Title
Simulating Political Polarization as a Function of Uncertain Inference and Signaling of Moral Values

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Pedersen, Julie Maria Ejby
Moore, Adam

Publication Date
2023

Peer reviewed

Simulating Political Polarization as a Function of Uncertain Inference and Signaling of Moral Values

Julie M. E. Pedersen (s1917169@ed.ac.uk)
Department of Psychology, University of Edinburgh,
7 George Square, EH8 9JZ, Edinburgh, UK

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

Abstract

Political polarization is driven by many factors, but the role of moral values as both a signal of political identity and a source of internal conflict is understudied. We report an agent-based computational model of polarization that fills this gap. Agents seek to differentiate in- and outgroup neighbors with a slight preference for the former. However, they must do so by inferring neighbors' identities from visible but transient moral signals. Moreover, agents experience conflicts within their own values, and if difficult to resolve internally, can copy the values of their ingroup or disengage (i.e., act immorally). Results show that liberals form larger, more homogeneous clusters, are happier, and experience less moral conflict than conservatives. Conservatives experience more and higher levels of conflict and morally disengage significantly more often than liberals.

Keywords: agent-based model; morality; political polarization; uncertain inference

Introduction

The rise of ideological/political polarization (Doherty et al., 2022; Yudkin et al., 2019) has already led to multiple occasions of violence (e.g., the recent insurrection of the US Capitol) and threatens to disrupt political consensus and effective responses to serious societal threats such as climate change or the recent Covid-19 pandemic. Previous research has explored separate factors related to polarization at the individual level (e.g., cognitive limitations; Singer et al., 2019) and macro-level (e.g., homogeneity of social networks; Tokita et al., 2021). However, a formal model combining these could provide important implications for the understanding of polarization (Kashima et al., 2021; van Baar & FeldmanHall, 2021).

Agent-based modelling allows for such interactions between individual-level factors and macro-level phenomena to emerge while leveraging simple agent-level assumptions. Such formal models have been applied to polarization of political elites (Macy et al., 2021) and of voters (i.e., mass polarization; Axelrod et al., 2021), in both cases with agents motivated to *avoid dissimilar others* while se*eking out similar others* under some level of tolerance. This process can lead to a polarization point of no return, even in the face of seemingly unifying external threats (e.g., pandemics).

Most modelling of political polarization has assumed that agents know, or can correctly identify, their political in- and outgroup, or sources of information/media relevant to these groups (e.g., Theodoropoulos, 2020; Tokita et al., 2021). Yet this assumption is not necessarily realistic; most voters have a robust political *identity*, but they do not have a coherent political *ideology* (Kalmoe, 2020; see also, Macy et al., 2021). Rather, ideological content seems to accrue via a preference for ingroup members/content and a dislike for the outgroup (Baldassarri & Page, 2021; Iyengar et al., 2019). Thus, people must deal with two simultaneously interacting layers of uncertainty linked to their identity: first, when identifying in- and outgroup members and, second, when they infer the prototypical values of their ingroup from the expressed/observed values of other ingroup members. This process is vital, functioning to reduce subjective uncertainty about the self (cf., Hogg, 2000) and provide critical information about community norms. Indeed, when signaling their own political identity, people cannot always reliably link their political identity to specific, static policy positions. Instead, they tend to signal identity by expressing outrage at, or endorsement of, issues inferred to differentiate the in- and outgroups (Funkhouser, 2020) and identify these issues via shared values of ingroup members. Hence, uncertain inference regarding ingroup membership necessarily introduces uncertainty into the inference of the ingroup's core values. This also suggests that the perceived values of the political ingroup may, over time, influence an individual to change their own moral values to reaffirm their political identity, a process similar to how pluralistic ignorance can change social behaviors (cf., Prentice & Miller, 1996).

Empirical evidence seems to support the idea that moral values are a medium by which political identity is signaled and inferred. Liberals and conservatives have distinct patterns of moral concern according to Moral Foundations Theory (Graham et al., 2013), which posits five distinct domains of morality: care, fairness, ingroup loyalty, respect for authority, and sexual/spiritual purity. While conservatives rate each foundation as approximately equally important/relevant to them according to the Moral Foundations Questionnaire (MFQ), liberals tend to rank fairness and care highest, with lower consideration/emphasis given to respect for authority, ingroup loyalty, and purity, respectively (Graham et al., 2009, 2011; Milesi, 2016, 2017; van Leeuwen & Park, 2009). These distinctive moral value distributions could identify political ingroup members as a function of the similarity with one's own values. Indeed, the

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

use of moral-emotional language on Twitter increased the likelihood of political messages spreading through social networks (Brady et al., 2017). This spread of moral content was typically larger within political ingroup networks than between, although this was moderated by contextual factors such as the specific topic of the post. Thus, the dual drivers to signal one's group membership and to infer the group membership of others may push people to express moral values as a signal of political identity, but also to adjust their own moral values as a function of uncertain inference about what constitutes the 'true' values of their ingroup.

Moral Values Conflict and Polarization

Previous modelling work has explored some consequences of differing moral values on political polarization. Moral values can conflict not only between individuals but within a given individual confronted with a moral dilemma. Cognitive dissonance is the negative motivational state that arises from behaviours that are incongruent with one's beliefs, or when one simultaneously holds incompatible beliefs. It typically leads to changing beliefs about the world or one's own actions to minimize the dissonance (Festinger, 1962; Harmon-Jones & Mills, 2019). This includes a situation where different values from two (or more) moral foundations may prescribe opposite actions. Moral Disengagement Theory (e.g., Bandura, 1999) posits that one way people react to this is by cognitively reframing the situation to render conflicting moral considerations irrelevant. This reduces cognitive dissonance and allows people to act in ways that elide/avoid consideration of moral conflicts. Theodoropoulos (2020) demonstrated that, among agents attempting to self-select into interactive clusters of like-minded others (i.e., homophily; cf., Axelrod, 1997), the moral foundations typical for conservatives drove less clustering than those associated with liberals and higher rates of internal moral conflict and, thus, higher rates of moral disengagement on average compared to liberals. This conforms with findings that political conservatism and correlated measures such as right-wing authoritarianism and social dominance orientation are positively associated with the endorsement/use of moral disengagement strategies (Devereux et al., 2021; Jackson & Gaertner, 2010; Wilson & Collins, 2019).

However, this model did not account for uncertainty related to the individual-level processes of inferring ingroup identity of others or moral values inference/updating. In a sense, given that morality was static, it was a foregone conclusion that conservatives would experience more moral conflict and, therefore, disengage more. It remains unclear if polarisation (i.e., self-selected segregation into non-heterogeneous subgroups) will occur when identity must be inferred from noisy moral signals and each agent's moral values can shift over time to match their perceived ingroup.

The Present Research

The need to maintain socio-political identity is deeply rooted, but existing theoretical work on political polarization has not yet simultaneously addressed uncertainty in ingroup membership inference and inference regarding ingroup values, nor connected these to the need of individuals to align their moral values to perceived group norms. How these processes interact with more well-explored variables such as homophily and (in)tolerance of different others to drive polarization also remains unclear. Extending previous modelling efforts, we built an agent-based model to simulate these processes at the individual level of moral conflict, decision-making, signaling moral information, and (uncertain) inference of the political identity of adjacent agents while leveraging macro-level assumptions of homophily and tolerance for dissimilar others.

Model Definition

Our agent-based model consists of two stages: firstly, assigning agents initial properties and, secondly, iteratively updating these via a three-step algorithm of identifying ingroup members among adjacent agents via inference from moral signals; dealing with potential intra-agent moral conflict and subsequent updating of moral values; and determining potential movement from the outcomes of the previous steps. The model code and additional information are available here: osf.io/gye7t/.

Initial Agent Properties

N agents are initially randomly distributed in a 51x51 square space (one agent per 1x1 space; wrapped horizontally and vertically) with *N* determined by a density parameter (75% for the present study; see Table 1). Each agent is randomly assigned a fixed label for a binary political identity ('liberal' or 'conservative')[1], which determines the Gaussian distributions from which one weight for each moral foundation is randomly drawn. That is, an agent *i* has a set of five weights, MF weights or $W_{i,t}$ where *t* is the number of iterations, $|W_{i,t}| = 5$ and $\dot{w}_{i,m,t} \in W_{i,t}$ representing the weight for a moral foundation, *m*. The means and standard deviations of these distributions are derived from MFQ responses by self-identified liberals and conservatives (Graham et al., 2011). The weights and distributions are bounded between 0-5 preserving the scale of the original data. Secondly, an agent, *i*, has a set of up to eight agents on adjacent spaces, which varies by iteration, *t*, i.e., its neighborhood set: $A_{i,t}$ where $0 \leq |A_{i,t}| \leq 8$.

Updating Algorithm

The updating steps are interdependent, e.g., past outcomes of moral choices become moral signals taken as input during ingroup inference. The initial 10 iterations are, therefore,

[1] We have higher political identity resolution, but we focus on our findings for binary identity assignment here.

'burn-in' trials to supply agents with moral signals for social inference. Each subsequent iteration, each agent completes (1) an ingroup identification and social influence procedure; (2) a moral decision-making, conflict detection, and moral values updating procedure; and (3) a movement procedure.

Navigating a Social World. During the ingroup identification and social influence procedure, on iteration *t*, an agent, *i*, identifies a set of ingroup agents, $L_{i,t}$ where $0 \leq |L_{i,t}| \leq |A_{i,t}|$, from the set of neighbouring agents, $A_{i,t}$. Agent *i* iterates over each member of $A_{i,t}$, *a*, and extracts its set of observable moral signals, $C_{a,t}$ where $0 \leq |C_{a,t}| \leq 5$ and $a \in A_{i,t}$ (most recent moral choice outcomes; see Navigating a Moral World). Agent *i* computes the Levenshtein Distance (i.e., edit distance; see Doan et al., 2012) between this set and agent *i*'s moral signals set ($C_{i,t}$ where $0 \leq |C_{i,t}| \leq 5$), which we call $ld(C_{i,t}, C_{a,t})$. This is used to compute the Levenshstein Similarity Ratio in Equation 1, $sim(C_{i,t}, C_{a,t})$, an operationalized index of perceived shared moral values between agents *i* and *a*:

$$sim(C_{i,t}, C_{a,t}) = \frac{|C_{i,t}| + |C_{a,t}| - ld(C_{i,t}, C_{a,t})}{|C_{i,t}| + |C_{a,t}|} \quad (1)$$

These ratios are evaluated against the fixed dissimilarity tolerance parameter, *DT*, (see Table 1 for settings), which represents the amount of moral dissimilarity agents will tolerate while perceiving another agent as part of their political ingroup. Thus, agent *i* adds *a* to their ingroup set, $L_{i,t}$, if the following condition holds:

$$sim(C_{i,t}, C_{a,t}) \geq 1 - DT \quad (2)$$

Once each neighbor has been evaluated, each agent compares their L_t to L_{t-1} to identify ingroup members, *g*, who have been ingroup for at least two iterations, i.e., $G_{i,t} = L_{i,t} \cap L_{i,t-1}$ where $g \in G_{i,t}$. This set of consistent ingroup members exerts social influence on agent *i*'s moral values via a revision of agent *i*'s MF weights for each moral foundation, *m*, $\dot{w}_{i,m,t}$, giving rise to new weights, $\ddot{w}_{i,m,t}$:

$$\ddot{w}_{i,m,t} = \dot{w}_{i,m,t} + rev(\dot{w}_{i,m,t}) \cdot S \cdot \left(\frac{freq(m_{G,t}) - freq(m_{i,t})}{5} \right) \quad (3)$$

Where $rev(\dot{w}_{i,m,t})$ is a revision to the current MF weight associated with *m* estimated as the impact of choosing *m* over another moral foundation (see Equation 11). This is moderated by the interaction between the social influence strength parameter, *S*, (see Table 1 for settings) and the difference between the frequency of *m* in agent *i*'s moral signals, i.e., $freq(m_{i,t}) = |\{c \in C_{i,t} \mid \mathrm{c} = \mathrm{m}\}|$ and in the moral signals of members of $G_{i,t}$ on average, i.e., $freq(m_{G,t}) = \frac{\sum_{g=1}^{|G_{i,t}|} |\{c \in C_{g,t} \mid \mathrm{c} = \mathrm{m}\}|}{|G_{i,t}|}$ proportioned by the maximum cardinality of moral signals (set to 5).

Navigating a Moral World. Moral dilemmas occur to agents randomly, subject to three constraints: choice prevalence, choice proportion, and choice invariance (see Settings in Table 1). Choice prevalence indicates the proportion of iterations where moral choices occur. Choice proportion indicates the proportion of randomly selected agents assigned a moral choice on these iterations. These two parameters reflect our assumption that ordinary moral conflict is occasional and does not target an entire population simultaneously. Choice invariance determines how agents assigned a moral dilemma perceive that dilemma (i.e., what percentage of those faced with a moral choice perceive it framed as a choice between values A and B, versus some other combination of values). This reflects the sense of shared reality that agents have when, for example, reacting to similar news stories or current events. For the remainder, moral choice constituent foundations are randomized individually, reflecting ordinary moral choices that occur in daily life (Hoffman et al., 2014).

Agent *i*'s moral choice, *D*, on iteration *t* is a set of two moral foundations, $D_{i,t} = \{f_1, f_2\}$, with the corresponding MF weights $\ddot{w}_{i,f_1,t}$ and $\ddot{w}_{i,f_2,t}$. Agent *i* is conflicted if its tolerance for the proximity of moral values in a choice is breached, defined by the conflict range parameter, *CR* (setting in Table 1):

$$|\ddot{w}_{i,f_1,t} - \ddot{w}_{i,f_2,t}| \leq CR \quad (4)$$

If conflicted, agent *i* updates a moral conflict tracking variable, $k_{i,t} = 1$, and activates a set of alternative conflict resolution procedures. It also adds this to a count of consecutive conflicts when choosing, $cc_{i,t}$ where $0 \leq cc_{i,t} \leq 5$. If not conflicted, i.e., $k_{i,t} = 0$, the agent resets/empties its conflict count, $cc_{i,t} = 0$, and completes its moral decision by choosing the moral foundation with the higher corresponding MF weight (e.g., if $\ddot{w}_{i,f_1,t} > \ddot{w}_{i,f_2,t}$, then f_1 is chosen over f_2), and the chosen foundation is added to *i*'s moral signals on the following iteration, $C_{i,t+1}$. If $|C_{i,t}| = 5$, the chosen moral foundation replaces the oldest choice outcome.

Conflicted agents can either copy the moral signals of ingroup members, $L_{i,t}$, disengage from the choice, which will be signaled to adjacent agents, or pick one foundation at random. Probabilities for each strategy are generated and fed to a weighted draw algorithm. We assume that ingroup copying is favoured (causes the least cognitive dissonance; Hogg, 2000). Therefore, the probability $P(co)_{i,t}$ that a conflicted agent, *i*, on iteration *t* will copy their perceived ingroup set, $L_{i,t}$, is determined first as a function of the number of ingroup members among adjacent agents, i.e., $\frac{|L_{i,t}|}{|A_{i,t}|}$, and the proportion of those ingroup members' choices that are relevant (i.e., either of the moral foundations of the choice or disengagement, *d*):

$$P(co)_{i,t} = \frac{\sum_{l=1}^{|L_{i,t}|} |\{c \in C_{l,t} \mid c = f_1 \vee c = f_2 \vee c = d\}|}{\sum_{l=1}^{|L_{i,t}|} |C_{l,t}|} \cdot \frac{|L_{i,t}|}{|A_{i,t}|} \quad (5)$$

The remaining probability space is divided between random choosing and disengagement as a function of the average of $\ddot{w}_{i,f_1,t}$ and $\ddot{w}_{i,f_2,t}$ proportioned by the maximum

weight, i.e., $\bar{w}_{i,f,t} = \frac{\ddot{w}_{i,f_1,t} + \ddot{w}_{i,f_2,t}}{10}$, representing the degree of care for the dilemma, and the proportion of agent i's disengagement outcomes in the moral signals set, i.e., $dp_{i,t} = \frac{|\{c \in C_{i,t} \mid c = d\}|}{|C_{i,t}|}$, reflecting past propensity to disengage. If $0 < dp_{i,t} < 1$, agent i will disengage with probability:

$$P(dis)_{i,t} = \left(1 - P(co)_{i,t}\right) \cdot \bar{w}_{i,f,t} \cdot \left(\frac{dp_{i,t}}{\bar{w}_{i,f,t}} - dp_{i,t} + 1\right) \quad (6)$$

The probability of choosing an element of $D_{i,t}$ at random, $P(ran)_{i,t}$, is then the remainder of the probability space:

$$P(ran)_{i,t} = 1 - P(co)_{i,t} - P(dis)_{i,t} \equiv \left(1 - P(co)_{i,t}\right) \cdot \left(1 - \bar{w}_{i,f,t}\right) \cdot \left(1 - dp_{i,t}\right) \quad (7)$$

Thus, higher moral care for the choice and/or higher propensity to disengage will increase the probability of disengagement while reducing the probability of choosing randomly.

If an agent i has not disengaged within their last five choices, i.e., $dp = 0$, $P(dis)_{i,t}$ is only moderated as an increasing function of moral care for the choice:

$$P(dis)_{i,t} = \left(1 - P(co)_{i,t}\right) \cdot \bar{w}_{i,f,t} \quad (8)$$

Whereas if agent i has only disengaged within the last five choices, i.e., $dp_{i,t} = 1$, $P(dis)_{i,t}$, disengagement takes up the remaining probability space while $p(ran)_{i,t} = 0$:

$$P(dis)_{i,t} = \left(1 - P(co)_{i,t}\right) \quad (9)$$

These probabilities are fed to a weighted draw algorithm. If disengagement is drawn, agent i updates their moral signals for the following iteration, $C_{i,t+1}$, to reflect this. If random choosing is drawn, a random member of $D_{i,t}$ is determined and used to update $C_{i,t+1}$. If copying ingroup members is drawn, the proportion of moral signals of all ingroup members corresponding to each possible decision, i.e., $\{f_1, f_2, d\}$, e.g., $freq(f_{1,L}) = \sum_{l=1}^{|L_{i,t}|} |\{c \in C_{l,t} \mid c = f_1\}|$, is transformed by a soft-max function and inputted to a weighted draw algorithm to determine the outcome, for example:

$$P(f_1) = \frac{e^{freq(f_{1,L})}}{e^{\left(freq(f_{1,L}) + freq(f_{2,L}) + freq(d_L)\right)}} \quad (10)$$

Finally, the implications on moral values of making a moral trade-off are accounted for by updating the associated MF weights of the choice to $\ddot{w}_{i,m,t}$ for $m = f_1$ and $m = f_2$. The weight is decremented if the moral foundation m was not chosen and m is not in $C_{i,t}$. *If* m was chosen, the corresponding weight is incremented. Disengagement decrements MF weights for f_1 and f_2. Both increments (+) and decrements (–) to the weight corresponding to moral foundation m of agent i are determined by equation 11, which computes the new weight, $\ddot{w}_{i,m,t}$, as a change from $\dot{w}_{i,m,t}$ equivalent to sigmoidal growth, i.e., larger changes for medium-sized weights, but smaller changes for larger and smaller weights:

$$\ddot{w}_{i,m,t} = \dot{w}_{i,m,t} \pm \frac{\partial}{\partial \dot{w}_{i,m,t}} \frac{1}{1 + e^{-2.5(\dot{w}_{i,m,t} - 2.5)}} \quad (11)$$

Evaluating Spatial Position At the end of iteration t, agent i may move to an unoccupied space via a random-walk algorithm as a function of dissatisfaction, cognitively (i was conflicted during t) or socially (i had too few adjacent ingroup members during t). Specifically, to stay, an agent i *must* have a proportion of ingroup members that meets the homophily constraint determined by the fixed similarity needed parameter (see Table 1 for setting), *SN:*

$$\frac{|L_{i,t}|}{8} \geq SN \quad (12)$$

Agents who fulfil this condition without experiencing conflict will stay in their current location. On iterations with no choices, only this condition determines movement.

An agent i, conflicted from a choice, stays if the homophily condition is met *and* as a probabilistic function of the moral care for its choice, $\bar{w}_{i,f,t}$, the moderating effect of implemented conflict resolution strategy, r, (where $mod(co) = 0, mod(ran) = .5,\ mod(dis) = 1$), and the proportion of the last five choices that they were consecutively conflicted, $\frac{cc_{i,t}}{5}$, i.e., the persistence of conflict:

$$P(mov)_{i,t} = \left(\frac{1}{4} \cdot \bar{w}_{i,f,t} + mod(r_{i,t})\right) \cdot \left(1 + \frac{cc_{i,t}}{5}\right) \quad (13)$$

As other factors impacting happiness are not modelled, we equate movement to unhappiness and staying to happiness.

Summary. Following burn-in trials, agents conduct social inference on each iteration, and, if faced with a moral dilemma, resolve it. Depending on these processes, agents may move or stay. Agents infer identity by observing recent moral choice outcomes of neighbours (i.e., the moral foundations recently chosen and disengagement) and determine ingroup members by similarity to their own choice outcomes constrained by tolerance for dissimilar others. The subset of ingroup members, evaluated as such for at least two consecutive iterations, exert social influence on an agent's MF weights as a function of the discrepancy between the agent's own moral signals and those of this subset, the influence strength parameter, and the MF weight revision function. Agents choose between two moral foundations constrained by choice prevalence, choice proportion, and choice invariance. MF weights associated with this choice determine if one moral foundation can be chosen over another, or the agent experiences moral conflict and must use an alternative choice strategy: copy the perceived ingroup's moral signals, disengage from the moral choice, or pick one of the moral foundations randomly. A choice initiates an MF weight revision. Finally, as a function of moral conflict and sufficient ingroup member density (determined by the similar-needed parameter), agents may stay in their current location or move to a new randomly determined space.

Data Simulation and Collection

Simulations varied four parameters (choice prevalence, choice invariance, dissimilarity tolerance, and influence strength) with a total of 96 setting combinations (see Table 1)

while keeping the remainder fixed. These settings were determined by pre-testing to find realistic conditions, e.g., an amount of similarity needed that was obtainable for most agents.

Table 1: Model Parameters and Settings Simulated.

Parameter	Settings simulated
Density	75 %
Choice prevalence	33 %, 66 %
Choice proportion	75 %
Choice invariance	33 %, 66 %, 99 %
Conflict range	0.3 (out of 5)
Similarity needed	62.5 %
Dissimilarity tolerance	10 %, 30 %, 60 %, 80 %
Influence strength	25 %, 50 %, 75 %, 100 %

Each parameter permutation was simulated with approximately 1,900 agents[2] for 100 iterations[3] repeated 20 times to mitigate stochastic elements. At each iteration, we collected each agent's political identity, happiness, conflict state, number of perceived and actual (matching identity label) adjacent ingroup members, total adjacent agents, MF weights, moral choice (if any), choice outcome, and choice resolution strategy. The 10 initial iterations of each simulation were excluded as 'burn-in' trials.

Results

We report only results for dissimilarity tolerance at 30% (the level at which agents were the most accurate in inferring the political identity of adjacent agents), choice prevalence at 33% (results are virtually identical for the 66% setting), choice invariance at 99% (see discussion for the latter point) and collapsed across levels of social influence.

Liberals experienced less conflict (see Figure 1) and were happier, perceived their ingroup to be larger, and formed larger and more homogeneous groups than conservatives (one-sided Bayes Factors (BFs) = Inf). This was true even when comparing agents who experienced a similar number of conflicts (see Figure 1; all BFs > 11.20). Conservatives also experienced more conflicts than liberals (more conservatives in low, moderate, and high conflict bins; BFs > 2.50e58; but similar numbers in no/very conflict; BF = 0.03).

We also found that liberals resolved moral choices by clearly favouring one value over another more often than conservatives (BF = 423.57). When faced with conflicting moral dilemmas, conservatives copied their ingroup or chose randomly less often than liberals and disengaged more often (all $BFs_{10} \geq 466.54$; see Figure 2).

Discussion

When agents must signal and infer political identity via non-stationary moral choices, we observe identity-based polarization. That is, compared to conservatives, liberals formed larger, more homogeneous clusters of identity-congruent agents, found moral choices less conflicting, and when faced with dilemmas that could not easily be resolved, copied from their ingroup neighbors more and morally disengaged less. This is particularly notable because the fixed parameter settings used constitute conditions that seem to favor moral convergence (and thus, depolarization); one where moral challenges are not overly frequent and nearly all agents perceive moral challenges identically, contact with outgroup members is common and their moral choices can

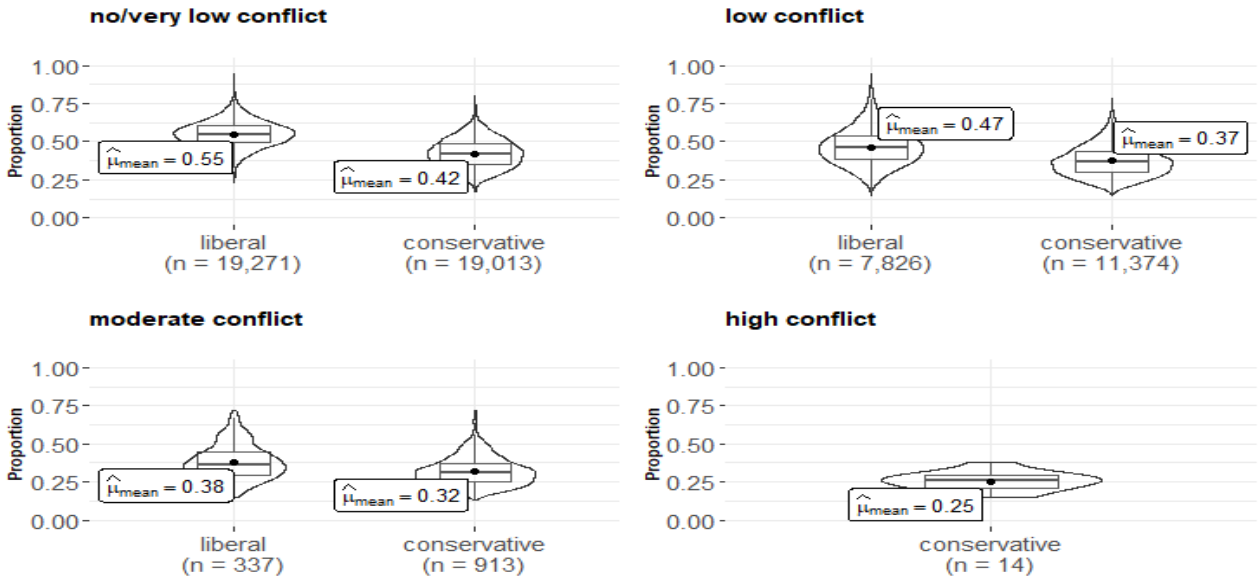

Figure 1: Box-violin plots of the proportion of rounds that agents were happy (i.e., did not move), by political identity and internal conflict level (binned proportion of choices that made an agent conflicted).

[2] Density had some random noise to reduce computational power

[3] Increasing run length to 300 iterations did not change the results.

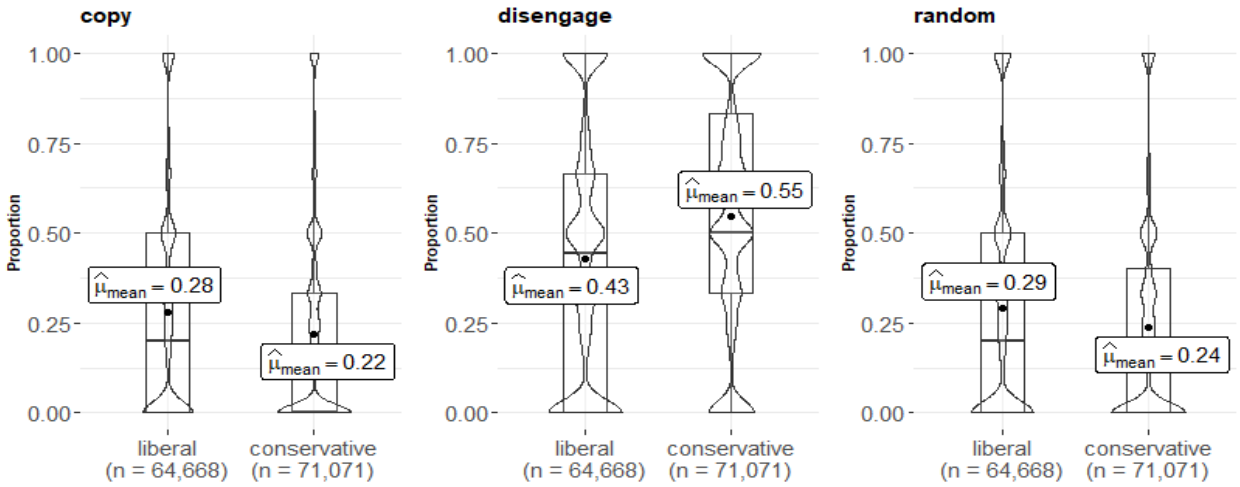

Figure 2: Box-violin plots of the proportion of moral conflicts resolved by copying ingroup values, disengaging from moral values, and randomly selecting, split by political identity.

influence other agents' values across political identities. We find the opposite; polarization occurs anyway, emerges relatively fast, and remains stable.

We treated identity as a fixed latent variable that cannot be directly observed, only inferred from observing others' moral choices (e.g., 'Are they outraged about what I'm outraged about?'). This is analogous to debates about who is a 'real' member of a particular group. This process seemed to drive differing patterns of moral choice and moral disengagement that can serve as reliable signals of underlying identity. Indeed, moral disengagement served as an identity signal just like moral choices, albeit more so for conservatives than liberals. This is consistent with research suggesting that political identity is antecedent to moral values (Hatemi, et al., 2019), that moral values can be temporally unstable (Smith et al., 2016), and that they are a type of motivated social cognition that, particularly for conservatives, serve system-justification and uncertainty or threat reduction needs (Strupp-Levitsky et al., 2020).

For political elites, moralized language (i.e., giving moral signals) is a reliable and effective way to distinguish themselves as leaders (or 'true' members) of their respective ingroups (Bos & Minihold, 2021) and can result in substantial and relatively immediate alterations to the ideological policy positions of fellow ingroup members, even when new positions are the opposite of old ones (Jung, 2019; Slothuus & Bisgaard, 2020). We interpret this as a product of the uncertainty of the 'true' ingroup beliefs/values. Our model predicts this for moral values, but insofar as policy positions are also interpreted as group membership signals, they should be equally malleable.

Interestingly, the model predicts that greater levels of moral disengagement follow from having higher concern for more moral values (cf. Theodoropoulos, 2020), consistent with existing data (Devereux et al., 2021; Jackson & Gaertner, 2010). Another empirically supported prediction is that liberals should form more ideologically homogenous clusters than conservatives (Bakshy et al., 2015; Eady et al., 2019; Wu & Resnick, 2021). Such informational echo chambers may accelerate political polarization perhaps by narrowing available moral signals (e.g., Colley et al., 2022).

Future work could fruitfully explore the limitations of our assumptions, bringing them closer to moral, social, and political reality. The moral dilemmas that our agents could encounter were unsystematic, which is unlikely to be the case. Indeed, if people are self-sorting into groups who largely share moral values, then polarization may accelerate as they encounter moral challenges presented via their ingroup 'lens'. This ties into another assumption and direction for future work – our agents all perceived moral issues in the same way, i.e., high choice invariance. Realistically, perceptions of what moral values are at stake in a given problem may be decoupled (e.g., biased framing from information sources/social media etc.). There is also wide scope for examining social network structure in this model (i.e., varying agent-specific social influence), given that different network topologies have substantially different effects on cognition (for review see Momennejad, 2021). Similarly, individual differences in tolerance, homophily, or resistance to moral conflict remain to be explored.

Conclusion

Our model combines theoretical insights from moral conflict and political polarization to investigate the viability of moral values as the basis for identity signaling and inference. We show broad consistency between our simulations and existing literature on differences between liberals and conservatives in terms of the role of moral values in political communication, group polarization, and the tendency to morally disengage. The model makes interesting predictions, e.g., disengagement functions to signal identity just as moral values do. Future directions include network dynamics interacting with different information streams creating echo chambers that appeal to certain combinations of moral values.

Acknowledgments

This research was supported by an Undergraduate Vacation Scholarship from the Carnegie Trust for the Universities of Scotland (VAC011919).

References

Axelrod, R. (1997). The Dissemination of Culture: A Model with Local Convergence and Global Polarization. *Journal of Conflict Resolution*, *41*(2), 203–226. https://doi.org/10.1177/0022002797041002001

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

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

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

Bandura, A. (1999). Moral Disengagement in the Perpetration of Inhumanities. *Personality and Social Psychology Review*, *3*(3), 193–209. https://doi.org/10.1207/s15327957pspr0303_3

Bos, L. & Minihold, S. (2021). 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

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

Colley, T. P., Granelli, F., & Althuis, J. (2020). Disinformation's Societal Impact: Britain, COVID and Beyond. *Defence Strategic Communications, 8*(1), 89–140.

Devereux, P. G., Miller, M. K., & Kirshenbaum, J. M. (2021). Moral disengagement, locus of control, and belief in a just world: Individual differences relate to adherence to COVID-19 guidelines. *Personality and Individual Differences*, *182*, Article 111069. https://doi.org/10.1016/j.paid.2021.111069

Doan, A., Halevy, A., & Ives, Z. (2012). 4—String Matching. In A. Doan, A. Halevy, & Z. Ives (Eds.), *Principles of Data Integration*. Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-416044-6.00004-1

Doherty, C., Kiley, J., Asheer, N., & Price, T. (2022). *As Partisan Hostility Grows, Signs of Frustration With the Two-Party System*. Pew Research Center. https://www.pewresearch.org/politics/wp-content/uploads/sites/4/2022/08/PP_2022.09.08_partisan-hostility_REPORT.pdf

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. *SAGE Open*, *9*(1), Article 2158244019832705. https://doi.org/10.1177/2158244019832705

Festinger, L. (1962). Cognitive Dissonance. *Scientific American*, *207*(4), 93–106.

Funkhouser, E. (2020). A tribal mind: Beliefs that signal group identity or commitment. *Mind & Language*, *37*(3). https://doi.org/10.1111/mila.12326

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In P. Devine & A. Plant (Eds.), *Advances in Experimental Social Psychology* (Vol. 47). Academic Press. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

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., 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

Harmon-Jones, E., & Mills, J. (2019). An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In E. Harmon-Jones (Ed.), *Cognitive dissonance: Reexamining a pivotal theory in psychology (2nd ed.)*. American Psychological Association. https://doi.org/10.1037/0000135-001

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

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*(6202), 1340–1343. https://doi.org/10.1126/science.1251560

Hogg, M. A. (2000). Subjective Uncertainty Reduction through Self-categorization: A Motivational Theory of Social Identity Processes. *European Review of Social Psychology, 11*(1), 223–255. https://doi.org/10.1080/14792772043000040

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

Jackson, L. E., & Gaertner, L. (2010). Mechanisms of moral disengagement and their differential use by right-wing authoritarianism and social dominance orientation in support of war. *Aggressive Behavior*, *36*(4), 238–250. https://doi.org/10.1002/ab.20344

Jung, J. H. (2020). The Mobilizing Effect of Parties' Moral Rhetoric. *American Journal of Political Science*, *64*(2), 341–355. https://doi.org/10.1111/ajps.12476

Kalmoe, N. P. (2020). Uses and Abuses of Ideology in Political Psychology. *Political Psychology*, *41*(4), 771–793. https://doi.org/10.1111/pops.12650

Kashima, Y., Perfors, A., Ferdinand, V., & Pattenden, E. (2021). Ideology, communication and polarization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1822), Article 20200133. https://doi.org/10.1098/rstb.2020.0133

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), Article e2102144118. https://doi.org/10.1073/pnas.2102144118

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

Momennejad, I. (2022). Collective minds: social network topology shapes collective cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences, 377(1843)*, Article 20200315. https://doi.org/10.1098/rstb.2020.0315

Prentice, D. A., & Miller, D. T. (1996). Pluralistic Ignorance and the Perpetuation of Social Norms by Unwitting Actors. Advances in Experimental Social Psychology, 28, 161–209. https://doi.org/10.1016/S0065-2601(08)60238-5

Singer, D. J., Bramson, A., Grim, P., Holman, B., Jung, J., Kovaka, K., Ranginani, A., & Berger, W. J. (2019). Rational social and political polarization. *Philosophical Studies*, *176*(9), 2243–2267. https://doi.org/10.1007/s11098-018-1124-5

Slothuus, R., & Bisgaard, M. (2021). How Political Parties Shape Public Opinion in the Real World. *American Journal of Political Science*, *65*(4), 896–911. https://doi.org/10.1111/ajps.12550

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

Strupp-Levitsky, M., Noorbaloochi, S., Shipley, A., & Jost, J. T. (2020). Moral "foundations" as the product of motivated social cognition: Empathy and other psychological underpinnings of ideological divergence in "individualizing" and "binding" concerns. *PLoS One*, *15*(11), Article e0241144. https://doi.org/10.1371/journal.pone.0241144

Theodoropoulos, N. C. (2020). Computational modelling of social cognition and behaviour [Doctoral dissertation, University of Edinburgh].

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), Article e2102147118. https://doi.org/10.1073/pnas.2102147118

van Baar, J. M., & FeldmanHall, O. (2021). The polarized mind in context: Interdisciplinary approaches to the psychology of political polarization. *American Psychologist*, *77*(3), 394–408. https://doi.org/10.1037/amp0000814

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

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *ELife*, *8*, Article e49547. https://doi.org/10.7554/eLife.49547

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.

Yudkin, D., Hawkins, S., & Dixon, T. (2019). *The Perception Gap: How False Impressions are Pulling Americans Apart* (Hidden Tribes Project). More in Common. https://doi.org/10.31234/osf.io/r3h5q