JASSS
THE JOURNAL OF ARTIFICIAL SOCIETIES AND SOCIAL SIMULATION

TITLE

[Anonymised review copy]

Abstract: In this article we present the results of agent-based simulations on an extended Hegselmann-Krause (HK) model with social identity dynamics. We aimed to study the role of socially relevant variables on opinion polarization, such as open-mindedness (traditionally known as ε or uncertainty in the HK literature), and social identity. We modeled the emergence of identity groups as a dynamic process dependent on the state of the opinion space at any given time. Identity effects are modeled as an additional layer of boundedness - agents only update their opinions based on incoming ideas from other members of their own identity group at the time. We also define a parameter termed Salience of Proximity in Identity-Relevant Opinions (SPIRO), that controls the influence of opinion clustering on the identity groups detected by an observer. SPIRO influences the number and granularity of the identity groups so formed, and can take on different values for different observing agents. We expand upon some known results in opinion dynamics such as the consensus-driving tendency of heterogeneity in ε, and also present results on how identity influences polarization. A key result we find is that SPIRO modulates the influence of ε on polarization - as one raises the granularity of the identity clusters, the system dynamics transition from being mainly driven by opinions to mainly driven by identity.

Keywords: Polarization, Identity, Social Identity Theory, Bounded-confidence model, Heterogeneity, Openness

Introduction

1.1 As modern communication technologies like social media continue to transform the landscape of human communication, scholars have called for an urgent and extensive investigation of how these influence the functioning and health of societies (Bak-Coleman et al. 2021). The increased accessibility of human opinions on online platforms has also been followed by concerns about widespread societal consequences because of the potential for misuse at scale. Some of the worrying online behaviors include the spread of misinformation (Vosoughi et al. 2018; Tran et al. 2021) and hate speech (Matamoros-Fernández & Farkas 2021), cyberbullying and cybergrooming (Mladenović et al. 2021; Machimbarrena et al. 2018), and the emergence and incubation of conspiracy theories and extremist ideologies (Dow et al. 2021; Kim & Kim 2023), all of which appear to thrive in the new communication landscape afforded by social media. Another worrying issue is regarding the rise of non-human agents which can be deployed en masse as sources of propaganda, and are known to influence online human behavior (Bessi & Ferrara 2016; Shao et al. 2018). Therefore it is both necessary and urgent to attempt to understand the nature of political communication in societies with the aim of developing generalizable insights about collective human behavior and political communication.

1.2 One insidious consequence of the ubiquity of new media appears to be the exacerbation of political polarization (Wilson et al. 2020; Kubin & von Sikorski 2021). "Polarization" as a term is open to interpretation, but generally implies a state of heightened animosity between partisan camps that each consist of like-minded individuals (Bauer 2019). While opinion diversity is considered a desirable trait in a healthy democracy, extreme polarization is detrimental to the health of democracies since it limits the value of democratic debate and can lead to irreconcilable differences between factions (Macy et al. 2021).

1.3 In this modeling study we further analyze an opinion dynamics model we had previously developed and discussed in Kalvas et al. (2023), which was a modified Hegselmann-Krause (HK) model (Hegselmann & Krause 2002) that allowed for dynamic emergence of identity groups as per Social Identity Theory (SIT, Tajfel et al. (1979)). We build upon our previous results in this article by analyzing the influence of identity and open-mindedness on polarization. We also help link our implementation of social identity in our existing model to the theoretical principle of meta-contrast from Self-Categorization Theory, as it is to our knowledge the only opinion dynamics model of meta-contrast that does not assume repulsive forces between individuals or identity groups.

Doi: 10.18564/jasss.xxxx

1.4 This study is a part of a larger project with an overarching goal of understanding the influence of different communication landscapes on opinion dynamics using Agent-Based Models (ABM's). We plan to incrementally build towards a dynamic conceptual model characterized by a positive feedback loop between media use and media-influenced behavior, known as the Reinforcing Spirals Model (RSM, Slater (2007)). The strength of this feedback loop in RSM is moderated by openness of the system or subsystem the individual is situated in. We interpret openness in our opinion dynamics model as referring to the open-mindedness (ε) variable of the HK model, which can be allowed to vary across agents (Fu & Zhang 2014; Lorenz 2007). In our operationalization, the openness of the whole system refers to the distribution mean of ε. RSM also conceptualizes the role of social identity as an exacerbating factor for the spiral that can drive societies towards inter-group conflict.

1.5 In Kalvas et al. (2023) we sought to model social identity as a higher-order process which exerts its own influence over and above that of interpersonal communication, consistent with SIT and RSM. We also presented regression results that broadly suggested that higher average open-mindedness alleviates polarization and that identity modulates this relationship. In this article we present a detailed examination of the macro-behavior of our model, with a focus on identity dynamics and how it influences political polarization. We also present some results highlighting the role of other key variables in RSM, such as the distribution parameters of open-mindedness. Along the way, we hope to contextualize our model in the social simulation literature by presenting a review of the underlying social theories, related empirical findings, and some of the existing ABM studies of opinion dynamics in the presence of social identity; along with a discussion about how our model incorporates specific theoretical features of the Social Identity Approach.

1.6 The rest of this article is structured as follows: the remainder of section 1 covers relevant literature from different fields - here we discuss some aspects of polarization, Social Identity Theory, and related Agent-Based Models in the literature. In Section 2 we present a description of our model, along with rationale for our modeling choices and an overview of the simulation procedure. We discuss our simulation results in Section 3, and interpret our findings in Section 4.

Background

Opinion Polarization in Agent-Based Models

1.7 Polarization as a social phenomenon has been studied in a variety of contexts and using a variety of study methods. We define polarization as individuals being grouped into clusters based on one or more similar attitudes and beliefs (Esteban & Ray 1994). It is increasingly understood that polarization may jeopardize the political process in the United States (Heltzel & Laurin 2020) and the European Union (Apergis & Pinar 2023) for example, which is encouraging scholars from a variety of disciplines to consider issues related to polarization. Therefore, there are numerous literatures that address issues of polarization including Political Science (Levin et al. 2021), Communication (Bolsen & Shapiro 2017), Sociology (Downey 2022), Economics (Montalvo & Reynal-Querol 2003) and other areas.

1.8 Though there is some conceptual consensus, there is no single widely accepted method to define polarization operationally, partially due to the diversity of researchers and disciplinary approaches existing in this space. This is true even in the ABM literature. Most measures of polarization typically measure one or more of the following three characteristics of the opinion distribution: (a) the spread of the opinion distribution (Biondi et al. 2023); (b) the distribution of extremists, typically simplified as a ratio or difference Li & Xiao (2017); (c) the clustering characteristics of the opinion distribution (Schweitzer et al. 2020). Additionally, some studies take the distance of opinion between pre-determined factions of agents (Diep et al. 2023).

1.9 As we wished to adopt a more nuanced measure of polarization that accounts for multiple characteristics of the opinion distribution we chose to adapt the Equal-Size Binary Grouping (ESBG) algorithm (Tang et al. 2022) as our primary polarization metric. The ESBG algorithm can be applied to a continuous opinion space, and outputs a polarization metric in [0, 1], where 0 indicates perfect consensus and 1 indicates perfect polarization (two perfectly tight clusters of equal size that together contain all agents in the population, at maximum possible distance from each other). As intermediate values of this polarization metric are harder to interpret, we also supplement ESBG with more traditional spread-based measures of polarization.

Agent-Based Models of Social Influence

1.10 Many models of opinion dynamics assume that conformity to socially-derived consensus at least one of the central mechanisms behind the shaping of public opinion. This has its empirical roots in the early studies of

social conformity, which established that people can be made to doubt their judgement on ambiguous as well as unambiguous situations when they are faced with a dissenting majority (Sherif (1936); Asch (1955), for review see Flache et al. (2017)). Interestingly this effect is believed to require the subject to self-identify as being similar to the prospective influencer (David & Turner 2001). Besides the assumption of social influence, many contemporary models of opinion dynamics typically also assume two more dynamic principles - (a) homophily, i.e., similar agents are more likely to influence one another; and (b) when an agent is influenced by another, the magnitude of influence scales with the magnitude of dissimilarity between the agents (Flache 2018).

1.11 The HK model is a case of Bounded-Confidence models, which form a class of models that emerge from the formalization of theories of social influence. In such models, agents have a (usually homogeneous and stable) tolerance for opinion dissimilarity which is used to filter out prospective influencers. In the HK literature, this variable is typically denoted as ε, and is variably referred to as "confidence level", "uncertainty", "boundary", or "open-mindedness". While more is known about the behavior of HK models where all the agents share the same ε parameter due to them being more analytically tractable (Su et al. 2017), some studies have also looked at HK models with heterogenous ε (Lorenz 2009; Kou et al. 2012; Han et al. 2019),.and they typically studies assume that the ε value of agents is drawn from a small pre-determined set. The dearth of models that assume more naturalistic distributions of ε is particularly noteworthy when viewed through the lens of the social judgment theory of attitude change (Sherif & Hovland 1961), which assumes that individuals have a range of attitudes that are unacceptable to them in a given time and context, and that this range is variable across individuals (Sherif et al. 1973).

The Social Identity Approach

1.12 The term 'Social Identity Approach' is conventionally used to refer to the broader theoretical framework comprising two closely related theories in social psychology - Social Identity Theory (SIT) and the subsequent Self-Categorization Theory (Hornsey 2008). SIT was based on experiments where inter-group discrimination effects such as in-group favoritism were observed even when the groups were externally formed by the experimenter on completely arbitrary criteria, and group members were not allowed to interact with one another (Tajfel et al. 1971; Billig & Tajfel 1973). These 'Minimal Group Paradigm' experiments helped establish that group phenomena such as partisanship and mob violence cannot be entirely explained as being emergent from interpersonal dynamics in the context of shared (or conflicting) interests between members of the same (different) groups or from competition for resources. Rather, it is the very act of being categorized into a group that can cause individuals to discriminate between in-group and out-group members (Billig & Tajfel 1973). Thus self-identification with the group is the critical variable for inter-group discrimination.

1.13 According to SIT, individuals are driven to favor their groups not merely because they rationally view participation in group politics as favorable to them in securing their personal goals ('Realistic conflict'), but because identification with a social group can satisfy one's deeper psychological needs for a positive self-concept in relation to others (Turner 1975). The extent to which individuals favor their in-groups can vary over time and depends on how salient one's identity is in the present social context at any given time (Tajfel et al. 1979).

1.14 SIT can explain group-level dynamics, but the question of what psychological drives at the level of individuals lead to identification with a social group and the consequent behaviors that serve the group's interests was formalized in the subsequent Self-Categorization Theory (SCT) (Turner et al. 1987). According to SCT, a single individual can identify with several hierarchically related groups, the highest (in hierarchy) of which is the group of all humans, and the lowest of which is the individual themself. Therefore a single individual could identify as a human, an Italian, a Sicilian, a female, a botanist, or an individual with a specific name, at different points in time. Importantly, behaviors driven by association with each possible identity group are not exhibited all at once - rather, social context at any given time makes one or more identities of an individual more salient than their other possible identities. SCT proposes that a person's expressed social group identification at any given time exists in a 'functional antagonism' with their other social identities.

1.15 While in this view the person's self-categorization as an individual is simply one among many possible self-categorizations that exist in mutual tension, for modeling purposes we will be simplifying this principle to assume that there are at least two self-categorizations that may influence one's behavior. As we will elaborate in Section 2, our model assumes that one can exhibit behaviors as an individual and as a member of a larger social group in parallel.

1.16 Self-Categorization Theory also provides insights on how people classify a set of individuals into different identity groups. The principle of meta-contrast proposed as part of SCT states that as individuals classify others into social groups they optimize for two things: one is maximizing the intra-group difference with respect to

some socially relevant set of traits, while the other is minimizing the in-group heterogeneity. While some ABM's of social identity have directly operationalized this principle to the study of opinion dynamics (Salzarulo 2006; Weimer et al. 2022), these models have assumed that the groups emerging from meta-contrast will also exert repulsion or negative influence between each other. In our model, discussed in Section 2, meta-contrast is implicit in our usage of a graphical community detection algorithm to detect identity groups, but does not lead to negative influence between detected identity groups.

Social Identity in Political Communication

1.17 Group-level dynamics play an important role in political cognition. Although cognitive biases such as motivated reasoning is commonly observed in politically relevant judgements (Druckman & Bolsen 2011), political identities can further strongly bias beliefs in favor of partisan in-groups (for review see Leeper & Slothuus (2014)). For example, people's judgements of policies are strongly biased by the position of their favored political party without their conscious awareness (Cohen 2003). Similarly, support for conspiratorial rhetoric by political leaders is moderated by political identity (Dow et al. 2023). These partisan biases are more likely a consequence of motivated effortful reasoning than blind conformity to party positions (Petersen et al. 2013), and are sensitive to salience of partisan identities (Bolsen et al. 2014). Strikingly, people can vote against their personal interest in favour of their identity group even in minimal group settings with randomized group assignments (Bassi et al. 2011).

1.18 Recent studies in cognitive science and neuroscience have shed more light into the processing of political identity. Studies with Implicit Attitude Tests show an automatic preference for policy proposals from in-group members (Smith et al. 2012). Judgement of visually observed stimuli can be biased by one's political identity (Kahan et al. 2012). Neuroimaging data suggests that individual brains have a representation of "in-groupness" that is common across multiple group identification criteria. A classifier trained to decode in-groupness in experimental minimal group settings from neural data could also predict in-groupness of real political stimuli (Cikara et al. 2017). Neural responses to in-groupness in minimal group experimental settings (where the minimal group is salient) are similar to the responses to racial in-groups in neuroimaging studies of racial membership representations (where race is salient, Van Bavel et al. (2008)).

1.19 Of additional interest to us is the existence of social identities that are purely based on political opinion - as opposed to other identity-relevant variables such as ethnicity, gender, or socio-economic status. While much of the social identity literature treats political and non-political social identity similarly - reflecting the broad applicability of the theory - some authors have explicitly referred to opinion-based identities. Bliuc et al. (2007) draws a clear distinction between identity groups centered on opinions, and those constructed around other features, and found that opinion-based group membership can act over and above the opinions themselves in predicting political action. Smith et al. (2015) proposed that the formation of novel collective action groups - that were not perfectly aligned to pre-existing societal fault lines - can be explained by individual agents communicating their grievances to others and renegotiating preferred norms.

1.20 The key takeaways for us is that opinion-based social identities exist, and can be thought of as a dynamical component that operates over and above interpersonal dynamics. These theoretical considerations prompted us to model identity as a dynamic feedback process of its own that both depends on one's opinion, and influences the evolution of future opinions.

Model Description

Rationale - Integrating Social Identity with Opinion Dynamics

2.1 Our model was built with a few objectives in mind: firstly we wanted to capture the effect of identity as a component distinct from interpersonal dynamics - meaning we wanted to take an existing model of opinion dynamics based on interpersonal interaction and extend its updating rule with identity dynamics. Secondly, while social identity can arise in a number of contexts, we are specifically interested in one's political identity, which we assume is a function of one's opinion in relation to the opinions held by others - in other words we apply the principle of meta-contrast to the dynamic opinion space to derive dynamic identity assignments. Thirdly, we assume that perceived social identities are dependent on the observer - some people may see more identity groups than others. Fourthly we assume that the effect of identity on communication between any two individuals is similar to an all-or-nothing filter - a perceived identity mismatch between the listener and the speaker

leads to the listener ignoring the speaker's opinion entirely. Finally, we have tried to keep the model as general as possible within the scope of our research questions, by allowing heterogeneity in several agent properties.

2.2 We built five variants of the Hegselmann-Krause (HK) bounded-confidence model (Hegselmann & Krause 2002), two of which contain identity dynamics, to enable comparison of the influence of different dynamical processes on polarization. The basic HK model assumes that individuals update their opinion by moving towards the average opinion of all other agents who meet a selection criteria. Typically a speaking agent's opinion is taken into consideration as long as it is within a dissimilarity limit of the listener's opinion, known commonly as 'uncertainty' in the HK literature and denoted as ε. All our models exhibited this basic behavior, but differed in whether identity effects were present and whether there was heterogeneity in agent variables.

2.3 We integrated identity dynamics into the HK model by keeping the effect of identity mismatches qualitatively similar to the effect of opinion mismatches in the classical model. At every time step agents classify themselves and others into identity groups based on the distribution of opinions at that time via the Louvain Community Detection algorithm (Blondel et al. 2008). After an agent assigns all agents to some identity group, they filter out incoming opinions from agents that are not perceived to be in the same identity group as themself. Therefore, the listener is selective of incoming information not only based on its content, but also based on the social identity of the speaker.

2.4 An important parametrization for Louvain Community Detection is to do with the resolution of the communities found, since the original algorithm is biased to some sizes relative to the network size (Jones et al. 2016). We resolve this by defining a parameter we term as the Salience of Proximity in Identity-Relevant Opinions (SPIRO). SPIRO determines how sensitive an observer is to the tightness of possible identity clusters in the opinion space. We also show that this parameter controls the granularity of identity groups so detected. We interpret SPIRO as a socially relevant variable that is allowed to vary across observing individuals. Thus, it is possible that two agents observing the same opinion landscape may differ in their judgment as to how many identity groups exist and which agent falls into which group.

2.5 With our long-term goals in mind, we also took a more naturalistic approach to our modelling and interpretation. We use approximate Gaussian distributions for key variables such as openness, conformity, and SPIRO, in the model variants for which these variables were heterogeneous. We also explicitly interpret ε as openness or open-mindedness to differences in opinion, which is an interpretation present in the HK literature (Fu & Zhang 2014; Lorenz 2009).

Model Components

2.6 We developed and studied five hierarchically related models with the idea of understanding the influence of different dynamic components with polarization. These five variants were built incrementally, each subsequent model building on top of the components of the previous. Thus the first of these models is closest to the classical HK model with its minimal features, while last of these models is the richest and has several variables related to identity. In this section we will describe the different model components and relevant experimental variables, and then highlight which dynamic components are present in each model. The reader is directed to Kalvas et al. (2023) for a more detailed description of all models and model elements.

2.7 The model broadly has two interacting dynamic components: the HK bounded confidence rule for updating opinions and the implementation of social identity as an additional boundary every agents communication. The application of social identity itself is a two-step process - first an agent must assign every agent including itself to an identity group based on the distribution of the opinion space, and its own sensitivity to opinion clustering. Then the agent must take action based on the identity assignments - in our model, this translates to ignoring agents belonging to different identity groups. We break down these components below:

Hegselmann-Krause with conformity

2.8 An agent's opinion is a real number between -1 and $+1$ that is updated according to the conformity-modified Hegselmann-Krause rule, where an agent's future opinion is their current opinion shifted towards the neighborhood average opinion. The magnitude of opinion shift of an agent i towards its neighborhood average opinion is obtained by weighting the difference of opinion between the agent and its neighborhood average by the conformity parameter, α_i.

Social Identity

2.9 For the variants of the model which have identity, identity groups are implemented as discrete tags associated with every agent which are updated at every time step.

2.10 Our implementation of social identity is related to the meta-contrast principle of Self-Categorization Theory, which is in practice also achieved in community detection algorithms in Network Theory. We thus used Louvain Community Detection as a proxy for an agent's innate identity categorization algorithm that it applies to categorize itself and other agents into discrete identities. In the model variants with identity dynamics, the agent only listens to other agents that are perceived to be within its own identity group, besides satisfying the modified HK criterion.

2.11 Our implementation of identity is parametrized by the experimental variable, SPIRO - the Salience of Proximity in Identity-Relevant Opinions. The value of SPIRO controls the granularity of the discovered identity groups by filtering out agents that have too few neighbours that are similar in opinion to themselves, prior to Louvain Community Detection. Higher values of SPIRO lead to detection of more number of more tightly clustered identity groups.

2.12 While we were interested in studying the influence of diversity in perception of identity groups, having every agent perform an Louvain Community Detection on the entire opinion space at every time step would be computationally very expensive. We therefore constrained the possible values of SPIRO to a fixed set of 8 values. In the model variants where we allow SPIRO to vary across agents, at the beginning of the simulation each agent is assigned one of the eight possible SPIRO values. In this way we could study the influence of diverse perceptions of identity in the same simulation while keeping the number of Louvain Community Detection passes low.

Model Variants

2.13 We studied the influence of identity in the presence of other important variables by incrementally building towards the richest variant of the model which has variable identity perception. A description of the models is as follows:

Deterministic Start HK Model (DHK) Initial distribution of opinions is evenly spaced. All agents have the same openness ε. Agents have individual values for conformity (α_i), parameterized at the population level with clipped normal distribution[1] parameters μ_α (mean) and σ_α (standard deviation). No identity dynamics.

Randomized Start HK Model (RHK) Initial distribution of opinions is uniformly distributed. All agents have the same openness ε. All agents have the same openness ε. Agents have individual values for conformity (α_i), parameterized at the population level with clipped normal distribution[1] parameters μ_α (mean) and σ_α (standard deviation). No identity dynamics.

Heterogeneous Boundary Model (VB) Initial distribution of opinions is done as in DHK in a half of simulations and as in RHK in the other half. Agents have individual values for openness (ε_i), parameterized at the population level with clipped normal distribution[1] parameters μ_ε (mean) and σ_ε (standard deviation). Conformity is distributed in the same fashion. No identity dynamics.

Heterogeneous Boundary with Identity (VBI) Initial distribution of opinions is done as in DHK in a half of simulations and as in RHK in the other half. Agents have individual values for openness (ε_i for the i^{th} agent), parameterized at the population level with clipped normal distribution[1] parameters μ_ε (mean) and σ_ε (standard deviation). Conformity is distributed in the same fashion. Identity dynamics are present but homogeneous, i.e. all agents have a common SPIRO.

Heterogeneous Boundary with Heterogeneous Identity (VBVI) Initial distribution of opinions is done as in DHK in a half of simulations and as RHK in the other half. Agents have individual values for openness (ε_i for the i^{th} agent), parameterized at the population level with clipped normal distribution[1] parameters μ_ε (mean) and σ_ε (standard deviation). Conformity is distributed in the same fashion. Identity dynamics are present and heterogeneous - agents have variable SPIRO, parameterized at the population level with μ_{SPIRO} (mean) and σ_{SPIRO} (standard deviation) [2].

[1] Values drawn from beyond the variable's range are resampled

[2] To reduce the number of Louvain Community Detection passes per time step, the drawn sample is replaced by its closest value from a set of 8 possible values of $\{0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85\}$

Experimental Design

2.14 In all we ran 2,504,964 simulations in NetLogo spanning the five model variants and different combinations of experimental variables. Every simulation has either 100 or 101 agents, depending on the value of the evenness variable (described below). Each simulation was run for a maximum of 365 time steps. In case consensus was reached before 365 time steps, the simulation was terminated at that point. We used 60 random seeds for each combination of parameters using NetLogo's 6.3.0.

Independent Variables

2.15 We had 8 IV's in all: $\mu_\alpha, \sigma_\alpha, \mu_\varepsilon, \sigma_\varepsilon, \mu_{SPIRO}, \sigma_{SPIRO}$, Evenness of population size, Randomness of initial opinion distribution. Not all IV's are applicable to all models, although all of them are applicable to the VBVI Model.

- μ_α**:** Mean conformity
- σ_α**:** Standard deviation of conformity
- μ_ε**:** Mean openness
- σ_α**:** Standard deviation of openness
- μ_{SPIRO}**:** Mean SPIRO
- σ_{SPIRO}**:** Standard deviation of SPIRO
- **Evenness of population size:** Whether the population size is 100 (even/yes) or 101 (odd/no)
- **Randomness of initial opinion distribution (Random_Start?):** Whether the initial opinion distribution was random-uniformly drawn from [-1, 1], or evenly spaced in [-1, 1]

Dependent Measures

2.16 We used three dependent measures to get aggregate characterizations of the opinion space after each simulation.

Equal-Size Binary Grouping (ESBG) Polarization: We adapted the the ESBG algorithm by (Tang et al. 2022) which provides a computational logic for measuring polarization in a real-valued opinion space. The value of ESBG polarization is 1 when there are exactly two perfectly tight equally-sized clusters of agents in the opinion space that are as far away from each other in opinion as possible, and zero when there is perfect consensus among all agents anywhere in the opinion space. Details of implementation in our simulation can be found in (Kalvas et al. 2023). Given that ESBG is non-linearly related to some other measures of consensus eg, diversity, we conduct our analyses after decomposing our experimental variables into categorical factors to avoid problems with assumptions about monotonicity.

Extremeness: Extremeness is the summed magnitudes of all opinions from the centre of the opinion space.

Diversity: Diversity is the overall standard deviation of the opinions, ie the square-root of mean squared distances from the mean opinion.

Results

Identity worsens polarization, but heterogeneity keeps it in check

3.1 A negative relationship between average agent openness and ESBG polarization was expected, and is consistently found across all our models (Figure 1). HK models with homogenous ε of greater than approximately 0.2 tend towards consensus (Lorenz 2007), which corresponds to an ESBG polarization value of 0. Our models showed the same trend, but the strength of the relationship varies across the models. Our basic model with randomized initial opinions (RHK) maintains higher polarization on average in comparison to the model with heterogenous openness (VB). This is consistent with prior work that suggested a strong consensus-driving effect of variance in open-mindedness (Lorenz 2009).

3.2 Both models with identity (VBI and VBVI) maintain significantly higher polarization levels for higher average openness than VB, displaying the polarizing role of identity. Taking grand averages across all simulations conducted for each model, we observe that mean polarization across models varies as per VB < DHK < RHK < VBVI < VBI (Table 1). This is consistent with our general conclusion that identity drives polarization, but heterogeneity in openness has a mitigating effect.

3.3 In Kalvas et al. (2023) we performed a multiple regression to study the relationship between our different experimental variables and ESBG polarization. Here we take our analysis one step further by performing a step-wise regression to help identify the relative importance of our experimental variables in our richest dynamical model, VBVI (Table 2)[3]. We coded our independent variables as categorical variables to avoid making assumptions about monotonicity. We observe that polarization has a negative relationship with all levels of μ_ε, and σ_ε, and a positive relationship with all levels of μ_{SPIRO}. We will di scuss in more detail our interpretation of the influence of these variables in the coming sections. We also report weak or absent effects of our other experimental variables - $\sigma_{SPIRO}, \mu_\alpha, \sigma_\alpha$, evenness of population size, and non-randomness of the initial opinion distribution.

3.4 The step-wise regression helped us understand the explanatory power of our experimental variables. Considering only main effects, we found that including $\mu_\varepsilon, \sigma_\varepsilon, \mu_{SPIRO}. \sigma_{SPIRO}, \mu_\alpha, \sigma_\alpha$, evenness of population size, and non-randomness of the initial opinion distribution in the regression model in total explained about 22.4% of the variance in ESBG polarization at the end of each simulation. Including only μ_α, σ_α, evenness of population size, and non-randomness of the initial opinion distribution as independent variables in the regression explained just 0.4% of the variance but adding μ_ε raised the explained variance to 4.9%. Adding σ_ε to this model gives us the largest increase in R^2 in our step-wise regression model, while further adding μ_{SPIRO} gives us the second largest jump in R^2, corresponding to explained variance of about 16.3% and 22% respectively. Adding σ_{SPIRO} to this model weakly raises the explained variance to 22.4% in the full regression model with only main effects.

Table 1: Descriptive statistics for ESBG polarization at end of each simulation by model

Model	N	Min	Max	IQR	Median	Mean	SD	SE	CI
DHK	84	0	0.419	0.361	0.282	0.199	0.177	0.019	0.038
RHK	5040	0	0.534	0.371	0.304	0.242	0.167	0.002	0.005
VB	80640	0	0.872	0.251	0.026	0.114	0.157	0.001	0.001
VBI	483840	0	0.937	0.154	0.408	0.378	0.177	0.000	0.000
VBVI	1935360	0	0.940	0.208	0.405	0.354	0.195	0.000	0.000

Heterogeneous perceptions of identity have only weak effects on the dynamics

3.5 We introduced variability in perception of social identity in our model, which follows from Self-Categorization Theory. While we found a weak relationship between σ_{SPIRO} and polarization in our multiple regression (Table 2), the computational cost of introducing multiple runs of community detection in every time-step would make heterogeneity in SPIRO an unreasonably expensive inclusion to our model unless it strongly influenced dynamics in some way. The weak effects of σ_{SPIRO} on the dynamics of our model are also apparent from observing that the effects of μ_ε and μ_{SPIRO} on ESBG (Figure 1) are qualitatively similar in VBI and VBVI.

SPIRO modulates the influence of open-mindedness on polarization

3.6 Open-mindedness naturally promotes consensus and mitigates polarization, as is consistently observed in our data. However, our implementation of social identity assumes an additional layer of closure as a consequence of opinion-dependant identity perceptions. In model VBI this translates to the population being divided at any time into mutually closed identity groups that do not communicated with one another irrespective of how open-minded the constituent agents are. The corresponding dynamics in VBVI are slightly more complicated because an agent can perceive itself to be in the same identity group as another agent that perceives the former agent to be in a different group. However, in both models we expected SPIRO to play a modulatory role in the

[3]For this analysis we reduced the parameter space along μ_ε to have 5 values evenly distributed between the minimum and maximum μ_ε instead of the 21 possible values used in the simulations.

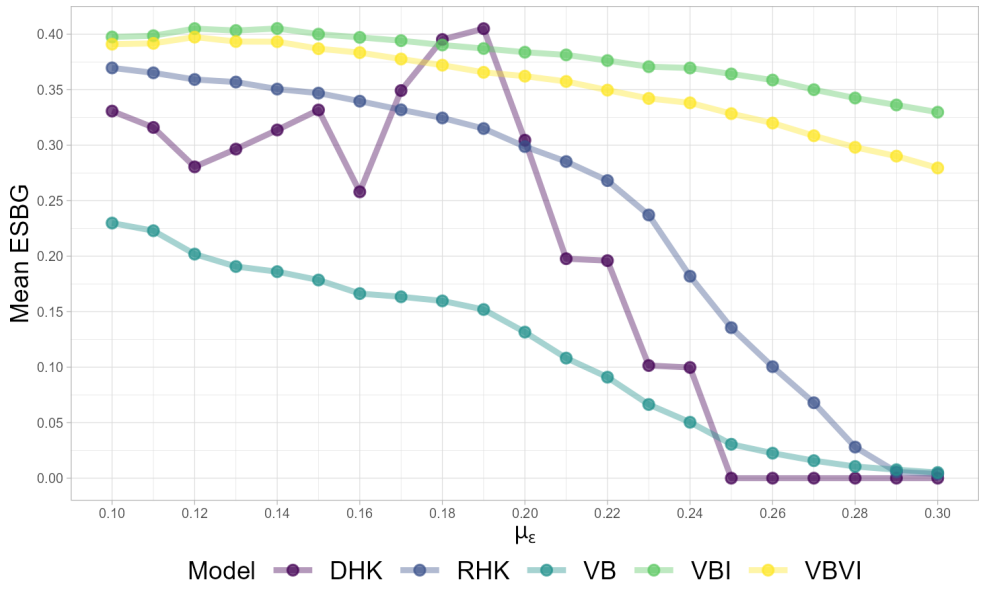

Figure 1: Average Polarization plotted against μ_ε for all models. Every data point is pooled across all combinations of other experimental variables.

relationship between ε and ESBG because of the additional communication barriers introduced by having more identity groups.

3.7 We observe that in the parameter region studied, the dependency of ESBG on ε is weakest for the highest values of μ_{SPIRO} (Figure 2). The system is likely to be fractured at the end of the simulation in this parameter region, with many tight opinion clusters that arise because of the emergence of a large number of small identity groups. Thus as one increases μ_{SPIRO}, the system dynamics can be thought of as transitioning from opinion-driven to identity-driven.

Polarization, Consensus, Fracturing, and Transition Regions

3.8 The opinion space at the end of the simulation tended to be in consensus, polarization, or in a fractured state with several tight opinion clusters. Figure 2 shows the variation of average ESBG at the end of VBI simulation runs with μ_{SPIRO}, μ_ε, and σ_ε. A consistent finding in both our models with identity is that polarization is related in a non-linear fashion with μ_{SPIRO} - polarization initially increases with μ_{SPIRO} and peaks for simulations with μ_{SPIRO} at 0.61, but decreases and saturates for the highest SPIRO values. This is to be expected as higher SPIRO values would lead to more number of tighter identity clusters that are closed to other groups. Moderate SPIRO values thus are most likely to support a bi-polarized opinion space, while the highest values would fracture the opinion space into many clusters. This is also supported by the observations that simulations run in the highest SPIRO values were the least sensitive to μ_ε and σ_ε (Figure 2), and that unlike ESBG, neither extremeness nor diversity - measures which do not privilege bi-polarized states over fractured states - decreased at the highest values (Figures 4 and 5).

3.9 The consensus-driving effect of higher σ_ε values can be seen in Figure 2. Despite the consensus region expanding with higher σ_ε, the system is robustly fractured at the highest two μ_{SPIRO} values. Higher σ_ε values also appear to weaken the dependency of polarization (and diversity and extremeness; Figures 4, 5) on μ_ε. For the highest σ_ε value of 0.15, the state of the system appears much more strongly driven by μ_{SPIRO} than by μ_ε.

3.10 Observing the variance in ESBG at the end of the simulation across different initial conditions reveals that system dynamics are more predictable in some regions of the parameter space than in others (Figure 3). The high variance regions typically occur at intermediate parameter values between high and low polarization causing regions (Figures 2 and 3). For example when $\mu_{SPIRO} = \sigma_\varepsilon = 0$, the higher variance region (Figure 3) occurs at a μ_ε of about 0.23, which is where the ESBG mean indicates a shift to consensus (Figure 2). These can be thought of as transition regions where system behavior sensitive to initial conditions because of shifting dynamics as one moves through the parameter space. Figure 3 suggests a widening in the transition region along the σ_ε axis in the parameter region being studied. Interestingly, we do not observe transition regions close to

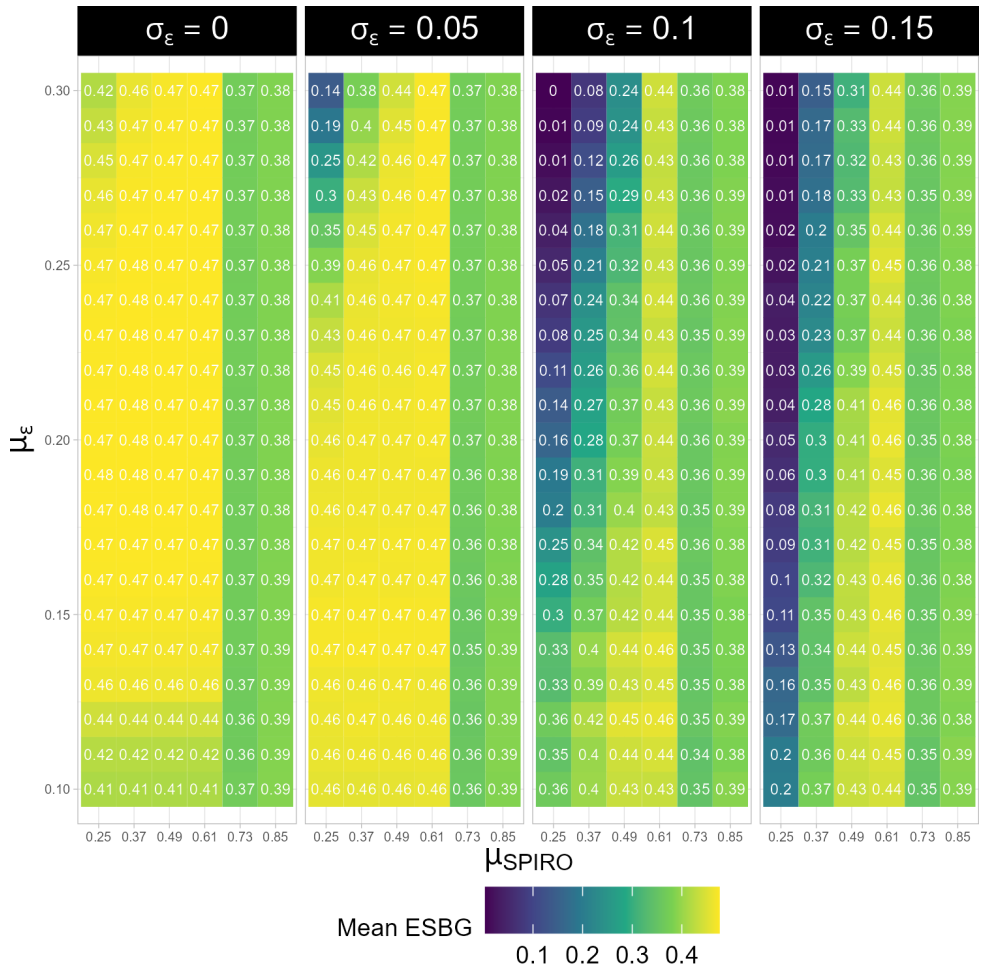

Figure 2: Average polarization across simulations in VBI as a function of experimental variables. Each panel is for a value of σ_ε, and ESBG mean is represented on a color scale for specific values of μ_{SPIRO} and μ_ε within each panel

the highest μ_{SPIRO} values as the system dynamics changes from polarized to fractured, suggesting that the system changes sharply as one goes from moderate to high SPIRO values.

Discussion

4.1 To our knowledge our work extending the HK model to include social identity group effects is the first study to integrate the Social Identity Approach with a bounded-confidence model of opinion dynamics. In Kalvas et al. (2023) we provided a technical description of the model and covered some basic results on how our experimental variables influence polarization. In this article we placed our model in the context of the social psychology literature on social identity, and discuss the results of our model in detail, and help lay the foundation of future work which aims to test ideas from social science and communication sciences via our extended HK model.

4.2 We analyzed our HK model with emergent identity dynamics by examining the effect of our experimental variables on the final state of the system. Specifically, we studied how different parametrizations of open-mindedness, identity detection, and conformity led to different opinion distributions characterized by their ESBG polarization - a metric that emphasizes bi-polarized and distant opinion clusters.

4.3 Our key findings can be summarized as follows: firstly we find that including identity dynamics in the model dramatically increases polarization. Secondly, SPIRO, which is our operationalization for an agent's sensitivity to opinion clustering while assigning identity groups, modulates the influence of open-mindedness on polarization. In simulations where agents were highly sensitive and perceived many identity clusters, open-mindedness did not exert as strong a negative influence on polarization. Finally, among all our experimental variables polarization was most strongly predicted by the heterogeneity in open-mindedness.

Table 2: Step-wise linear regression (N=460,800)

	ESBG				
	(1)	(2)	(3)	(4)	(5)
SPIRO_STD0.05					-0.015^{***}
					(0.001)
SPIRO_STD0.1					-0.027^{***}
					(0.001)
SPIRO_STD0.15					-0.036^{***}
					(0.001)
SPIRO_Mean0.37				0.057^{***}	0.057^{***}
				(0.001)	(0.001)
SPIRO_Mean0.49				0.110^{***}	0.110^{***}
				(0.001)	(0.001)
SPIRO_Mean0.61				0.150^{***}	0.150^{***}
				(0.001)	(0.001)
SPIRO_Mean0.73				0.076^{***}	0.076^{***}
				(0.001)	(0.001)
SPIRO_Mean0.85				0.091^{***}	0.091^{***}
				(0.001)	(0.001)
Boundary_STD0.05			-0.021^{***}	-0.021^{***}	-0.021^{***}
			(0.001)	(0.001)	(0.001)
Boundary_STD0.1			-0.129^{***}	-0.129^{***}	-0.129^{***}
			(0.001)	(0.001)	(0.001)
Boundary_STD0.15			-0.151^{***}	-0.151^{***}	-0.151^{***}
			(0.001)	(0.001)	(0.001)
Boundary0.15		-0.004^{***}	-0.004^{***}	-0.004^{***}	-0.004^{***}
		(0.001)	(0.001)	(0.001)	(0.001)
Boundary0.2		-0.029^{***}	-0.029^{***}	-0.029^{***}	-0.029^{***}
		(0.001)	(0.001)	(0.001)	(0.001)
Boundary0.25		-0.063^{***}	-0.063^{***}	-0.063^{***}	-0.063^{***}
		(0.001)	(0.001)	(0.001)	(0.001)
Boundary0.3		-0.112^{***}	-0.112^{***}	-0.112^{***}	-0.112^{***}
		(0.001)	(0.001)	(0.001)	(0.001)
HK_distribution	0.022^{***}	0.022^{***}	0.022^{***}	0.022^{***}	0.022^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N101	-0.008^{***}	-0.008^{***}	-0.008^{***}	-0.008^{***}	-0.008^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Conformity_STD0.1	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Conformity0.8	-0.005^{***}	-0.005^{***}	-0.005^{***}	-0.005^{***}	-0.005^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
R^2	0.004	0.049	0.163	0.220	0.224
Note:					$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Doi: 10.18564/jasss.xxxx

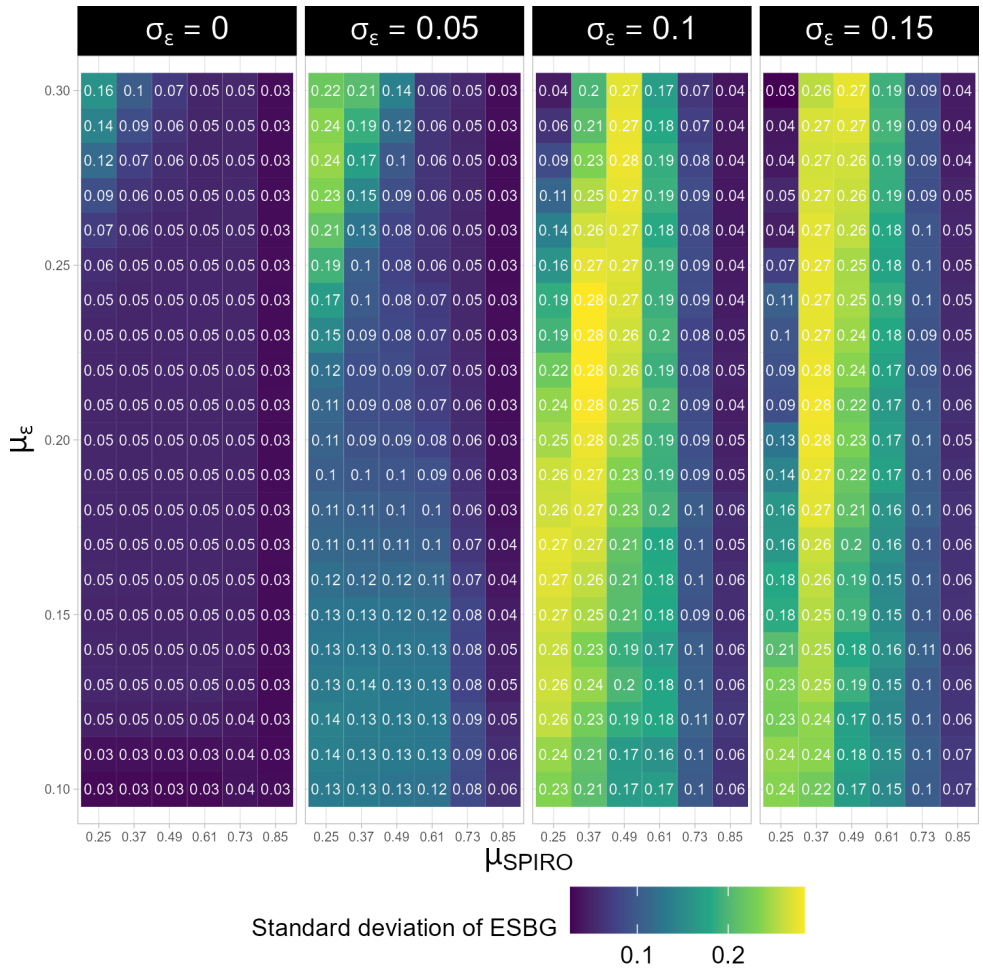

Figure 3: Standard Deviation of polarization across simulations as a function of experimental variables. Each panel is for a value of σ_ε, and ESBG standard deviation is represented on a color scale for specific values of μ_{SPIRO} and μ_ε within each panel

4.4 Our results add to a body of work that demonstrate that it is possible for bi-polarization to emerge even without repulsion between dissimilar individuals (Mäs & Flache 2013; Kurahashi-Nakamura et al. 2016; Flache et al. 2017). The implementation of the meta-contrast principle of Self-Categorization Theory in our model took into account dissimilarity only while assigning group identities, but did not assume that these assignments manifest as repulsion or negative influence. The additional communication barriers introduced due to social identity are sufficient to drive polarization as well as formation of multiple clusters or a "fracturing" of the opinion space.

4.5 We also looked at how varying our identity parameter SPIRO influences polarization. From our color plots it is evident that SPIRO modulates the influence of ε, which is the main parameter of the HK model, on the system dynamics. Since higher SPIRO values lead to a larger number of identity groups being detected, the highest SPIRO values fracture the opinion space beyond bi-polarization independently of the average ε.

4.6 This effect of higher SPIRO values driving the system to behave more in accordance with identity dynamics can be seen as paralleling the sociological phenomenon of deindividuation (S. D. Reicher & Postmes 1995) in the context of social identity, but with the key difference that the extent of deindividuation is not modulated by social context. With this in mind, we recognize that future iterations of our model ought to include elements representing social context, so that the balance between interpersonal and identity-driven dynamics is more realistically mediated by the environment.

4.7 We also find it interesting that modeling social identity as an observer-dependent phenomenon did not exert as much influence on the dynamics as we had expected. This can be seen from both our regression results and the aggregate behavior of our models (Figure 1) which suggests qualitatively similar model behavior between VBI and VBVI. While our stepwise regression revealed a significant effect of σ_{SPIRO}, the magnitude of this effect is far weaker than our expectation. This is particularly noteworthy in recognition of the fact that letting different agents use different criteria to determine co-members is computationally very expensive. However,

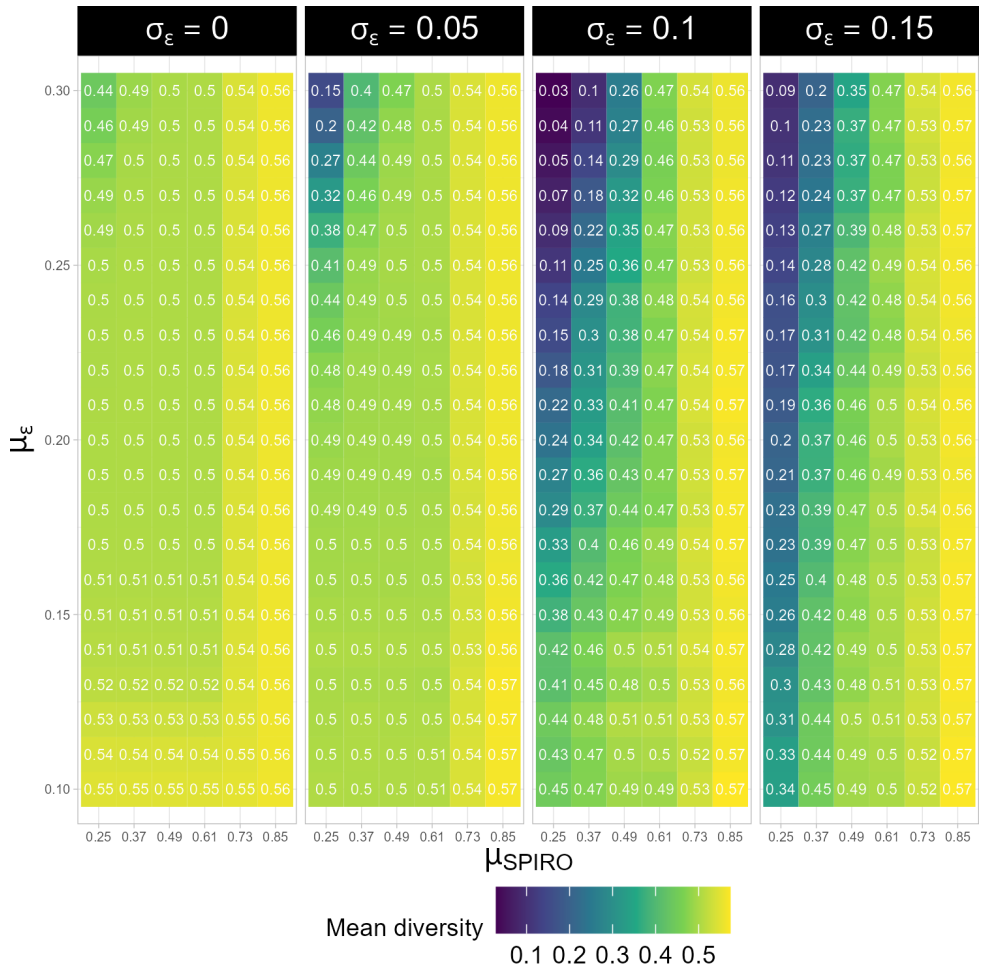

Figure 4: Average diversity at the end of the simulation across simulations as a function of experimental variables. Each panel is for a value of σ_ε, and ESBG mean is represented on a color scale for specific values of μ_{SPIRO} and μ_ε within each panel

we also recognize the need for caution with this inference as there may be subtle effects of letting social identity memberships to vary across observers that our macro analysis may not capture.

4.8 Despite the presence of identity dynamics which clearly drive polarization, the strongest predictor of polarization in our full model is the heterogeneity in open-mindedness. The strong consensus-driving effect of variance in ε has previously been studied by Lorenz (2009). It was very interesting to us to observe that the strength of this effect was even higher than that of our identity variables.

4.9 The effect of heterogeneity on alleviating polarization can be understood through at least two computational possibilities. Firstly, having high variance in naturalistic distributions of openness such as in our model means that in any given simulation there will likely be some outlier agents with high open-mindedness that can act as 'bridging agents'. Essentially open-minded agents may more strongly influence dynamics than close-minded agents, and thus exert a disproportionate consensus-driving effect. This speculated mechanism is also consistent with the observation that moderately high heterogeneity makes the system behave more sensitively to its initial conditions (Figure 3). Specifically, in any given simulation the initial opinion positions and corresponding sampled ε values would determine the effectiveness of bridging agents. This effect also predicts that increasing heterogeneity even further ought to collapse the transition region since having more outliers should make it more likely that there are enough bridging agents to promote consensus.

4.10 A distinct and more intriguing effect of heterogeneity however was detected in Jan Lorenz's work (Lorenz 2009), which effectively showed that heterogeneity can promote consensus even in the absence of extraordinarily open-minded agents. This effect is emergent from the complex temporal dynamics driven by the interplay between agents with different ε values. Put perhaps a little simplistically, open- and close-minded agents appear to play distinct roles in promoting consensus, and thus having the presence of both types of agents can drastically drive consensus even in comparison to scenarios with only open-minded agents. While our model does not

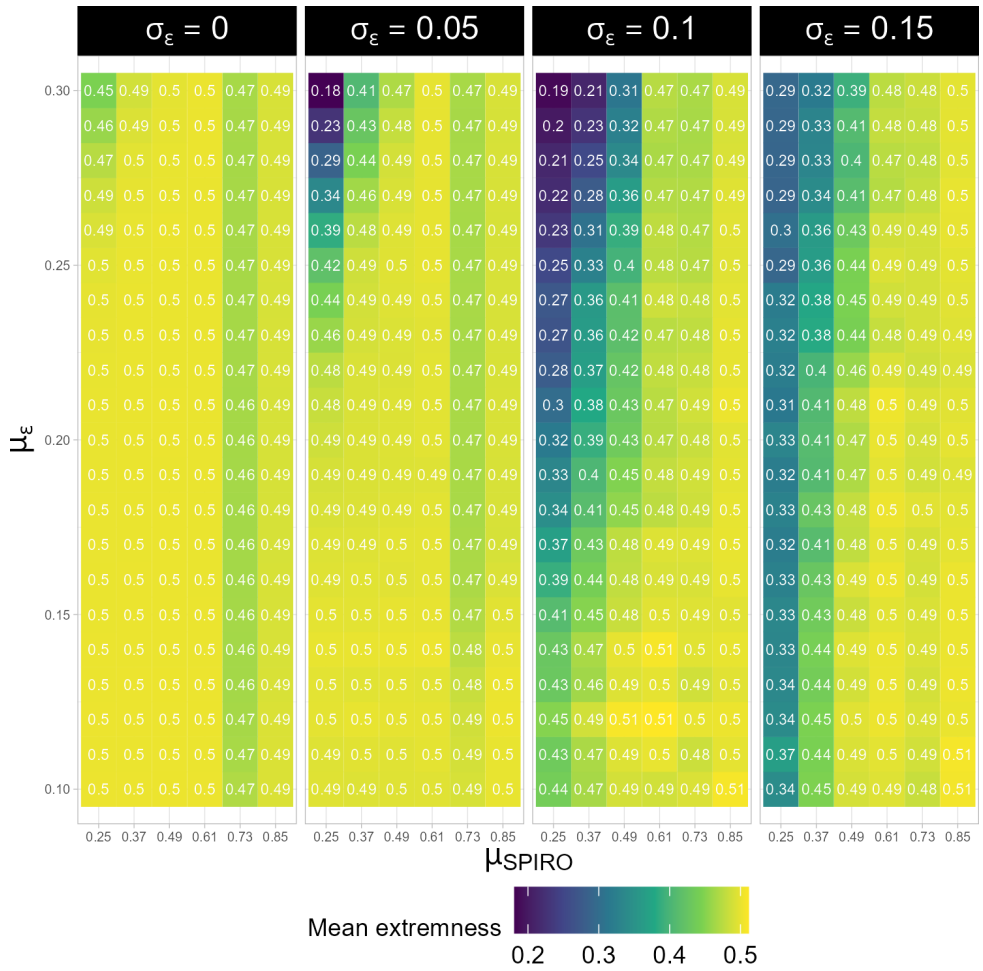

Figure 5: Average extremeness at the end of the simulation across simulations as a function of experimental variables. Each panel is for a value of σ_ε, and ESBG mean is represented on a color scale for specific values of μ_{SPIRO} and μ_ε within each panel

directly lend itself to such argumentation owing to our choice of adopting a naturalistic distribution rather than a few pre-determined values for openness, we expect that a similarly subtle and complex interplay between the influence of agents with different ε values may go further in explaining the strong effect of heterogeneity in our model.

Model characteristics

4.11 Our richest model VBVI captures the dynamics of social identity in the context of opinion formation and dissemination mediated by social influence. We take a classical Hegselmann-Krause bounded confidence model and include emergent opinion-dependant social identity dynamics between agents such that agents are only influenced by others that satisfy an opinion-based criteria as well as an identity-based criteria. The opinion-based criteria refers to the classical HK rule that thresholds opinion distances, while the identity criteria requires the speaking agent to be perceived as being in the same identity group as the listener. Moreover we assume naturalistic distributions for agent variables such as conformity, openness, and SPIRO.

4.12 In contrast with some of the existing ABM formulations of social identity in literature, we make simplified assumptions about the origin of social identity by choosing not to model all of the complex psychological processes that give rise to social identification such as social context, normative fit, accessibility, and the interplay of personal as opposed to social identity (Wijermans et al. 2023). Instead, we assume that identification can occur on the basis of opinions alone, and via a mechanism that conceptually parallels the idea of comparative fit, specifically the principle of meta-contrast. This was an appropriate modeling choice for our purposes, since the main variables in our model are also the main variables of interest to us, namely openness, social identity,

and the associated distribution parameters. That is of course not to say that these aforementioned features of SIA that our model ignores or takes for granted are not important. Through simulations we were also able to establish some of our other model parameters - mean and variance of conformity, and the variance of SPIRO - did not have a strong influence on dynamics, which allows us to focus our efforts on studying our aforementioned variables of interest.

4.13 While most ABM's of social identity in the context of opinion dynamics assume static and often *a priori* identity assignments (Scholz et al. 2023), the dynamic social identity component in our identity models operates by assigning identity group tags to agents based on clustering patterns in the opinion space at every time step. Moreover, we assumed observer-dependent identity group assignments, consistent with the theoretical formulation of SCT which has always implied that group membership is perceived from the perspective of an individual: '*A.7.1 ... any collection of stimuli is more likely to be categorized as an entity... to the degree that the differences between those stimuli on relevant dimensions... are perceived as less than the differences between that collection and other stimuli*' (Turner et al. 1987).

Model limitations

4.14 Here we discuss some aspects of SIA or opinion dynamics that are either not explicitly modeled or not fully captured by our models.

Identity Salience

4.15 An important construct in SIA is Salience - which determines which of the multiple possible identities associated with an individual influence their behavior most strongly at a given time. Salience is dependant on social and psychological context at any given time, such that the most salient identity can rapidly change for an individual, leading to shifts in behavior mirroring the shift in their currently most salient social identity.

4.16 Although SPIRO is conceptually related to Salience in SIA, it is a distinct concept. An agent with a higher SPIRO is likely to perceive a smaller, more fine-grained identity group and behave according to that characterization of the opinion space, analogous as per SCT to how an individual can be influenced most by one of their least populous, most immediate identity group when this identity becomes salient. The "Proximity" in the name "Salience of Proximity in Identity Relevant Opinions" refers to proximity in the opinion space. While our formulation of SPIRO hinged on our assumption that identity groups are formed based on opinion alone, conceptually SPIRO is generalizable to identity group assignments on the basis of any other continuous agent characteristic. In this study we modeled SPIRO as a static characteristic for an individual, and looked at the effect of varying SPIRO across different agents. One possibility for future work is to integrate our implementation of SPIRO with an explicit model of canonical salience, by modeling a dynamic social environment that cause SPIRO values to vary within an individual agent.

Prototypicality and Prototypical behaviors

4.17 According to SIA, one way through which social identity groups exert their influence on social dynamics is through prototypical behaviors or norms that characterize the group. One way to operationalize the influence exerted by a social identity group is to define a prototypical group characteristic - such as an average opinion - which is explicitly computed and adopted by identity group members. In our identity models we do not explicitly model prototypical opinions of an identity group - rather identity groups have aggregate characteristics that are dynamic but do not exert an influence of their own. Rather, identity groups function via their constituent agents that discriminate between in-group and out-group members in their self-categories. Moreover as our identity group assignments are based on opinions, their aggregate characteristics are also limited to the space of opinions. A richer model is conceivable where identity groups are not only assigned dynamically, but also exert their influence on the opinion space through explicitly defined prototypical opinions, or by having other prototypical identity group characteristics - or norms - such as a prototypical openness value. We discuss this possibility in the next subsection regarding future work.

Future Directions

4.18 As mentioned in the first section, we are aiming to build towards theories in communication research that can expand our understanding of polarization in the context of modern technology. such as the Reinforcing Spirals

Model (RSM) (Slater 2007). We are also interested in building models that can also produce system dynamics expected from more classical communication theories such as the Spiral of Silence (SoS) (Noelle-Neumann 1974). Simulating the behaviors expected to be observed by these theories would require us to build a more rich model of the environment the agents are embedded in. While we have currently assumed a complete network, as is common in simulation studies with bounded-confidence models, we will in the future incorporate geographical constraints, differential broadcasting power, and traditional and social media agents into our model.

4.19 RSM also defines characteristic norms, especially with regards to openness and closedness of identity groups. Although we define the distribution of the openness of the population, the average openness of identity groups is an endogenous, emergent variable. In our model openness is a stable attribute of individuals that are not influenced by the average openness of their social groups. However, there are at least two possible directions future work can take to study the openness of identity groups. Firstly, an *effective openness* can be computed for any agent sub-population which can be taken as the ratio of average number of communication partners in the sub-population to expected number of communication partners. Secondly, while we don't explicitly impose constraints on the openness of agents in specific identity groups, it is still possible that the emergent groups may differ in their openness because of the dynamics of the model.

4.20 Our simulation work presented in this article aims to bring us closer to operationalizing these conceptual models into computational models. Formalizng and simulating these conceptual models can deepen their understanding since we may encounter unexpected behaviors and emergent phenomena that would have been hard to predict without simulations.

References

Apergis, N. & Pinar, M. (2023). Corruption and partisan polarization: evidence from the European Union. *Empirical Economics*, *64*(1), 277–301. doi:10.1007/s00181-022-02247-z

Asch, S. E. (1955). Opinions and social pressure. *Scientific American*, *193*(5), 31–35

Bak-Coleman, J. B., Alfano, M., Barfuss, W., Bergstrom, C. T., Centeno, M. A., Couzin, I. D., Donges, J. F., Galesic, M., Gersick, A. S., Jacquet, J., Kao, A. B., Moran, R. E., Romanczuk, P., Rubenstein, D. I., Tombak, K. J., Bavel, J. J. V. & Weber, E. U. (2021). Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, *118*(27), e2025764118. doi:10.1073/pnas.2025764118

Bassi, A., Morton, R. B. & Williams, K. C. (2011). The effects of identities, incentives, and information on voting. *J. Polit.*, *73*(2), 558–571

Bauer, P. C. (2019). Working paper conceptualizing and measuring polarization: A review. https://files.osf.io/v1/resources/e5vp8/providers/osfstorage/5d7b7939710c95001c598c03?format=pdfaction=downloaddirec Accessed: 2023-11-10

Bessi, A. & Ferrara, E. (2016). Social bots distort the 2016 u.s. presidential election online discussion. *First Monday*, *21*(11). doi:10.5210/fm.v21i11.7090

Billig, M. & Tajfel, H. (1973). Social categorization and similarity in intergroup behaviour. *Eur. J. Soc. Psychol.*, *3*(1), 27–52

Biondi, E., Boldrini, C., Passarella, A. & Conti, M. (2023). Dynamics of opinion polarization. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *53*(9), 5381–5392. doi:10.1109/tsmc.2023.3268758

Bliuc, A.-M., McGarty, C., Reynolds, K. & Muntele, D. (2007). Opinion-based group membership as a predictor of commitment to political action. *European Journal of Social Psychology*, *37*(1), 19–32. doi: https://doi.org/10.1002/ejsp.334

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.*, *2008*(10), P10008

Bolsen, T., Druckman, J. N. & Cook, F. L. (2014). The influence of partisan motivated reasoning on public opinion. *Political Behavior*, *36*(2), 235–262

Bolsen, T. & Shapiro, M. (2017). The us news media, polarization on climate change, and pathways to effective communication. *Environmental Communication*, *12*, 1–15. doi:10.1080/17524032.2017.1397039

Cikara, M., Van Bavel, J. J., Ingbretsen, Z. A. & Lau, T. (2017). Decoding “us” and “them”: Neural representations of generalized group concepts. *J. Exp. Psychol. Gen.*, *146*(5), 621–631

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *J. Pers. Soc. Psychol.*, *85*(5), 808–822

David, B. & Turner, J. C. (2001). Majority and minority influence: A single process self-categorization analysis. *Group consensus and minority influence: Implications for innovation.*, *324*, 91–121

Diep, H. T., Kaufman, M. & Kaufman, S. (2023). An agent-based statistical physics model for political polarization: A monte carlo study. *Entropy*, *25*(7). doi:10.3390/e25070981

Dow, B. J., Johnson, A. L., Wang, C. S., Whitson, J. & Menon, T. (2021). The covid-19 pandemic and the search for structure: Social media and conspiracy theories. *Social and Personality Psychology Compass*, *15*(9), e12636. doi:https://doi.org/10.1111/spc3.12636

Dow, B. J., Wang, C. S. & Whitson, J. A. (2023). Support for leaders who use conspiratorial rhetoric: The role of personal control and political identity. *Journal of Experimental Social Psychology*, *104*, 104403. doi: https://doi.org/10.1016/j.jesp.2022.104403

Downey, D. J. (2022). Polarization and persuasion: Engaging sociology in the moral universe of a divided democracy. *Sociological Perspectives*, *65*(6), 1029–1051. doi:10.1177/07311214221124443

Druckman, J. N. & Bolsen, T. (2011). Framing, Motivated Reasoning, and Opinions About Emergent Technologies. *Journal of Communication*, *61*(4), 659–688. doi:10.1111/j.1460-2466.2011.01562.x

Esteban, J.-M. & Ray, D. (1994). On the measurement of polarization. *Econometrica*, *62*(4), 819–851

Flache, A. (2018). Between Monoculture and Cultural Polarization: Agent-based Models of the Interplay of Social Influence and Cultural Diversity. *Journal of Archaeological Method and Theory*, *25*(4), 996–1023. doi: 10.1007/s10816-018-9391-1

Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S. & Lorenz, J. (2017). Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, *20*(4), 2. doi: 10.18564/jasss.3521

Fu, G. & Zhang, W. (2014). Opinion dynamics of modified Hegselmann-Krause model with group-based bounded confidence. *IFAC Proceedings Volumes*, *47*(3), 9870–9874

Han, W., Huang, C. & Yang, J. (2019). Opinion clusters in a modified hegselmann–krause model with heterogeneous bounded confidences and stubbornness. *Physica A: Statistical Mechanics and its Applications*, *531*, 121791. doi:https://doi.org/10.1016/j.physa.2019.121791

Hegselmann, R. & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, *5*(3), 1–2

Heltzel, G. & Laurin, K. (2020). Polarization in america: two possible futures. *Current Opinion in Behavioral Sciences*, *34*, 179–184. doi:https://doi.org/10.1016/j.cobeha.2020.03.008. Political Ideologies

Hornsey, M. J. (2008). Social identity theory and self-categorization theory: A historical review. *Soc. Personal. Psychol. Compass*, *2*(1), 204–222

Jones, I., Wang, R., Han, J. & Liu, H. (2016). Community cores: Removing size bias from community detection. *ICWSM*, *10*(1), 603–606

Kahan, D. M., Hoffman, D. A., Braman, D., Evans, D. & Rachlinski, J. J. (2012). “they saw a protest”: Cognitive illiberalism and the speechconduct distinction. *Stanford Law Rev.*, *64*(4), 851–906

Kalvas, F., Ramaswamy, A. & Slater, M. (2023). *Identity Drives Polarization: Advancing the Hegselmann-Krause Model by Identity Groups*, (pp. 249–262). doi:10.1007/978-3-031-34920-1_20

Kim, S. & Kim, J. (2023). The information ecosystem of conspiracy theory: Examining the qanon narrative on facebook. *Proc. ACM Hum.-Comput. Interact.*, *7*(CSCW1). doi:10.1145/3579626

Kou, G., Zhao, Y., Peng, Y. & Shi, Y. (2012). Multi-level opinion dynamics under bounded confidence. *PLoS One*, *7*(9), e43507

Kubin, E. & von Sikorski, C. (2021). The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, *45*(3), 188–206. doi:10.1080/23808985.2021.1976070

Kurahashi-Nakamura, T., Mäs, M. & Lorenz, J. (2016). Robust clustering in generalized bounded confidence models. *Journal of Artificial Societies and Social Simulation*, *19*(4), 7. doi:10.18564/jasss.3220

Leeper, T. J. & Slothuus, R. (2014). Political parties, motivated reasoning, and public opinion formation. *Polit. Psychol.*, *35*, 129–156

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

Li, J. & Xiao, R. (2017). Agent-based modelling approach for multidimensional opinion polarization in collective behaviour. *Journal of Artificial Societies and Social Simulation*, *20*(2), 4. doi:10.18564/jasss.3385

Lorenz, J. (2007). Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C*, *18*, 1819–1838

Lorenz, J. (2009). Heterogeneous bounds of confidence: Meet, discuss and find consensus! *Complexity*, (pp. NA–NA)

Machimbarrena, J. M., Calvete, E., Fernández-González, L., Álvarez Bardón, A., Álvarez Fernández, L. & González-Cabrera, J. (2018). Internet risks: An overview of victimization in cyberbullying, cyber dating abuse, sexting, online grooming and problematic internet use. *International Journal of Environmental Research and Public Health*, *15*(11). doi:10.3390/ijerph15112471

Macy, M. W., Ma, M., Tabin, D. R., Gao, J. & Szymanski, B. K. (2021). Polarization and tipping points. *Proceedings of the National Academy of Sciences*, *118*(50), e2102144118. doi:10.1073/pnas.2102144118

Matamoros-Fernández, A. & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, *22*(2), 205–224. doi:10.1177/1527476420982230

Mladenović, M., Ošmjanski, V. & Stanković, S. V. (2021). Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. *ACM Comput. Surv.*, *54*(1). doi:10.1145/3424246

Montalvo, J. G. & Reynal-Querol, M. (2003). Religious polarization and economic development. *Economics Letters*, *80*(2), 201–210

Mäs, M. & Flache, A. (2013). Differentiation without distancing. explaining bi-polarization of opinions without negative influence. *PLOS ONE*, *8*(11), 1–17. doi:10.1371/journal.pone.0074516

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of communication*, *24*(2), 43–51

Petersen, M. B., Skov, M., Serritzlew, S. & Ramsøy, T. (2013). Motivated reasoning and political parties: Evidence for increased processing in the face of party cues. *Political Behavior*, *35*(4), 831–854

S. D. Reicher, R. S. & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology*, *6*(1), 161–198. doi:10.1080/14792779443000049

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

Scholz, G., Wijermans, N., Paolillo, R., Neumann, M., Masson, T., Chappin, Ã., Templeton, A. & Kocheril, G. (2023). Social agents? a systematic review of social identity formalizations. *Journal of Artificial Societies and Social Simulation*, *26*(2), 1–6

Schweitzer, F., Krivachy, T. & Garcia, D. (2020). An Agent-Based Model of Opinion Polarization Driven by Emotions. *Complexity*, *2020*, 5282035. doi:10.1155/2020/5282035. Publisher: Hindawi

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A. & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 4787. doi:10.1038/s41467-018-06930-7

Sherif, C. W., Kelly, M., Rodgers Jr., H. L., Sarup, G. & Tittler, B. I. (1973). Personal involvement, social judgment, and action. *Journal of Personality and Social Psychology*, *27*(3), 311–328. doi:10.1037/h0034948. Place: US Publisher: American Psychological Association

Sherif, M. (1936). The psychology of social norms. *https://psycnet.apa.org › recordhttps://psycnet.apa.org › record*, *210*

Sherif, M. & Hovland, C. I. (1961). *Social judgment: Assimilation and contrast effects in communication and attitude change*, vol. 218. Oxford, England: Yale Univer. Press Social judgment

Slater, M. D. (2007). Reinforcing spirals: The mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. *Communication Theory*, *17*(3), 281–303

Smith, C. T., Ratliff, K. A. & Nosek, B. A. (2012). Rapid assimilation: Automatically integrating new information with existing beliefs. *Soc. Cogn.*, *30*(2), 199–219

Smith, L. G. E., Thomas, E. F. & McGarty, C. (2015). "we must be the change we want to see in the world": Integrating norms and identities through social interaction. *Political Psychology*, *36*(5), 543–557. doi: https://doi.org/10.1111/pops.12180

Su, W., Gu, Y., Wang, S. & Yu, Y. (2017). Partial convergence of heterogeneous Hegselmann-Krause opinion dynamics. *Science China Technological Sciences*, *60*(9), 1433–1438. doi:10.1007/s11431-016-0615-x

Tajfel, H., Billig, M. G., Bundy, R. P. & Flament, C. (1971). Social categorization and intergroup behaviour. *Eur. J. Soc. Psychol.*, *1*(2), 149–178

Tajfel, H., Turner, J. C., Austin, W. G. & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, *56*(65), 9780203505984–9780203505916

Tang, T., Ghorbani, A., Squazzoni, F. & Chorus, C. G. (2022). Together alone: a group-based polarization measurement. *Quality & Quantity*, *56*(5), 3587–3619. doi:10.1007/s11135-021-01271-y

Tran, T., Valecha, R., Rad, P. & Rao, H. R. (2021). An Investigation of Misinformation Harms Related to Social Media during Two Humanitarian Crises. *Information Systems Frontiers*, *23*(4), 931–939. doi:10.1007/s10796-020-10088-3

Turner, J. C. (1975). Social comparison and social identity: Some prospects for intergroup behaviour. *Eur. J. Soc. Psychol.*, *5*(1), 1–34

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D. & Wetherell, M. S. (1987). Rediscovering the social group: A self-categorization theory. *https://psycnet.apa.org › recordhttps://psycnet.apa.org › record*, *239*

Van Bavel, J. J., Packer, D. J. & Cunningham, W. A. (2008). The neural substrates of in-group bias: a functional magnetic resonance imaging investigation. *Psychol. Sci.*, *19*(11), 1131–1139

Vosoughi, S., Roy, D. & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. doi:10.1126/science.aap9559

Weimer, C. W., Miller, J. O., Hill, R. R. & Hodson, D. D. (2022). An opinion dynamics model of meta-contrast with continuous social influence forces. *Physica A: Statistical Mechanics and its Applications*, *589*, 126617

Wijermans, N., Scholz, G., Neumann, M., Paolillo, R., Templeton, A., Netshandama, V. & Neuberger, D. (2023). Editorial: Social identity modelling. *Journal of Artificial Societies and Social Simulation*, *26*(3), 15. doi: 10.18564/jasss.5188

Wilson, A. E., Parker, V. A. & Feinberg, M. (2020). Polarization in the contemporary political and media landscape. *Current Opinion in Behavioral Sciences*, *34*, 223–228. doi:https://doi.org/10.1016/j.cobeha.2020.07.005. Political Ideologies