

Political Polarization in Algorithmic Environments: A Hybrid Mathematical Model of Opinion Dynamics with Large Language Model-Augmented Agents

Dr. Hafid MRHIZOU

PhD in Legal and Political Sciences, Sidi Mohamed Ben Abdellah

University – Fez

Morocco

Abstract

Digital environments increasingly shape political opinion through recommender systems and large language models (**LLMs**). This study proposes a hybrid mathematical model in which polarization emerges from interactions among social influence, bounded confidence, biased assimilation, algorithmic personalization, and generative responses. The model distinguishes structural from effective influence networks and treats **LLM** output as a stochastic component requiring independent calibration. It supports comparative simulation, sensitivity analysis, and validation. The central argument is that personalization is not inherently polarizing, but may reinforce polarization when combined with selective exposure, narrow confidence bounds, and biased assimilation. **LLM** agents are therefore modeled as generative components rather than substitutes for human citizens.

Keywords: political polarization; opinion dynamics; large language models; algorithmic personalization; recommender systems.

Table of Principal Symbols and Parameters

Symbol	Meaning
N	Number of actors in the synthetic population
$x_i(t)$	Political position of actor i at time t
$A=[a_{ij}]$	Structural communication network or baseline social availability matrix
$W=[w_{ij}]$	Baseline matrix of social influence weights
$\mathcal{N}_i(t)$	Effective confidence neighborhood of actor i
$B(t)=[b_{ij}(t)]$	Time-varying effective influence network
θ_i	Confidence bound or maximum ideological distance compatible with influence
b_i	Biased-assimilation coefficient
$q_{ij}(t)$	Availability of source j 's content to user i 's recommender system
ρ	Intensity of algorithmic personalization
$R_i(t)$	Expected political position of algorithmically exposed content
$Y_i(t)$	Generated textual response of the agent
$L_i(t)$	Numerical stance extracted from the generated response
$\alpha_i, \beta_i, \gamma_i, \kappa_i$	Weights of social, algorithmic, linguistic, and external influence
λ_i	Total susceptibility of the actor's position to updating
P_D, P_E, P_N	The fragmentation and polarization of opinions, and the disconnection of the network.

Introduction

Digital platforms are no longer merely channels that transmit political discourse; they have become part of the infrastructure that determines what enters a user's field of political visibility in the first place. What appears on a home feed or in a recommendation list is not a neutral reflection of all available content, but the result of ranking and selection processes that draw on prior behavior, network structure, predicted interaction probabilities, and algorithmic estimates of content value. The visibility, repetition, and ordering of political content have therefore become part of the causal environment within which political attitudes are formed.

This transformation becomes more consequential with the rise of large language models. Algorithms that once primarily ranked human-produced discourse can now generate text, summarize it, reformulate it, and adapt it to different profiles and interaction contexts. The result is a composite environment that cannot be reduced to a direct relationship between a “user” and a “platform,” because political opinion may simultaneously be shaped by network structure, user-level selectivity, recommender-system selectivity, and stochastic generative outputs.

Models of opinion dynamics have long shown that attitudes can be represented as states that change over time under social influence, while bounded-confidence models add the important restriction that actors do not take all opinions into account, but primarily those that fall within an acceptable range of difference (Hegselmann & Krause, 2002, pp. 3–4). Yet network homophily alone does not fully explain persistent divergence. Dandekar, Goel, and Lee show that biased assimilation of new evidence constitutes an additional mechanism capable of altering the dynamics of opinion formation (Dandekar, Goel & Lee, 2013, p. 5791).

Exposure to opposing views also does not automatically produce moderation. In a field experiment on Twitter users, Republican participants became more conservative after structured exposure to liberal counter-attitudinal content, while the corresponding effect among Democrats was weaker (Bail et al., 2018, p. 9216). This finding underscores that the distinction between exposure and influence is not a minor conceptual detail: the effect of a message depends on its distance from an actor’s prior position, the identity of its source, and the way it is interpreted within a particular political context.

LLM-based simulation introduces a further methodological layer. Argyle and colleagues demonstrate that conditioning language models on different social and demographic profiles can reproduce some response patterns observed in human samples, which they conceptualize in terms of algorithmic fidelity rather than equivalence between the model and human cognition (Argyle et al., 2023, p. 337). Conversely, Bisbee and colleagues show that similarity in mean responses can conceal differences in variance and in relationships among variables, limiting the

validity of treating synthetic data as a direct substitute for human data (Bisbee et al., 2024, p. 401).

The central problem addressed by this study therefore lies in the gap between conventional opinion-dynamics models, which offer clarity and interpretable mechanisms but necessarily simplify contextual complexity, and **LLM**-augmented simulations, which permit richer linguistic responses but make calibration, validation, and causal attribution more difficult. This gap is compounded by the fact that digital interaction is not direct: recommender systems intervene before influence occurs by changing the probability that particular content will be seen.

Accordingly, the study asks: under what conditions, and through what interacting mechanisms, can algorithmic personalization and a generative linguistic component become part of a trajectory that amplifies or reduces political polarization? The central hypothesis is that polarization is an emergent property of nonlinear interaction among communication structure, tolerance for disagreement, biased assimilation, algorithmic selectivity in exposure, and the context-dependent response of an **LLM**-augmented agent. No single component is assumed to produce polarization on its own.

The study adopts a theoretical-modeling approach. It does not claim empirical results prior to implementing the simulation. Instead, it specifies a programmable mathematical model, defines its variables and mechanisms, establishes its initial conditions and measures, and sets out a validation strategy. Its contribution lies in explicitly separating the structural network from the effective influence network, actor position from content position, and algorithmic exposure from linguistic response, before integrating these levels into a single update equation that can be subjected to ablation and isolation tests. This caution is particularly important given recent reviews that identify validation as a central challenge in generative social simulation (Larooij & Törnberg, 2025).

Section I. Theoretical and Mathematical Foundations of Political Opinion Dynamics

The model proceeds from individual states to collective outcomes. It does not assume polarization at the outset. Instead, it specifies how each actor's position changes, who can influence whom, and what makes an existing social tie become an effective channel of political influence. Aggregate patterns are then allowed to emerge from repeated interaction over time.

1. Political Opinion as a Dynamic State

Assume a political community composed of N actors. For purposes of the baseline model, each actor's position is represented on a one-dimensional political continuum. The political position of actor i at time t is denoted by:

$$(1) \quad \mathbf{x}_i(t) \in [-1, 1], i = 1, \dots, N$$

Values close to -1 represent one pole of the political space, values close to $+1$ represent the opposing pole, and values near **zero** represent relatively centrist positions. These values are not assigned a fixed ideological meaning; the substantive interpretation of the two poles depends on the issue to which the model is applied. The purpose of the numerical representation is to measure distances between positions and track their evolution, not to reduce complex political identity to a single dimension.

The positions of all actors can be assembled into a state vector describing the political configuration of the system at time t :

$$(2) \quad \mathbf{x}(t) = [\mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_N(t)]^T$$

The vector $\mathbf{x}(t)$ provides a system-level snapshot of the distribution of opinions. When the system moves from t to $t+1$, one can assess whether positions converge, diverge, or form stable clusters. This representation belongs to the tradition of continuous-opinion models, in which states are updated over successive time steps (Hegselmann & Krause, 2002, p. 3).

2. Structural Network and Potential Social Influence

Political actors do not interact in a social vacuum. The model therefore distinguishes between the existence of a communication channel and the strength of influence transmitted through that channel. The baseline structural communication network is represented by an adjacency matrix $\mathbf{A}$, where a positive entry indicates the presence of a potential social or informational connection between actors:

$$(3) \quad \mathbf{A} = [\mathbf{a}_{ij}]_{N \times N}, \mathbf{a}_{ij} \in \{0, 1\}$$

Within the baseline simulation, $\mathbf{A}$ is treated as fixed unless a specific experiment is designed to study tie formation and dissolution. This decision is important because the model should not attribute structural change to a network for which no rewiring rule has been specified. What does change over time is the effective influence network, defined below.

We also define a social-weight matrix $\mathbf{W}$, where each element specifies the baseline weight that actor i assigns to actor j 's position:

$$(4) \quad \mathbf{W} = [\mathbf{w}_{ij}]_{N \times N}$$

$$(5) \quad w_{ij} \geq 0, \quad \sum_{j=1}^N w_{ij} = 1$$

Thus, $\mathbf{W}$ is a row-stochastic matrix. If $\mathbf{w}_{ij} = 0$, actor j 's position does not enter actor i 's baseline influence average, while $\mathbf{w}_{ii}$ denotes the self-weight assigned to the actor's own prior position. In the simplest linear specification, social updating can be written as:

$$(6) \quad x_i(t+1) = \sum_{j=1}^N w_{ij} x_j(t)$$

$$(7) \quad \mathbf{x}(t+1) = \mathbf{W} \mathbf{x}(t)$$

This formulation captures the weighted-averaging logic associated with DeGroot-type models, but it is insufficient for the present study because it assumes

that every tie with a positive weight remains capable of transmitting influence. In polarized environments, a social connection may persist while its capacity to transmit political influence declines as ideological distance increases. The model therefore moves from the structural network to an effective influence network conditioned by confidence.

3. Bounded Confidence and the Effective Influence Network

For actor i at time t , define the set of acceptable opinions as the actors whose positions fall within the confidence bound θ_i :

$$(8) \quad \mathcal{N}_i(t) = \{ j : |x_j(t) - x_i(t)| \leq \theta_i \}, \quad \theta_i > 0$$

Actors outside this set do not disappear from the social network; rather, their ties become ineffective for political updating during that round. The effective influence network $\mathbf{B}(t)$ is therefore defined by combining structural connectivity with the confidence condition:

$$(9) \quad b_{ij}(t) = a_{ij} \cdot \{ |x_j(t) - x_i(t)| \leq \theta_i \}$$

This formulation resolves an important methodological problem. The structural communication network A may remain fixed while the effective influence network $\mathbf{B}(t)$ fragments or reconnects as opinions change. Interaction alters ideological distances, and the resulting distances determine who can influence whom in the next round. This is an endogenous feedback process that does not require assuming that friendship or structural connectivity itself is rewired.

After inactive ties are excluded, the social signal received by actor i is computed as a weighted average of opinions within the confidence neighborhood:

$$(10) \quad S_i(t) = \frac{\sum_{j \in \mathcal{N}_i(t)} w_{ij} x_j(t)}{\sum_{j \in \mathcal{N}_i(t)} w_{ij}}$$

The denominator is assumed to be positive. In practice, this can be ensured by assigning a positive self-weight $w_{ii} > 0$, since an actor always lies within their own confidence neighborhood. As θ_i increases, a wider range of opinions can influence the actor; as it decreases, the influence neighborhood becomes more homogeneous. This is the central logic of bounded-confidence models (Hegselmann & Krause, 2002, p. 4).

This construction separates exposure from influence. Counter-attitudinal content may reach a user because it is available through the communication network or recommendation layer, yet it need not enter the opinion-update process. Thus, $\mathbf{A}$ defines the possibility of connection, whereas $\mathcal{N}_i(t)$ and $\mathbf{B}(t)$ determine whether that connection is politically effective.

4. Biased Assimilation and the Reweighting of Acceptable Opinions

Crossing the confidence threshold does not imply that all acceptable opinions receive equal weight. An actor may assign greater importance to positions closer to their prior view than to more distant positions, even when all of them remain within the confidence neighborhood. This idea corresponds to biased assimilation. Dandekar, Goel, and Lee show that network homophily alone does not necessarily explain increasing polarization, and that the way new evidence is processed can constitute an independent mechanism in the dynamics of divergence (Dandekar, Goel & Lee, 2013, p. 5791).

To operationalize this mechanism in the present model, define a coefficient $b_i \geq 0$ that discounts baseline social weights as ideological distance increases:

$$(11) \quad \tilde{w}_{ij}(t) = \frac{w_{ij} \exp[-b_i |x_i(t) - x_j(t)|]}{\sum_{k \in \mathcal{N}_i(t)} w_{ik} \exp[-b_i |x_i(t) - x_k(t)|]}$$

$$(12) \quad \sum_{j \in \mathcal{N}_i(t)} \tilde{w}_{ij}(t) = 1, \quad \tilde{w}_{ij}(t) = 0 \quad \text{if } j \notin \mathcal{N}_i(t)$$

When $b_i=0$, ideological distance no longer contributes to the redistribution of weights, and the model reduces to the normalized social weights within the confidence neighborhood. Higher values of b_i make influence decay more rapidly with distance. The exponential function is an operational specification proposed in this study rather than a literal reproduction of the mathematical formulation used by Dandekar and colleagues. Their work supports the theoretical role of biased assimilation; the exponential weighting function used here is an independent modeling choice.

The resulting adjusted social signal is:

$$(13) \quad S_i^*(t) = \sum_{j \in N_i(t)} \tilde{w}_{ij}(t) x_j(t)$$

The baseline social mechanism is now complete: the structural network determines who can potentially connect to whom, bounded confidence determines which ties remain effective channels of influence, and biased assimilation redistributes importance among the opinions that remain admissible. Keeping these stages separate makes each mechanism independently testable rather than subsuming them under a single, non-isolable concept.

5. Polarization as a Multidimensional Emergent Property

Repeated operation of the preceding mechanisms can generate system-level outcomes that cannot be reduced to any single actor's position. Polarization is not equivalent to the mere presence of actors at the extremes of the political spectrum. Opinions may diverge while networks remain connected, and social groups may become structurally separated without reaching maximal ideological extremity. The literature likewise distinguishes among dispersion, bimodality, ideological sorting, and intergroup hostility (DiMaggio, Evans & Bryson, 1996, p. 690; Iyengar, Sood & Lelkes, 2012, p. 405; Mason, 2015, p. 128).

Because the present model does not include an independent affective state, it does not claim to measure affective polarization. That dimension is treated as a possible extension requiring a separate emotional state variable and appropriate

validation data. The current specification focuses on three dimensions that can be measured directly from the model:

$$(14) \quad \mathbf{P}(t) = [\mathbf{P}_D(t), \mathbf{P}_E(t), \mathbf{P}_N(t)]^T$$

Here, $\mathbf{P}_D$ denotes opinion dispersion, $\mathbf{P}_E$ denotes extremity, and $\mathbf{P}_N$ denotes segregation within the effective influence network. These dimensions are not mechanically aggregated into a single index because differences in their temporal trajectories are themselves part of the phenomenon to be explained. This design makes it possible to test, for example, whether a mechanism increases network fragmentation without producing a comparable increase in ideological extremity.

Section II. Algorithmic Personalization and the Construction of the LLM-Augmented Agent

The digital environment introduces a layer between an actor and the actor's network: the recommender system. The mere existence of content in the network does not imply that the user will see it, just as exposure does not imply influence. The model therefore separates three stages: content availability, algorithmic ranking, and subsequent processing by the agent. This separation prevents the conflation of communication structure, exposure structure, and generative response.

1. The Hybrid Political Agent

The agent is a modeling unit that combines a dynamic political position, parameters governing selectivity in influence, and a memory that retains part of the interaction history. Its dynamic state at time t is defined as:

$$(15) \quad \mathbf{z}_i(t) = [\mathbf{x}_i(t), \boldsymbol{\theta}_i, \mathbf{b}_i, \mathbf{M}_i(t)]$$

Here, $\mathbf{x}_i(t)$ is the current political position, $\boldsymbol{\theta}_i$ is the confidence bound, $\mathbf{b}_i$ governs biased reweighting, and $\mathbf{M}_i(t)$ is the agent's memory. More stable characteristics—such as demographic attributes, identity strength, or social background—are stored in the profile $\mathbf{P}_i$ rather than in the dynamic state vector unless the design explicitly intends to update them over time.

This distinction matters because the agent is not identical to the **LLM**. The agent consists of a state, memory, parameters, and update rules, while the language model performs a specific generative function within that architecture. The logic is consistent with generative-agent approaches that link responses to memory, context, and interaction history rather than treating each message as completely independent (Park et al., 2023). Yet such organization does not make the agent equivalent to a human actor. The concept of algorithmic fidelity proposed by Argyle and colleagues concerns the reproduction of certain response patterns, not equivalence of underlying cognitive processes (Argyle et al., 2023, p. 337).

2. Content Position and Political Similarity

To avoid equating an actor's political position with the stance of every item of content produced by that actor, the model defines a political position for candidate content. If $\mathbf{c}_j(t)$ denotes the position of content produced by source j on the same political continuum, the baseline model may adopt the simplifying assumption:

$$(16) \quad \mathbf{c}_j(t) = \mathbf{x}_j(t)$$

This assumption is stated explicitly because it is a modeling simplification rather than a claim about political communication. Richer applications may allow $\mathbf{c}_j(t)$ to differ from $\mathbf{x}_j(t)$, or may define multiple posts for each source, with each post assigned its own political position.

Because both $\mathbf{x}_i(t)$ and $\mathbf{c}_j(t)$ belong to $[-1,1]$, the maximum possible distance between them is 2. Political distance can therefore be converted into a normalized similarity score:

$$(17) \quad \mathbf{s}_{ij}(t) = 1 - |\mathbf{x}_i(t) - \mathbf{c}_j(t)| / 2, \quad 0 \leq \mathbf{s}_{ij}(t) \leq 1$$

The value of $\mathbf{s}_{ij}$ approaches one when the content is close to the user's position and approaches **zero** when ideological distance is maximal. The study does not claim that real platforms use this exact formula. It is used as an experimental device that makes the degree of algorithmic preference for political similarity explicit and controllable.

3. Probability of Algorithmic Exposure

The model distinguishes the structural network A from content availability to the recommender system. Let $q_{ij}(t) \geq 0$ denote the baseline availability weight of source j 's content in user i 's candidate set. This allows availability to reflect social-network ties, search, suggested content, or other channels without equating it directly with the communication matrix.

The probability that content from source j is shown to user i is then defined through a **softmax** function that depends on political similarity and the personalization-intensity parameter ρ :

$$(18) \quad p_{ij}(t; \rho) = \frac{q_{ij}(t) \exp[\rho s_{ij}(t)]}{\sum_{k=1}^N q_{ik}(t) \exp[\rho s_{ik}(t)]}$$

The denominator is assumed to be positive, which requires at least one item to be available for recommendation. When $\rho=0$, the ranking function grants no additional preference to politically similar content within the candidate set. As ρ increases, differences in s_j exert greater influence on ranking. In the model, the algorithm therefore changes the distribution of exposure opportunities rather than directly changing political opinion.

This distinction is consistent with the analysis by Bakshy, Messing, and Adamic, who distinguish between content available through a user's network, content that passes through News Feed ranking, and content the user subsequently chooses to engage with. They show that all three stages contribute, to different degrees, to exposure to ideologically diverse content (Bakshy, Messing & Adamic, 2015, pp. 1130–1131).

The political environment produced by the recommender system can be summarized by the expected position of exposed content:

$$(19) \quad R_i(t) = \sum_{j=1}^N p_{ij}(t; \rho) c_j(t)$$

$\mathbf{R}_i(t)$ is an expected signal rather than necessarily an actual feed. In an implemented simulation, K items can be sampled from the distribution p_{ij} to form the realized exposure set $\mathbf{E}_i(t)$; K should either be fixed or treated as a separate experimental factor. This preserves the distinction between $\mathbf{S}_i^*(t)$, which represents social influence after bounded confidence and biased assimilation, and $\mathbf{R}_i(t)$, which captures the ideological tendency of the algorithmically displayed environment.

4. Linguistic Response and Stance Coding

Once the displayed content has been determined, the language model enters as a generative component. Because language generation is stochastic, the response is not represented as a deterministic function but as a draw from a probability distribution conditioned on the agent profile, memory, exposure set, and political context:

$$(20) \quad \mathbf{Y}_i(t) \sim p_{\theta}(\cdot \mid \mathbf{P}_i, \mathbf{M}_i(t), \mathbf{E}_i(t), \mathbf{C}(t))$$

$\mathbf{P}_i$ denotes the agent profile, $\mathbf{M}_i(t)$ the memory state, $\mathbf{E}_i(t)$ the realized exposure set, and $\mathbf{C}(t)$ the external political context, while θ denotes the language-model parameters and generation settings. This formulation prevents the generated text from being treated as a fixed consequence of identical inputs and makes model version and sampling parameters part of the reproducibility requirements.

Because $\mathbf{Y}_i(t)$ is open-ended text while the opinion model operates on a numerical state variable, the response must be converted into a numerical political stance through an independent coding function:

$$(21) \quad \mathbf{L}_i(t) = \phi(\mathbf{Y}_i(t)), \quad \mathbf{L}_i(t) \in [-1, 1]$$

The function ϕ should not be equated with sentiment analysis. A text may have a positive tone while opposing a political target, or a negative tone while supporting a policy direction. Stance-detection research distinguishes the position taken toward a target from general affective polarity (Mohammad et al., 2016, p. 31). The function ϕ therefore requires independent validation on human-labeled data, including appropriate accuracy metrics and, where relevant, inter-coder agreement.

This precaution is particularly important in light of the findings reported by Bisbee and colleagues: **LLM**-generated responses may match human data in average values while exhibiting lower variance and different relationships among variables (Bisbee et al., 2024, p. 401). $L_i(t)$ is therefore treated as a model-based measurement requiring calibration, not as the “true opinion” of a virtual citizen.

5. The Political-Opinion Update Equation

The central update equation integrates the sources of influence while keeping them separable and open to ablation. To maintain a direct relationship between overall susceptibility to updating and component weights, define:

$$(22) \quad \lambda_i = \alpha_i + \beta_i + \gamma_i + \kappa_i, \quad 0 \leq \lambda_i \leq 1$$

Here, α_i is the weight of social influence, β_i the weight of the algorithmic environment, γ_i the weight of the political stance extracted from the linguistic response, and κ_i the weight of the external political signal $X_i(t)$. The quantity $1-\lambda_i$ therefore represents inertia in the actor's previous position.

$$(23) \quad \begin{aligned} \mathbf{x}_i(t+1) = \Pi_{[-1,1]} [& (1-\lambda_i)\mathbf{x}_i(t) + \alpha_i \mathbf{S}_i^*(t) + \beta_i \mathbf{R}_i(t) + \gamma_i \mathbf{L}_i(t) \\ & + \kappa_i \mathbf{X}_i(t) + \boldsymbol{\varepsilon}_i(t)] \end{aligned}$$

$$(24) \quad \boldsymbol{\varepsilon}_i(t) \sim \mathbf{N}(\mathbf{0}, \sigma_{\varepsilon}^2)$$

The projection operator $\Pi_{[-1,1]}$ keeps the updated position within the predefined interval. The equation permits any pathway to be disabled independently. For example, setting $\beta_i=0$ isolates the model from algorithmic personalization, while $\gamma_i=0$ removes the linguistic component. $\mathbf{R}_i(t)$ and $\mathbf{L}_i(t)$ must also be interpreted as distinct pathways to avoid double-counting: the former captures properties of the displayed informational environment, whereas the latter captures the stance generated after the agent processes that environment. Under a “full mediation” specification, β_i can be set to **zero** so that only $\mathbf{L}_i$ transmits the algorithmic-exposure effect; under a “direct-plus-indirect effect” specification, both pathways can remain active and be evaluated through ablation tests.

Simulation cycle at time t : the agent's current state is read; the candidate content set is constructed; the recommendation distribution is computed; the realized exposure set is sampled; the textual response is generated; the response stance is coded; and the opinion state and memory are updated. Confidence neighborhoods and the effective influence network are then recomputed before the next round.

Section III. Experimental Design, Measurement of Polarization, and Validation

Mathematical equations do not become political knowledge merely by being specified. The value of the model depends on whether the experimental design can distinguish the effects of initial conditions from the effects of the mechanisms under study, whether results are sufficiently stable across independent runs, and whether model outputs can be compared with suitable external criteria. This section therefore defines the synthetic population, network structure, comparative model variants, polarization measures, and the strategy for sensitivity analysis and validation.

1. Synthetic Population and Initial Conditions

Multiple population sizes are tested so that results are not tied to one small synthetic system:

$$(25) \quad N \in \{100, 250, 500, 1000\}$$

In an initially non-polarized scenario, political positions are generated from a truncated normal distribution over $[-1, 1]$:

$$(26) \quad \mathbf{x}_i(0) \sim \text{TN}_{[-1,1]}(0, \sigma^2)$$

For a scenario with pre-existing division, a bimodal mixture is used:

$$(27) \quad \mathbf{x}_i(0) \sim \pi \text{TN}_{[-1,1]}(-\mu, \sigma^2) + (1-\pi) \text{TN}_{[-1,1]}(\mu, \sigma^2)$$

Separating these scenarios prevents polarization that was inserted into the initial conditions from being misattributed to the model itself. Network structure should likewise not be treated as a neutral assumption. Robustness tests should therefore

be run on more than one network family—random, small-world, and community-block networks—while holding mean degree as constant as practicable across comparisons.

$$(28) \quad \mathbf{A}(0) \sim \mathcal{G}_g(\eta_g), \quad g \in \{\mathbf{Random}, \mathbf{Small-World}, \mathbf{Block}\}$$

Equation (28) does not imply that the three network families are socially equivalent; rather, it treats network type as an experimental factor. $\mathbf{A}$ remains fixed within the baseline run, whereas $\mathbf{B}(t)$ changes as confidence neighborhoods change. If later work seeks to study tie formation and dissolution, a separate rewiring rule should be introduced rather than attributing unspecified change to $\mathbf{A}$.

2. Comparative Models and Isolation of Mechanisms

The contribution of each component is evaluated through nested model variants, each differing from the preceding specification by one identifiable layer:

$$(29) \quad \mathcal{M}_0 = \mathbf{Social\ Influence}$$

$$(30) \quad \mathcal{M}_1 = \mathcal{M}_0 + \mathbf{Bounded\ Confidence} + \mathbf{Biased\ Assimilation}$$

$$(31) \quad \mathcal{M}_2 = \mathcal{M}_1 + \mathbf{Algorithmic\ Recommendation}$$

$$(32) \quad \mathcal{M}_3 = \mathcal{M}_2 + \mathbf{LLM-Augmented\ Generation}$$

The difference between $\mathcal{M}_1$ and $\mathcal{M}_2$ estimates the added contribution of algorithmic mediation under the experimental conditions, whereas the difference between $\mathcal{M}_2$ and $\mathcal{M}_3$ tests whether introducing the linguistic component changes outcomes or merely adds complexity without systematic effect. Within $\mathcal{M}_1$, biased assimilation can be ablated by setting $\mathbf{b=0}$ while holding $\boldsymbol{\theta}$ fixed. Within $\mathcal{M}_3$, the full-mediation specification can likewise be compared with the direct-plus-indirect effect specification.

This design is consistent with purpose-based validation: greater model complexity is not evidence of model quality in itself. Each component should be justified by an additional explanatory or empirical contribution that can be tested (Sargent, 2013, pp. 12–13).

3. Measuring Polarization

Polarization is measured along three independent dimensions. First, compute the mean political position:

$$(33) \quad \bar{x}(t) = \frac{1}{N} \sum_{i=1}^N x_i(t)$$

Opinion dispersion around the mean is then measured as:

$$(34) \quad P_D(t) = \frac{1}{N} \sum_{i=1}^N [x_i(t) - \bar{x}(t)]^2$$

Extremity is measured by the mean absolute distance from the center of the ideological interval:

$$(35) \quad P_E(t) = \frac{1}{N} \sum_{i=1}^N |x_i(t)|$$

Network segregation is not measured using only the proportion of within-group ties, because that proportion depends on group sizes. The observed cross-group tie rate in the effective influence network $\mathbf{B}(t)$ is therefore computed first:

$$(36) \quad p_{\text{cross}}(t) = \frac{E_{\text{cross}}(t)}{E_{\text{total}}(t)}$$

The expected cross-group rate under random mixing is then computed from ideological group shares $\mathbf{q}_g(t)$:

$$(37) \quad p_{\text{cross}}^0(t) = 1 - \sum_g q_g^2(t)$$

The corrected segregation index is defined as:

$$(38) \quad P_N(t) = 1 - \frac{p_{\text{cross}}(t)}{p_{\text{cross}}^0(t)}$$

P_N is approximately **zero** when cross-group mixing is close to what would be expected under random mixing with the same group sizes, approaches one under complete segregation, and may become negative when cross-group ties exceed the random-mixing benchmark. Modularity or assortativity should also be reported as robustness checks rather than silently substituted for the primary measure. P_D , P_E , and P_N are not combined into a single score because divergence among these dimensions is itself an outcome of substantive interest.

4. Hypotheses and Sensitivity Analysis

The main simulation parameters are collected in an experimental vector that is distinct from θ , which denotes the language-model parameters:

$$(39) \quad \psi = (\rho, \theta, \mathbf{b}, \alpha, \beta, \gamma, \kappa, \sigma_\epsilon)$$

The first hypothesis is that the effect of algorithmic personalization on polarization need not be linear. The second is that the effect of ρ becomes stronger as θ decreases or $\mathbf{b}$ increases. The expected interaction between personalization and biased assimilation for polarization dimension q at final time T can be written as:

$$(40) \quad \frac{\partial^2 P_q(T)}{\partial \rho \partial \mathbf{b}} > 0, \quad q \in \{D, E, N\}$$

This condition is not assumed to hold over the entire parameter space. Rather, it defines a testable hypothesis over specified parameter ranges. No fixed directional hypothesis is imposed on γ , because the effect of the linguistic component depends

on the agent profile, exposure content, memory, generation settings, and stance-coding procedure.

Because the model is partly stochastic, each parameter configuration is run R times using independent random seeds. The mean and standard deviation of each polarization dimension are then computed:

$$(41) \quad \bar{P}_q(T) = \frac{1}{R} \sum_{r=1}^R P_q^{(r)}(T)$$

$$(42) \quad s_q = \sqrt{\frac{\sum_{r=1}^R [P_q^{(r)}(T) - \bar{P}_q(T)]^2}{R - 1}}$$

R should not be chosen arbitrarily. It should be increased until estimated means and uncertainty intervals stabilize or the Monte Carlo error falls below a prespecified threshold. This prevents a single simulation run from being treated as a general property of the system.

5. Operational Validation and Reproducibility

The design distinguishes among code verification, internal validation, and external validation. Code verification asks whether the implementation executes the equations as specified. Internal validation examines limiting cases, the stability of results, and sensitivity to parameters and random seeds. External validation compares model-generated patterns with independently observed indicators in relevant human or digital data.

The simulation literature emphasizes that model validity is relative to the purpose for which a model was built rather than to comprehensive correspondence with reality (Sargent, 2013, pp. 12–13). The problem becomes more demanding in generative models because of stochasticity, black-box architecture, cultural biases, and limited access to underlying mechanisms. Linguistic plausibility is therefore insufficient, and reproducing one empirical indicator does not establish general validity (Larooij & Törnberg, 2025).

The study by Tomašević and colleagues illustrates this logic. They compare independent **LLM**-agent simulation runs with data from an actual digital forum and identify similarities on some dimensions alongside systematic differences on others, showing why success on one property cannot justify claims of comprehensive validity (Tomašević et al., 2026, PDF p. 1). Validation criteria should therefore match the type of claim: opinion distributions for claims about attitudes, network structure for claims about segregation, and validated performance of ϕ for claims based on linguistic responses.

Reproducibility requires preserving the **LLM** version and access date, prompts, sampling settings and temperature, random seeds, memory rule, feed size K , recommender configuration, stance-coding function, parameter values, and the number of simulation rounds and independent runs. These details are part of the method rather than a technical appendix, because synthetic outputs can change under limited modifications to prompting or model configuration (Bisbee et al., 2024, p. 412).

Section IV. Political Implications and Algorithmic Governance

The value of the model extends beyond describing polarization. It can be used to examine the conditions under which personalization becomes part of a trajectory that reinforces closure or, alternatively, limits it, and to test alternative interventions in a controlled environment before considering their deployment in human systems. Such use remains exploratory: simulation can compare mechanisms and scenarios, but outputs from a synthetic society cannot be converted into general policy rules without independent validation.

1. Personalization and the Threshold Hypothesis

The study does not assume a linear relationship between personalization intensity and polarization. An increase in ρ may have little effect while the effective influence network remains relatively open, then produce a stronger response after the system crosses a level at which feedback among prior opinion, similar-content exposure, and subsequent updating intensifies. A critical-threshold hypothesis can be expressed operationally as:

$$(43) \quad \frac{\partial P_q(T)}{\partial \rho} \approx 0 \quad \text{if } \rho < \rho^*, \quad \frac{\partial \mathbb{E}[P_q(T)]}{\partial \rho} > 0 \quad \text{if } \rho > \rho^*$$

ρ^* is not a universal constant. It depends on θ , b , network structure, and the weights in the update equation. This framing avoids describing all personalization as polarizing by definition. Bojić and colleagues report that higher levels of personalization in their simulation are associated with unequal changes across network structure and several polarization-related indicators, consistent with the possibility that effects are configurational rather than uniform across dimensions (Bojić et al., 2025, pp. 21–22).

The political question therefore becomes the point at which personalization ceases to function merely as a ranking mechanism and begins to support a feedback loop: the current political position affects what is more likely to be shown; displayed content changes the informational environment; and that environment contributes to updating the position that will later re-enter the recommendation process.

2. Limits of Exposure to Opposing Views

Increasing exposure to counter-attitudinal content may appear to offer a direct remedy for informational bubbles, yet the empirical evidence does not support the assumption that such exposure automatically produces moderation. Bail and colleagues show that structured exposure to messages from the opposing political camp was associated, among participating Republicans, with movement toward more conservative positions rather than convergence (Bail et al., 2018, p. 9216).

The model therefore proposes testing gradual disagreement rather than moving directly to highly distant content. An experimental band can be defined immediately outside the actor's current confidence neighborhood:

$$(44) \quad \theta_i < |c_j(t) - x_i(t)| \leq \theta_i + \delta, \quad \delta > 0$$

Equation (44) does not assume that this band is necessarily preferable. Instead, it turns the broad idea of “diversity” into a testable experimental condition. Multiple values of δ can be compared with similarity-based recommendation and with

exposure to very distant content, while also examining whether the effect of disagreement varies with identity strength in the agent profile or with prior interaction history.

3. Testing Recommendation–Diversification Policies

The simulation can compare a baseline scenario I_0 with interventions I_k , such as lowering ρ , imposing a minimum share of dissimilar content, or modifying the candidate set q_{ij} to increase the opportunity for cross-cluster content to appear. The effect of intervention k is measured separately for each polarization dimension:

$$(45) \quad \Delta P_{q,k} = P_{q,I_k}(T) - P_{q,I_0}(T)$$

A negative value means that the intervention reduced dimension q relative to the baseline. It does not, however, establish that the intervention was successful overall: network segregation may decline while extremity remains unchanged, or polarization may decline at the cost of engagement. Bojić and colleagues provide an example of this intervention logic in simulation, where introducing a share of ideologically dissimilar content reduced network modularity without an obvious collapse in interaction (Bojić et al., 2025, p. 23). This finding should not be read as identifying an optimal proportion for real platforms, but as evidence that alternative design choices can be compared under controlled conditions.

Diversification interventions should also be examined through parallel ablation tests. If an intervention succeeds only when θ is high, then changes in display policy alone may be insufficient in environments where dissimilar opinions are rejected before they can enter the influence process. If the intervention fails when b is high, the core mechanism may lie in reweighting after exposure rather than in insufficient diversity of exposure itself.

4. Limits of LLM Agents for Political Inference

The ability of an **LLM** to produce coherent political discourse does not establish that it is an empirically valid representation of a citizen. Argyle and colleagues show that language models can reproduce some differences associated with social

backgrounds (Argyle et al., 2023, p. 337), whereas Bisbee and colleagues show that synthetic data can differ from human data in variance and in relationships among variables even when means are similar (Bisbee et al., 2024, p. 401).

The model therefore adopts a clear interpretive rule: an **LLM**-augmented agent is not a human representative but a contextual generative component within a formal model. Its role is to test whether adding a memory- and context-conditioned linguistic response changes system dynamics relative to an otherwise comparable model without that component. This preserves the value of **LLMs** as simulation tools without treating linguistic fluency as evidence of behavioral validity.

These limits are consistent with recent critical reviews arguing that **LLM** integration may intensify rather than solve validation challenges, because outputs are stochastic, sources of bias are opaque, and generative mechanisms are difficult to specify in the causal terms used in social science (Larooij & Törnberg, 2025).

5. From Personalization to Democratic Governance

The model highlights that algorithmic governance concerns not only the content of political messages, but also the rules that allocate attention, visibility, and opportunities for interaction. An algorithm need not impose a particular political position to have political consequences. It is sufficient for some messages and ties to become systematically more likely to appear, thereby altering the environment in which opinion formation takes place.

A recommender system should therefore not be evaluated solely by its capacity to maximize engagement. Engagement may rise while informational diversity declines or fragmentation increases. The resulting trade-off can be expressed in general terms without presupposing a ready-made normative welfare function:

$$(46) \quad \mathcal{W} = f(\text{Eng}, \text{Div}, \text{Pol})$$

Eng denotes engagement, **Div** informational diversity, and **Pol** the polarization measures. The equation does not imply that these objectives can be neutrally

collapsed into a single number; assigning weights among them is a normative and political decision. Simulation can show how outcomes change when one objective is given greater priority, but it cannot determine the optimal form of the democratic public sphere.

The political function of simulation is therefore to make trade-offs visible and testable. What happens to network segregation when personalization intensity is reduced? Does diversity increase at the cost of interaction? Do the same interventions work across different network structures? Do unintended effects emerge? Normative judgments about which outcomes are acceptable remain the responsibility of political institutions and society rather than of the computational model.

Conclusion

The study answers its central question by arguing that political polarization in algorithmic environments is not produced automatically by personalization, network structure, or large language models considered separately. When polarization emerges, it does so as an outcome of interaction over time among prior positions, the effective influence network, confidence bounds, biased reweighting, algorithmic selectivity in exposure, generative responses, and the external political context.

This answer implies that personalization should not be described as intrinsically polarizing. Its effect may be limited when the effective influence network remains relatively open and confidence bounds are wide, yet it may become part of a reinforcing feedback loop when narrow confidence bounds are combined with strong biased assimilation. In that case, prior positions affect which content is more likely to be displayed; displayed content changes the signal that reaches the actor; and the resulting position enters the next round of ranking and influence.

The principal methodological contribution lies in separating levels that are often conflated in explanations of digital polarization. The structural network is not the same as the effective influence network; an actor's political position is not necessarily identical to the stance of the content they produce; exposure is not equivalent to influence; and the displayed information environment is not identical to the response an agent generates after processing it. The linguistic component is likewise not treated as a human representative, but as a stochastic generative process that requires stance coding and external calibration.

This separation makes the model amenable to isolation and falsification. A system based solely on social influence can be compared with one that adds bounded confidence and biased assimilation; the effect of personalization can be measured before the linguistic component is introduced; and the model can then test whether adding an **LLM** provides substantive explanatory value. Measuring dispersion, extremity, and network segregation separately also reveals cases in which different

dimensions of polarization move in different directions rather than being collapsed into a single index.

The value of the model nevertheless remains conditional on validation. Linguistic realism does not establish behavioral realism, similarity in selected means is insufficient, and a single simulation run cannot support a general conclusion. The study therefore does not move from theoretical specification to empirical claims until the model is implemented, calibrated, repeatedly simulated, subjected to sensitivity analysis, externally validated, and documented in a form that permits reproduction.

At the level of governance, this framework shifts the discussion from the simplified question of whether algorithms “cause” polarization to the more precise question of the conditions under which algorithmic mediation becomes part of a trajectory that closes or opens the effective influence network. Simulation can be used to test alternative recommendation designs, gradual exposure strategies, and trade-offs among engagement, diversity, and polarization before interventions are considered in real human environments. It cannot, however, confer on the model the authority to choose the democratic values that should be prioritized.

The next stage of the project is therefore empirical and computational: implementing the model, specifying the initial network structures, calibrating ϕ and the generation parameters, running the comparative scenarios, and testing the hypotheses and robustness of the results. Only after this stage can the project move from a falsifiable theoretical framework to empirical findings that can be evaluated, criticized, and reproduced.

References

- Argyle, Lisa P.; Busby, Ethan C.; Fulda, Nancy; Gubler, Joshua R.; Rytting, Christopher; & Wingate, David. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3).
- Bail, Christopher A.; Argyle, Lisa P.; Brown, Taylor W.; Bumpus, John P.; Chen, Haohan; Hunzaker, M. B. Fallin; Lee, Jaemin; Mann, Marcus; Merhout, Friedolin; & Volfovsky, Alexander. (2018). Exposure to Opposing Views on Social Media Can Increase Political Polarization. *Proceedings of the National Academy of Sciences*, 115(37). <https://doi.org/10.1073/pnas.1804840115>
- Bakshy, Eytan; Messing, Solomon; & Adamic, Lada A. (2015). Exposure to Ideologically Diverse News and Opinion on Facebook. *Science*, 348(6239). <https://doi.org/10.1126/science.aaa1160>
- Bisbee, James; Clinton, Joshua D.; Dorff, Cassy; Kenkel, Brenton; & Larson, Jennifer M. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4).
- Bojić, Ljubiša; Ilić, Vladimir; Prodanović, Vladimir; & Vuković, Vuk. (2025). An Agent-Based Simulation of Politicized Topics Using Large Language Models: Algorithmic Personalization and Polarization on Social Media. *Chinese Political Science Review*. <https://doi.org/10.1007/s41111-025-00326-x>
- Cinelli, Matteo; De Francisci Morales, Gianmarco; Galeazzi, Alessandro; Quattrociocchi, Walter; & Starnini, Michele. (2021). The Echo Chamber Effect on Social Media. *Proceedings of the National Academy of Sciences*, 118(9). <https://doi.org/10.1073/pnas.2023301118>
- Dandekar, Pranav; Goel, Ashish; & Lee, David T. (2013). Biased Assimilation, Homophily, and the Dynamics of Polarization. *Proceedings of the National Academy of Sciences*, 110(15). <https://doi.org/10.1073/pnas.1217220110>
- DeGroot, Morris H. (1974). Reaching a Consensus. *Journal of the American Statistical Association*, 69(345).
- DiMaggio, Paul; Evans, John; & Bryson, Bethany. (1996). Have Americans' Social Attitudes Become More Polarized? *American Journal of Sociology*, 102(3).

- Hegselmann, Rainer; & Krause, Ulrich. (2002). Opinion Dynamics and Bounded Confidence Models, Analysis, and Simulation. *Journal of Artificial Societies and Social Simulation*, 5(3).
- Iyengar, Shanto; Sood, Gaurav; & Lelkes, Yphtach. (2012). Affect, Not Ideology: A Social Identity Perspective on Polarization. *Public Opinion Quarterly*, 76(3).
- Larooij, Maik; & Törnberg, Petter. (2025). Validation Is the Central Challenge for Generative Social Simulation: A Critical Review of LLMs in Agent-Based Modeling. *Artificial Intelligence Review*. <https://doi.org/10.1007/s10462-025-11412-6>
- Mason, Lilliana. (2015). “I Disrespectfully Agree”: The Differential Effects of Partisan Sorting on Social and Issue Polarization. *American Journal of Political Science*, 59(1).
- Mohammad, Saif M.; Kiritchenko, Svetlana; Sobhani, Parinaz; Zhu, Xiaodan; & Cherry, Colin. (2016). SemEval-2016 Task 6: Detecting Stance in Tweets. *Proceedings of the 10th International Workshop on Semantic Evaluation*. <https://aclanthology.org/S16-1003/>
- Park, Joon Sung; O’Brien, Joseph; Cai, Carrie Jun; Morris, Meredith Ringel; Liang, Percy; & Bernstein, Michael S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.
- Sargent, Robert G. (2013). Verification and Validation of Simulation Models. *Journal of Simulation*, 7(1).
- Tomašević, Aleksandar; Cvetković, Dimitrije; Major, S.; Maletić, Slobodan; et al. (2026). Towards Operational Validation of LLM-Agent Social Simulations: A Replicated Study of a Reddit-Like Technology Forum. *EPJ Data Science*. <https://doi.org/10.1140/epjds/s13688-026-00674-x>
- Törnberg, Petter. (2022). How Digital Media Drive Affective Polarization through Partisan Sorting. *Proceedings of the National Academy of Sciences*, 119(42). <https://doi.org/10.1073/pnas.2207159119>