio et al., 1996). Such multimodality can affect concrete democratic outcomes such as voter turnout, electoral volatility, satisfaction with democracy, and the likelihood of political or even violent conflict (Dalton, 2008; Esteban & Schneider, 2008; Hoerner & Hobolt, 2020). Understanding the antecedents, mechanisms, and consequences of opinion polarization is thus crucial for diagnosing democratic vulnerabilities and designing policy responses that reinforce democratic resilience. We aim to extend this understanding by linking polarization to individual-level psychological characteristics and embedding these relationships within a dynamic simulation model.

Theoretical Foundations

Agent-Based Modeling and the Integration of Social Science Domain Knowledge

There are several reasons for adding more psychological characteristics to dynamic simulation models. The growing complexity of societal polarization has rendered overly simplistic, physics-inspired (spin-like) models of social behavior insufficient for capturing the richness of human interaction (Franceschetti et al., 2022; Sobkowicz, 2018; Squazzoni, 2023). As Schweighofer et al. (2020) emphasized, many models in sociophysics rely on ad-hoc assumptions that fail to account for the multifaceted drivers of individual and group-level decision-making. To address this limitation, the integration of social science domain knowledge, such as empirical findings from sociology, psychology, political science, and economics, into agent-based modeling (ABM) has become essential for predicting, steering, and modifying collective behavior (Alizadeh et al., 2015; Kalvas et al., 2022).

Including social science domain knowledge in generative models improves the ecological validity of, for example, generative mechanisms and agent heterogeneity. As an example, simulating fear-driven behavior during crises can be more credibly executed using psychological constructs rather than arbitrary rules (Horned & Vanhée, 2023; Kish Bar-On et al., 2024). This increased ecological validity of generative processes makes model processes and predictions easier to translate to real-world scenarios, which is helpful in cases in which models claim to model real-world human interactions, as in the example of fear-driven behavior. Basing generative mechanisms on social science domain knowledge is particularly relevant when causal processes involve social science variables such as preferences, normative constructs, or socio-cognitive biases (Renzini et al., 2023). Adding social science domain knowledge to dynamic simulation models can also be a way to realistically model heterogeneity. ABMs informed by domain knowledge allow for heterogeneous agent modeling, incorporating traits such as cognitive biases, identity-based motivations, and emotion-driven decision-making. Implementing such traits leads to greater behavioral realism, which is critical for explaining and intervening in social phenomena.

The utility of ABMs informed by social science domain knowledge can also be considered from a social science perspective: It extends to data-poor or ethically sensitive contexts where direct experimentation is infeasible. In such settings, social science informed models provide a rigorous framework for exploring policy alternatives, such as in health access or alcohol regulation scenarios (Renzini et al., 2023; Sgouros, 2023). Importantly, these models do not only apply existing theory, they also generate new theoretical insights. Approaches such as inverse Generative Social Science (iGSS) illustrate how empirically calibrated simulations can help revise and refine behavioral theories (Plähn et al., 2023).

Finally, social science insights are indispensable for addressing structural limitations in current modeling paradigms. For example, models of opinion dynamics that rely solely on attraction/repulsion mechanics often fail to replicate the multimodal distribution of real-world political opinions (Squazzoni, 2023). By incorporating constructs like network topology, identity formation, and differentiated influence mechanisms, ABMs can more accurately represent complex political landscapes (Chattoe-Brown, 2023). Thus, integration of social science domain knowledge greatly advances scientific robustness and policy relevance in computational modeling (Secchi, 2023). Recent publications demonstrate a shift in the modeling community towards investigating the role of personality traits in shaping susceptibility to polarization. For example, Lou and Xu (2025) use agents embodying the Big Five trait of personality (e.g., openness, conscientiousness, extraversion, agreeableness, and neuroticism) to model disinformation spread, while Muhammad and Kasahara (2024) explore the interaction between personality, trust, and information dissemination. Both demonstrate that polarization is deeply rooted in psychological variation. Similarly, We show how incorporating psychological constructs from justice research into a Hegselmann-Krause model improves the model's ecological validity.

Justice Research and Personality in Polarization Dynamics

In fact, justice-related emotions and cognitions are central to affective polarization, defined as negative emotional responses to outgroups (Gollwitzer & van Prooijen, 2016; Kish Bar-On et al., 2024; Peter et al., 2013; Saveski et al., 2022). The perception of injustice triggers strong emotional reactions, such as anger, moral indignation, or resentment, and motivates efforts to restore justice (Gollwitzer & van Prooijen, 2016; Peter et al., 2013), all of which are linked to increased affective polarization (Crockett, 2017; Dagnes, 2019; Jahnke et al., 2020). Not only do justice-related traits relate to affective polarization, there also is first evidence linking these traits to issue polarization. Cole et al. (2025) report that issue polarization is shaped by differing moral foundations: conservatives often emphasize binding values (loyalty, authority), whereas liberals emphasize individualizing values (care, fairness). Given these findings regarding justice-related traits and (mostly affective) polarization, we postulate that similar associations occur for issue polarization. Moreover, by modeling the traits' influence on polarization in a Hegselmann-Krause model, we can effectively calibrate the model to better match empirical polarization patterns.

Modeling Justice-Relevant Traits in Polarization

We propose a modified Hegselmann-Krause model in which two central parameters, confidence bound (c) and self-influence (m), are calibrated by regressing them on four justice-relevant personality traits: empathy, ambiguity tolerance, justice sensitivity, and moral indignation. These traits determine how open an agent is to opposing viewpoints (as displayed by parameter c) and how easily their opinions are susceptible to influence (as displayed by parameter m).

Empathy

is defined as the ability to vicariously understand and emotionally resonate with another's internal state (Davis, 1983). Perspective-taking, the cognitive component of empathy, enhances reflective political reasoning and openness to differing opinions (Muradova, 2021). Empathetic listening increases attitudinal clarity and reduces prejudice during disagreements (Weinstein & Itzchakov, 2025). Empathetic concern, the affective component, is also a strong predictor of tolerance, particularly in controversial or emotionally charged dialogues (Butrus & Witenberg, 2013). We therefore predict that both cognitive and affective empathy are positively associated with openness for differing opinions, i.e., parameter c.

Ambiguity Tolerance

describes comfort with uncertainty and novel information (Budner, 1962; Litman, 2010). Higher tolerance is associated with openness to differing opinions and reduced cognitive rigidity (Luzsa & Mayr, 2019; Salvi et al., 2023). In contrast, low ambiguity tolerance correlates with black-and-white thinking and heightened xenophobic attitudes (Furnham & Marks, 2013). We therefore predict that ambiguity tolerance also are positively associated with openness for differing opinions, i.e., parameter c.

Justice Sensitivity

refers to heightened sensitivity to perceived injustice, distinguishable across victim, observer, beneficiary, and perpetrator orientations (Schmitt et al., 2010). Observer sensitivity promotes prosocial behavior and openness to differing opinions (Baumert et al., 2022; Bondü et al., 2022), while victim sensitivity is often linked to defensiveness, social withdrawal, and low receptivity (Bondü et al., 2022). Following previous findings, we assume that specifically justice sensitivity from the observer and victim perspectives can help predict agents' behavior. We predict that high justice sensitivity from the observer perspective is positively associated with openness for differing opinions, i.e., parameter c, whereas justice sensitivity from the victim perspective is negatively associated with the degree of susceptibility to influence (and therefore positively to the self-influence parameter m).

Moral Indignation

is defined as strong emotional arousal in response to perceived moral violations (Haidt et al., 2003). In addition to being defined as a state variable, Disposition for moral indignation can also be measured as a personality trait (Bernhardt, 2000). In this case, it reflects how quickly people tend to become morally indignant in ambiguous situations. While appropriate in response to dignity violations, excessive or mismatched moral indignation can fuel polarization and reduce conversational openness (Lück & Nardi, 2019; Neuhäuser, 2023; Nyhan & Reifler, 2010). Given these findings, we predict that moral indignation is negatively associated with the openness for differing opinions, i.e., parameter c, reinforcing opinion rigidity and resistance to disconfirming evidence.

Thus, higher empathy and ambiguity tolerance increase c, making agents more open to diverse inputs. Justice sensitivity modulates both c and m, with orientation-specific effects, and moral indignation typically reduces c.

To test the applicability of a model enriched in this way, we compare the model's prediction with four established findings in research on polarization.

The Four Findings

Moscovici and Zavalloni (1969)

Conducted with small discussion groups in late-1960s France, participants privately rated attitude targets (e.g., public figures, out-groups), discussed them, and then rated the targets again. Polarization was operationalized as group extremization: post-discussion means shifted further away from the neutral point, and a higher share of individuals moved toward the group's prior direction.

Lord, Ross and Lepper (1979)

Initially pro- and anti–capital-punishment participants were exposed to mixed evidence (both supportive and critical studies). Polarization was operationalized as divergent change: proponents became more favorable while opponents became less favorable, assessed with mean difference scores and corresponding t-tests by initial stance.

Taber and Lodge (2006).

In the mid-2000s U.S. political context, participants evaluated balanced pro and con arguments on controversial issues while their prior attitudes were recorded. Polarization was operationalized as biased assimilation: post-treatment extremity increasing with a regression slope of post on pre extremity exceeding 1.

Bail et al. (2018)

U.S. partisans were incentivized to follow a Twitter bot that retweeted messages from ideological opponents for about a month. Polarization was operationalized as movement in issue positions relative to baseline: cross-cutting exposure produced "backfire," i.e., shifts away from the out-party and toward more extreme in-party–consistent views.

Hypotheses

Synthesizing the theoretical rationale and the mechanisms distilled from classic polarization studies, we posit that justice-related dispositions systematically modulate the two core parameters of the Hegselmann–Krause model, agents' confidence bound c (openness to differing opinions) and self-influence m (inverse to susceptibility to influence). If these traits shape who listens to whom (c) and how strongly opinions are updated once interaction occurs (m), then trait-informed ABMs should reproduce empirical polarization signatures more faithfully than a baseline model without trait influences. We therefore state and test the following hypotheses:

H1 There are empirical associations between justice-related traits and agent parameters c and m.

(a) Empathy is positively associated with willingness to talk to people with differing opinions

(b) Ambiguity tolerance is positively associated with willingness to talk to people with differing opinions

(c) Justice sensitivity from the observer perspective is positively associated with willingness to talk to people with differing opinions

(d) Justice sensitivity from the victim perspective is positively associated with the degree of self-influence in an interaction

(e) Disposition for moral indignation is negatively associated with willingness to talk to people with differing opinions

H2 When entering empirically found associations into a Hegselmann-Krause model, this model then predicts classic polarization phenomena better than a Hegselmann-Krause model without the trait influences.

Study 1: Empirical Data

Sample, Procedure, Measures

The survey experiment was built so that it followed the mechanisms of a single updating step in the opinion-dynamics ABM. The design and analysis plan were preregistered on Oct 16, 2023 at the OSF, (https://doi.org/10.17605/OSF.IO/SK274). In addition to assessing relevant personality traits, the experiment included an intervention in which self-influence and external influence were simulated. We used the data from the experiment to determine the range and distributions of the relevant agent traits, agent' confidence bounds, as well as their interaction with each other. This data was then used to calibrate the ABM so that the baseline distributions of all character traits were realistic and that the individual part of agents' parameters c and m could be predicted by their traits. The online survey was conducted from 17.01.24 to 05.02.24 via the panel provider Bilendi and included $N = 1804$ complete datasets.

Sample

The sample was stratified to be roughly representative of the German population. Relevant sociodemographics are reported in Table 1.

Procedure and Measures

Figure 1 illustrates the procedure. Participants first filled out the questionnaires for the personality traits. Participants rated all personality traits on a 6-point answering scale with the endpoints labeled as *does not apply at all* and *fully applies.* The exact and original German wording of all items can be found in the online supplemental material.

We measured empathy with the short version of the Basic Empathy Scale (BES; original scale in Jolliffe and Farrington, 2006, short version in Oliva Delgado et al., 2011, in the German translation of Heynen et al., 2016). The short version of this scale consists of nine items, of which four items measure affective empathy and five items measure cognitive empathy. The scale showed acceptable internal consistency, $\alpha_{cognitive} = .725$ and $\alpha_{affective} = .723$.

Ambiguity tolerance was measured with a modified scale by Maes et al. (1996). In the original version, this scale consists of five items which describe ambiguity *intolerance*, in our study we extended the scale by five inversely formulated items measuring ambiguity *tolerance.* The subscales showed acceptable internal consistency, $\alpha_{intolerance} = .63$ and $\alpha_{tolerance} = .68$ and small correlation between the subscales $r = .06, p = .011$.

Justice sensitivity from the observer and victim perspective were measured by the corresponding items by Schmitt et al. (2010), with ten items each. The scales showed excellent internal consistencies, $\alpha_{victim} = .91$ and $\alpha_{observer} = .93$ and correlated moderately $r = .57, p < .001$.

Disposition for moral indignation was a self-constructed scale, which we validated in three previous studies. The scale consists of 17 items, five of which ask for general self-assessment regarding disposition for moral indignation, and 12 which assess thoughts and behavior that could be indicative of moral indignation in reference to a specific situation (e.g., a stranger disregarding commonly known rules). In the three validation studies ($n_1 = 96$, $n_2 = 147$, $n_3 = 197$), the entire scale showed excellent internal consistencies $\alpha_1 = 0.82$, $\alpha_2 = .89$ and $\alpha_3 = .92$. The internal consistency in this study was also excellent, $\alpha = .85$.

The measures for participants' traits were followed by a rating of an opinion on a controversial topic and opinion certainty. In order to have participants' opinions on a real-world issue, we asked them to rate their opinion on migration in Germany. This opinion was measured on a 10-point scale with the endpoints labeled as *I am someone who rejects migration in Germany* and *I am someone who supports migration in Germany*, in addition, all gradations on the scale were numbered from 1 to 10. To quantify c, i.e., the participants' confidence bound, participants were shown the opinion scale again and asked to select all those opinions that they would be willing to have a conversation with. c was then recorded as the percentage of the scale selected (i.e., only one number selected = 10% or 0.1, all numbers selected = 100% or 1)

The experiment then entailed an intervention in two parts: In the first part of the intervention, we induced the self-influence part of a typical ABM by asking participants to please elaborate on their opinion. Participants had to write at least 30 characters before they could continue with the survey. The parameter m, i.e., participant's individual self-influence, was derived from the texts that participants wrote, by sorting them onto the opinion scale the same way as the previously created texts that they interacted with. The self-influence was then estimated as the difference between their previously stated opinion and the value of their written text, mapped onto a scale from 0 to 1.

In the second part of the intervention, participants were shown a text with a differing opinion. We had prepared ten texts in advance, which were rated in a pre-study (n = 629).We then sorted each text to one of the ten steps of the opinion scale, so that we had a text illustrating each possible opinion. Each participant then was shown a text corresponding to the position that had the largest distance to the participant's own opinion while still lying within their confidence bound.

After they were presented with the text, participants were again asked to state their own opinion and opinion certainty. The study ended with a couple of questions as intervention checks ("Did you read the text you were shown?", "Did writing a short statement change your opinion?", "Did reading the other person's statement change your opinion?").

Data Analysis

To test hypothesis H1, we performed two regression analyses in which we predicted the confidence bound c and the self-influence m using the personality traits. Since both c and m were not normally distributed but bottom-heavy (in case of c) or top-heavy (in case of m), we decided to estimate a generalized mixed model (GLM) with gamma distribution and logistic link-function for each dependent variable. Table 2 displays regression weights, p-values and which effect we hypothesized.

Results

Table 2 shows that only a subset of the theorized links between justice-relevant traits and the model parameters is supported in the expected direction. Most notably, cognitive empathy positively predicts a wider confidence bound c (greater openness to differing opinions), and moral indignation negatively predicts c (narrower openness), consistent with our theorizing that empathetic perspective taking facilitates receptivity while indignation tends to heighten rigidity and resistance to counter-attitudinal input (Davis, 1983; Muradova, 2021; Neuhäuser, 2023; Nyhan & Reifler, 2010). These two effects are theoretically crisp and empirically robust in our data, and they provide the clearest leverage for calibrating agent heterogeneity in the ABM. By contrast, several predictors, especially those posited to shape the confidence bound c, did not emerge in the hypothesized direction, whereas all predictors for the self-influence parameter m behaved as expected.

Study 2: Agent-Based Model

Model Specification

Next, we fed the associations from the Study 1 into our agent-based model so that, instead of drawing the parameters c (confidence bound) and m (self–influence) at random, agents first drew *character traits* from empirically informed distributions and then computed c and m from a regression whose nonzero coefficients matched Table 2: cognitive empathy, ambiguity tolerance, justice sensitivity (victim perspective), and moral indignation for predicting c, and justice sensitivity (victim) for predicting m. After visual inspection and posterior predictive checks, trait distributions were approximated either by a Normal or a zero–one–inflated Beta (ZOIB); Normal provided the better fit for all traits except ambiguity tolerance (Figure 2).[1]

Concretely, for each agent i with traits $\mathbf{X}_i$, we used the Gamma–log GLMs estimated on the survey data:

$$\log \mathbb{E}[c \mid \mathbf{X}] = -1.926 + 0.301\, ce - 0.465\, a + 0.240\, vjs - 0.560\, mi$$

$$\log \mathbb{E}[m \mid \mathbf{X}] = -0.198 + 0.044\, vjs$$

. We set $c_i = \exp(\cdot)$ as the agent-specific confidence radius (in the same units as opinions; see below), constrained to be nonnegative. Because m_i serves as a *weight* in the convex combination below, we bounded it to $[0, 1]$ (clipping values from the GLM if necessary) so that updates remained averages.

The model comprises $N = 1,000$ agents connected by a Watts–Strogatz small-world graph with mean degree $k = 2$ and rewiring probability $p = .02$ (Watts & Strogatz, 1998). This captures short paths with nontrivial clustering while avoiding complete regularity. Agent attributes used in the update (traits, c, m, and current opinion) are summarized in Table 3.

Details of the model specification as suggested by the ODD-protocol (Grimm et al., 2020) are given in Appendix A.

Apart from these empirically informed parameters, the interaction rule follows the classic Hegselmann–Krause (HK) bounded-confidence model (Hegselmann & Krause, 2002). Agents hold a continuous opinion $x_i(t)$ and update *synchronously.* At time t, agent i averages the opinions of those j whose opinions are within c_i of its own:

$$\mathcal{N}_i(t) \;=\; \{\, j : |x_j(t) - x_i(t)| \le c_i \,\}, \qquad \bar{x}_i(t) \;=\; \frac{1}{|\mathcal{N}_i(t)|} \sum_{j \in \mathcal{N}_i(t)} x_j(t).$$

To allow persistence of prior beliefs, we blend the HK average with a self–weight m_i:

$$x_i(t{+}1) \;=\; m_i\, x_i(t) \;+\; \big(1 - m_i\big)\, \bar{x}_i(t),$$

[1] ZOIB was used after rescaling to $[0, 1]$ and back-transforming to $[-1, 1]$

with all agents updating in parallel. Opinions are represented on the $[1, 10]$ survey scale and truncated to that range after each update. The HK model with homogeneous c is known to produce consensus for large c and multiple opinion clusters for smaller c; heterogeneous confidence bounds (as used here) are a common extension (Blondel et al., 2009; Lorenz, 2007).

In order to represent both the process and the results of the four findings as accurately as possible, we added additional parameters. All of these parameters were selected so that they would qualitatively represent the findings in a baseline model without social science calibration. In a second step, we then tested whether social science calibration improved the proximity to the empirical findings.All of these parameters, their definitions and their values are displayed in Table **??**.

We use optional biases to form weights for the neighbors' opinions, which is relevant in the models mimicking the experiments by Taber and Lodge (2006) and Lord et al. (1979), additionally both Bail et al. (2018) and Moscovici and Zavalloni (1969) report that reaction to given information differed depending on whether the source of that information was considered a member of the ingroup or of an outgroup. To allow for that effect, we introduced the weights g_{in} and g_{out}, so that the updating process reads

$$w_{ij}(t) \;=\; \exp\bigl(d_{ij}(t)/c_i\bigr) \times \underbrace{\begin{cases} w_+ & \text{if } i, j \text{ on same side of the midpoint } (5) \\ w_- & \text{otherwise} \end{cases}}_{\text{confirmation bias}} \times \underbrace{\begin{cases} g_{\text{in}} & \text{if } i, j \text{ share group} \\ g_{\text{out}} & \text{otherwise} \end{cases}}_{\text{ingroup weighting}} .$$

The (possibly biased) local mean is $\tilde{x}_i(t) = \frac{\sum_{j \in \mathcal{N}_i} w_{ij}(t)\, x_j(t)}{\sum_{j \in \mathcal{N}_i} w_{ij}(t)}$.

To allow for selective exposure as reported by Taber and Lodge (2006), we introduced a parameter κ, which further limits the number of neighbors that can influence an agent c.

In the model reported by Bail et al. (2018), interaction can lead to repulsion as well as convergence. To allow for that mechanism, we added a parameter $\delta_{repulse}$ to define the distance to an agent's own opinion that would lead to repulsion and the repulsion step size ρ which quantified the degree of repulsion if it occured.

Both Taber and Lodge (2006) and Moscovici and Zavalloni (1969) report a tendency towards more extreme opinions, which we ensured by adding a drift towards a more extreme opinion, α. This drift was added to each interaction step:

$x_i(t{+}1) \leftarrow 5 + (1 + \alpha)\bigl(x_i(t{+}1) - 5\bigr)$,

The simulation can be terminated in two ways. First, we define the stopping criterion `stop.crit`, which was set to .001 for all runs. If the average change in all opinions becomes smaller than stop.crit, the simulation is terminated. If this is not achieved in 500 simulation steps, option 2 for termination applies, namely reaching `stop.max` $= 500$. Thus the model runs until

$\frac{1}{N}\sum_i |x_i(t{+}1) - x_i(t)| <$ `stop.crit` or $t =$ `stop.max`.

Experiments

To test the predictive power of using informed self-influence and confidence bound, we replicated four opinion polarization experiments, using both a Hegselmann-Krause model (with the modification parameters $w_+, w_-, \kappa, \rho, \delta_{repulse}, \alpha, g_{in}, g_{out}$ but fixed parameters of $c = 0.2$ and $m = 0.5$) and our informed variation (with the modification parameters $w_+, w_-, \kappa, \rho, \delta_{repulse}, \alpha, g_{in}, g_{out}$ and with parameters c and m predicted by traits) and compared which model was better in reproducing the original findings. That is, we had four model variations that qualitatively replicated the procedures of each original experiment. To compare the simulations to the original findings, we computed a study-specific statistic for each dataset (empirically informed model vs. a baseline model) and mapped it to a unit-interval score via monotone transforms. All four scores were computed separately for each source (empirically informed model vs. baseline); larger values indicate better qualitative agreement with the original finding.

Lord et al. (1979): mixed evidence and biased assimilation

The first version uses some modifications to better fit the experiment described by Lord et al. (1979). They had participants with strong prior attitudes evaluate mixed pro– and counter–attitudinal evidence; participants found congenial studies more convincing and became *more* polarized after exposure. To mimic simultaneous exposure to pro- and counter-attitudinal evidence, we inject two fixed *evidence* agents at the extremes (IDs 9,000,001 and 9,000,002) with $x = -1$ and $x = +1$ in the internal scale (mapped to $[1, 10]$), set to $m = 1$ and $c = 0$ (they never change and never listen). We connected them to all other agents. Other parameters remained at baseline (confirmation bias is neutral at $w_+ = w_- = 1$). Operationally, each agent's neighborhood now always included both extreme sources if within c_i; the relative pull of the extremes depended on initial location and c_i, reproducing the classic pattern that people weight congenial information more strongly and polarize relative to their priors when mixed evidence is present.

Lord et al. (1979) signed-change score

Within each source, agents were split at t=0 into a lower vs. higher initial opinion group (relative to the source-specific mean), which mimiced the identification as either opponents or proponents of capital punishment in the original study.

Within each source, we defined Proponents as $x_0 > 5$ and Opponents as $x_0 < 5$.

Since the original study reported t-tests between pre- and post-interaction opinions, we did the same and compared our t-values to those reported in the original findings as the comparison score:

Let $D = x_{t\_end} - x_{t\_0}$. We computed one-sample t-statistics for $D > 0$ in Proponents (t_{pro})

and $D < 0$ in Opponents (t_{opp}). The score rewarded the expected signs in both groups:

$$S_{\text{Lord}} \;=\; \tfrac{1}{2}\Big[\operatorname{logit}^{-1}(t_{\text{pro}}) \;+\; \operatorname{logit}^{-1}(-t_{\text{opp}})\Big].$$

Moscovici and Zavalloni (1969): group polarization via intragroup amplification

In the second finding, performed by Moscovici and Zavalloni (1969), small groups of participants first reported individual attitudes, discussed as a group, and then reported again; group discussion shifted the mean and post–discussion individuals toward more *extreme* positions consistent with the group's initial tendency. We encode group membership once at $t = 0$ by the side of the midpoint: $\text{group}_i = 0$ if $x_i(0) < 5$, else 1. We then (i) overweight ingroup opinions ($g_{\text{in}} = 1.3$, $g_{\text{out}} = 1$), and (ii) included a small extremizing drift around 5 ($\alpha = 0.08$). Confirmation bias was neutral ($w_+ = w_- = 1$), and selective exposure was off ($\kappa = 0$). These changes kept the updating mechanism averaging locally but slightly privileged same-group influence and added a mild outward push once a within-group consensus formed, reproducing the empirical finding that group discussions shift the mean further toward the pre-discussion inclination.

Moscovici and Zavalloni (1969) group-polarization score

Polarization in the original finding was assessed by an increase of extremity of opinion over time. We defined extremity as the distance from the neutral point of the opinion scale, 5.0: individual extremity is

$$E_t \;=\; |x_t - 5|.$$

For each source we summarized (a) the mean change in extremity $\overline{\Delta E} = \overline{E_{t\_end} - E_{t\_0}}$, and (b) the proportion p of agents whose extremity increased ($\Delta E > 0$). Let $s_{\Delta E}$ be the standard deviation of ΔE. The score combined effect size and majority support:

$$S_{\text{Mosc}} \;=\; \tfrac{1}{2}\Big[\operatorname{logit}^{-1}\big(\overline{\Delta E}/s_{\Delta E}\big) \;+\; \operatorname{logit}^{-1}\big((p-0.5)/0.1\big)\Big],$$

where the constant 0.1 set the steepness of the mapping around a 50% majority.

Taber and Lodge (2006): motivated reasoning, confirmation bias and selective exposure

The third finding, by Taber and Lodge (2006), also showed opinion extremization: In a lab experiment with policy arguments, participants preferentially sought congenial information and evaluated it more favorably; motivated reasoning increased attitude polarization, especially among strong partisans. We implement two mechanisms from the experiment: (i) *biased assimilation* via asymmetric weights $w_+ > 1$ for congenial content and $w_- \leq 1$ (possibly <0 for counter-arguing) for uncongenial content; and (ii) *selective exposure* by sampling a similarity-biased subset of accepted neighbors with probabilities proportional to $\exp\{\kappa(1 - d_{ij}/c_i)\}$ (we used $\kappa = 4$ in the main runs; higher values increase homophily of exposure). All other parameters were set to their baseline value. In our code this means the neighbor set $\mathcal{N}_i$ was first defined by bounded confidence, then thinned stochastically toward similar alters, and finally aggregated with congeniality-dependent weights, capturing the documented tendency to prefer agreeable information and to assimilate it more strongly.

Taber and Lodge (2006) polarization slope

Polarization was assessed by whether post-exposure extremity grew more than one-for-one with baseline extremity. We defined extremity as the distance from the neutral point of the opinion scale, 5.0: individual extremity was

$$E_t \;=\; |x_t - 5|.$$

For each source we fitted

$$E_T \;=\; \beta_0 + \beta_1\, E_0 + \epsilon,$$

and transformed the distance of β_1 from the theoretical threshold 1 using its standard error:

$$S_{\text{Taber}} \;=\; \text{logit}^{-1}\!\left(\frac{\beta_1 - 1}{\text{SE}(\beta_1)}\right).$$

Bail et al. (2018): cross-cutting exposure and backfire

Finally, Bail et al. (2018) conducted a field experiment on Twitter (now X). In this experiment, partisans were paid to follow bots retweeting opposing–party elites; Republicans became more conservative on average, while Democrats showed weak or no movement—evidence for backfire under cross–cutting exposure. We randomly assigned half the population to a *treated* condition. Two fixed "bot" agents were added at the extremes ($x = 1$ and $x = 10$ on the survey scale) with $m = 1$, $c = 0$. Each treated agent was linked to the bot representing the *opposing* side of its initial position (cross-cutting exposure). We activated a repulsion step with threshold δ and step size ρ:

$$\text{if } \exists j \in \mathcal{N}_i(t) \text{ with } d_{ij}(t) > \delta \text{ (or on the opposite side of 5), then } x_i(t{+}1) \leftarrow x_i(t){+}\rho\,\text{sign}\big(x_i(t){-}x_j(t)\big)\,|x_i(t){-}x_j(t)|.$$

Otherwise the standard convex update (with m_i) is used. We set $(\rho, \delta) = (0.10, 0.6)$, $w_+ = w_- = 1$, $\alpha = 0$, and $g_{\text{in}} = 1.2$, $g_{\text{out}} = 1.0$. This produced the backfire pattern: for treated agents, sustained exposure to far opposing content increases extremity on average, while control-group agents did not show this backfire effect.

Bail et al. (2018) backfire score

Within each source, agents were split at t=0 into a lower vs. higher initial opinion group (relative to the source-specific mean), which mimicked the association with either democrats or republicans in the original study. We first calculated the mean opinion shifts from t_0 to t_end for each opinion groups x treatment/control group (i.e., four opinion shifts) and then the difference between treatment and control shifts for each opinion group:

Let t_end be the last recorded tick, and for group $g \in \{\text{low}, \text{high}\}$ define the difference-in-differences

$$\Delta_g \;=\; \big(\bar{x}^{\text{Trt}}_{g,t\_end} - \bar{x}^{\text{Trt}}_{g,t\_0}\big) \;-\; \big(\bar{x}^{\text{Ctrl}}_{g,t\_end} - \bar{x}^{\text{Ctrl}}_{g,t\_0}\big).$$

For the score we relativized this difference using the standard deviation of individual pre–post differences $(x_T - x_0)$ within group g and then transformed this value to lie between 0 and 1:

Let s_g be the standard deviation of individual pre–post differences $(x_{t\_end} - x_{t\_0})$ within group g. The score rewarded backfire in the initially higher group and (near-)zero change in the other group:

$$S_{\text{Bail}} \;=\; \tfrac{1}{2}\Big[\operatorname{logit}^{-1}\big(\Delta_{\text{high}}/s_{\text{high}}\big) \;+\; \exp\big\{-\big(\Delta_{\text{low}}/s_{\text{low}}\big)^2\big\}\Big].$$

Results

The replication scores (0–1; higher = better fit) showed a mixed pattern across the four benchmarks, illustrated in Figure 3. For Bail et al. (2018), the baseline model and the empirically informed model were virtually identical (0.75 vs. 0.74), indicating that the observed backfire contrast can largely be matched without the model's social science calibration. In reproducing Lord et al. (1979), the empirically informed model clearly outperformed the baseline (0.50 vs. 0.26), capturing the expected signed change with proponents shifting up and opponents shifting down. The strongest gain was for Moscovici and Zavalloni (1969), where the empirically informed model achieved a very high score (0.90) relative to the baseline model (0.31), consistent with robust group polarization in extremity. By contrast, for Taber and Lodge (2006) both models scored 0, meaning neither reproduced a slope greater than 1 from pre-exposure to post-exposure extremity. Overall, the empirical informed model substantially improved fit for classic polarization and signed-change effects (Lord et al., 1979; Moscovici & Zavalloni, 1969), did not add explanatory power for the backfire pattern (Bail et al., 2018), and both empirical informed model and baseline model failed to generate the superlinear polarization predicted by Taber and Lodge (2006).

Discussion

Summary of Results

Study 1 yields a mixed but informative pattern. Only a subset of the theorized links between justice-relevant traits and the model's parameters is supported in the expected direction (Table 2). Most notably, *cognitive empathy* reliably predicts a wider confidence bound, c (greater openness to differing opinions), whereas *moral indignation* predicts a narrower c (reduced openness), precisely in line with the idea that perspective taking facilitates receptivity while indignation fosters rigidity and resistance to disconfirming information (Davis, 1983; Muradova, 2021; Neuhäuser, 2023; Nyhan & Reifler, 2010).

By contrast, several other predictors, especially those posited to shape the confidence bound c, did not emerge in the hypothesized direction or were weak. In comparison, *all* predictors for the self-influence parameter m behaved as expected (Table 2). This pattern diverges from much of the theorized evidence base, which typically examines sustained interpersonal dialogue, deliberation, or choice-of-exposure contexts (Baumert et al., 2022; Bondü et al., 2022; Butrus & Witenberg, 2013; Luzsa & Mayr, 2019; Salvi et al., 2023).

Turning to model performance, using empirically informed c and m improved the reproduction of classic polarization findings (e.g., stronger fit for the Moscovici and Zavalloni (1969) shifts and the signed changes in Lord et al. (1979), but it did not capture certain backfire patterns (Bail et al., 2018; Taber & Lodge, 2006).

Implications

The fact that not all associations between traits and model parameters c and m were found as predicted in Study 1 leads to interesting theoretical implications. Our single-step design fixes exposure and strips away many situational affordances (relationship cues, platform features, strategic avoidance) that likely amplify trait effects on *openness for differing opinions* per se. Thus, the unexpected results are informative in that they indicate that empathy and indignation exert fast, gating-level shifts in c, whereas ambiguity tolerance and justice-sensitivity orientations may manifest more strongly when agents can select interlocutors, curate information, or interact repeatedly, which are the conditions that classic studies use. This helps refine theory by locating when and where trait to c pathways are most active, while confirming that trait to m pathways are already operative at the granularity of a single opinion-update.

Substantively, the moral-indignation result is especially interesting. Trait-like readiness to feel morally outraged is a relatively new addition to opinion-dynamics modeling, yet here it shows a clear, directional association with narrower c. This finding aligns with work on motivated reasoning and moralized discourse that links indignation to lower receptivity and greater polarization (Lück & Nardi, 2019; Nyhan & Reifler, 2010). In modeling terms, allowing indignation to shrink c generates more selective neighborhoods and, in turn, steeper effective confirmation dynamics—an empirically grounded pathway from moral emotion to macro-level opinion patterns.

The mixed success in reproducing patterns from four classic polarization findings in Study 2 also leads to several implications. They suggest that trait-based heterogeneity is helpful in reproducing certain polarization scenarios. Specifically backfire effects likely also depend on other variables such as message framing, identity threat, or networked exposure asymmetries that are not fully represented by

c and m alone. Future models could experiment with adding these influences in addition or instead of our traits. More broadly, our results support calls to enrich ABMs with psychologically meaningful variables rather than relying solely on stylized heuristics (see, e.g., Sobkowicz, 2009), and they highlight justice-focused constructs, particularly empathy and moral indignation, as promising levers for bridging micro-level measurement and macro-level dynamics.

Limitations

Our choice of traits is necessarily selective. A more *global* set (e.g., Big Five) or a more *justice-centric* set (e.g., belief in a just world, justice centrality) might explain additional variance in c and, especially, m. The current model also leaves some empirical patterns unexplained, likely because it omits mechanisms such as explicit social identity, group norms, dynamic homophily, framing effects, or asymmetrical processing of confirmatory vs. disconfirmatory evidence. As a next step we plan to introduce social groups and identity-weighted influence, and to test interaction terms (e.g., indignation and group salience). Finally, modeling is a perennial trade-off between realism and parsimony. Naturally, the model could be made either more parsimonious or more realistic, but we deliberately aim for a pragmatic middle ground that preserves the parsimony of Hegselmann–Krause dynamics while adding real-world inter-agent heterogeneity by anchoring key parameters in justice related traits.

Conclusion

Classical averaging models become more informative when underpinned by domain knowledge from social science. Our findings identify where justice-related traits most reliably enter the opinion-updating pipeline: cognitive empathy and moral indignation act quickly on openness c, while broader trait influences on c likely unfold in settings that permit selective exposure and repeated contact; trait effects on m are already visible after one update. Empirically calibrated parameters thus move the model closer to observed polarization phenomena, even as they spotlight additional mechanisms needed to capture backfire. Taken together, this marks a step toward a pragmatic “sweet spot”—retaining the parsimony of Hegselmann–Krause dynamics while grounding key parameters in measurable social science constructs.

References

Alizadeh, M., Cioffi-Revilla, C., & Crooks, A. (2015). The effect of in-group favoritism on the collective behavior of individuals' opinions. *Advances in Complex Systems*, *18*(01n02), 1550002. https://doi.org/10.1142/S0219525915500022

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. G., 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*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Baumert, A., Adra, A., & Li, M. (2022). Justice sensitivity in intergroup contexts: A theoretical framework. *Social Justice Research*, *35*(1), 7–32. https://doi.org/10.1007/s11211-021-00378-9

Bernhardt, K. (2000). *Ein kognitives Trainingsprogramm zur Steuerung von Empörung* [Doctoral dissertation, Universität Trier]. https://web.archive.org/web/20200322104749id_/https://ubt.opus.hbz-nrw.de/opus45-ubtr/frontdoor/deliver/index/docId/12/file/20010214.pdf

Berntzen, L. E., Kelsall, H., & Harteveld, E. (2024). Consequences of affective polarization: Avoidance, intolerance and support for violence in the united kingdom and norway. *European Journal of Political Research*, *63*(3), 927–949. https://doi.org/10.1111/1475-6765.12623

Blondel, V. D., Hendrickx, J. M., & Tsitsiklis, J. N. (2009). On Krause's multi-agent consensus model with state-dependent connectivity. *IEEE Transactions on Automatic Control*, *54*(11), 2586–2597. https://doi.org/10.1109/TAC.2009.2031211

Bondü, R., Holl, A. K., Trommler, D., & Schmitt, M. J. (2022). Responses toward injustice shaped by justice sensitivity–evidence from Germany. *Frontiers in Psychology*, *13*, 858291. https://doi.org/10.3389/fpsyg.2022.858291

Bowes, S. M., & Tasimi, A. (2025). How intellectual humility relates to political and religious polarization. *The Journal of Positive Psychology*, *20*(4), 569–581. https://doi.org/10.1080/17439760.2024.2394447

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

Budner, S. (1962). Intolerance of ambiguity as a personality variable. *Journal of Personality*, *30*(1), 29–50. https://doi.org/10.1111/j.1467-6494.1962.tb02303.x

Butrus, N., & Witenberg, R. T. (2013). Some personality predictors of tolerance to human diversity: The roles of openness, agreeableness, and empathy. *Australian Psychologist*, *48*(4), 290–298. https://doi.org/10.1111/j.1742-9544.2012.00081.x

Cakanlar, A. (2024). Breaking climate change polarization. *Journal of Public Policy & Marketing*, *43*(4), 276–294. https://doi.org/10.1177/07439156241244737

Chattoe-Brown, E. (2023). All the right moves? Systematically exploring the effects of random movement in agent-based models. *Conference of the European Social Simulation Association*, 553–565. https://doi.org/10.1007/978-3-031-34920-1_44

Cole, J. C., Gillis, A. J., van der Linden, S., Cohen, M. A., & Vandenbergh, M. P. (2025). Social psychological perspectives on political polarization: Insights and implications for climate change. *Perspectives on Psychological Science*, *20*(1), 115–141. https://doi.org/10.1177/17456916231186409

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771. https://doi.org/10.1038/s41562-017-0213-3

Dagnes, A. (2019). Us vs. them: Political polarization and the politicization of everything. In *Super mad at everything all the time: Political media and our national anger* (pp. 119–165). Springer. https://doi.org/10.1007/978-3-030-06131-9_4

Dalton, R. J. (2008). The quantity and the quality of party systems: Party system polarization, its measurement, and its consequences. *Comparative Political Studies*, *41*(7), 899–920. https://doi.org/10.1177/0010414008315860

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113. https://doi.org/10.1037/0022-3514.44.1.113

Devia, C. A., Giordano, G., et al. (2023). Classification-based opinion formation model embedding agents' psychological traits. *JASSS*, *26*(3). https://doi.org/10.18564/jasss.5058

DiMaggio, P., Evans, J., & Bryson, B. (1996). Have American's social attitudes become more polarized? *American Journal of Sociology*, *102*(3), 690–755. https://doi.org/10.1086/230995

Esteban, J., & Schneider, G. (2008). Polarization and conflict: Theoretical and empirical issues. *Journal of Peace Research*, *45*(2), 131–141. https://doi.org/10.1177/0022343307087168

Franceschetti, M., Herpson, C., & Kant, J.-D. (2022). How beliefs on food and climate change impact the dietary adoption? An agent-based approach. *Conference of the European Social Simulation Association*, 499–510. https://doi.org/10.1007/978-3-031-34920-1_40

Furnham, A., & Marks, J. (2013). Tolerance of ambiguity: A review of the recent literature. *Psychology*, *4*(09), 717–728. https://doi.org/10.4236/psych.2013.49102

Gesellschaft für integrierte Kommunikationsforschung (GIK). (2022). Best for planning: Germany (synopsis). https://fipp.s3.amazonaws.com/media/documents/best%20for%20planning%20Germany%20synopsis.pdf

Gollwitzer, M., & van Prooijen, J.-W. (2016). Psychology of justice. In C. Sabbagh & M. Schmitt (Eds.), *Handbook of social justice theory and research* (pp. 61–82). Springer. https://doi.org/10.1007/978-1-4939-3216-0_4

Grimm, V., Railsback, S. F., Vincenot, C. E., Berger, U., Gallagher, C., DeAngelis, D. L., Edmonds, B., Ge, J., Giske, J., Groeneveld, J., et al. (2020). The ODD protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism. *JASSS*, *23*(2). https://doi.org/10.18564/jasss.4259

Haidt, J., et al. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). Oxford: Oxford University Press.

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence: Models, analysis and simulation [Article 2]. *JASSS*, *5*(3). https://www.jasss.org/5/3/2.html

Heynen, E., Van der Helm, G., Stams, G., & Korebrits, A. (2016). Measuring empathy in a German youth prison: A validation of the German version of the Basic Empathy Scale (BES) in a sample of incarcerated juvenile offenders. *Journal of Forensic Psychology Practice*, *16*(5), 336–346. https://doi.org/10.1080/15228932.2016.1219217

Hoerner, J. M., & Hobolt, S. B. (2020). Unity in diversity? Polarization, issue diversity and satisfaction with democracy. *Journal of European Public Policy*, *27*(12), 1838–1857. https://doi.org/10.1080/13501763.2019.1699592

Horned, A., & Vanhée, L. (2023). From threatening pasts to hopeful futures. A review of agent-based models of anxiety. *Conference of the European Social Simulation Association*, 139–152. https://doi.org/10.1007/978-3-031-34920-1_12

Jahnke, S., Schröder, C. P., Goede, L.-R., Lehmann, L., Hauff, L., & Beelmann, A. (2020). Observer sensitivity and early radicalization to violence among young people in Germany. *Social Justice Research*, *33*(3), 308–330. https://doi.org/10.1007/s11211-020-00351-y

Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the Basic Empathy Scale. *Journal of Adolescence*, *29*(4), 589–611. https://doi.org/10.1016/j.adolescence.2005.08.010

Kalvas, F., Ramaswamy, A., & Slater, M. D. (2022). Identity drives polarization: Advancing the Hegselmann-Krause Model by identity groups. *Conference of the European Social Simulation Association*, 249–262. https://doi.org/10.1007/978-3-031-34920-1_20

Kish Bar-On, K., Dimant, E., Lelkes, Y., & Rand, D. G. (2024). Unraveling polarization: Insights into individual and collective dynamics. *PNAS nexus*, *3*(10), 426. https://doi.org/10.1093/pnasnexus/pgae426

Kreiss, D., & McGregor, S. C. (2024). A review and provocation: On polarization and platforms. *New Media & Society*, *26*(1), 556–579. https://doi.org/10.1177/14614448231161880

Litman, J. A. (2010). Relationships between measures of I-and D-type curiosity, ambiguity tolerance, and need for closure: An initial test of the wanting-liking model of information-seeking. *Personality and Individual Differences*, *48*(4), 397–402. https://doi.org/10.1016/j.paid.2009.11.005

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

Lorenz, J. (2007). Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C*, *18*(12), 1819–1838. https://doi.org/10.1142/S0129183107011789

Lou, Q., & Xu, W. (2025). Personality modeling for persuasion of misinformation using ai agent. *arXiv preprint arXiv:2501.08985*. https://arxiv.org/pdf/2501.08985

Lück, J., & Nardi, C. (2019). Incivility in user comments on online news articles: Investigating the role of opinion dissonance for the effects of incivility on attitudes, emotions and the willingness to participate. *SCM Studies in Communication and Media*, *8*(3), 311–337. https://doi.org/10.5771/2192-4007-2019-3-311

Luzsa, R., & Mayr, S. (2019). Links between users' online social network homogeneity, ambiguity tolerance, and estimated public support for own opinions. *Cyberpsychology, Behavior, and Social Networking*, *22*(5), 325–329. https://doi.org/10.1089/cyber.2018.0550

Maes, J., Schmitt, M., & Schmal, A. (1996). Gerechtigkeit als innerdeutsches Problem: Machiavellismus, Dogmatismus, Ambiguitätstoleranz, Toleranz und Autoritarismus als Kovariate[justice as an internal German problem: Machiavellianism, dogmatism, tolerance of ambiguity, tolerance, and authoritarianism as covariates]. *Berichte aus der Arbeitsgruppe Verantwortung, Gerechtigkeit, Moral*, (98).

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, *12*(2), 125–135. https://doi.org/10.1037/h0027568

Muhammad, R. F., & Kasahara, S. (2024). Agent-based simulation of fake news dissemination: The role of trust assessment and big five personality traits on news spreading. *Social Network Analysis and Mining*, *14*(1), 75. https://doi.org/10.1007/s13278-024-01235-8

Muradova, L. (2021). Seeing the other side? Perspective-taking and reflective political judgements in interpersonal deliberation. *Political Studies*, *69*(3), 644–664. https://doi.org/10.1177/0032321720916605

Neuhäuser, C. (2023). Meinungsfreiheit und Moralismus [Freedom of speech and moralism]. *Deutsche Zeitschrift für Philosophie*, *71*(4), 510–537. https://doi.org/10.1515/dzph-2023-0041

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Oliva Delgado, A., Antolín Suárez, L., Pertegal Vega, M. Á., Ríos Bermúdez, M., Parra Jiménez, Á., Hernando Gómez, Á., & Reina Flores, M. d. C. (2011). Instrumentos para la evaluación de la salud mental y el desarrollo positivo adolescente y los activos que lo promueven [instruments for the evaluation of mental health and positive adolescent development and the assets that promote it]. *Consejería de Salud*. https://www.formajoven.org/AdminFJ/doc_recursos/201241812465364.pdf

Pereira, C., da Silva, R., & Rosa, C. (2025). How to measure political polarization in text-as-data? A scoping review of computational social science approaches. *Journal of Information Technology & Politics*, *22*(2), 172–185. https://doi.org/10.1080/19331681.2024.2318404

Peter, F., Donat, M., Umlauft, S., & Dalbert, C. (2013). Einführung in die Gerechtigkeitspsychologie [Introduction to justice psychology]. In C. Dalbert (Ed.), *Gerechtigkeit in der Schule* (pp. 11–32). Springer. https://doi.org/10.1007/978-3-531-93128-9_1

Plähn, J., Bellora-Bienengräber, L., Mertens, K. G., & Meyer, M. (2023). Combining experiments with agent-based modeling: Benefits for experimental management accounting research. *Conference of the European Social Simulation Association*, 371–382. https://doi.org/10.1007/978-3-031-34920-1_30

Renzini, F., Debernardi, C., Bianchi, F., Cremonini, M., & Squazzoni, F. (2023). The new frontiers of social simulation in the data science era: An introduction to the proceedings. *Conference of the European Social Simulation Association*, 1–10. https://doi.org/10.1007/978-3-031-34920-1_1

Salvi, C., Iannello, P., Cancer, A., Cooper, S. E., McClay, M., Dunsmoor, J. E., & Antonietti, A. (2023). Does social rigidity predict cognitive rigidity? Profiles of socio-cognitive polarization. *Psychological Research*, *87*(8), 2533–2547. https://doi.org/10.1007/s00426-023-01832-w

Saveski, M., Gillani, N., Yuan, A., Vijayaraghavan, P., & Roy, D. (2022). Perspective-taking to reduce affective polarization on social media. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*, 885–895. https://ojs.aaai.org/index.php/ICWSM/article/download/19343/19115

Schmitt, M., Baumert, A., Gollwitzer, M., & Maes, J. (2010). The justice sensitivity inventory: Factorial validity, location in the personality facet space, demographic pattern, and normative data. *Social Justice Research*, *23*, 211–238. https://doi.org/10.1007/s11211-010-0115-2

Schweighofer, S., Garcia, D., & Schweitzer, F. (2020). An agent-based model of multi-dimensional opinion dynamics and opinion alignment. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *30*(9). https://doi.org/10.1063/5.0007523

Secchi, D. (2023). A simple model of citation cartels: When self-interest strikes science. *Conference of the European Social Simulation Association*, 23–32. https://doi.org/10.1007/978-3-031-34920-1_3

Sgouros, N. M. (2023). Embedding social simulation in the design of wine pricing policies. *Conference of the European Social Simulation Association*, 397–406. https://doi.org/10.1007/978-3-031-34920-1_32

Smaldino, P. (2023). Modeling social behavior: Mathematical and agent-based models of social dynamics and cultural evolution.

Sobkowicz, P. (2009). Modelling opinion formation with physics tools: Call for closer link with reality. *JASSS*, *12*(1), 11–25. https://www.jasss.org/12/1/11.html

Sobkowicz, P. (2018). Opinion dynamics model based on cognitive biases of complex agents. *JASSS*, *21*(4). https://doi.org/10.18564/jasss.3867

Squazzoni, F. (2023). *Advances in social simulation*. Springer.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Wagner, M. (2024). Affective polarization in Europe. *European Political Science Review*, *16*(3), 378–392. https://doi.org/10.1017/S1755773923000383

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Weinstein, N., & Itzchakov, G. (2025). Empathic listening satisfies speakers' psychological needs and well-being, but doesn't directly deepen solitude experiences: A registered report. *Journal of Experimental Social Psychology*, *117*, 104716. https://doi.org/10.1016/j.jesp.2024.104716

Table 1

Sociodemographic Characteristics of Empirical Study

Variable		Number	%	% in Population*
Gender				
	male	840	46.56	50.44
	female	9544	52.88	49.53
	other/prefer not to say	10	0.55	<1.00
Age(years)				
	< 30	332	18.40	20.68
	30 − 39	330	18.29	19.10
	40 − 49	328	18.18 ´	17.71
	50 − 59	443	24.55	23.53
	> 60	371	20.56	18.97
Education				
	low	496	27.49	26.51
	mid	588	32.59	32.69
	high	720	39.91	40.80
Net Household Income				
	< 1000€	362	20.06	29.88
	1000–1999€	750	41.57	31.81
	2000–2999€	443	24.56	23.94
	> 2999€	249	13.80	14.35
Employment				
	student	135	7.48	8.79
	employed	1295	71.78	72.06
	unemployed	213	11.81	7.25
	retired	161	8.92	11.91

note. * as reported by the b4p study, (Gesellschaft für integrierte Kommunikationsforschung (GIK), 2022)

Table 2

Result Regression Analyses

predicting confidence bound c			
	Estimate	p-Value	Expected Direction
(Intercept)	-1.926	$<$ .001	
Affective Empathy	0.009	.887	+
Cognitive Empathy	0.301*	$<$.001	+
Ambiguity Tolerance	-0.465	$<$.001	+
Justice Sensitivity (Observer)	-0.047	.625	+
Justice Sensitivity (Victim)	0.240	.009	0
Moral Indignation	-0.560*	$<$ .001	−
predicting self-influence m			
(Intercept)	-0.198	$<$ .001	
Affective Empathy	0.000*	.977	0
Cognitive Empathy	-0.016*	.182	0
Ambiguity Tolerance	0.011*	.593	0
Justice Sensitivity (Observer)	-0.014*	.344	0
Justice Sensitivity (Victim)	0.044*	.002	+
Moral Indignation	-0.033*	.138	0

note. Weights that are in the predicted direction are marked with an *

Table 3

Class Attributes

Attribute	**Label**	**Type**	Values
opinion	o	vector{float}	$X \sim N(5,2)$
cognitive empathy	ce	float	$X \sim N(0.38, 0.35)$
ambiguity tolerance	a	float	$X \sim ZOIB(0.05, 2.28, p[0] < 0.001, p[1] = 0.004), X \in [-1, 1]$
justice sensitivity (Victim)	vjs	float	$X \sim N(0.37, 0.34)$
Disposition for moral indignation	mi	float	$X \sim N(-0.10, 0.19)$
confidence bound	c	float	$\log \mathbb{E}[c \mid \mathbf{X}] = -1.926 + 0.301\, ce - 0.465\, a + 0.240\, vjs - 0.560\, mi$
self-influence	m	float	$\log \mathbb{E}[m \mid \mathbf{X}] = -0.198 + 0.044\, vjs$

Table 4
Additional Parameters and Their Values in Each Model Variation

Variable	Definition	standard HK	Bail et al (2018)	Taber & Lodge (2006)	Lord et al. (1979)	Moscovici & Zavalloni (1969)
w_+	weight for opinion in same direction as self	1.0	1.0	2.0	1.6	1.0
w_+	weight for opinion in different direction as self	1.0	1.0	0.0	0.6	1.0
κ	selective exposure: sample a subset biased to similarity	0.0	0.0	4.0	0.0	0.0
ρ	repulsion step size	0.0	0.1	0.0	0.0	0.0
$\delta_{repulse}$	disagreement for repulsion	10	6	10	10	10
α	extremizing drift size for extremizing drift away from mid-point	0.0	0.0	0.2	0.0	0.08
g_{in}	ingroup neighbor weight multiplier	1.0	1.2	1.0	1.0	1.8
g_{out}	outgroup neighbor weight multiplier	1.0	1.0	1.0	1.0	0.8

note. HK = Hegselmann-Krause model

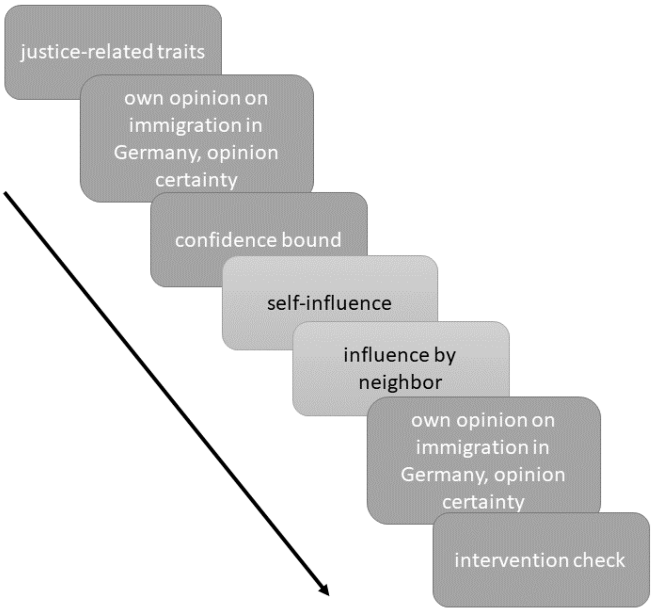
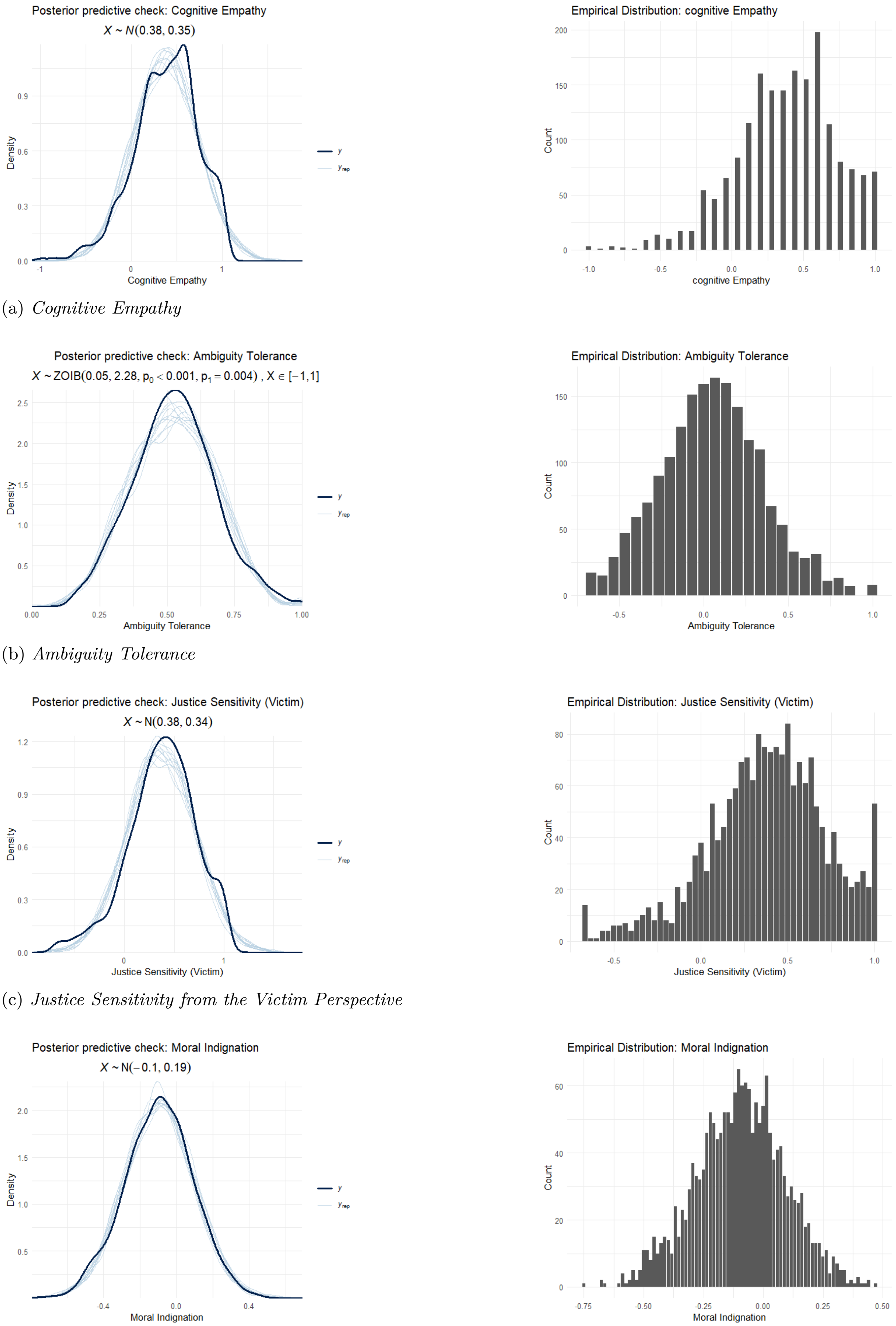

justice-related traits
own opinion on immigration in Germany, opinion certainty
confidence bound
self-influence
influence by neighbor
own opinion on immigration in Germany, opinion certainty
intervention check

Figure 1
Experiment Procedure



(a) *Cognitive Empathy*

(b) *Ambiguity Tolerance*

(c) *Justice Sensitivity from the Victim Perspective*

(d) *Moral Indignation*

Figure 2
Posterior predictive checks and empirical frequency distributions

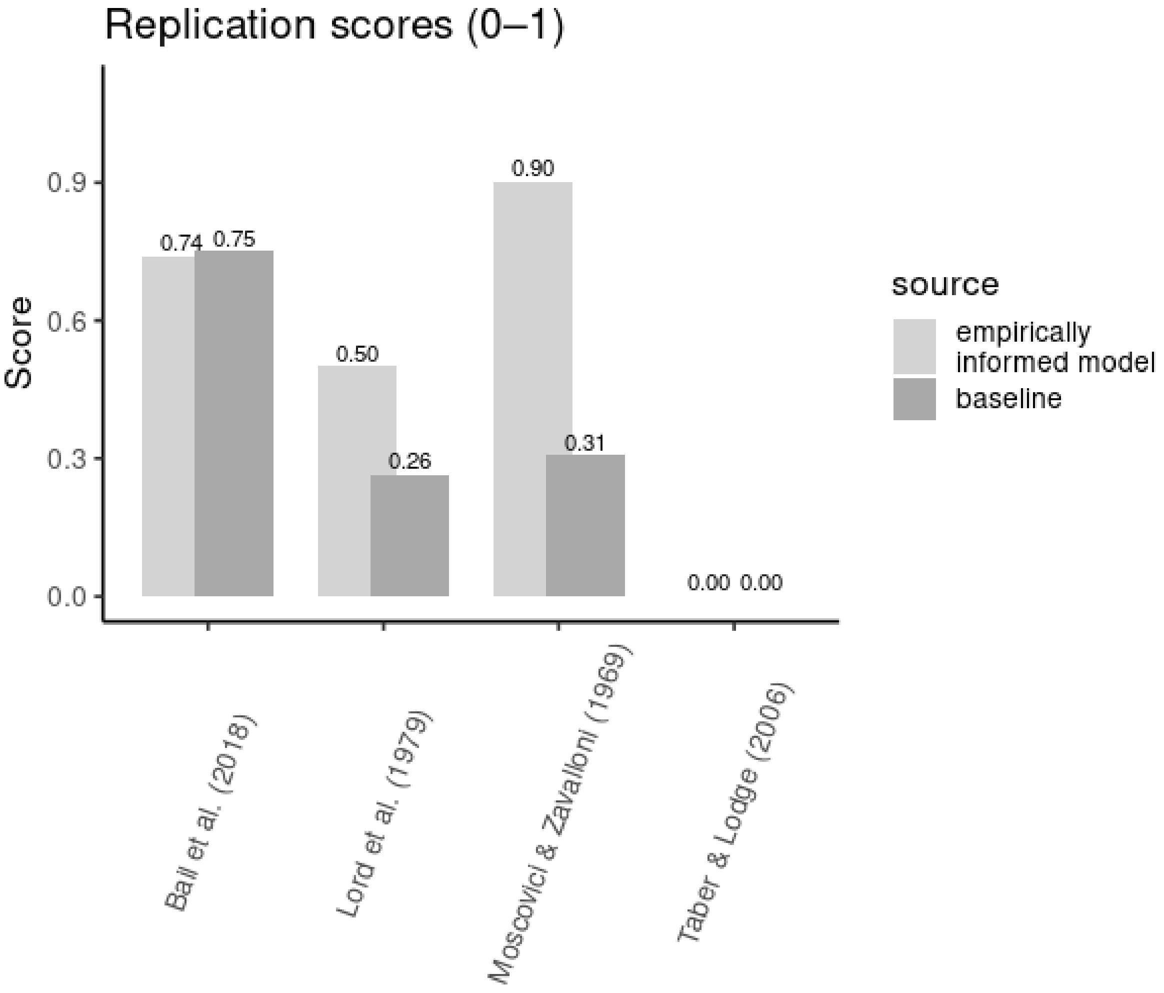


Figure 3
Replication Scores

Appendix

Model Specification Details

ODD Protocol

Purpose

We test whether a classic bounded–confidence ABM can be improved by *empirically informing* two key parameters—confidence bound c (openness to differing opinions) and self-influence m (weight on one's own prior)—with *justice-related personality traits* (empathy, ambiguity tolerance, justice sensitivity, moral indignation). We also ask whether this empirically tuned model better reproduces empirical polarization phenomena from *Bail et al.* (2018), *Taber & Lodge* (2006), *Lord et al.* (1979), and *Moscovici & Zavalloni* (1969), relative to a baseline ("null") model.

Entities, state variables, and scales

- **Entities:** Agents (people). Some scenarios include **bot / evidence nodes** with fixed opinions.
- **State (per agent i):** scalar opinion $x_i \in [1, 10]$; optional group label $g_i \in \{0, 1\}$ (assigned only in Bail and Moscovici scenarios); confidence bound $c_i > 0$; self-influence $m_i \in [0, 1]$; justice-related traits $\{\mathrm{Emp}_i, \mathrm{AT}_i, \mathrm{JS}_i^{(\mathrm{victim})}, \mathrm{MI}_i\}$; and, where applicable, a `treated` flag.
- **Interaction topology:** Watts–Strogatz small-world network with $N = 1000$, neighborhood $k = 2$, rewiring $p = 0.2$ (seed $= 42$).
- **Temporal scale:** discrete ticks; at each tick agents update *synchronously* on one random topic (the code samples a component of a topic vector; here we write x_i for the active component).

Process overview and scheduling

At initialization the network is created (or read), agents are loaded, traits are drawn from empirical posteriors, and (c_i, m_i) are set via trait–parameter regressions. The scheduler runs each tick:

1. `log_agents` (ticks $0, 1, 2, \dots$);
2. `step` (opinion update; once per tick);
3. `check_criterium` (early stop if mean change $<$ threshold);
4. `at_end` on termination.

Updates are **synchronous**: before updating, each agent stores x_i^{old}.

Design concepts (short)

- **Basic principles:** Hegselmann–Krause bounded confidence with psychologically motivated modifiers: biased assimilation, selective exposure, repulsion, extremizing drift, and group weighting.

- **Emergence:** opinion clustering, extremization, and backfire patterns from local rules.
- **Adaptation / Objectives / Learning / Prediction:** none; agents are reactive (no learning or optimization).
- **Sensing:** an agent samples neighbors' previous opinions x_j^{old} and whether they are on the same side of the midpoint 5 (congenial vs. uncongenial).
- **Interaction:** network neighbors (plus bot / evidence nodes in some scenarios).
- **Stochasticity:** initial traits and opinions, Watts–Strogatz rewiring, random topic choice, selective-exposure sampling, and (in Bail) random treatment assignment.
- **Collectives:** fixed groups g_i (assigned only at scenario setup; not updated over time).
- **Observation:** at each tick, log $\{x_i\}$, traits, and scenario flags; downstream analysis computes replication scores per study.

Initialization

- **Opinions:** $x_i^{(0)} \sim \text{clip}\big(\mathcal{N}(5, 2^2),\ 1, 10\big)$.
- **Traits:** drawn from fitted posteriors (files provided; cf. posterior predictive checks).
- **Trait → (c, m) mappings (log-link; deterministic in code):**

$$\log c_i \;=\; 0.7096 + 0.251374\,\text{Emp}_i + 0.023577\,\text{AT}_i + 0.0146663\,\text{JS}_i^{(\text{victim})} - 0.317839\,\text{MI}_i, \tag{A1}$$

$$\log m_i \;=\; -0.2247 + 0.0355525\,\text{JS}_i^{(\text{victim})}. \tag{A2}$$

We set $c_i = \exp(\cdot) > 0$ and $m_i = \min\{1, \max\{0, \exp(\cdot)\}\}$.

Input data

A YAML file per scenario provides global parameters; pickled objects store posterior trait distributions; the network is generated via Watts–Strogatz or read from file.

Submodels (equations)

Let $d_{ij} = |x_j^{\text{old}} - x_i^{\text{old}}|$ on the active topic and $\varepsilon_i = c_i$. Define the logistic $\sigma(z) = \frac{1}{1+e^{-z}}$.

1. **Acceptance (bounded confidence):**

$$j \in N_i \iff d_{ij} \le \varepsilon_i.$$

2. **Distance kernel (similarity weighting; $\beta > 0$):**

$$\mathrm{base}_{ij} = \exp\left[-\left(\frac{d_{ij}}{\varepsilon_i}\right)^{\beta}\right].$$

3. **Congenial vs. uncongenial (same side of midpoint 5):**

$$\mathrm{cong}_{ij} \iff \mathrm{sign}(x_j^{\mathrm{old}} - 5) = \mathrm{sign}(x_i^{\mathrm{old}} - 5).$$

4. **Confirmation-bias multiplier:**

$$\mathrm{cb}_{ij} = \begin{cases} w_+ & \text{if } \mathrm{cong}_{ij}, \\ w_- & \text{otherwise.} \end{cases}$$

5. **Ingroup / outgroup weight:**

$$\mathrm{gb}_{ij} = \begin{cases} g_{\mathrm{in}} & \text{if } g_j = g_i, \\ g_{\mathrm{out}} & \text{if } g_j \neq g_i. \end{cases}$$

6. **Final neighbor weight:** $w_{ij} = \mathrm{base}_{ij}\,\mathrm{cb}_{ij}\,\mathrm{gb}_{ij}$.

7. **Selective exposure (optional, $\kappa_{\exp} \geq 0$):** sample a subset of accepted neighbors with probabilities

$$\Pr(j \in S_i) \;\propto\; \exp\bigl(\kappa_{\exp}(1 - d_{ij}/\varepsilon_i)\bigr),$$

taking $k = \max\{1, \lfloor 0.6\,|N_i| \rceil\}$ neighbors if $N_i \neq \varnothing$; otherwise $S_i = \varnothing$.

8. **Neighbor aggregate (weighted mean):**

$$\bar{x}_{N_i} = \begin{cases} x_i^{\mathrm{old}}, & S_i = \varnothing, \\ \dfrac{\sum_{j \in S_i} w_{ij}\, x_j^{\mathrm{old}}}{\sum_{j \in S_i} w_{ij}}, & \text{otherwise.} \end{cases}$$

9. **Repulsion (backfire) if any $d_{ij} > \delta_{\mathrm{repulse}}$:** letting x_f^{old} be the farthest neighbor,

$$x_i^{\mathrm{rep}} = x_i^{\mathrm{old}} + \rho\, \mathrm{sign}\bigl(x_i^{\mathrm{old}} - x_f^{\mathrm{old}}\bigr)\,\bigl|x_i^{\mathrm{old}} - x_f^{\mathrm{old}}\bigr|.$$

10. **HK update with self-influence m_i:**

$$x_i^{\mathrm{hk}} = m_i\, x_i^{\mathrm{old}} + (1 - m_i)\, \bar{x}_{N_i}.$$

11. **Combine (repulsion takes precedence):**

$$x_i^* = \begin{cases} x_i^{\mathrm{rep}}, & \text{if repulsion triggered,} \\ x_i^{\mathrm{hk}}, & \text{otherwise.} \end{cases}$$

12. **Extremizing drift (optional, $\alpha_{\mathrm{ext}} \geq 0$):**

$$x_i^{\mathrm{drift}} \;=\; 5 + \big(1 + \alpha_{\mathrm{ext}}\big)\big(x_i^* - 5\big).$$

13. **Clipping and commit:**

$$x_i^{t+1} \;=\; \min\{10, \max\{1,\, x_i^{\mathrm{drift}}\}\}.$$

14. **Stopping rule (average movement):**

$$\mathrm{crit} \;=\; \frac{1}{N}\sum_{i=1}^{N} \big|x_i^{t+1} - x_i^t\big|.$$

Stop if crit $< 10^{-3}$ or t hits the scenario maximum (500 in all but Bail; Bail runs 54 ticks).

Scenario variants

Bail (2018).

$$w_+ = w_- = 1, \quad \kappa_{\mathrm{exp}} = 0, \quad \rho = 0.10, \quad \delta_{\mathrm{repulse}} = 0.6, \quad \alpha_{\mathrm{ext}} = 0, \quad g_{\mathrm{in}} = 1.2,\; g_{\mathrm{out}} = 1.0.$$

Two bots at $x = 0$ and $x = 10$; $\approx 50\%$ of agents are randomly treated and cross-connected to the opposing bot. Groups via initial side of 5. Run $T = 54$ ticks.

Taber & Lodge (2006).

Base HK with psychological weights; no extra nodes (slope-based validation).

Lord et al. (1979).

Two evidence nodes at $\{-1, +1\}$ (fixed: $m = 1$, $c = 0$) fully connected to all agents.

Moscovici & Zavalloni (1969).

Assign groups by initial side of 5; otherwise base HK.

Null (baseline)

Same network and scheduling; trait effects on (c, m) are neutralized (fixed values) and psychological modifiers set to neutral unless required by the scenario.

Observation and study scores

At each tick we log states. Replication is quantified per study. Let $\epsilon > 0$ be small, and define the Gaussian kernel $G(z; s) = \exp\big(-(z/s)^2\big)$.

Bail (backfire, DiD).

Using group×treatment differences DiD and denoting the initially higher group as the *target*:

$$\text{score}_s = \tfrac{1}{2}\left[\sigma\!\left(\frac{\text{DiD}_{\text{target}}}{\text{SD}_{\text{target}} + \epsilon}\right) + G(\text{DiD}_{\text{other}};\ \text{SD}_{\text{other}} + \epsilon)\right].$$

Taber–Lodge (polarization slope).

If $\hat{\beta}$ is the OLS slope of $\text{extremism}_{\text{end}}$ on extremism_0 with $\text{SE}(\hat{\beta})$,

$$z = \frac{\hat{\beta} - 1}{\text{SE}(\hat{\beta}) + \epsilon}, \qquad \text{score}_s = \sigma(z).$$

Lord (directional updating).

With one-sample t-tests per side, t_{pro} (Proponents) and t_{opp} (Opponents),

$$\text{score}_s = \tfrac{1}{2}\big[\sigma(t_{\text{pro}}) + \sigma(-t_{\text{opp}})\big].$$

Moscovici–Zavalloni (shift to extremity).

With $\Delta|x - 5|$ and share p becoming more extreme,

$$\text{score}_s = \tfrac{1}{2}\left[\sigma\!\left(\frac{\mathbb{E}[\Delta|x-5|]}{\text{SD}[\Delta|x-5|] + \epsilon}\right) + \sigma\!\left(\frac{p - 0.5}{0.1}\right)\right].$$

Defaults & seeds

Unless overridden in YAML: $w_+ = w_- = 1$, $\kappa_{\text{exp}} = 0$, $g_{\text{in}} = g_{\text{out}} = 1$, $\rho = 0$, $\delta_{\text{repulse}} = 1.1$, $\alpha_{\text{ext}} = 0$, $\beta = 2$, max $t = 500$. Network seed = 42; other random draws use fixed seeds where noted.