Political Polarization and Fractionalisation from Rational Values-Based Inference in an Agent-Based Graph Network

Nicolas Navarre* **(nnavarre@ed.ac.uk)**
School of Informatics, University of Edinburgh
10 Crichton St, Edinburgh, EH8 9AB UK

Julie M. E. Pedersen* **(j.pedersen@ed.ac.uk)**
Department of Psychology, University of Edinburgh
7 George Square, Edinburgh, EH8 9JZ UK

Adam B. Moore (amoore23@ed.ac.uk)
Department of Psychology, University of Edinburgh
7 George Square, Edinburgh, EH8 9JZ UK

Abstract

The rise in political polarization disrupts political consensus and causes individual harm. We build on a theoretical framework of political polarization that emerges from uncertain political identity inference and signaling mediated by moral values. The current computational model extends this framework with rational inference tools and graph theory to better capture the complex dynamics of value-based inference and group formation. We find that minimally constrained signaling and promiscuous inference and updating of moral values leads to general network homogeneity. This contrasts with previous models using the same overarching theoretical framework and highlights the influence of model implementation, which should be further explored to triangulate the necessary causes of polarization. We discuss future extensions to the model to explore what facilitates political polarization as found in previous studies and the real world.

Keywords: Political polarization; Moral values; Agent-based modeling; Graph clustering; Rational inference

Introduction

Politics and voters around the world continue to become increasingly polarized (Boxell et al., 2023; Dimock et al., 2014). This poses a wide range of threats including increased rates of stochastic violence, lower efficacy for crisis management, and endorsement of authoritarian leaders. While theories on political polarization are promoting a social identity approach (Iyengar et al., 2012, 2019), computational modeling has remained largely focused on polarization of attitudes/beliefs (i.e., issue polarization) in single agents (Kvam et al., 2022) or networks (Baumann et al., 2020; Hahn et al., 2018; Olsson, 2013; Young et al., 2025), or groups of agents being influenced uni-directionally by media or other top-down inputs (Tokita et al., 2021). Pedersen and Moore, 2023 and Pedersen, 2024 (P&M hereafter) proposed a heuristics-driven agent-based model of political polarization grounded in social identity approaches and linking to belief change dynamics via moral values as an identity signaling and social inference mechanism. Here, we present a new model of political polarization grounded in the same theoretical framework but implemented in a graphical network model with information theoretical updating and inference principles. This approach allows us to begin to disentangle the relative influences of specific model implementation and the overarching theoretical framework proposed in previous work, which lead to polarization. It also naturally provides graph-based network analyses of clustering and ingroup formation dynamics, which may inform future research on measuring polarization in computational models.

Political Identity Polarization and Moral Values

While political and social scientists are largely regarding political polarization as a social phenomenon, computational models of polarization typically operate at the level of individual agents' beliefs and how they update in response to evidence (Jern et al., 2014; Kvam et al., 2022; Tokita et al., 2021) or other agents (Baumann et al., 2020). The former presupposes that identity processes precede and moderate beliefs while the latter often assumes the inverse. Empirical evidence largely supports identity-based political voter polarization. That is, an individual has a social identification with a political group (e.g., a political party or a political movement), which shapes their political attitudes/beliefs and how, or if, they are updated (Macy et al., 2019; Malka & Lelkes, 2010; Yudkin et al., 2019); e.g., attitudes are only polarized on policy issues with clear disagreement between the parties (Dias & Lelkes, 2022). In turn, increasing polarization is the result of increased tensions / hate between political groups (Baldassarri & Page, 2021; Iyengar et al., 2019) that causes, rather than being driven by, the divergence of beliefs.

A satisfactory model of social identity-based political polarization must capture, among other factors, ingroup homophily preferences: people prefer to seek out and interact with others who share their political identity group, i.e., the ingroup, over those that do not, i.e., the outgroup (Axelrod, 1997; Hogg & Reid, 2006; Hogg & Turner, 1985). Schelling (1971) showed that this homophily assumption leads to segregation, or in the context of political polarization, self-selected echo chambers (Theodoropoulos, 2022). However, most such models assume that agents possess perfect information about others' political identity, despite individuals seemingly more often signaling rather than disclosing their political identity (Settle & Carlson, 2019; Smaldino, 2022; van der Does et al., 2022). Consequently, people face a dual problem: they need to simultaneously generate signals that effectively demonstrate their social/political identity to other agents and decode other agents' signals to infer their ingroup status. But uncertainty arises as to what signals should be sent and what received signals mean. Resolving this requires agents to use their own values as a proxy for their in-group's true values and generate signals from these representations (Hogg & Turner, 1985). An additional problem arises due to the lack of

*Equal contribution.

grounding for political identities (Kinder & Kalmoe, 2017); an agent should update their internal in-group values representations as a function of the signals emitted from perceived ingroup members (i.e., those inferred to be from the ingroup based on their signals) to keep signals in line with other ingroup members' expectations (Axelrod, 1997) (see Pedersen (2024) for a full theoretical motivation and outline; see also Rosario et al. (2024) and Levine et al. (2023) for similar theoretical work).

P&M implemented a heuristics-based agent-based model of this process where moral values acted as the representation of ingroup norms, i.e., as the interface for generating signals (moral expressions) and inferring ingroup membership from others' signals, updated as a function ingroup members' signals. The use of moral values was supported by multiple empirical findings necessary to a candidate mechanism: moral values are distinct between political identities (see Figure 2) (Franks & Scherr, 2015; Graham et al., 2009, 2011, 2012; Milesi, 2016, 2017; Nilsson & Erlandsson, 2015; van Leeuwen & Park, 2009; Waytz et al., 2019; Womick et al., 2024), change over time (Hatemi et al., 2019; Smith et al., 2017) and in response to perceived ingroup moral values (Ciuk, 2018), and are used by both political leaders and voters to signal political groups (Atikcan & Hand, 2024; Bos & Minihold, 2022; Kidd & Vitriol, 2022), accelerating information spread within partisan social networks (Brady et al., 2017, 2019). In short, malleable moral values are a promising yet understudied pathway for modeling political polarization as a function of social identity motivations.

A new Model for Identity Inference and Signaling via Moral Values

In this study, we present a new model using tools from information and graph theory to model identity inference and signaling with moral values. Previous studies have applied these tools to model complex dynamics of belief updating with argument exchange and self-censoring (Assaad et al., 2023; Schöppl & Hahn, 2024). Thus, by using these tools, we aim to expand the theoretical framework of previous computational modeling work on identity inference and signaling via moral values. We now outline critical, theoretically important, differences between the current work and previous models of these processes.

P&M modeled agents in a grid space with latent political identities (liberal/conservative) and a set of internal weights representing the importance of different moral values. On each time step, agents engaged in three principal actions. 1) they attempted to infer the political identity of their neighbors in a binary fashion (ingroup/not ingroup) using a limited set of past signals. This prompts an update to their own moral values based on differences in their inferred in-group's and their own signals (i.e., social influence). 2) they generated new moral signals as a choice between two conflicting moral values resolved with the weights corresponding to the choice. 3) They evaluate and decide whether to alter their position in the grid based on the number of perceived ingroup members around them (i.e., homophily preference). In this formulation, echo chambers or clusters were conceptualized based on a heuristic: non-moving groups of agents with largely the same latent identity.

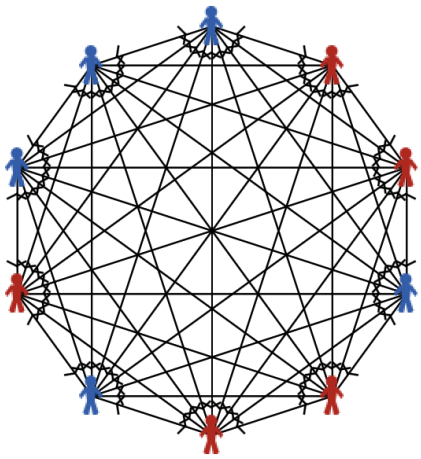

Figure 1: Example fully connected directed network of 10 agents.

Turning to the new model of moral signaling, we start with the inference of other agents' moral foundations. Agents in this model aim to infer the latent moral values of other agents rather than infer whether other agents should be in an ingroup. This allows us to use a Bayesian belief updating approach where moral values are represented as a distribution and signals are treated as evidence to update that distribution. Furthermore, the grid dynamics of the previous model are replaced by a graphical representation where each agent is connected to one another in a fully connected graph (cf. Figure 1).

This allows for the connections in the network to be weighted and mediated by the perceived similarity between agents, leaving room for graph theoretical clustering tools to analyze the formation of clusters and echo chambers. Moreover, this relaxes the assumption made by P&M that agents could only perceive a very limited number of other agents at any given time (see Discussion). Thus, inferring other agents' moral values as well as updating individual values within this network is also determined by the strength of connections in the graph which represents the degree of social influence between agents.

The Present Research

While the advance of political polarization and its consequences continue world-wide, recent computational modeling work and theoretical advances on the underlying psychological processes remain disconnected. In this paper, we implement the identity-based approach to political polarization via moral values, but deviating from the original work by using more common modeling tools in cognitive science: information theoretical belief updating and inference in a graph-based agent network. This captures the originally pro-

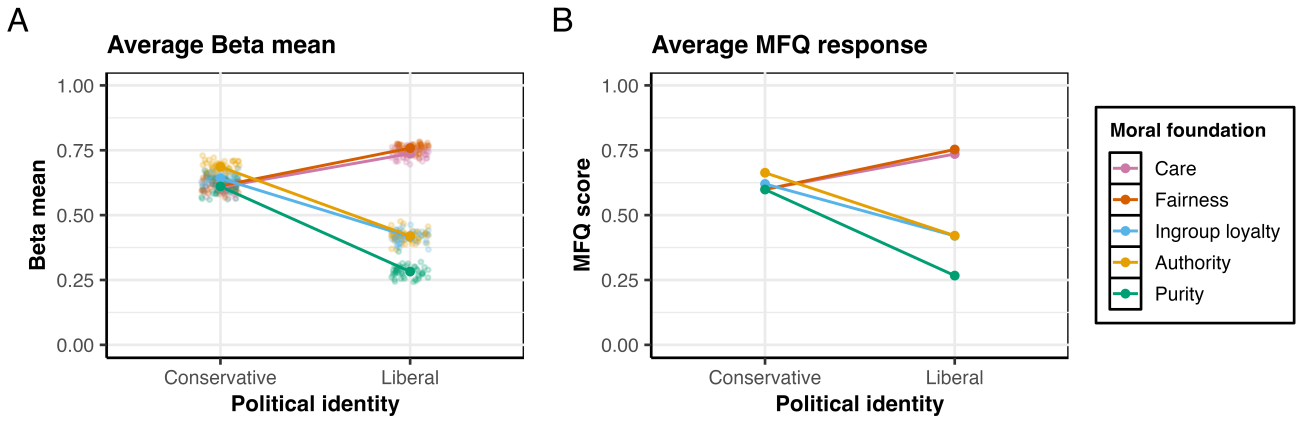

Figure 2: (A) Average simulated beta means and (B) mean MFQ responses from (Graham et al., 2011). Faint points in (A) are simulation-level average means.

posed psychological process by which identity is signaled and inferred using internal moral values while (1) exploring model predictions of previous work beyond specific modeling choices, and (2) providing insights into and future directions on small-scale community formation and its role in political polarization. These alterations could improve how polarization can be measured, thus, expanding validation methods and accuracy for model predictions against empirical data.

Model Specification

Following P&M, we built an agent-based graph model where agents with latent political identities iteratively (1) produce a moral signal, (2) infer moral values similarity with connected agents, and (3) update their own moral values as a function of perceived similarity (Axelrod, 1997). All model code, simulation data, and analysis and supplementary materials are available on GitHub.[1]

Initialization and Representation

We initiated the model with N agents in a fully connected, directed graph. N was sampled from a normal distribution to reduce the influence of sample size on results (see Similarity). Each agent was randomly assigned a latent political identity (liberal/conservative) and had five beta-distributed moral values, each capturing the relevance to a moral value, operationalized as the moral foundation from Moral Foundations Theory (MFT; Graham et al.,2014): care, fairness, ingroup loyalty, respect for authority, and purity. Based on identity, each agent's moral value distributions was parameterized with a randomly sampled empirical response to the Moral Foundations Questionnaire (US-based participants; N = 20,057; $N_{liberals}$ = 16,621; $N_{conservatives}$ = 3,436) [2], originally collected and presented by Graham et al. (2011). As seen in Figure 2, we successfully reproduce the structure of the complete empirical dataset at the group-level. This parametrization approach also captures inter-dependencies between moral values. For an agent, i, at time t, we denote the relevance distribution for moral value m as Equation 1.

$$v_{i,\,m,\,t} \sim Beta(\alpha,\,\beta) \tag{1}$$

Since the network is fully connected, all agents performed similarity inference on all other agents in the network. Thus, each agent was also initiated with $N-1$ sets of moral value distributions for all other agents in the network. Hence, agent i at time t has an inferred distribution for any other arbitrary agent a for each moral value m as shown in Equation 2. We initialized these inferred distributions with random parameters.

$$v_{i\,\rightarrow\,a,\,m,\,t} \sim Beta(\alpha,\,\beta) \tag{2}$$

Similarity

At the start of a time step, all agents compute a perceived similarity index for each of its connected agents. This perceived similarity between an agent and one of its connections ($w_{i\rightarrow a,\,t}$ in Equation 3) is given by the sum of the Kullback–Leibler divergence between that agent's moral values and its belief about its connection's moral values:

$$w_{i\rightarrow a,\,t} = \sum_{m} D_{KL}(v_{i\rightarrow a,\,m,\,t} || v_{i,\,m,\,t}) \tag{3}$$

For analysis purposes, we also computed the same similarity index for the actual moral values each pair of agents:

$$w_{i,a,\,t} = \sum_{m} D_{KL}(v_{i,a,\,m,\,t} || v_{i,\,m,\,t}) \tag{4}$$

Moral Value Inference

Each agent performs a two-step inference based on the moral signals emitted on the previous time step (see signaling): 1)

[1] Code link: https://github.com/navarrenicolas/MoralABM/.

[2] Simulations contained an approx. equal number of liberals and conservatives, thus, not reflecting the imbalance of the empirical data.

inferring other agents' moral values and 2) inferring personal moral values that will best align with agents that are perceived similar (i.e., social influence). Following methodlogy in Fränken et al., 2024, the first step applies a beta-binomial Bayesian update to the representation of a connected agent's moral values based on their signal, $S_{a,\,t}$; increasing the likelihood for the distribution corresponding to the signal (see Equation 5) and integrating the uncertainty about which other moral value was defeated by the chosen signal (see Equation 6):

$$\alpha_{i\to a,\,m,\,t+1} = \begin{cases} \alpha_{i\to a,\,m,\,t} + 1, & \text{if } S_{a,\,t} = m \\ \alpha_{i\to a,\,m,\,t}, & \text{otherwise} \end{cases} \tag{5}$$

$$\beta_{i\to a,\,m,\,t+1} = \begin{cases} \beta_{i\to a,\,m,\,t} + \frac{1}{4}, & \text{if } S_{a,\,t} \neq m \\ \beta_{i\to a,\,m,\,t}, & \text{otherwise} \end{cases} \tag{6}$$

The personal moral values inference step operates on the same principles as the inference of other agents' moral values. However, it incorporates the signals from all other agents in the network and moderates their influence on the agent's moral values with their associated similarity index:

$$\alpha_{i,\,m,\,t+1} = \alpha_{i,\,m,\,t} + \sum_a \begin{cases} \frac{1}{N-1} e^{-w_{i\to a,\,m,\,t}}, & \text{if} S_{a,\,t-1} = m \\ \delta_{S_{a,\,t-1}=m}, & \text{if } a = i \end{cases} \tag{7}$$

$$\beta_{i,\,m,\,t+1} = \beta_{i,\,m,\,t} + \sum_a \begin{cases} \frac{1}{4(N-1)} e^{-w_{i\to a,\,m,\,t}}, & \text{if} S_{a,\,t-1} \neq m \\ \frac{1}{4}\delta_{S_{a,\,t-1}\neq m}, & \text{if } a = i \end{cases} \tag{8}$$

The moderation by perceived similarity in Equations 7 and 8 captures our assumption that an agent updates their moral values such that the signals of those it believes to be similar influences it strongly, but those who are dissimilar have hardly any influence.

Signal Production

At the end of each time step, all agents emit a signal representing one of five moral values. We assume signaling is constrained beyond individual agents, leading to a limited set of options for a given signal. To capture this, each agent faced a moral choice pitting a randomly sampled pair of moral values against one another (m_1, m_2). Each agent determined its signal from its pair using a soft-max decision rule, which took a value for each competing moral value, $v_{i,m_1,t}$ and $v_{i,m_2,t}$, sampled from its current moral values distributions (see Equation 1).

Data Simulation Methods and Analysis

We simulated the model as specified above for 50 independent simulations of 1,500 time steps with $N \sim N(\mu = 100, \sigma = 15)$ for each simulation.

Clustering Measures

The graphical representation of the system allows us to extract multiple measures of clustering (Ng et al., 2001; von Luxburg, 2007). We use these to measure how agents would cluster as function of actual and perceived similarity indices at each time step for a given simulation. We can compare these to the initial latent identity groups to extract the identity homogeneity of clusters and accuracy of perceived similarity.

The model provides two bi-directional graph networks: the perceived moral values graph, G_b and the actual moral values graph, G_a. In the perceived moral values graph, connections represent agents' perceived similarity weighted as in Equation 3. Meanwhile, in the actual moral values graph connections are weighted from the derived actual similarity between agent's latent moral values as in Equation 4 [3]. Following methodology in von Luxburg, 2007 we find an optimal number of clusters to perform spectral clustering . Finally, we use scikit-learn's `SpectralClustering` tool to compute the clusters (Pedregosa et al., 2011). We defer to the supplementary materials for more detailed discussion about these methods.

Cluster Identity Homogeneity Given the latent identities of agents, we first estimate an index of the identity homogeneity of the estimated clusters. That is, to what extent do agents manage to form clusters of mostly only their ingroup based on their perceived moral values about others. For each agent, i and its current cluster, $C_{i,\,t}$, this homogeneity index is the number of agents in their current cluster of the same political identity, *ingroup*, relative to the total number of agents in the simulation:

$$H_{i,\,t} = \frac{1}{N} ingroup(C_{i,\,t}) \tag{9}$$

Moral Values Inference Accuracy Additionally, we compute an index of how accurately agents' perceived moral values maps to the actual moral values. In this analysis, we form clusters based on the agents' beliefs (G_b) and evaluate how close those agents truly are to each-other looking at the similarities between each-other in the actual graph (G_a).

$$A_{i,\,t} = \frac{|C_{i,\,t}|}{N} \sum_{a \in C_{i,\,t}} G_a(i,a) \tag{10}$$

Results

To determine whether the network polarizes based on agents' latent identities, we examine the change in their moral values over time and the change in the clustering measures over time. Figure 3 shows the change in the means of moral values [4] across simulations. Clearly, both liberals and conservatives converge on the same moderate moral values across the simulations. This shift in centrality of agents' moral values is further corroborated by reduced uncertainty [5] (initial

[3] Both graphs represent the KL-divergence between two nodes, i.e., a measure of distance. We normalize this into a similarity matrix with bounded weights (see supplementary materials for details).

[4] i.e., the averaged means of $v_{i,m,t}$: $\frac{\alpha_{i,m,t}}{\alpha_{i,m,t}+\beta_{i,m,t}}$

[5] the averaged sd of $v_{i,m,t}$: $\sqrt{\frac{\alpha_{i,m,t}\beta_{i,m,t}}{(\alpha_{i,m,t}+\beta_{i,m,t})^2(\alpha_{i,m,t}+\beta_{i,m,t}+1)}}$

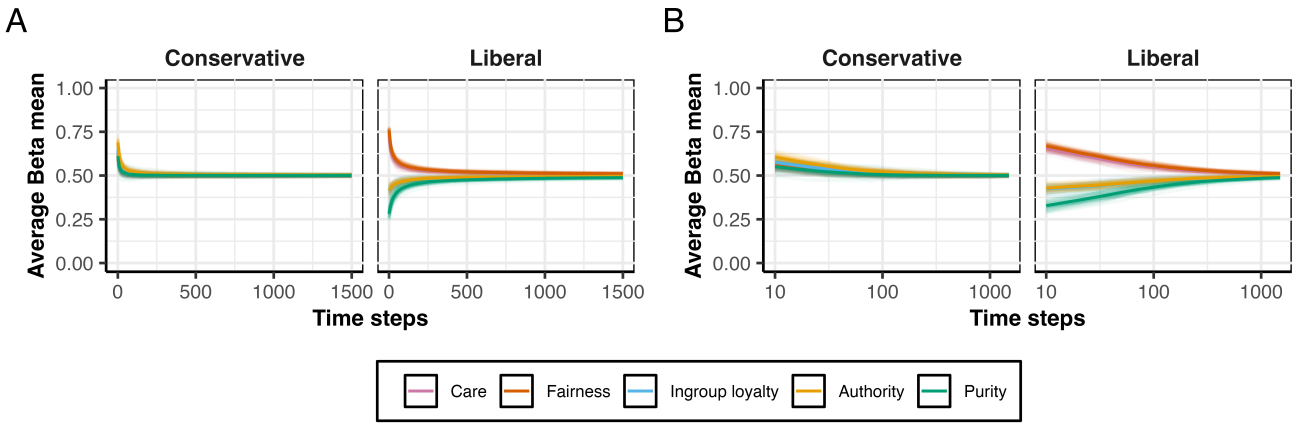

Figure 3: Mean agent moral values by identity over (A) time and (B) log(time). Thick lines indicate the trend across simulations, and faint lines represent individual simulations.

mean $\sigma_{liberals} < 0.18$ and mean $\sigma_{conservatives} < 0.18$; final mean $\sigma_{conservatives} < 0.02$ and mean $\sigma_{liberals} < 0.02$). Thus, the original moral distinctiveness of each group is erased as a result of the simulations.

Indeed, as shown in Table 1, agents formed multiple smaller and largely heterogeneous clusters, a trend that increased over time (see Supplementary materials). In line with this result, both liberals and conservatives had consistently near-zero accuracy (see Table 1), indicating that the perceived similarity index did not capture real differences between agents.

Table 1: Homogeneity and accuracy averaged across simulations (standard error).

Time step	Homogeneity		Accuracy	
	Conservative	Liberal	Conservative	Liberal
1	0.44 (0.01)	0.42 (0.02)	0.15 (0.01)	0.15 (0.01)
500	0.26 (0.03)	0.24 (0.03)	0.07 (0.01)	0.07 (0.01)
1000	0.25 (0.02)	0.25 (0.02)	0.06 (0.01)	0.06 (0.01)
1500	0.26 (0.03)	0.26 (0.03)	0.07 (0.01)	0.07 (0.01)

We also see no major trend between the size of cluster formed and the proportion of political identities as shown in Figure 4. This indicates that agents had little preference with regards to political identity when forming clusters, regardless of its size. In turn, agents seemingly fail to make inferences that distinguish them based on political identity which corroborates the convergent homogeneity in Figure 3.

These results were robust against normalization of moral values between time steps (see supplementary materials).

Discussion

We presented a computational implementation of political identity inference and signaling via moral values, which built on previous modeling efforts and a large empirical literature

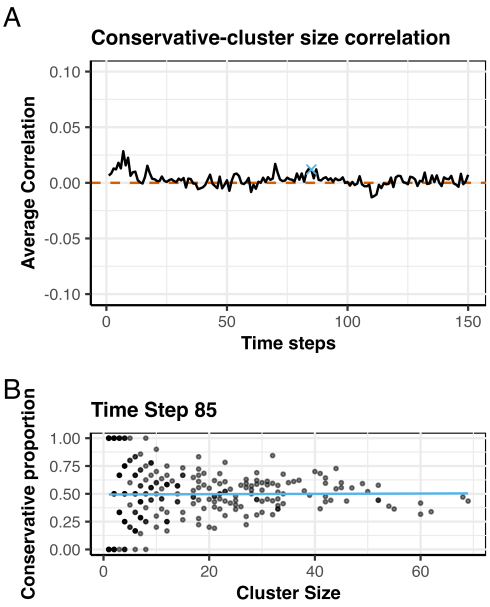

Figure 4: (A) Correlation between cluster size and proportion of conservatives over time averaged over simulations. (B) Example time step showing how correlations were derived. The blue x in (A) indicates the time step shown in (B).

indicating the importance of *uncertainty* in these processes. In this model, however, we found no evidence of political polarization or identity asymmetries. Rather, the simulations indicated that liberals and conservatives symmetrically converged on a homogeneous and moderate set of moral values. Locally, agents formed heterogeneous clusters with no relationship to the original identity categories.

This symmetric convergence of liberals and conservatives is in contrast to previous modeling work using the same theo-

retical framework and to existing empirical data. P&M found echo chambers and moral convergence only among liberals. Likewise, data on social media use and social connections indicate that liberals and conservatives form and interact as distinct groups (Eady et al., 2019; Wu & Resnick, 2021; Yudkin et al., 2019). Given modifications from P&M's model, there are many explanations for these divergent results. Firstly, we assumed resource-rational processes without capacity or visibility constraints. That is, agents integrated information from all other agents and updated their moral values via a continuous similarity representation including themselves. However, both assumptions are unrealistic. Moreover, the current approach fails to capture the binary and polar nature of political/social identity in/out-group representations (Iyengar et al., 2019; Puryear et al., 2024; Tajfel, 1978; Turner et al., 1987). Agents were influenced in their beliefs proportional to similarity, but at no point did they exclude any agents from influencing them, which could lead to the observed value homogenization. This is in stark contrast to human reactions to perceived out-group signals (Ditto & Lopez, 1992; Moore et al., 2021; Ostrom et al., 1993; Thürmer & McCrea, 2018). Individuals are not only attempting to infer and signal in line with ingroup norms under multiple sources of uncertainty, but also avoiding signaling as an out-group member (Berger & Heath, 2008; Schöppl & Hahn, 2024). Alternatively, self-influence could be a source of to convergence. Currently, while social influence is weighted by the number of agents in the network, self-influence is not. That is, agents may overfit to their personal choices, which are largely constrained by uniform noise from the signal production process. We discuss how these misaligned assumptions are easily rectified in the Future Directions section. Our results may indicate that resource-rational agents would not polarize from uncertain ingroup inference and signaling via moral values without some or all of: representing ingroup membership as (a) binary, (b) the inverse to the out-group, (c) limitations on how many other agents a given agent can perceive and update beliefs on, and/or (d) more proportional self-to-social-influence. Further model development is needed to fully explore these possibilities.

Most models that attempt to capture polarization discover it often as strong bifurcation (Bramson et al., 2017). This challenges our ability to validate models of polarization and disentangle the factors with real empirical implications or predictive power from model artifacts. By providing a model framework that with permissive assumptions does not produce polarization, we can conduct systematic interventions on agent processes and conditions to test what combination of factors may predict polarization in a way that complies with empirical literature. The graphical representations in our models in particular may be suited to model polarization as a nuanced rather bifurcated process.

Future Directions

There are multiple developments of our modeling framework to reduce the dissonance with empirical and modeling findings on political polarization. Any development should, however, preserve the fundamental theoretical assumptions whereby moral values serves identity motivations by functioning as an interface to signal and infer political identity.

Based on the presented simulations, agents clearly did not succeed in signaling their identity distinctively from the out-group. We consider multiple alterations that may reduce the noise around these inferences. One such development would be to incorporate a binary ingroup/out-group inference of other agents leveraging existing representations about other agents' moral values (see discussion for theoretical motivation). Following previous models of signaling and polarization (Schöppl & Hahn, 2024; van der Does et al., 2022), this would lead to signaling dynamics where agents must engage in more complex decisions to maximize their chances of finding ingroup members while minimizing the risk of getting ousted by other members. This complexity may be important factor in political polarization not captured by our implementation. This would require further decision-theoretic commitments to the agents' signal inference and production processes. This may also include counter-signaling by which moral values degrade if signaled by highly dissimilar agents. Additionally, further resource-rational constraints could be imposed on the agents' signal and inference processes including constraints on the representations of other agents' moral values or limiting the number of agents that an agent can observe. This could be implemented through a process of connection decay and discovery.

Furthermore, we could capture hierarchies in social networks and how these moderate and inform signaling (Jost et al., 2022). There are multiple empirical findings on political leaders'/influencers' social media behavior that is congruent with guiding/disambiguating (moral) signaling, e.g., more extreme beliefs (DeSilver, 2022), less signal variability (Zhang et al., 2023) and faster signal spread (Brady et al., 2019). In addition to the factors discussed above, these hierarchical features may contribute to distinct signaling between political groups, amplifying polarization. The graphical model representation could be used to extract such hierarchical dependencies with agglomerative clustering methods that capture the structure of nodes with larger influence (Müllner, 2011). However, this development would require additional theoretical and potentially empirical work on leadership representations, and how they interact with the existing theoretical assumptions.

Conclusion

We presented an initial graph agent-based model of political identity inference and signaling via moral values. With the graphical representation, we provided measures of polarization sensitive to small-level communities and dynamics that may inform future modeling efforts on polarization. The resulting network diverged from previous model instantiations of the same theoretical framework, warranting and providing a testing ground for further study to disentangle implementation artifacts from actual drivers of political polarization.

Acknowledgments

We thank Tadeg Quillien for his valuable inputs into the model conceptualization. This work was supported in part by the Economic and Social Research Council through the Scottish Graduate School of Social Sciences [ESRC Grant number: ES/P000681/1]; and the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh.

References

Assaad, L., Fuchs, R., Jalalimanesh, A., Phillips, K., Schoeppl, L., & Hahn, U. (2023). A bayesian agent-based framework for argument exchange across networks. https://doi.org/10.48550/ARXIV.2311.09254

Atikcan, E. Ö., & Hand, K. (2024). Moral framing and referendum politics: Navigating the empathy battlefield. *Political Psychology*, *45*(1), 193–210. https://doi.org/10.1111/pops.12921

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

Baldassarri, D., & Page, S. E. (2021). The emergence and perils of polarization [Publisher: National Academy of Sciences Section: Perspective]. *Proceedings of the National Academy of Sciences*, *118*(50), e2116863118. https://doi.org/10.1073/pnas.2116863118

Baumann, F., Lorenz-Spreen, P., Sokolov, I. M., & Starnini, M. (2020). Modeling echo chambers and polarization dynamics in social networks [Publisher: American Physical Society]. *Physical Review Letters*, *124*(4), 048301. https://doi.org/10.1103/PhysRevLett.124.048301

Berger, J., & Heath, C. (2008). Who drives divergence? identity signaling, outgroup dissimilarity, and the abandonment of cultural tastes [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *95*(3), 593–607. https://doi.org/10.1037/0022-3514.95.3.593

Bos, L., & Minihold, S. (2022). The ideological predictors of moral appeals by european political elites; an exploration of the use of moral rhetoric in multiparty systems. *Political Psychology*, *43*(1), 45–63. https://doi.org/10.1111/pops.12739

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2023). Cross-country trends in affective polarization. https://doi.org/10.3386/w26669

Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T., & Van Bavel, J. J. (2019). An ideological asymmetry in the diffusion of moralized content on social media among political leaders [Place: US Publisher: American Psychological Association]. *Journal of Experimental Psychology: General*, *148*(10), 1802–1813. https://doi.org/10.1037/xge0000532

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Flocken, C., & Holman, B. (2017). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science*, *84*(1), 115–159. https://doi.org/10.1086/688938

Ciuk, D. J. (2018). Assessing the contextual stability of moral foundations: Evidence from a survey experiment. *Research & Politics*, *5*(2), 2053168018781748. https://doi.org/10.1177/2053168018781748

DeSilver, D. (2022). *The polarization in today's congress has roots that go back decades* [Pew research center].

Dias, N., & Lelkes, Y. (2022). The nature of affective polarization: Disentangling policy disagreement from partisan identity [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12628]. *American Journal of Political Science*, *66*(3), 775–790. https://doi.org/10.1111/ajps.12628

Dimock, M., Doherty, C., Kiley, J., & Oates, R. (2014). Political polarization in the american public.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *63*(4), 568–584. https://doi.org/10.1037/0022-3514.63.4.568

Eady, G., Nagler, J., Guess, A., Zilinsky, J., & Tucker, J. A. (2019). How many people live in political bubbles on social media? evidence from linked survey and twitter data [Publisher: SAGE Publications]. *SAGE Open*, *9*(1), 2158244019832705. https://doi.org/10.1177/2158244019832705

Fränken, J.-P., Valentin, S., Lucas, C. G., & Bramley, N. R. (2024). Naïve information aggregation in human social learning. *Cognition*, *242*, 105633. https://doi.org/https://doi.org/10.1016/j.cognition.2023.105633

Franks, A. S., & Scherr, K. C. (2015). Using moral foundations to predict voting behavior: Regression models from the 2012 u.s. presidential election. *Analyses of Social Issues and Public Policy*, *15*(1), 213–232. https://doi.org/10.1111/asap.12074

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Graham, J., Nosek, B. A., & Haidt, J. (2012). The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum. *PLOS ONE*, *7*(12), e50092. https://doi.org/10.1371/journal.pone.0050092

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847

Hahn, U., Hansen, J. U., & Olsson, E. J. (2018). Truth tracking performance of social networks: How connectivity and clustering can make groups less competent. *Synthese*, *197*(4), 1511–1541. https://doi.org/10.1007/s11229-018-01936-6

Hatemi, P. K., Crabtree, C., & Smith, K. B. (2019). Ideology justifies morality: Political beliefs predict moral foundations. *American Journal of Political Science*, *63*(4), 788–806. https://doi.org/10.1111/ajps.12448

Hogg, M. A., & Reid, S. A. (2006). Social identity, self-categorization, and the communication of group norms. *Communication Theory*, *16*(1), 7–30. https://doi.org/10.1111/j.1468-2885.2006.00003.x

Hogg, M. A., & Turner, J. C. (1985). Interpersonal attraction, social identification and psychological group formation. *European Journal of Social Psychology*, *15*(1), 51–66. https://doi.org/10.1002/ejsp.2420150105

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, *76*(3), 405–431. https://doi.org/10.1093/poq/nfs038

Jern, A., Chang, K.-m. K., & Kemp, C. (2014). Belief polarization is not always irrational [Place: US Publisher: American Psychological Association]. *Psychological Review*, *121*(2), 206–224. https://doi.org/10.1037/a0035941

Jost, J. T., Baldassarri, D. S., & Druckman, J. N. (2022). Cognitive–motivational mechanisms of political polarization in social-communicative contexts [Number: 10 Publisher: Nature Publishing Group]. *Nature Reviews Psychology*, *1*(10), 560–576. https://doi.org/10.1038/s44159-022-00093-5

Kidd, W., & Vitriol, J. A. (2022). Moral leadership in the 2016 u.s. presidential election. *Political Psychology*, *43*(3), 583–604. https://doi.org/10.1111/pops.12782

Kinder, D. R., & Kalmoe, N. P. (2017). *Neither liberal nor conservative: Ideological innocence in the american public*. University of Chicago Press.

Kvam, P. D., Alaukik, A., Mims, C. E., Martemyanova, A., & Baldwin, M. (2022). Rational inference strategies and the genesis of polarization and extremism. *Scientific Reports*, *12*(1), 7344. https://doi.org/10.1038/s41598-022-11389-0

Levine, S., Chater, N., Tenenbaum, J., & Cushman, F. (2023, May 5). Resource-rational contractualism: A triple theory of moral cognition. https://doi.org/10.31234/osf.io/p48t7

Macy, M. W., Deri, S., Ruch, A., & Tong, N. (2019). Opinion cascades and the unpredictability of partisan polarization. *Science Advances*, *5*(8), eaax0754. https://doi.org/10.1126/sciadv.aax0754

Malka, A., & Lelkes, Y. (2010). More than ideology: Conservative–liberal identity and receptivity to political cues. *Social Justice Research*, *23*(2), 156–188. https://doi.org/10.1007/s11211-010-0114-3

Milesi, P. (2016). Moral foundations and political attitudes: The moderating role of political sophistication. *International Journal of Psychology: Journal International De Psychologie*, *51*(4), 252–260. https://doi.org/10.1002/ijop.12158

Milesi, P. (2017). Moral foundations and voting intention in italy. *Europe's Journal of Psychology*, *13*(4), 667–687. https://doi.org/10.5964/ejop.v13i4.1391

Moore, A., Hong, S., & Cram, L. (2021). Trust in information, political identity and the brain: An interdisciplinary fMRI study. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1822), 20200140. https://doi.org/10.1098/rstb.2020.0140

Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. https://doi.org/10.48550/ARXIV.1109.2378

Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* (Vol. 14). MIT Press. https://proceedings.neurips.cc/paper_files/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf

Nilsson, A., & Erlandsson, A. (2015). The moral foundations taxonomy: Structural validity and relation to political ideology in sweden. *Personality and Individual Differences*, *76*, 28–32. https://doi.org/10.1016/j.paid.2014.11.049

Olsson, E. J. (2013). A bayesian simulation model of group deliberation and polarization. In F. Zenker (Ed.), *Bayesian argumentation: The practical side of probability* (pp. 113–133). Springer Netherlands. https://doi.org/10.1007/978-94-007-5357-0_6

Ostrom, T. M., Carpenter, S. L., Sedikides, C., & Li, F. (1993). Differential processing of in-group and out-group information [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *64*(1), 21–34. https://doi.org/10.1037/0022-3514.64.1.21

Pedersen, J. M. E. (2024). *A formalised theory of political polarisation as uncertain identity inference and signalling supported by moral values* [MSc dissertation]. Psychology Department, School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Scotland.

Pedersen, J. M. E., & Moore, A. (2023). Simulating political polarization as a function of uncertain inference and signaling of moral values. *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*. https://escholarship.org/uc/item/0d143859

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Puryear, C., Kubin, E., Schein, C., Bigman, Y. E., Ekstrom, P., & Gray, K. (2024). People believe political opponents accept blatant moral wrongs, fueling partisan divides. *PNAS Nexus*, *3*(7), pgae244. https://doi.org/10.1093/pnasnexus/pgae244

Rosario, K. d., Bavel, J. J. V., & West, T. (2024). What does my group consider moral?: How social influence shapes moral expressions. https://doi.org/10.31234/osf.io/dwzq2

Schelling, T. C. (1971). Dynamic models of segregation† [Publisher: Routledge _eprint: https://doi.org/10.1080/0022250X.1971.9989794]. *The Journal of Mathematical Sociology*, *1*(2), 143–186. https://doi.org/10.1080/0022250X.1971.9989794

Schöppl, K., & Hahn, U. (2024). Exploring effects of self-censoring through agent-based simulation [Peer-Reviewed]. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*. https://escholarship.org/uc/item/9b32v6xc

Settle, J. E., & Carlson, T. N. (2019). Opting out of political discussions [Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2018.1561563]. *Political Communication*, *36*(3), 476–496. https://doi.org/10.1080/10584609.2018.1561563

Smaldino, P. E. (2022). Models of identity signaling. *Current Directions in Psychological Science*, *31*(3), 231–237. https://doi.org/10.1177/09637214221075609

Smith, K. B., Alford, J. R., Hibbing, J. R., Martin, N. G., & Hatemi, P. K. (2017). Intuitive ethics and political orientations: Testing moral foundations as a theory of political ideology. *American Journal of Political Science*, *61*(2), 424–437. https://doi.org/10.1111/ajps.12255

Tajfel, H. (1978). *Differentiation between social groups: Studies in the social psychology of intergroup relations*. Academic Press.

Theodoropoulos, N. C. (2022). *Computational modelling of social cognition and behaviour* [Doctoral dissertation]. Psychology Department, School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Scotland.

Thürmer, J. L., & McCrea, S. M. (2018). Beyond motivated reasoning: Hostile reactions to critical comments from the outgroup [Place: US Publisher: Educational Publishing Foundation]. *Motivation Science*, *4*(4), 333–346. https://doi.org/10.1037/mot0000097

Tokita, C. K., Guess, A. M., & Tarnita, C. E. (2021). Polarized information ecosystems can reorganize social networks via information cascades. *Proceedings of the National Academy of Sciences*, *118*(50). https://doi.org/10.1073/pnas.2102147118

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. W., & S., W., M. (1987). *Rediscovering the social group: A self-categorization theory*. Basil Blackwell.

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

van Leeuwen, F., & Park, J. H. (2009). Perceptions of social dangers, moral foundations, and political orientation. *Personality and Individual Differences*, *47*(3), 169–173. https://doi.org/10.1016/j.paid.2009.02.017

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416. https://doi.org/10.1007/s11222-007-9033-z

Waytz, A., Iyer, R., Young, L., Haidt, J., & Graham, J. (2019). Ideological differences in the expanse of the moral circle. *Nature Communications*, *10*(1), 4389. https://doi.org/10.1038/s41467-019-12227-0

Womick, J., Goya-Tocchetto, D., Ochoa, N. R., Rebollar, C., Kapsaskis, K., Pratt, S., Payne, K., Vaisey, S., & Gray, K. (2024). Moral disagreement across politics is explained by different assumptions about who is most vulnerable to harm. https://doi.org/10.31234/osf.io/qsg7j

Wu, S., & Resnick, P. (2021). Cross-partisan discussions on YouTube: Conservatives talk to liberals but liberals don't talk to conservatives. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 808–819. Retrieved July 13, 2022, from https://ojs.aaai.org/index.php/ICWSM/article/view/18105

Young, D. J., Madsen, J. K., & de-Wit, L. H. (2025). Belief polarization can be caused by disagreements over source independence: Computational modelling, experimental evidence, and applicability to real-world politics. *Cognition*, *259*, 106126. https://doi.org/https://doi.org/10.1016/j.cognition.2025.106126

Yudkin, D., Hawkins, S., & Dixon, T. (2019, September 14). *The perception gap: How false impressions are pulling americans apart*. More in Common. https://doi.org/10.31234/osf.io/r3h5q

Zhang, Y., Chen, F., & Lukito, J. (2023). Network amplification of politicized information and misinformation about COVID-19 by conservative media and partisan influencers on twitter [Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2022.2113844]. *Political Communication*, *40*(1), 24–47. https://doi.org/10.1080/10584609.2022.2113844