A WEIGHTED BALANCE MODEL OF OPINION HYPERPOLARIZATION

A PREPRINT

Simon Schweighofer
Medical University of Vienna
Complexity Science Hub Vienna
Vienna, Austria
schweighofer@csh.ac.at

Frank Schweitzer
ETH Zurich
Zurich, Switzerland
Complexity Science Hub Vienna
Vienna, Austria
fschweitzer@ethz.ch

David Garcia
Medical University of Vienna
Complexity Science Hub Vienna
Vienna, Austria
garcia@csh.ac.at

December 25, 2019

ABSTRACT

Polarization is threatening the stability of democratic societies. Until now, polarization research has focused on opinion extremeness, overlooking the correlation between different policy issues. We explain the emergence of hyperpolarization, i.e. the combination of extremeness and correlation between issues, by developing *Weighted Balance Theory (WBT)*, a new theory of opinion formation. WBT extends Heider's cognitive balance theory to encompass multiple weighted attitudes. We validate WBT on empirical data from the 2016 National Election Survey. Furthermore, we develop an opinion dynamics model based on WBT, which, for the first time, is able to generate hyperpolarization and to explain the link between affective and opinion polarization. In addition, our theory encompasses other phenomena of opinion dynamics, including mono-polarization and backfire effects.

Keywords Polarization | Balance Theory | Opinion Dynamics | Agent-Based Modeling

1 Introduction

Political polarization has increased steeply over recent years in many democratic societies, up to the point of posing a threat to political stability [1, 2]. If we want to explain the continuing surge of polarization, it is crucial to understand the psychological and social mechanisms that generate it and the circumstances under which they operate. Our aim with this study is to identify a minimal set of mechanisms that can generate two essential aspects of polarization: *opinion extremeness* (how far positions are from the center) and *issue constraint* (how positions on different issues correlate).

So far, the literature on polarization has largely focused on *opinion extremeness*. Opinion extremeness is quantified by how much the distribution of positions on various policy issues (such as abortion, immigration, or cannabis legalization) is concentrated on both fringes of a policy dimension, as opposed to the center [3]. The emergence of opinion extremeness poses a theoretical problem: Most psychological research on social influence has produced examples of assimilative influence, in which individuals' opinions become more similar [see 4, 5]. But if social influence is only assimilative, it can be shown that all individuals in a society would sooner or later converge to a complete consensus on all issues (as long as they are at least indirectly connected; [6, 7]). However, in reality we do not observe societies in a complete consensus state, suggesting that social influence cannot be exclusively assimilative. And indeed, several studies have found that social influence can also be repulsive, meaning interacting individuals become more dissimilar [8, 9, 10]. This phenomenon, named the backfire effect, is thought to occur between individuals whose opinions are already quite dissimilar. Of course, such a repulsive force could explain how opinions diverge into opposite extremes. The problem is that the number of studies that find a backfire effect is matched by an equal number of studies that do not find it [11, 12, 13]. So far, no explanation for this apparent contradiction has been put forward.

While explaining opinion extremeness is an important problem, political scientists have stressed that extreme opinions are necessary, but not sufficient for a society to be polarized. Positions on different policy issues also need to be *correlated* with each other - a property known as *issue constraint* [14, 15]. Low issue constraint means that the positions of political actor on any one issue are independent from their positions on other issues. Thus, in situations with low issue constraint, political actors can have any combination of issue positions and none of these combinations will be substantially more frequent than any other in the population. For political actors who want to form alliances, this poses a problem: They would rarely encounter other actors with whom they agree on many issues, as most actors would agree and disagree with each other on roughly half of the issues. This makes the formation of stable political alliances very difficult, since political actors who are, for example, jointly supporting cannabis legalization, will likely find themselves on different sides of the next issue, e.g. abortion. Thus, without issue constraint, opinion extremeness alone does not lead to polarization, but to political fragmentation into many small groups that can hardly form stable alliances.

In political systems with high issue constraint, a multitude of issue positions can be described by a position on a single ideological dimension with negligible loss of information. For example, instead of describing certain political actor as 'in favor of gun control, cannabis legalization, and against increased defense spending, etc.', we can just describe it with the word 'leftist'. Most political systems are characterized by such an ideological dimension, usually labeled 'left-right' or 'liberal-conservative' [16]. Poole [17] highlights the origin of opinion constraint as a major open research question in political science (see also [1]). He notes that "this bundling [of issue dimensions] does not have to be a function of a logically consistent philosophy" [17, p. 204]. This means that often there is only a very distant, if any, logical connection between issue positions, such as gay marriage and corporate tax. Of course, one can always construct an ad hoc logical connection between two given issues, but finding a consistent mechanism that explains issue constraint is still an open question. It remains to be tested whether issue constraint can emerge from the micro-level interactions between individuals without the need to assume a preexisting complex structure, such as logical links between issues [18].

To clarify that polarization is not just opinion extremeness, we define *hyperpolarization* as the coexistence of opinion extremeness and issue constraint in a multidimensional opinion space. A metric of hyperpolarization of a political system must be maximal if 1) the political system is divided into two blocks, each encompassing half of the population, 2) each of these two blocks has perfect internal consensus on all relevant issues, and 3) the blocks are in total disagreement with each other on all relevant issues [see also 19, 20]. A metric of hyperpolarization must be low if there are more or less than two political blocks, if the size difference between blocks is large, if there is disagreement within blocks, or if there is agreement (at least on some issues) between different blocks. Consequently, hyperpolarization is zero if a political system is in a state of complete consensus on all relevant issues.

Our goal is to explain the emergence of hyperpolarization from the interactions between individuals without having to assume complex social or logical structures. To do so, we develop a theory of opinion change that, when formulated as a computational model, simultaneously generates both aspects of hyperpolarization: opinion extremeness and issue constraint. In Section 2, we give a brief overview of polarization models in the opinion dynamics literature, highlighting the models that generate some aspects of hyperpolarization. In Section 3, we present *Weighted Balance Theory (WBT)* and empirically test some of its propositions against data from the American National Election Survey (ANES). In Section 4, we develop an opinion dynamics model based on WBT and our empirical analyses. We introduce a metric to quantify hyperpolarization from the multidimensional issue positions of agents. We apply this metric to show the hyperpolarization outcomes of this model under a wide range of circumstances and compare the model with a benchmark of previous opinion dynamics models (see Appendix A).

2 Literature on Opinion Dynamics Models of Polarization

Opinion dynamics models typically encompass a number of agents characterized by issue positions on one or several, discrete or continuous opinion dimensions. These agents influence each other's opinions over time, following specified rules of interaction that produce different opinion distributions.

2.1 One-Dimensional Opinion Dynamics Models

Most conventional opinion dynamics models have focused on the extremeness aspect of polarization [for reviews, see 21, 4], and have treated the existence of a single ideological dimension as given instead of as an emergent phenomenon in need of explanation. As described by Mäs and Flache [11] and Flache et al. [4], one-dimensional models of continuous opinions can be categorized into three classes: 1) models with only positive influence between agents always create consensus [6, 7], 2) *bounded confidence* models, in which agents only interact with similar others, can create multiple opinion clusters [22, 21], and 3) models with repulsion between dissimilar agents can create bi-polarization [23, 24]. Model classes 2 and 3 can, for certain parameter values, replicate the extremeness aspect of hyperpolarization. However, when extended to multidimensional opinion spaces, they do not generate the issue constraint necessary for hyperpolarization, as we show in Appendix A.

2.2 Multidimensional Models of Hyperpolarization

Only few opinion dynamics models operate in multidimensional opinion spaces. To our knowledge, three of these multidimensional opinion dynamics models generate a form of hyperpolarization under special conditions:

Huet and Deffuant [25] propose a model with two opinion dimensions, in which the dynamics on the second dimension is determined by the state of the first dimension. If two agents are close together on the first dimension, they will attract each other on both dimensions. If they are far apart on the first dimension, they do not exchange opinions on the first dimension (i.e., bounded confidence), and move further apart on the second dimension (rejection). Under certain parameter configurations, the model produces a hyperpolarized state with two clusters in opposite corners of the opinion space, and a third cluster in the middle, and thus a high degree of issue constraint and an intermediate degree of opinion extremeness. This dependence between dimensions is a way to encode issue constraint, not an attempt to generate issue constraint from interaction between individuals.

Flache and Mäs [26] present a multidimensional model containing both opinion and demographic dimensions. While agents' issue positions are continuous and change over time, demographic attributes are binary and immutable. Agents' distance in the combined opinion-demographic space determines whether they approach or repulse each other. This model can generate hyperpolarization if the demographic dimensions are highly correlated. While this outcome is certainly interesting, it means that hyperpolarization in this model does not emerge, but is induced by design through the correlation of demographic attributes. Without demographic attributes the model is reduced to a multidimensional repulsion model that does not generate hyperpolarization (see Appendix A).

Finally, Flache and Macy [19] explore a multidimensional opinion dynamics model that combines attraction and repulsion mechanisms with a caveman social network structure with densely connected clusters. While this model is able to create an intermediate degree of hyperpolarization for two opinion dimensions, hyperpolarization declines rapidly when they add more than two opinion dimensions. Therefore, besides requiring complex social structures, this model does not reproduce hyperpolarization for a realistic number of opinion dimensions. To sum up, hyperpolarization has not been shown to emerge from standard modeling assumptions without additional social structures, like the caveman network, or correlated demographic dimensions from which issue constraint trivially follows.

3 Weighted Balance Theory

3.1 Extending Balance Theory

Most opinion dynamics models assume that the opinions of one person *directly* influence the opinions of another person. In contrast, our theory is based on the assumption that the social influence that an individual j exerts on another individual i is moderated by the interpersonal attitude of i towards j (i.e., to what degree i likes or dislikes j). In other words, our theory combines cognitive (issue positions) and affective components (interpersonal attitudes) to explain opinion change. We postulate that issue positions and interpersonal attitudes influence each other in a dynamic way: Interpersonal attitudes are influenced by issue positions – agreement fosters liking and disagreement fosters disliking. Conversely, issue positions are adapted to interpersonal attitudes – human beings want to agree with others they like and want to contradict who they dislike.

We formalize the rules of this mutual adaptation of issue positions and interpersonal attitudes in an extended version of Balance Theory, which we call *Weighted Balance Theory* (WBT). Balance Theory was developed by Heider [27] to explain the cognitive organization of attitudes, and later expanded to social networks in the form of Structural Balance Theory by Cartwright and Harary [28]. According to Heider, attitudes can have positive or negative valence, and be directed to objects, ideas, events, or other individuals. Configurations of attitudes can be either balanced or imbalanced, and human beings strive to increase balance in their cognitive organization.

Heider specifically focuses on triads consisting of an ego i, an alter j, and an object d. In the context of this paper, d is a particular policy issue. In the following, we will denote the attitude of an individual i to a political issue d as $\mathbf{o}_d^i$, with $\mathbf{o}^i$ being the opinion vector of individual i, representing i's attitudes towards all D policy issues under consideration. Each of these opinions has a sign, denoting whether the individual is in favor or against a certain issue. We denote the interpersonal attitude of i towards j as $\mathcal{A}(i, j)$. Note that $\mathbf{o}_d^j$ is i's perception of j's attitude towards d, and not necessarily the actual opinion of j. Heider postulates that an i-j-d triad is in balance either if i has a positive attitude towards j, and i and j agree in their attitudes towards d (i.e., their attitudes towards d have the same sign), or if i has a negative attitude towards j and they disagree about d (their attitudes have different signs). Generally speaking, in a balanced i-j-d triad, the sign of each attitude relation must be the product of the signs of the other two relations [28].

Modeling attitudes as purely binary, i.e. either positive or negative, is an oversimplification. In reality, attitudes do not only have a sign, but also a certain strength or *weight*: One can be more or less in favor of or against something, or neutral towards it. We define this weight to be a real number between 0 and 1. Thus, signed and weighted attitude relations can be represented by a real number between -1 and 1. The necessity of expanding balance theory by including attitude weights was already acknowledged by [28].

To extend Balance Theory to include weights, we require a rule to compute the weight and sign of relations between individuals [see also 29, 30]. Let us assume we have perfectly balanced i-j-d triad with signed and weighted relations between the three elements. If we only know the signs and weights of two of the three relations in the triad, how can we determine the third relation? We postulate two basic requirements: 1), if the weight of any of the first two attitude relations in a i-j-d triad is zero, the third relation must be zero as well, in order to obtain a balanced triad. In other words, if i does not care about d either way, i will also not care about j's attitude towards d, and i's resulting attitude towards j will be neutral. 2), the weight of the third attitude relation should be between the weights of the first two relations.

Simply taking the product of the weights would satisfy requirement 1, but not requirement 2: If both i and j have a positive attitude of weight 0.5 towards d, the product rule would predict an attitude relation $\mathcal{A}(i,j)$ of just 0.25. Thus, to determine the attitude weight, we instead use the geometric mean, i.e., the square root of the product. In combination with the product rule for the sign of the attitude relation, this gives us a function that we call *signed geometric mean* (SGM):

$$\mathrm{SGM}(x_1, ..., x_n) = \prod_{i=1}^{n} \mathrm{sign}(x_i) \left(\prod_{i=1}^{n} |x_i| \right)^{\frac{1}{n}} \tag{1}$$

Consequently, in a perfectly balanced i-j-d triad, each of the three attitude relations is given by the SGM of the other two relations.

3.2 Determining Interpersonal Attitudes

If the attitude of i and of j towards issue d is known, we can apply the SGM to $\mathbf{o}_d^i$ and $\mathbf{o}_d^j$, in order to determine the interpersonal attitude $\mathcal{A}(i,j)$. Figure 1a shows the interpersonal attitude $\mathcal{A}(i,j)$ resulting from the position of i and of j on issue d. Like in classical balance theory, the relation between i and j is positive if their attitudes towards d have the same sign, and negative otherwise. The intensity of their positive or negative relation is proportional to the intensity of their attitudes towards d.

To model polarization in multidimensional opinion spaces, we have to define $\mathcal{A}(i,j)$ for cases where i and j have attitudes towards many different issues. In a first step, for each issue $d = 1, ..., D$, we compute a separate $\mathrm{SGM}(\mathbf{o}_d^i, \mathbf{o}_d^j)$, to then combine them to determine $\mathcal{A}(i,j)$. As an initial approximation, we assume that each issue contributes equally to $\mathcal{A}(i,j)$. We choose this as a parsimonious assumption, as in practice some issues might have higher weight than others.

$$\overline{\mathrm{SGM}}(i,j) = \frac{1}{D} \sum_{d=1}^{D} \mathrm{SGM}(\mathbf{o}_d^i, \mathbf{o}_d^j) \tag{2}$$

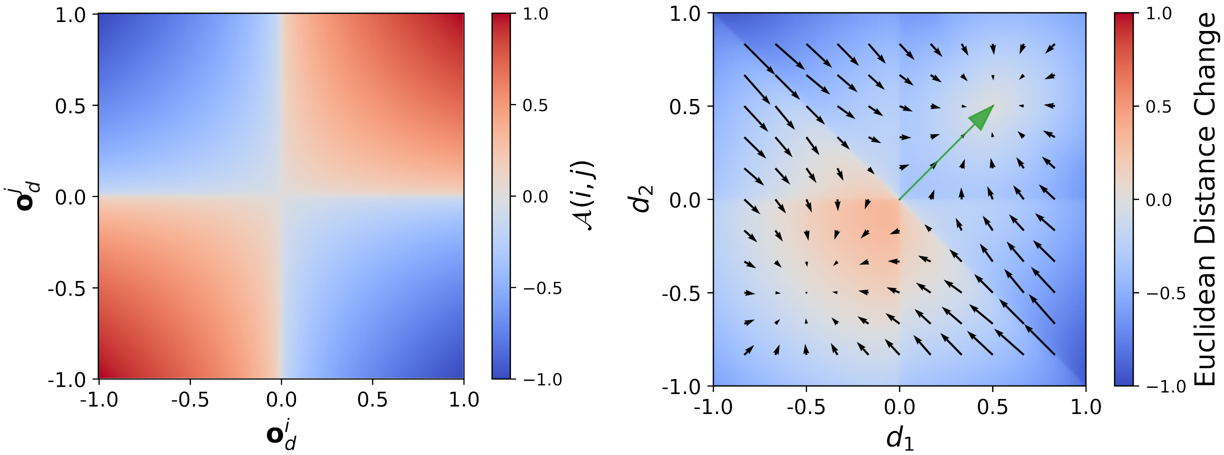
We calculate $\mathcal{A}(i,j)$ as the result of applying a monotonously increasing function $f(x)$ of the arithmetic mean of the SGMs:

$$\mathcal{A}(i,j) = f\left(\overline{\mathrm{SGM}}(i,j)\right) \tag{3}$$

For now, we assume $f(x)$ to be the identity function. In Section 4.2 we provide empirical evidence of the shape of $f(x)$ and in Appendix B we study the role of its shape in the outcomes of the opinion dynamics model presented below.

Figure 1b depicts the interpersonal attitude of i towards j based on their $2D$ opinion vectors. The axes represent the two opinion dimensions, d_1 and d_2 and the green arrow represents the opinion vector of individual j, which is set to $\mathbf{o}^i = [1, 1]$ for this example. The color indicates the interpersonal attitude $\mathcal{A}(i,j)$ that would result from j interacting with i with an opinion vector $\mathbf{o}^i$, corresponding to the colored position in the coordinate system. For example, the deep blue color in the bottom left corner tells us that an individual i with $\mathbf{o}^j = [-1, -1]$ has an interpersonal attitude of -1 to j.

Figure 1b illustrates two interesting properties of the SGM: First, there is a sharp change in interpersonal attitude between the sectors of the coordinate system. This means that i is very sensitive to whether j is on the same side of all issues. And second, we can see that the transition between positive and negative interpersonal attitudes happens for vectors at a 90 degree angle from $\mathbf{o}^i$. This is also true for opinion spaces with more than two dimensions and other



(a) (b)

Figure 1: **Issue positions and interpersonal attitude.** Attitude $\mathcal{A}(i,j)$, depending on the position of $\mathbf{o}^i$ and $\mathbf{o}^j$ on a single issue d (panel a), or on the position of $\mathbf{o}^i$ in a 2-dimensional issue space, with $\mathbf{o}^j$ fixed to $[1,1]$ (green arrow; panel b).

opinion vectors. The importance of being on the same side of an issue, as well as the relevance of the angle between opinion vectors, are major distinctive features between WBT and standard opinion dynamics models, which usually only take into account the distance between two agents on one or several opinion dimensions.

3.3 Balance Maximization through Opinion Adjustment

Our next step is to show how opinion vectors change according to interpersonal attitudes to increase balance. For example, if i has a positive attitude towards j, but still disagrees with j on issue d, the triangle i-j-d is imbalanced. To increase balance, i adapts its opinion on issue d to be in concordance with j's opinion. The reverse is true if i dislikes j: In this case, balance increases if i adapts its opinions in such a way as to increase the contradiction with j. Given $\mathbf{o}^j$ and $\mathcal{A}(i,j)$, we can define an *optimally balanced vector* $\mathbf{b}^{ij}$. This vector is generated by computing the signed geometric mean of $\mathcal{A}(i,j)$ with every component of $\mathbf{o}^j$:

$$\mathbf{b}^{ij}_d = \mathrm{SGM}(\mathbf{o}^j_d, \mathcal{A}(i,j)) \tag{4}$$

Based on this optimally balanced vector, we can measure balance as the proximity of $\mathbf{o}^i$ to the optimally balanced vector $\mathbf{b}^{ij}$, normalized between 0 and 1:

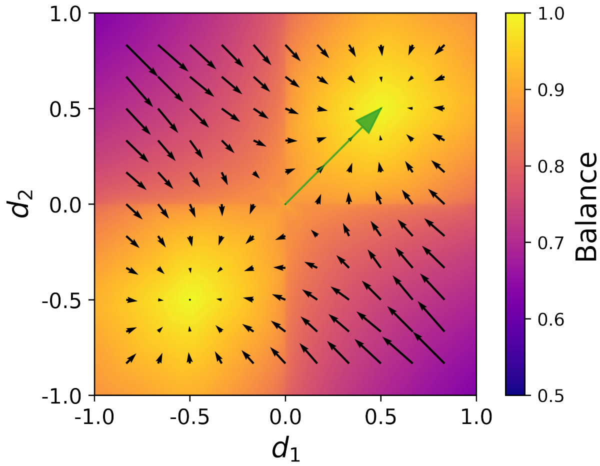
$$\mathcal{B}(i,j) = \frac{1}{2D} \sum_{d=1}^{D} 2 - |\mathbf{o}^i_d - \mathbf{b}^{ij}_d| \tag{5}$$

Heider's central tenet is that individuals strife to increase the balance of their cognitive organization [27]. Given our definition of balance, increasing balance is equivalent to approaching the maximally balanced vector $\mathbf{b}^{ij}$. However, it is reasonable to assume that opinions have a certain degree of inertia and do not change completely upon a single encounter with another person. Thus, we postulate that i does not reset its opinion vector $\mathbf{o}^i$ completely to the balanced vector $\mathbf{b}^{ij}$, but approaches $\mathbf{b}^{ij}$ by a fraction α:

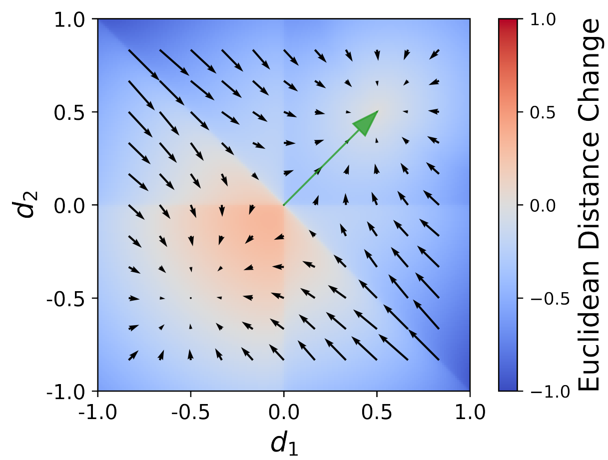
$$\Delta\mathbf{o}^i_d = \alpha(\mathbf{b}^{ij}_d - \mathbf{o}^i_d) \tag{6}$$

The arrows of the quiver plot in Figure 2a depict the direction of opinion changes of an individual i when interacting with j. Arrows are shows for several values of $\mathbf{o}_i$ when interacting with a fixed opinion $\mathbf{o}_j$ illustrated as the green arrow. The background color shows the balance $\mathcal{B}(i,j)$ between individuals before interacting. As we can see, if the angle between $\mathbf{o}^i$ and $\mathbf{o}^j$ is less than 90 degrees, $\mathbf{o}^i$ approaches $\mathbf{o}^j$. If no other reverse influences or interactions happen, $\mathbf{o}^i$ would converge to $\mathbf{o}^j$. In contrast, if the angle between $\mathbf{o}^i$ and $\mathbf{o}^j$ is larger than 90 degree, $\mathbf{o}^i$ would converge to $-\mathbf{o}^j$.

WBT encodes the phenomenon called 'backfire effect' (also known as 'repulsion' in the opinion dynamics literature). A backfire effect occurs if an attempt to persuade an individual of a certain issue position produces the opposite result, i.e., the individual moves even further away from this position than before. In this way, the attempt at persuasion 'backfired'.



(a)



(b)

Figure 2: **Quiver plots of opinion changes and balance in a 2D opinion space.** Opinion exchange between i and j as a function of $\mathbf{o}^i$, with $\mathbf{o}^j$ fixed (green arrow, $[.5, .5]$). Each black arrow represents the resulting change in $\mathbf{o}_i$ given the interaction between i and j: The basis of the arrow represents $\mathbf{o}^i$ before the interaction and the tip of the arrow is the position of $\mathbf{o}^i$ after interaction. The background color in (a) encodes the balance between i and j and in (b) shows the change in euclidean distance between $\mathbf{o}^i$ and $\mathbf{o}^j$ resulting from the interaction. The parameter α is set to 0.3 for both panels.

In the opinion dynamics literature, the backfire effect is usually conceptualized as a function of the distance between two agents in opinion space [4]: If two agents are more distant than a certain threshold value, they will repulse each other, i.e. become more distant after interaction (see Appendix A for simulations of this model). WBT offers an account of the backfire effect that differs from much of the opinion dynamics literature. In WBT, a backfire effect can occur, but only under certain conditions: First of all, the occurrence of a backfire effect is primarily determined by interpersonal attitudes, and not by issue positions. Issue positions are only relevant as far as they determine interpersonal attitudes. A negative interpersonal attitude is a necessary condition for the backfire effect.

However, it is not a sufficient condition. A negative interpersonal attitude $\mathcal{A}(i, j)$ motivates i to disagree with j to an extent warranted by their negative relation. If i already disagrees with j to this extent, i is in a balanced state and will remain there. If i disagrees with j more than warranted by $\mathcal{A}(i, j)$, i will approach j. A backfire effect only occurs if i feels too close to j, given its negative interpersonal attitude towards j. Consequently, in WBT, repulsion and attraction are non-linear functions of distance in opinion space.

Figure 2b shows the backfire effect in WBT with two opinion dimensions: The color in Figure 2b encodes the change in the Euclidean distance between $\mathbf{o}^i$ and $\mathbf{o}^j$. Red color indicates that this distance has increased, i.e., that a backfire effect has occurred. As we can see, a backfire effect only occurs if $\mathbf{o}^i$ and $\mathbf{o}^j$ have a different sign in at least one dimension, and is strongest if they have different signs in both dimensions. But even in this case, the backfire effect only occurs if i is less extreme in its opinions than j. Thus, different from other formulations of the backfire effect, WBT does not predict the strength of the backfire effect to be a linear function of the distance between two individuals in opinion space.

3.4 Empirical Test of Interpersonal Attitude Formation in Weighted Balance Theory

A novel assumption of WBT is that interpersonal attitudes are formed based on relative issue positions (Equation 3). In this section, we test this assumption and explore the shape of $f(x)$ in survey data from the American National Election Study (ANES). ANES is a nationwide representative survey of American voters, conducted before and after every presidential election. Among others, the 4270 respondents of the 2016 ANES survey were asked for their opinion on six different policy issues ranging from defense spending (increase vs decrease) to health insurance (government vs private) on 7-point rating scales. The respondents were also asked for their perception of the position of presidential candidates (Hillary Clinton and Donald Trump) on the same policy issues, again with 7-point scales. And finally, the respondents were asked to complete two affective thermometer scale items, on which they rated their subjective feelings towards each presidential candidate. These affective measures of attitudes towards candidates are measured with high resolution, from 0, meaning "very cold or unfavorable feeling" to 100, signifying "very warm or favorable feeling".

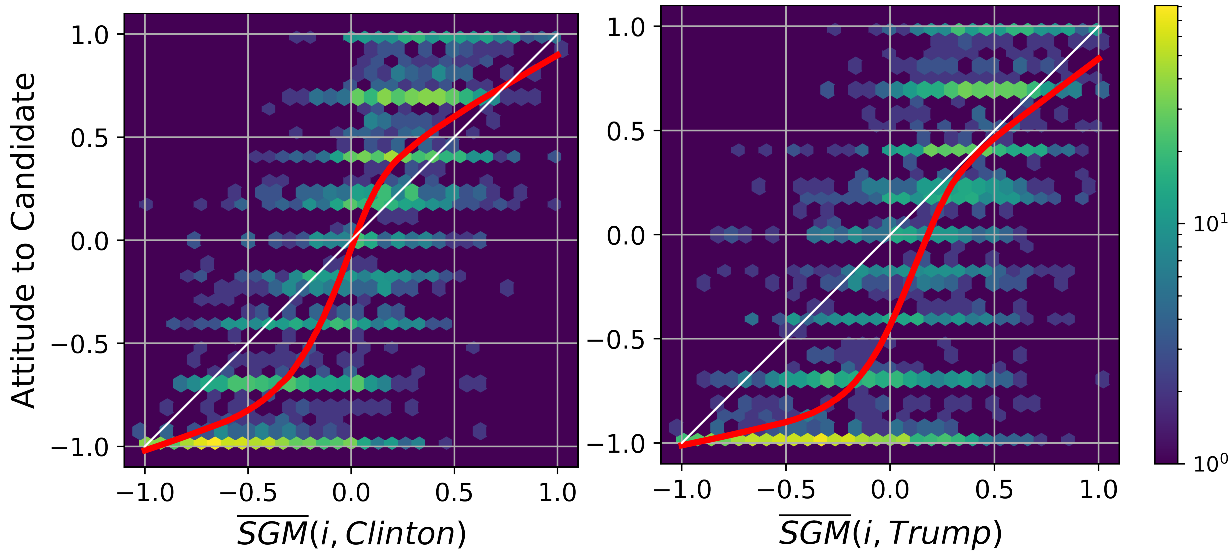


Figure 3: **Testing Weighted Balance Theory on ANES survey data.** Mean SGM between respondents and their perception of candidate positions (x-axis) and self-rated attitude (y-axis) towards Democratic (left) and Republican (right) presidential candidates. Red curves show LOESS fit.

We rescale both the policy questions and the thermometer items between -1 and 1, with 0 corresponding to a neutral position on the policy issues or a neutral attitude towards the presidential candidates. For every respondent, we construct:

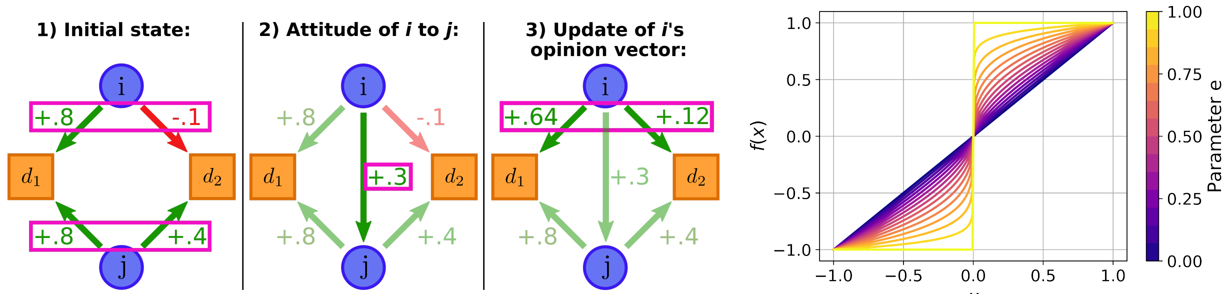
1. A 6-dimensional opinion vector $\mathbf{o}^i$ of the respondent's own issue positions
2. Two 6-dimensional opinion vectors of the respondent's perception of each presidential candidate's issue positions, $\mathbf{o}^{i,Trump}$ and $\mathbf{o}^{i,Clinton}$
3. The attitudes of the respondent towards each of the two presidential candidates ($\mathcal{A}(i, Trump)$ and $\mathcal{A}(i, Clinton)$), measured by the affective thermometer scales.

Due to respondents having missing values in at least one of the policy or thermometer items, our sample size is reduced to 2621 valid respondents for Hillary Clinton, and 2593 for Donald Trump. To predict each respondent's attitude towards each of the two presidential candidates, we apply Equation 3 to the respondent's own opinion vector $\mathbf{o}^i$ and the respondent's estimates of the opinion vectors of the candidates, $\mathbf{o}^{i,Trump}$ and $\mathbf{o}^{i,Clinton}$, respectively.

The x-axes in Figure 3 show the average SGM between the positions of each respondent and each presidential candidate, whereas the y-axes represent the re-scaled thermometer ratings, separate for Hillary Clinton (left panel) and Donald Trump (right panel). The color encodes the logarithm of the number of respondents in each bin of a 2D histogram. If $f(x)$ was the identity function, and we could perfectly predict the attitudes towards the presidential candidates, all respondents would lie on the diagonal. Clearly, this is not the case, but nevertheless the prediction with an identity function reaches R^2 values of 0.52 for Clinton and 0.46 for Trump.

However, the thermometer ratings are not deviating from the WBT based predictions in a random fashion, i.e., towards both sides of the diagonal. The red curves in Figure 3 represent locally weighted regression (LOESS) curves. As we can see, for both candidates the LOESS curves have a sigmoid shape, a monotonically increasing form of $f(x)$. For Hillary Clinton, if our prediction of the respondent's attitude towards the presidential candidate is positive, the actual attitude tends to be even more positive. If the prediction is negative, the actual attitude tends to be even more negative. In other words, the thermometer ratings of Hillary Clinton tend to be on average *more extreme* than our predictions. For Donald Trump, the LOESS curve has a sigmoid shape too, only that it is below the diagonal. This probably reflects other, non policy-related factors that cause respondents to judge him more negatively, and that are not captured by the ANES political questionnaire.

Thus, in their judgment of political figures, respondents seem to tend to a 'manichean', black-and-white world view, in which "whoever is not with me is against me" (Matthew 12:30). We call this tendency *evaluative extremeness*. In Section 4, we describe how we implement evaluative extremeness in a WBT based opinion dynamics model. The degree of evaluative extremeness will play a crucial role for the behavior of this model.



(a) (b)

Figure 4: **Opinion Exchange and Evaluative Extremeness.** Panel a) shows a schematic of opinion exchange: Agents i and j have attitudes to policy issues d_1 and d_2 (1), i creates a interpersonal attitude towards j based on these policy attitudes (2), i modifies its opinion vector to increase balance (3). Panel b) shows the form of the evaluative extremeness function $f(x)$, for varying parameter e.

4 Simulating Weighted Balance Theory

After describing the central tenets of Weighted Balance Theory, and empirically testing its predictions with regard to interpersonal attitudes, we now want to explore whether an opinion dynamics model based on WBT can generate hyperpolarization.

4.1 Notation and Starting Conditions

The code to replicate our model is stored on the CoMSES Computational Model Library under the following url: `https://www.comses.net/codebases/789bfc4e-a645-4b05-91f1-b91260e3576e/releases/1.0.0/`. Every instance of our WBT opinion dynamics model (as well as the other models in the benchmark of Appendix A) contains N agents. The issue positions of each of these agents are represented by D-dimensional opinion vectors $\mathbf{o}^1, ..., \mathbf{o}^N$. Each component represents the agent's attitude to a specific issue, quantified as a real number between -1 and 1. When a new simulation starts, the components of every opinion vector of each agent are initiated uniformly at random. The simulation then proceeds in discrete time steps $t = 1, ..., T$. The state of the model at time t can be characterized by a $N \times D$ opinion matrix $\mathbf{O}_t$, where the opinion vectors of all N agents are represented as row vectors.

At each time step, all agents are selected in random order. Each agent i then interacts with a randomly chosen other agent j. These interaction pairs are chosen purely at random, without assuming any underlying social network or neighborhood structure [as in 31]. We model interactions between agents as unilateral, meaning if agents i and j interact, agent i is influenced by j, but not vice versa (of course, influence in the reverse direction can occur at another time). We run each simulation until it converges to a stable state. This stable state is reached if, for five consecutive time steps, the changes in the opinion matrix $|\mathbf{O}_t - \mathbf{O}_{t-1}|$ are not larger than expected based on the noise level z (see following Section).

4.2 Opinion Exchange under Evaluative Extremeness

An interaction between two agents happens in three steps: First, agent i determines its attitude towards agent j, $\mathcal{A}(i, j)$, following Equation 3. Second, agent i adjusts its opinion vector $\mathbf{o}^i$ to increase its balance with $\mathcal{A}(i, j)$ and $\mathbf{o}^j$, following Equation 6. These two steps are shown in Figure 4a. In the third step, every agent's opinion vector is affected by a noise vector in which each entry is independently drawn from a normal distribution with mean zero and standard deviation z. This way, the parameter z controls the level of noise in the simulation. Adding certain amount of noise is important, since it has been found that, for bounded confidence models, polarized states that are stable in the absence of noise become unstable under low noise levels [32, 4]. By setting $z > 0$ and simulating the model several times, we ensure that our findings hold in realistic opinion dynamics scenarios affected by factors not covered by the model.

We implement evaluative extremeness in our model by changing the functional form of $f(x)$ in Equation 3. First, we use the sigmoid LOESS curve retrieved from our analysis of the ANES data as functional form of $f(x)$ (specifically the one for Hillary Clinton; see Section 3.4).

We then develop a stylized version of this LOESS function, which is able to encode varying degrees of evaluative extremeness:

$$f(x) = \text{sign}(x) \cdot |x|^{1-e} \tag{7}$$

where e is a free parameter between 0 and 1, quantifying the degree of evaluative extremeness. The function is monotonically increasing, and its range is confined to the interval $[-1, 1]$.

Figure 4b, shows the sigmoid shape of this function for various parameter values, resembling the empirical results of Section 3.4. If $e > 0$, the function transforms input values into more extreme output values: positive values x become more positive, and negative values more negative. If $e = 0$, there is no evaluative extremeness, making $f(x)$ the identity function. The larger the value of e, the more similar the transformation becomes to a step function.

4.3 Metrics

To quantify the different aspects of hyperpolarization, we will apply three different metrics to the opinion matrix $\mathbf{O}$: A metric of opinion extremeness, $E(\mathbf{O})$, a metric of issue constraint, $C(\mathbf{O})$, and a direct metric of hyperpolarization, $H(\mathbf{O})$, that we design to capture the coexistence of opinion extremeness and issue constraint.

First, we quantify the extremeness aspect of hyperpolarization, $E(\mathbf{O})$, as the standard deviation of issue positions. If there is more than one issue dimension, we compute the arithmetic mean of the standard deviations on all different issue dimensions, i.e., the columns of the opinion matrix $\mathbf{O}$. Second, we quantify issue constraint, $C(\mathbf{O})$, as the average inter-correlation between opinion dimensions. More precisely, we compute Pearson correlations between all pairs of opinion dimensions, then apply the Fisher Z transformation to the absolute correlation values, and finally compute the arithmetic mean of the transformed values. In the last step, we back-transform this average into a correlation value by applying the inverse Fisher Z transformation. This average correlation is 1 if there is maximal issue constraint, and near zero if opinion values are completely independent.

And third, we design a metric to directly measure hyperpolarization, $H(\mathbf{O})$, which takes into account both extremeness and constraint. If we compute all distances between unordered pairs of opinion vectors, it can be shown that the sum of these pairwise distances is maximal for the case of maximum polarization. Furthermore, if we square the pairwise distances between opinion vectors, their sum is *uniquely* maximal for the case of maximal polarization, i.e., there is no opinion matrix $\mathbf{O}$ that is not maximally polarized and still reaches the maximal value. The metric takes the following form:

$$H(\mathbf{O}) = \frac{1}{\delta_{max}^2} \frac{4}{N^2} \sum_{1=i<j}^{N} \delta(\mathbf{o}^i, \mathbf{o}^j)^2 \tag{8}$$

where $\delta(\mathbf{x}, \mathbf{y})$ is the (Euclidean or Manhattan) distance between vectors $\mathbf{x}$ and $\mathbf{y}$, and δ_{max} is the maximally possible distance between two vectors, which is $\delta_{max} = 2D$ for Manhattan, and $\delta_{max} = \sqrt{4D}$ for Euclidean distance. The first two terms on the right hand side serve to rescale $H(\mathbf{O})$ between 0 and 1.

$H(\mathbf{O})$ is sensitive to the number of internally unanimous, mutually opposed groups in a political system, i.e., it decreases if the number of groups increases above two. $H(\mathbf{O})$ is also sensitive to the relative size of groups, meaning it decreases if one group becomes bigger than the other. $H(\mathbf{O})$ is zero if and only if there is complete consensus in a political system. This metric captures the definition of hyperpolarization we gave in the introduction, where the polarization of a political system is maximal if the population is separated into two equal sized groups, who completely agree on all policy issues internally, but are maximally opposed to each other.

4.4 Simulation Outcomes

Figure 5 shows four snapshots of a simulation using the LOESS fit of attitudes towards Hillary Clinton as the functional form of $f(x)$. This simulation was run with six issue dimensions, and a moderate noise level ($z = 0.01$). The initial condition of a random distribution of opinions at $t = 1$ is followed by a transient state of consensus ($t = 10$). However, as time evolves, two groups emerge and start separating from each other ($t = 22$), converging to a point of near maximum hyperpolarization at $t = 60$.

We analyze the conditions in which our WBT opinion dynamics model produces hyperpolarization in an exhaustive exploration for several values of the number of dimensions D, of the evaluative extremeness parameter e, and of the noise parameter z. For all simulations, we set the number of agents in the system to $N = 500$, and the α-parameter, which controls the speed of opinion change and the timescale of the model, to 0.4.

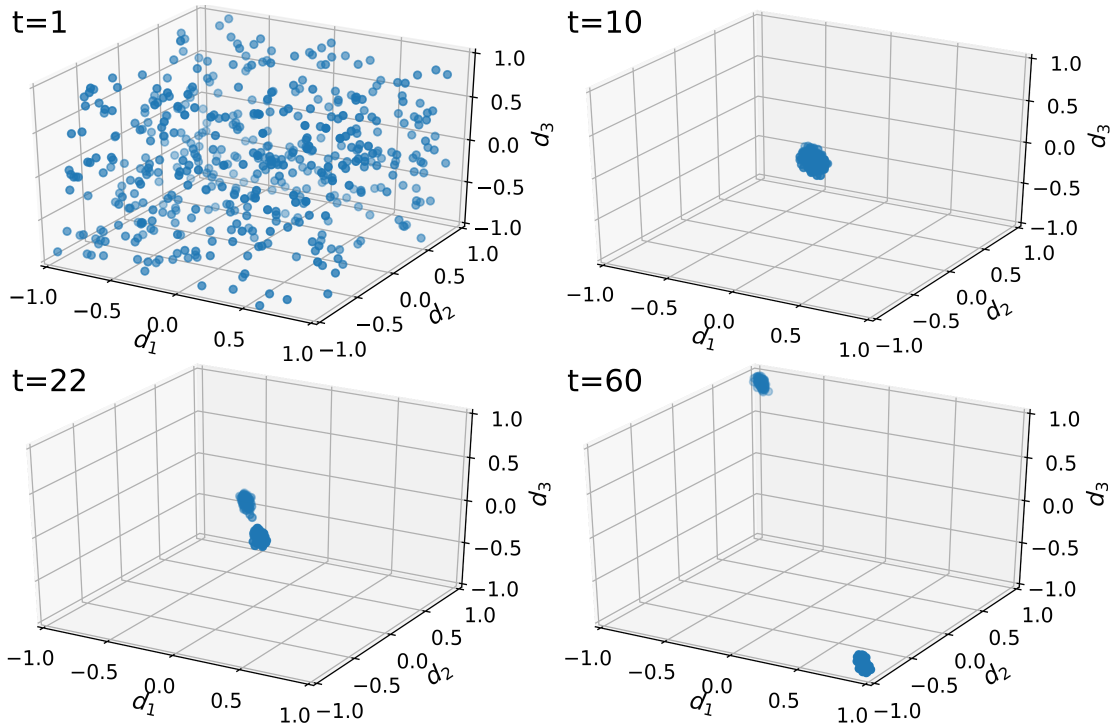


Figure 5: **Time Evolution of WBT-based Hyperpolarization Model.** Four snapshots in the evolution of a model with dimensionality $D = 3$, noise level $z = 0.05$, and evaluative extremeness $e = 0.3$. Blue dots represent endpoints of opinion vectors.

Figure 6a shows the median value of the hyperpolarization metric in 100 simulations (see Section 4.3) against the e-parameter value quantifying evaluative extremeness. If $z = 0$, i.e., if there is no noise, our model generates maximum hyperpolarization for any $e > 0$. This means that half the agents come to rest in each of two opposite corners of the opinion space, corresponding to two maximally opposed political camps, as shown in Figure 5 (lower right panel). In this case without noise, if $e \leq 0$, hyperpolarization abruptly collapses, and the model produces total consensus in the center of opinion space. Higher noise levels require higher values of e to generate hyperpolarization. For $z > 0$, increasing the number of opinion dimensions D has a similar effect: larger values of e are required to generate hyperpolarization for higher dimensionality D. It is worth mentioning that the model never produces states with an intermediate degree of hyperpolarization. In other words, either hyperpolarization emerges completely, or not at all. This appears as an abrupt jump in the median values of hyperpolarization, as can be seen in Figure 6a.

Figure 6b shows the time series of our three metrics in two exemplary model runs, one right below and one right above the critical value of e for the case of with $D = 12$ and $z = 0.01$ (dashed green in Figure 6a). In the long run, the case below the critical value ($e = 0.3$) does not produce hyperpolarization (dashed lines), while the case above the critical value ($e = 0.4$) leads to H near 1 (solid lines). In both simulation runs, the extremeness of issue positions, E, as well as the hyperpolarization measure, H, go down steeply within the first ten time steps. In the solid line scenario, the temporary low state of hyperpolarization coexist with increasing issue constraint, which can be seen in simulation runs as in Figure 5. Issue constraint grows steadily until it reaches its maximum value at $t = 25$. At this point, opinion extremeness also starts to grow, leading to a surge in hyperpolarization due to the existing issue constraint. The simulation run converges to a highly hyperpolarized state at around $t = 80$. In the dashed line scenario (subcritical e), issue constraint also increases over the first 10 time steps, but soon saturates and returns to a low level. This low level of constraint is not enough to trigger the increase of opinion extremeness. Thus, the dashed-line model perpetually remains in a state of minimal hyperpolarization (consensus).

In our model, hyperpolarization is possible when evaluative extremeness is high enough. We exhaustively analyzed previous opinion dynamics models in this multidimensional scenario, including attraction, repulsion, and bounded confidence. As shown in Appendix A, these models can generate consensus and fragmentation into several groups, but none of them generates the hyperpolarized state that can be observed in Figure 5. We further analyzed how this WBT opinion dynamics model, under the presence of a positivity bias, can generate mono-polarization, i.e., when agents reach consensus but very far from the center. We show these results in Appendix B.

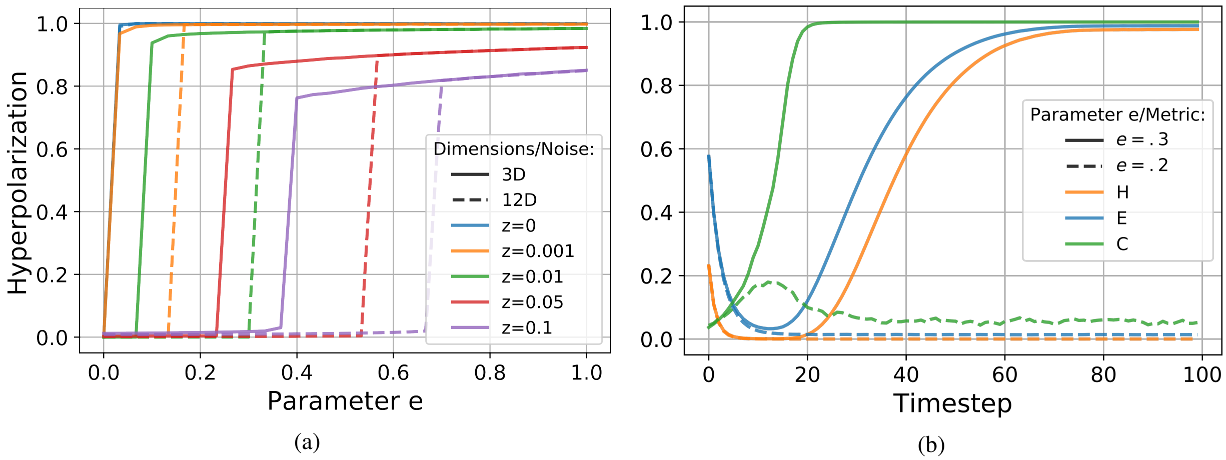


Figure 6: **Simulation Outcomes of WBT Model.** Panel a) shows the median hyperpolarization produced by our WBT-based models under varying degrees of evaluative extremeness (e), dimensionality (D), and noise level (z). Panel b) shows dynamics of hyperpolarization (H), extremeness (E), and constraint (C) for two separate model with $D = 12$ dimensions and a noise level of $z = 0.01$. The level of evaluative extremeness is $e = 0.3$ for the dashed lines and $e = 0.4$ for the solid lines.

5 Discussion

In this article, we present an opinion dynamics model that, based on the psychological principles of WBT, simultaneously generates extreme opinions and issue constraint, the two aspects of hyperpolarization. To our knowledge, this is the first model that generates hyperpolarization without introducing complex social structures or preexisting demographic or opinion correlates. Different from other multidimensional opinion dynamics models, our WBT-based model generates hyperpolarization without introducing already hyperpolarized demographic dimensions [26] and without effectively reducing the dimensionality of the model by introducing issue constraint by design [25]. While complex network structures can generate hyperpolarization in low dimensional spaces [see 19], our model shows that there is no need to assume any particular network structure to generate hyperpolarization, even for cases with many opinion dimensions an random noise.

A decisive determinant of whether our model produces hyperpolarization or consensus is the degree of evaluative extremeness. Even in a model without noise, a minimal degree of evaluative extremeness is necessary to produce hyperpolarization. If this minimal degree of evaluative extremeness is present, however, our WBT based model produces maximal hyperpolarization for arbitrary numbers of opinion dimensions D. This appears as a first-order phase transition with a sharp jump between consensus and hyperpolarization for different values of e. We leave the analytical treatment of this phase transition for further research that can show whether it fits on a wider class of dynamical systems.

Under the right conditions, our WBT-based model generates both opinion extremeness and issue constraint in the long run. As the exemplary model runs in Figure 6b show, the dynamics of extremeness and constraint are not parallel. Rather, opinion extremeness declines in the first part of the model run, while issue constraint rises. In the hyperpolarizing cases, opinion extremeness starts rising again once issue constraint has reached a high level. This poses a testable hypothesis: issue constraint starts to rise before opinion extremeness as part of the combination of balance dynamics and evaluative extremeness. Historical analyses using parliamentary or survey data could be used to test this prediction of the model.

The simulations of our model show the existence of a transient state of low opinion extremeness (consensus) but with rising issue constraint that eventually ends in hyperpolarization. This gives a new interpretation to the current high levels of hyperpolarization across societies: perhaps the rules of opinion dynamics and information spreading have not changed so much and we were on the road to hyperpolarization all along. Our model illustrates that there is no need for echo chambers to exist in order to generate hyperpolarization, it is enough to have cognitive balance dynamics with evaluative extremeness.

WBT encodes a formulation of the backfire effect in which the strength of the backfire is a nonlinear function of the distance between two individuals in opinion space. This nonlinearity could be one of the reasons why the backfire effect has not been found consistently across empirical studies [8, 9, 10, 11, 12, 13]. A second factor that could explain this discrepancy is based on non-political (positive and negative) influences on the quality of interpersonal attitudes. The fact that differences in political opinion do not always determine interpersonal attitudes might explain why the backfire effect does not reliably occur in empirical studies: For example, in a study where political arguments are supplied by the experimenter [9], it is quite unlikely that participants will dislike the experimenter just because they disagree with these arguments. Likewise, in a study where the issue under discussion is emotionally neutral [11], disagreement is unlikely to cause participants to develop negative attitudes towards each other. The lack of negative interpersonal attitudes might explain why these studies could not find a backfire effect. The situation in these studies is more likely to be governed by a general positivity bias, which we discuss in Appendix B. Thus, WBT poses an interplay between interpersonal affect and opinion changes that has the potential to unify the current evidence on the backfire effect.

The central role of interpersonal affect in WBT is supported by recent studies analyzing the emotional underpinnings of popular polarization in the US. Under the label of 'affective polarization', researchers have shown how in the current political climate, even small differences in opinions are mapped into very negative feelings [33, 34, 35].

These outcomes also fit to our observation of evaluative extremeness in attitudes towards US presidential candidates. Our simulations reveal that the degree of evaluative extremeness is key to the emergence of polarization. This of course points us to the question where evaluative extremeness originates from. The psychological literature offers two mechanisms that increase evaluative extremeness: First, it has been found that emotional arousal induces a tendency towards more extreme evaluations [36, 37]. Thus, factors that induce more arousal, especially if they are related to political information, should contribute to increasing polarization. We do not have to look far for potential sources of arousal: The rise of 'infotainment' over the last decades has turned the induction of emotions from a side effect into the main objective of television news programs [38]. Second, research in social identity theory has revealed a tendency to evaluate groups more extremely than individuals [39]. This is in line with studies showing a growing tendency towards tribalistic party identification in the US [40, 35]. However, so far little is known about the root causes of this tendency. One possible line of inquiry would be to enrich our model with a direct representation of affective dynamics [such as 41].

6 Conclusion

Explaining the emergence of opinion extremeness and issue constraint has long been a challenge to theorists in political science and opinion dynamics. In this article, we present a solution based on an extended version of Heider's cognitive balance theory [27], which we call *Weighted Balance Theory*. In contrast to other, more cognition-focused theories, WBT recognizes the importance of social emotions in explaining opinion dynamics. Whether two individuals' opinions become more similar or dissimilar though interaction is determined by whether they like or dislike each other. But adapting their opinions can in turn change individuals' interpersonal attitudes. Thus, in WBT, opinion change happens as a coevolution of interpersonal attitudes and opinions.

WBT explains 1) how individuals form weighted attitude relations towards each other based on their opinions on a variety of issues, and 2) how individuals increase the balance between their opinions and their interpersonal attitudes by adjusting the former to the latter. The driving force behind this coevolution is the need to reduce cognitive imbalance, which occurs when opinions and interpersonal attitudes are in conflict with each other, i.e., when individuals disagree with others they like, or agree with others they dislike.

We tested the first part of this theory, the relation between opinions and interpersonal attitudes, on data from the 2016 ANES survey. The results indicate that WBT can predict respondents' attitudes towards two presidential candidates very well, but that these attitudes tend to be more extreme than our predictions. At least in a political setting, there seems to be a tendency towards a black-and-white world view. We call this tendency *evaluative extremeness* and implement it in the form of a sigmoid reweighing function into a WBT-based opinion dynamics model. We show that this model can reproduce hyperpolarization, as long as a minimal degree of evaluative extremeness is present. Furthermore, it can do this for an arbitrary number of dimensions, and under a considerable degree of random noise. Thus, WBT can explain the emergence of hyperpolarization in a robust and stable way.

WBT also gives us a new interpretation of the so-called *backfire effect*, in the modeling literature also known as repulsion [23, 24]. The standard description of the backfire effect is that individuals with very dissimilar opinions move further away from each other when they interact, rather than approaching a consensus position. WBT suggests a different interpretation of the backfire effect: Under WBT, individuals' opinions only diverge if they dislike each other, and if their opinions are not different enough to be in balance with their negative interpersonal attitude.

In conclusion, WBT can offer a new perspective on the emergence of hyperpolarization, while at the same time integrating research strains from psychology, political science and opinion dynamics into an overarching theoretical framework.

7 Acknowledgments

This research has been funded by the Vienna Science and Technology Fund (WWTF) through project VRG16-005

References

[1] Christopher Hare and Keith T Poole. The Polarization of Contemporary American Politics. *Polity*, 46(3), 2014.

[2] Alan I Abramowitz and Kyle L Saunders. Is polarization a myth? *The Journal of Politics*, 70(2):542–555, 2008.

[3] Morris P Fiorina and Samuel J Abrams. Political polarization in the american public. *Annu. Rev. Polit. Sci.*, 11:563–588, 2008.

[4] Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4), 2017.

[5] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How Social Influence Can Undermine the Wisdom of Crowd Effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025, 2011.

[6] Robert P Abelson. Mathematical models of the distribution of attitudes under controversy. *Contributions to mathematical psychology*, 1964.

[7] Roger L Berger. A necessary and sufficient condition for reaching a consensus using degroot's method. *Journal of the American Statistical Association*, 76(374):415–418, 1981.

[8] Carl I Hovland, OJ Harvey, and Muzafer Sherif. Assimilation and contrast effects in reactions to communication and attitude change. *The Journal of Abnormal and Social Psychology*, 55(2):244, 1957.

[9] Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

[10] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

[11] Michael Mäs and Andreas Flache. Differentiation without distancing. explaining bi-polarization of opinions without negative influence. *PloS one*, 8(11):e74516, 2013.

[12] Károly Takács, Andreas Flache, and Michael Mäs. Discrepancy and disliking do not induce negative opinion shifts. *PloS one*, 11(6):e0157948, 2016.

[13] Thomas Wood and Ethan Porter. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 41(1):135–163, 2019.

[14] Philip E Converse. The nature of belief systems in mass publics (1964). *Critical review*, 18(1-3):1–74, 1964.

[15] Delia Baldassarri and Andrew Gelman. Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology*, 114(2):408–446, 2008.

[16] Kenneth Benoit and Michael Laver. *Party policy in modern democracies*. Routledge, 2006.

[17] Keith T Poole. *Spatial models of parliamentary voting*. Cambridge University Press, 2005.

[18] Frank Schweitzer. Sociophysics. *Physics Today*, 71(2):40–46, 2018.

[19] Andreas Flache and Michael W Macy. Small worlds and cultural polarization. *The Journal of Mathematical Sociology*, 35(1-3):146–176, 2011.

[20] Joan-Maria Esteban and Debraj Ray. On the measurement of polarization. *Econometrica: Journal of the Econometric Society*, pages 819–851, 1994.

[21] Jan Lorenz. Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C*, 18(12):1819–1838, 2007.

[22] Patrick Groeber, Frank Schweitzer, and Kerstin Press. How groups can foster consensus: The case of local cultures. *Journal of Aritifical Societies and Social Simulation*, 12(2):1–22, 2009.

[23] Wander Jager and Frédéric Amblard. Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory*, 10(4):295–303, 2005.

[24] Laurent Salzarulo. A continuous opinion dynamics model based on the principle of meta-contrast. *Journal of Artificial Societies and Social Simulation*, 9(1), 2006.

[25] Sylvie Huet and Guillaume Deffuant. Openness leads to opinion stability and narrowness to volatility. *Advances in Complex Systems*, 13(03):405–423, 2010.

[26] Andreas Flache and Michael Mäs. Why do faultlines matter? a computational model of how strong demographic faultlines undermine team cohesion. *Simulation Modelling Practice and Theory*, 16(2):175–191, 2008.

[27] Fritz Heider. Attitudes and cognitive organization. *The Journal of psychology*, 21(1):107–112, 1946.

[28] Dorwin Cartwright and Frank Harary. Structural balance: a generalization of heider's theory. *Psychological review*, 63(5):277, 1956.

[29] William M Wiest. A quantitative extension of heider's theory of cognitive balance applied to interpersonal perception and self-esteem. *Psychological Monographs: General and Applied*, 79(14):1, 1965.

[30] Norman T Feather. A structural balance model of communication effect. *Psychological Review*, 71(4):291, 1964.

[31] Patrick Groeber, Jan Lorenz, and Frank Schweitzer. Dissonance minimization as a microfoundation of social influence in models of opinion formation. *The Journal of Mathematical Sociology*, 38(3):147–174, 2014.

[32] F. Schweitzer and J. Hołyst. Modelling collective opinion formation by means of active Brownian particles. *European Physical Journal B*, 15(4):723–732, 2000.

[33] Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. Affect, not ideology. a social identity perspective on polarization. *Public opinion quarterly*, 76(3):405–431, 2012.

[34] Shanto Iyengar and Sean J Westwood. Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3):690–707, 2015.

[35] Lilliana Mason. I disrespectfully agree: The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, 59(1):128–145, 2015.

[36] Margaret S Clark, Sandra Milberg, and John Ross. Arousal cues arousal-related material in memory: Implications for understanding effects of mood on memory. *Journal of Verbal Learning and Verbal Behavior*, 22(6):633–649, 1983.

[37] Margaret S Clark, Sandra Milberg, and Ralph Erber. Effects of arousal on judgments of others' emotions. *Journal of Personality and Social Psychology*, 46(3):551, 1984.

[38] Daya Kishan Thussu. *News as entertainment: The rise of global infotainment*. Sage, 2008.

[39] David O Sears. The person-positivity bias. *Journal of personality and Social Psychology*, 44(2):233, 1983.

[40] Lilliana Mason. Distinguishing the polarizing effects of ideology as identity, issue positions, and issue-based identity. In *A paper presented at the Center for the Study of Democratic Politics Conference on Political Polarization: Media and Communication Influences. Princeton University*, 2015.

[41] Frank Schweitzer, Tamas Krivachy, and David Garcia. An agent-based model of opinion polarization based on emotional influence. *Complexity*, 2018. (under review).

[42] David G Myers and Helmut Lamm. The polarizing effect of group discussion. *American Scientist*, 1975.

[43] Cass R Sunstein. The law of group polarization. *Journal of political philosophy*, 10(2):175–195, 2002.

[44] Hermann Brandstätter. Social emotions in discussion groups. In Hermann Brandstätter, James H. Davis, and Heinz Schuler, editors, *Dynamics of Group Decisions*, chapter 6, pages 93–111. Sage Publications, 1978.

[45] Vijaya Venkataramani, Giuseppe Joe Labianca, and Travis Grosser. Positive and negative workplace relationships, social satisfaction, and organizational attachment. *Journal of applied psychology*, 98(6):1028, 2013.

[46] Jerry Boucher and Charles E Osgood. The pollyanna hypothesis. *Journal of verbal learning and verbal behavior*, 8(1):1–8, 1969.

[47] M Matlin and David Stang. The pollyanna principle: Selectivity in language, memory, and thought schenkman pub, 1978.

Appendices

A Higher Dimensional Models of Attraction, Bounded Confidence and Repulsion

In this Appendix Section, we explore how the three mechanisms described in Section 2 (attraction, bounded confidence, and repulsion) behave in higher dimensions. In particular, we want to find out whether bounded confidence and repulsion can generate hyperpolarized opinion distributions.

Attraction: We model the attraction mechanism as a direct approach in multidimensional Euclidean space:

$$\Delta \mathbf{o}^i = \alpha(\mathbf{o}^j - \mathbf{o}^i) + \mathbf{z} \tag{9}$$

where α determines the speed of attraction, and $\mathbf{z}$ is a noise vector drawn from a random normal distribution with mean zero and standard deviation z. Same as its one-dimensional equivalent, a multi-dimensional attraction model always generates complete consensus on the midpoint of the opinion space (Figure 7, top left panel).

Bounded confidence: We implement a multidimensional bounded confidence model by adding a threshold mechanism to the attraction model:

$$\Delta \mathbf{o}^i = \Theta(e - \delta(\mathbf{o}^i, \mathbf{o}^j)) \cdot \alpha(\mathbf{o}^j - \mathbf{o}^i) + \mathbf{z} \tag{10}$$

where Θ is the Heaviside function, $\delta(\mathbf{x}, \mathbf{y})$ is the Euclidean distance between two opinion vectors, and e is a threshold parameter. Depending on the value of e, the bounded confidence model will produce a higher or lower number of clusters in the opinion space, to which surrounding opinion vectors converge. However, as Figure 7 (top right panel) illustrates, these clusters are always symmetrically distributed around the midpoint of the opinion space. Thus, while the bounded confidence model prevents the formation of complete consensus, the model fails to reproduce opinion constraint and thus hyperpolarization. In addition, the multidimensional bounded confidence model is not robust against noise: Even small degrees of noise lead to a convergence of opinion vectors near the center of the opinion space.

Repulsion: Lastly, we implement a multidimensional repulsion model by replacing the threshold mechanism from the bounded confidence model with a term that, beyond a certain distance threshold $epsilon$, changes attraction into repulsion:

$$\Delta \mathbf{o}^i = (e - \delta(\mathbf{o}^i, \mathbf{o}^j)) \cdot \alpha(\mathbf{o}^j - \mathbf{o}^i) + \mathbf{z} \tag{11}$$

The repulsion mechanism can cause opinion vectors to leave the bounds of the opinion space. To prevent this, opinion vectors have to be confined artificially to the interval $[-1, 1]$. Even so, however, the repulsion model does not produce hyperpolarization. Depending on the value of e, the model either converges to complete consensus near the center of the opinion space, or to a state of political fragmentation, as illustrated in Figure 7 (bottom left panel). In this state, the agents' opinion vectors populate all corners of the opinion space in more or less equal numbers. While opinions in this state are maximally extreme, issue constraint is practically zero.

In conclusion, implementing standard opinion dynamics mechanisms in multidimensional spaces does not produce hyperpolarization, but either consensus or opinion fragmentation.

B A Weighted Balance Model of Mono-Polarization

While some degree of hyperpolarization seems to be universal in societies at large, this is not the case in smaller social units, such as court juries, management teams, or church groups. When these groups deliberate an issue, they tend **not** to fission into two opposed groups, but instead converge to a consensus position. This consensus, however, tends to be more extreme than the midpoint of the initial opinions of the groups' members – sometimes even more extreme than the most extreme initial position [42, 43]. It has been shown that this effect is stronger to the extent that group members have positive interpersonal attitudes towards each other [44].

Unfortunately, while qualitatively very different from hyperpolarization, this phenomenon has been dubbed 'group polarization'. To better distinguish the two phenomena, we call group polarization *mono-polarization*.

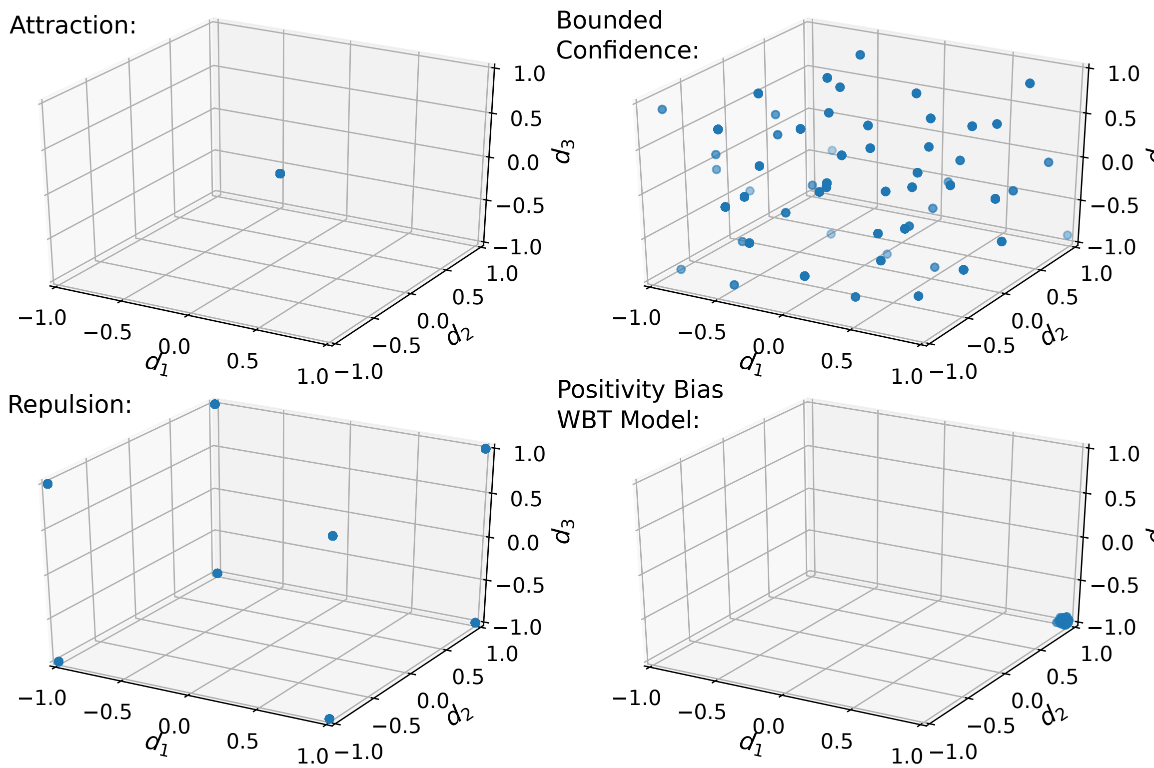


Figure 7: **Model outcomes under four different opinion exchange mechanisms.** $D = 3$ for all models. Noise level $z = 0$ for attraction, bounded confidence, and repulsion model, and $z = 0.01$ for positivity bias WBT model.

We quantify mono-polarization, $M(\mathbf{O})$, as the sum of the dot products between all pairs of opinion vectors, normalized between 0 and 1:

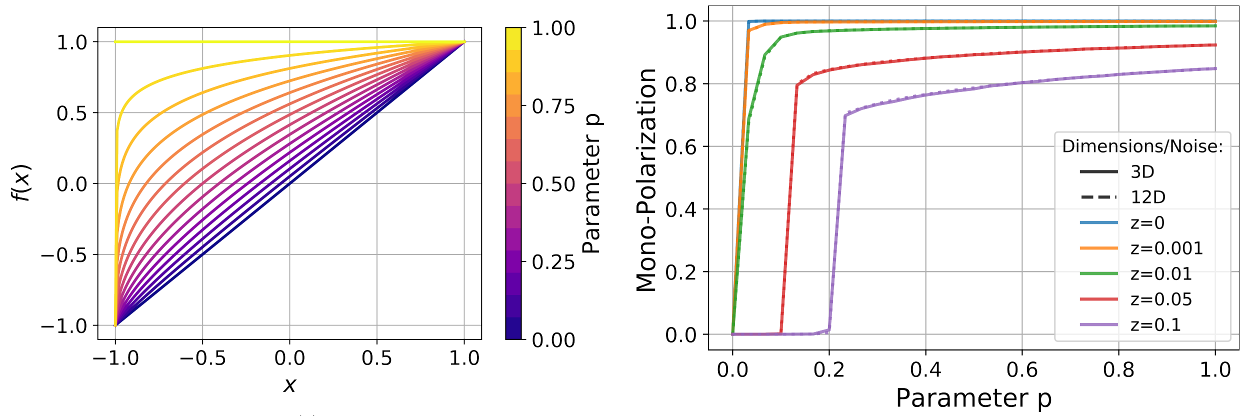
$$M(\mathbf{O}) = \frac{1}{D}\frac{2}{N(N-1)}\sum_{1=i<j}^{N} \mathbf{o}^i \cdot \mathbf{o}^j \tag{12}$$

The existence of mono-polarization poses a further challenge to theories of polarization: If deliberating groups usually come to an (extreme) consensus, why does this not happen in societies at large? Any theory of political polarization should also be able to explain this apparent discrepancy. In this section, we show that WBT can also account for mono-polarization.

Based on our empirical results from Section 3.4, we have assumed that individuals have a tendency towards evaluative extremeness. We implemented this tendency into our opinion dynamics model, and could show that it plays a crucial role in the emergence of polarization. However, there is reason to assume that evaluative extremeness is not universal in social relationships. Most social settings seem to be dominated by a bias towards positive evaluations of other individuals. For example, it has been found that employees expect to have positive relations with co-workers, and are strongly irritated by negative relations [45]. This preference for positive relations is part of a wider phenomenon called 'Pollyanna Principle', which describes a preference for positive content in memory, cognition, and language [46, 47]. Interestingly, it has been shown that positive relations among group members increase the strength of the mono-polarization effect [44, 43].

We implement this *positivity bias* into our model by replacing the functional form of $f(x)$ in Equation 3. Instead of the identity function or the evaluative extremeness function (Equation 7), we use the following equation:

$$f(x) = 2\left(\frac{x+1}{2}\right)^{1-p} - 1 \tag{13}$$



(a) (b)

Figure 8: **Positivity Bias WBT Model.** Panel a) shows the mono-polarization produced by WBT based models under varying degrees of positivity bias (p), dimensionality (D), and noise level(z). Panel b) shows dynamics of hyperpolarization (H), extremeness (E), and mono-polarization (M) for two separate model with $D = 12$ dimensions and a noise level of $z = 0.05$. The level of evaluative extremeness is $p = 0.1$ for the dashed lines and $p = 0.3$ for the solid lines.

where the parameter p controls the extent of positivity bias: The closer p is to 1, the higher the positivity bias. The effect of this function is illustrated in Figure 8a: While all interpersonal attitudes are made more positive, the positivity bias is strongest for negative attitudes, reflecting the tendency to avoid negative social relations. If $p = 1$, the output of the function is $f(x) = 1$, independent of the input.

Once we implement the positivity bias in our model, it ceases to generate hyperpolarized opinion configurations. Instead, we obtain mono-polarization: All opinion vectors end up in a single corner of the opinion space, reflecting a maximally extreme consensus position (see Figure 7, bottom right panel).

Figure 8b shows the degree of mono-polarization produced by simulation runs with varying positivity bias p.

If the positivity bias p is strong enough in relation to the noise level, the model produces maximal mono-polarization. Different from the case of hyperpolarization, this seems to be independent from the number of issue dimensions in the model.

In conclusion, a slight tweak of our WBT based model changes its dynamics radically: Instead of fissioning into two opposed groups, agents now converge to a consensus, and then move together towards more and more extreme positions – reflecting the phenomenon of group mono-polarization. This demonstrates that, based on empirically plausible changes in the underlying mechanism, WBT can replicate qualitatively very different empirical phenomena.