Explaining Attitudes from Behavior: A Cognitive Dissonance Approach[*]

Avidit Acharya[†] Matthew Blackwell[‡] Maya Sen[§]

May 28, 2015

Word Count: 10,491

Abstract

The standard approach in positive political theory posits that action choices are the consequences of attitudes. Could it be, however, that an individual's actions also affect her fundamental preferences? We present a broad theoretical framework that captures the simple, yet powerful, intuition that actions frequently alter attitudes as individuals seek to minimize cognitive dissonance. This framework is particularly appropriate for the study of political attitudes and enables political scientists to formally address questions that have remained inadequately answered by conventional rational choice approaches – questions such as "What are the origins of partisanship?" and "What drives ethnic and racial hatred?" We illustrate our ideas with three examples from the literature: (1) how partisanship emerges naturally in a two party system despite policy being multidimensional, (2) how ethnic or racial hostility increases after acts of violence, and (3) how interactions with people who express different views can lead to empathetic changes in political positions.

[*] Most of the results in our previous working paper titled "Attitudes Shaped by Violence" have been incorporated into this paper. We thank Joshua Kertzer, David Laitin, Ken Shepsle, and Paul Sniderman for helpful feedback. Ruxi Zhang and Alexandra Pagano provided outstanding research assistance.

[†] Assistant Professor of Political Science, Stanford University.

[‡] Assistant Professor of Government, Harvard University.

[§] Assistant Professor of Public Policy, Harvard University.

1 Introduction

What are the origins of ethnic hatred? How do young people become lifelong Republicans or Democrats? What causes people to change their most deeply held political preferences? These questions are the bedrock of many inquiries within political science. Numerous articles and books study the determinants of racial attitudes, why partisanship exists, and how political persuasion does or does not engender attitude change. Throughout, a theme linking these seemingly disparate literatures is the formation and evolution of political and social attitudes as an object of study.

Although the empirical literature in these areas is well developed, a general formal theory of attitude change, however, does not exist. This is in part because much of positive political theory has focused on traditional rational choice approaches, which derive the action choices of individuals from immutable preferences. In this paper, we outline a different approach to positive political theory that takes the perspective that attitudes are often the *consequence* of actions—the opposite of what is posited by standard rational choice theory. That is, actions do not necessarily reflect the fixed, immutable preferences of individuals; they instead may be chosen for a variety of reasons, including imitation, experimentation, and habit. Attitudes or preferences then adjust to justify the behaviors that were adopted.

Why would individuals change their attitudes in response to the actions they adopt? Our answer builds upon a rich literature beginning with Festinger (1957) that posits that actions affect attitudes primarily through a concept that social psychologists call *cognitive dissonance*. According to cognitive dissonance theory, an individual experiences a mental discomfort after taking an action that seems to be in conflict with his or her starting attitude. Individuals then change their attitudes to conform more closely with their actions, leading to an important source of attitude formation and change.

As we argue in this paper, this approach can be fruitfully applied to many settings in politics where individuals make choices or take actions and then later change their attitudes to be consistent with those choices. Because the theory views attitudes, preferences, and ideology as the consequences of actions, our approach is well suited to explore instances where actions and choices are the main independent variables and attitudes and preferences serve as the dependent variables. A vast subfield of political science—political behavior—is concerned with the origins of partisanship, ideology, racial attitudes, ethnic identification, etc. We demonstrate how a formal approach built upon the insights of cognitive dissonance theory can help us understand the sources of these attitudes. We show how the traditional rational choice approach can be extended in a straightforward way to incorporate the insights of cognitive dissonance theory.

We organize this paper as follows. First, we provide an overview of our approach in Section 2 through a simple stylized model. We then illustrate the applicability of our modeling approach via three more concrete examples. The first, presented in Section 3, demonstrates how the cognitive-dissonance based approach can explain the development of partisan affiliation. The second, presented in Section 4, shows how it can explain the emergence and persistence of ethnic or racial hostility from acts of violence. The third, presented in Section 5, demonstrates how individuals with differing attitudes but who feel empathy, or kinship, toward one another can find compromise by adjusting their preferences. We conclude in Section 6 with a discussion of other areas in which the approach might be fruitfully applied to further understand the politics of attitudes.

2 Actions Can Affect Attitudes

Incorporating the notion of cognitive dissonance into a positive formal theory of politics enables us to explicitly take account of the facts that (1) political preferences need not always be fixed and (2) actions can lead to changes in individual (and therefore group) preferences.

We justify our approach with evidence from the early days of cognitive dissonance theory. In a famous experiment Davis and Jones (1960) asked a treatment group of subjects (consisting of college students) to tell a fellow student that they were shallow and untrustworthy. They then asked both treated and control subjects to evaluate the character of the targeted student. They found that members of the treatment group were more likely to revise their opinions of the targeted student downward. Glass (1964) presents a similar experiment where subjects, who were all on record as being opposed to the use of electrical shocks in psychology studies were asked to shock another participant, who actually was a member of the research team. Subjects that engaged in the electric shocks negatively revised their opinions of those they shocked. In both of these studies, participants showed cognitive dissonance when engaging in an action (negative criticism or electric shocks) that is harmful to another person and, to alleviate such feelings, adopted a new, more negative attitude toward their victims. Here, the lowered opinion of the other participant is a *consequence* of the choice to harm them, developed as a result of trying to reduce the mental stress caused by engaging in behavior that is inconsistent with their self-image of being a good person.[1]

These illustrations are simple, but they demonstrate a basic human trait that has been well documented within the social psychology literature and increasingly so within behavioral economics: making a choice or undertaking an action—oftentimes exoge-

[1] Other famous examples include Festinger, Riecken and Schachter (1956) and Festinger and Carlsmith (1959). Festinger, Riecken and Schachter (1956) observed the actions of cult members who learned that, despite teachings to the contrary, the earth had not been destroyed. Instead of abandoning the cult, the members actually reformulated the teachings (i.e., changed their beliefs), pointing to their devotion as the reason for the earth's continued existence. Festinger and Carlsmith (1959) required study participants to engage in the boring task of turning pegs for an hour. Afterwards, participants were paid either $1 and $20 to convince another potential participant to undertake the exercise. The authors found that those paid $1 afterward rated turning the pegs as the most enjoyable. The explanation, the authors reasoned, lay in cognitive dissonance: rather than lie about something unenjoyable that was poorly remunerated, the participants convinced themselves more strongly of the pleasurableness of the experience. We should note, however, that some social psychologists have provided alternative theories to account for these types of examples, including Bem (1967)'s theory of self-perception and Cooper and Fazio (1984)'s controversial theory of aversive consequences.

nously, blindly, or even within a choice set involving comparable choices—can lead a person to develop an increased preference over time for the chosen alternative (Festinger and Carlsmith, 1959; Festinger, 1957; Brehm, 1956). We discuss specific political science examples below, but note here that the theory and related findings have extended to a variety of fields across business, economics, and sociology. For example, cognitive dissonance has explained *ex post* justifications of immoral or dangerous behavior (e.g., Akerlof and Dickens, 1982), resource allocation (e.g., Konow, 2000), and economic and business investments (e.g., Staw, 1976). In addition, the theory has been confirmed in experiments, including experiments involving young children, animals, and amnesiacs (Lieberman et al., 2001). Egan, Santos and Bloom (2007), for instance, document the fact that both children and monkeys choosing a certain kind of toy or color of candy would then, in the next round of experimentation, devalue other toys or colors, a finding replicated by Egan, Bloom and Santos (2010) when the subjects initially chose the objects blindly (see Chen and Risen (2010), however, for challenges). Neurological studies examining experimental subjects have also documented tangible physiological responses, suggesting that subjects form stronger commitments once they have selected from several choices (Sharot, De Martino and Dolan, 2009).

2.1 Stylized Model

We now formally illustrate the idea that a decision maker may adjust his attitude to conform more closely with an action that was chosen (or assigned to him). Consider a person with a starting attitude $x^o \in \mathbb{R}^n$, which is fixed. The individual takes an action a that maximizes a function $\pi(a, x^o)$ over a set $A \subseteq \mathbb{R}^n$. This action may be chosen by the individual, or it may be assigned. In the case that it is assigned, the action taken is exogenous to the individual's preferences, and the rest of the example applies.[2]

[2] What we mean is that the function $\pi(a, x^o)$ may not reflect the individual's *welfare*. It could, for example, be taken to reflect objectives of another person (e.g., the researcher) who is assigning the action to the individual.

Suppose that the maximizer of $\pi(a, x^o)$ over $A \subset \mathbb{R}^n$ is unique, and let a^* denote it. If $a^* \neq x^o$ then the decision-maker experiences a cognitive dissonance cost, and may want to change his attitude from x^o to a new attitude x^n chosen from a set $X \subseteq \mathbb{R}^n$. Changing his attitude is also costly. We denote the cost of changing the attitude by $\delta(x^n, x^o)$ and the cognitive dissonance cost by $d(a^*, x^n)$, where $d(\cdot, \cdot)$ and $\delta(\cdot, \cdot)$ are both metrics on $\mathbb{R}^n$. Substantively, this means that the intensity of this discomfort increases with the discrepancy between the individual's initial attitude and the behavior that he has chosen or has been assigned, consistent with the empirical literature. The individual's problem is then

$$\begin{aligned} &\max_{x^n \in X} - d(a^*, x^n) - \delta(x^n, x^o) \\ &\qquad \text{subject to } a^* = \arg\max_{a \in A} \pi(a, x^o) \end{aligned} \tag{1}$$

If the action a^* is assigned, rather than chosen by the individual, then a^* is determined exogenously and the decision-maker only faces the first line of (1).

For example, take $A = X = \mathbb{R}$ and $\pi(a, x^o) = (x^o + \rho) \log a - a$. The value of a that maximizes $\pi(a, x^o)$ is $a^* = x^o + \rho$. This is the value of a^* that the individual chooses. Thus, if $\rho \neq 0$ then $a^* \neq x^o$ and the decision maker suffers a positive cognitive dissonance cost if he doesn't change his attitude from x^o. Suppose that $d(a^*, x^n) = |a^* - x^n|$ and $\delta(x^n, x^o) = \frac{1}{2\kappa}(x^n - x^o)^2$ where $\kappa > 0$ is a parameter that weights the cost of changing one's attitude against the cognitive dissonance cost. Substituting the value of $a^* = x^o + \rho$ into the cognitive dissonance cost, the individual's problem then becomes

$$\max_{x^n \in \mathbb{R}} - |(x^o + \rho) - x^n| - \frac{1}{2\kappa}(x^n - x^o)^2$$

The value of x^n that solves this maximization problem is

$$x^n = \begin{cases} \min\{x^o + \kappa, x^o + \rho\} & \text{if } \rho \geq 0 \\ \max\{x^o - \kappa, x^o + \rho\} & \text{if } \rho < 0 \end{cases}$$

Thus, if ρ is a positive number then the individual revises up his attitude whereas if ρ is a negative number then he revises down her attitude. If κ is small in comparison to the magnitude of ρ, then the individual chooses to live with some cognitive dissonance since the cost of revising his attitude to eliminate all of the dissonance is too large in comparison to the distance between his initial attitude and action choice.

2.2 Comparison with Existing Approaches

Our approach builds on both the traditional rational choice approach as well as the traditional theory of cognitive dissonance developed in the social psychology literature. In this regard, four points are worth noting.

First, when an action that an individual chooses (or might choose) is in conflict with the individual's attitude, rational choice theory might predict that she will quit choosing the action (or avoid it). Depending on the individual's preferences, the assumption guiding the traditional rational choice approach is that preferences dictate actions, rather than vice versa.[3] In our case, actions also affect preferences.

Second, to the extent that individuals have different preferences, traditional rational choice theory tends to predict that it is because of structural factors, such as when high income earners oppose redistribution policy while low income earners support it—income being the structural factor here (e.g., Meltzer and Richard, 1981). Moreover, even when individuals have different beliefs, rational choice theory has insisted that it must be because they received different private information.[4] If all information were to be made public, then all non-structural disagreement would disappear. Our approach, however, opens the possibility that preferences may differ because past actions differ.[5]

[3] An example to the contrary is Dietrich and List (2011); see also Dietrich and List (2013) for a discussion of the issue within economics.

[4] This view is known as the *Harsanyi doctrine* (Harsanyi, 1967, 1968*a*,*b*).

[5] We should note, however, that in standard rational choice theory, an individual's preference over actions may change after an action is taken if the act of choosing leads to *learning*. For example, suppose an agent receives a noisy payoff from taking an action that leads her to update her beliefs about a payoff-relevant state of the world. If the agent's optimal action varies with the state, then beliefs about the state may drive her preferences over actions in future decisions. In this case, however, the agent's payoffs to

Third, the stylized example above also demonstrates how cognitive dissonance theory is closely related related to a broader interpretation of the rational choice approach, and may even be considered a part of it. After all, the decision maker in the example chooses attitudes to *optimize* an objective function, which can be interpreted as a *utility*. She seeks to minimize certain *costs*, which happen to be psychological rather than material. Our model uses the language of the rational choice approach—"optimize utility given costs"—to explain attitude change. The result, however, is that individuals do bring their attitudes to more closely align with their actions.

Fourth, it is worth noting that our paper is not a formalization of cognitive dissonance theory; rather, it is a formal theory of attitude change that is inspired by ideas that were developed in the cognitive dissonance literature. In particular, our approach includes the costs of attitude change and weighs these costs against the cognitive dissonance costs of not changing attitudes sufficiently. It also remains agnostic as to what drives agents to adopt actions that are in dissonance with their starting attitudes.

We now turn to demonstrating the similarities and differences between our approach and the traditional approaches more fully by applying cognitive dissonance based approach to three examples fundamental to the understanding of politics: (1) the development of partisanship, (2) the construction of ethnic or racial hostility through violence, and (3) empathetic changes in attitudes as a result of socialization.

3 Partisanship and Voting

We start by developing a theory of partisanship based on voters who experience psychological costs due to cognitive dissonance. The issue space is multidimensional and voter preferences are initially distributed across these multiple dimensions, a setting that has posed significant challenges in previous theoretical work (McKelvey,

various actions does not change, although her beliefs do. Our approach also allows the intrinsic payoffs to change even when there is no uncertainty as, for instance, the agent may grow a *taste* for a particular action after choosing it once.

1976; Plott, 1967). In our model, however, political competition between two policy-motivated parties endogenously produces an electorate that is essentially unidimensional in the sense that voter preferences become perfectly correlated across dimensions. This occurs because voters who want to minimize cognitive dissonance will adjust their policy preferences toward the platform of the party whom they support. This results in multidimensional policy preferences over time collapsing onto a single dimension, with partisanship emerging as a natural outcome of this process.

3.1 Model Setup

The electorate is a continuum of unit mass. Policy is two dimensional with a generic policy denoted $(x, y) \in \{0, 1\} \times \{0, 1\}$. The policy space, therefore, consists of four possible policies: $(0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$. We interpret 0 as the left-wing policy on each of the two issues and 1 as the right-wing policy. Voter ideal points are initially evenly distributed across the policy space so that, for each policy (x, y), exactly one quarter of the electorate has ideal point (x, y).

Two parties labeled L and R have ideal policies $(0, 0)$ and $(1, 1)$ respectively; in particular, if (x^w, y^w) denotes the policy chosen by the winning party that comes to power, then the payoff to each party L and R is given by

$$\begin{aligned} v^L(x^w, y^w) &= -\mathbb{1}_{\{x^w \neq 0\}} - \lambda^L \mathbb{1}_{\{y^w \neq 0\}} \\ v^R(x^w, y^w) &= -\mathbb{1}_{\{x^w \neq 1\}} - \lambda^R \mathbb{1}_{\{y^w \neq 1\}} \end{aligned}$$

where $\lambda^L > 0$ and $\lambda^R > 0$ are parameters that weight the second issue relative to the first, and $\mathbb{1}_{\{.\}}$ is the indicator function. Each party runs on one of the four possible policies. Denote by (x^L, y^L) the policy that party L runs on and (x^R, y^R) the policy that party R runs on. The parties are committed to the policies they run on, and need not run on their ideal points.

Although voters are initially evenly distributed across the four possible ideal points, each voter may change his or her ideal point in the course of deciding which party to support. A voter with initial ideal point (x^o, y^o) may change her ideal point to a new one (x^n, y^n). If $(x^n, y^n) = (x^o, y^o)$, then the voter has chosen not to change her initial ideal point. The voter's payoff from being a supporter of party $j = L, R$ and adopting the ideal point (x^n, y^n) is

$$u\left((x^n, y^n), (x^j, y^j) \mid (x^o, y^o)\right) = -\left(\mathbb{1}_{\{x^n \neq x^j\}} + \kappa \mathbb{1}_{\{x^n \neq x^o\}}\right) - \gamma\left(\mathbb{1}_{\{y^n \neq y^j\}} + \kappa \mathbb{1}_{\{y^n \neq y^o\}}\right)$$

where $\gamma > 0$ and $\kappa > 0$ are parameters. Therefore, if a voter supports a party j whose position differs from her on the first issue x, then she experiences a cognitive dissonance cost of -1. If her position differs on the second issue then she experiences a cognitive dissonance cost of $-\gamma$. κ is the psychological cost of changing her ideal point on the first issue and $\gamma\kappa$ is the analogous cost of changing her ideal point on the second issue.

Note that the two issues are separable and γ represents the salience of the second issue vis-à-vis the first. While we assume (for parsimony) that the parameter κ is constant across the electorate, voters may differ in the parameter γ, which varies on an interval $[\underline{\gamma}, \overline{\gamma}]$ with $0 < \underline{\gamma} < 1 < \overline{\gamma}$. Thus, some voters may consider the first issue more important than the second, while others consider the second issue to be more important than the first. We assume that $\kappa < 1$ which says that the cost of changing one's ideal point is small. This assumption enables us to focus on the interesting case where individuals change their preferences in response to the positions taken by the two parties. It is straightforward to solve the model without this assumption, but there would be several more cases to handle, some of which duplicate our main result and the rest of which are substantively uninteresting.

The joint distribution of γ and voter ideal points depends on a state variable, which we denote $\theta \in \{\theta_\ell, \theta_r\}$. For each θ the marginal distribution of γ is continuous on $[\underline{\gamma}, \overline{\gamma}]$. Given the state θ, let $\phi^-_{(x,y)}[\theta]$ denote the fraction of voters with ideal point (x, y) for whom $\gamma < 1$, and let $\phi^+_{(x,y)}[\theta]$ the fraction of voters with ideal point (x, y) for whom $\gamma > 1$. By the assumption that the marginal distribution of γ is continuous in both states, the fraction of voters for whom $\gamma = 1$ is always zero; therefore, $\phi^{-}_{(x,y)}[\theta] + \phi^{+}_{(x,y)}[\theta] = 1$ for all (x, y) and both realizations of θ. We assume that the two states θ_ℓ and θ_r are equally likely, and that:

(i) $\phi^{-}_{(0,1)}[\theta_\ell] + \phi^{+}_{(1,0)}[\theta_\ell] > 1$ and $\phi^{-}_{(0,1)}[\theta_r] + \phi^{+}_{(1,0)}[\theta_r] < 1$

(ii) $\phi^{-}_{(0,0)}[\theta_\ell] + \phi^{+}_{(1,1)}[\theta_\ell] > 1$ and $\phi^{-}_{(0,0)}[\theta_r] + \phi^{+}_{(1,1)}[\theta_r] < 1$

These assumptions simply imply that state θ_ℓ unambiguously favors party L and state θ_r unambiguously favors party R. Since we assumed that the two states are equally likely, neither party has an advantage *ex ante*.

The timing of events is as follows:

1. Parties announce their positions (x^L, y^L) and (x^R, y^R).

2. The state $\theta \in \{\theta_\ell, \theta_r\}$ is drawn, which determines the initial distribution of voter ideal points.

3. Each voter takes as given his or her initial ideal point (x^o, y^o) and decides what position to take on each of the two issues (x^n, y^n) and which party to support.

4. Voters vote expressively for the party that they support. That is, a voter with initial ideal point (x^o, y^o) votes for party L if
$$\max_{(x^n,y^n)} u\left((x^n, y^n), (x^L, y^L) \mid (x^o, y^o)\right) > \max_{(x^n,y^n)} u\left((x^n, y^n), (x^R, y^R) \mid (x^o, y^o)\right)$$
and votes for party R if the reverse inequality holds. Voters for whom the two sides are equal may vote for either of the two parties.

Previous models of voting have interpreted the cost to a voter from voting for a party whose position on an issue differs from her own position as an instrumental cost, as opposed to a psychological one. Our interpretation of this cost as being psychological is more in line with the assumption that voting is expressive—an assumption that is typical for voting models with a continuum of voters like this one (see, e.g., Roemer, 2000, on both these points).

3.2 Analysis

The following proposition states that under the setup of the previous section the two parties run on their ideal points, and voters sort into left partisans and right partisans.

Proposition 1 *There is a unique Nash equilibrium in which parties* L *and* R *run on their ideal policies* $(x^L, y^L) = (0,0)$ *and* $(x^R, y^R) = (1,1)$. *Voters with initial ideal points* $(0,0)$ *and* $(1,1)$ *do not change their ideal points. Voters with initial ideal point* $(0,1)$ *and preference parameter* $\gamma < 1$ *change their ideal points to* $(0,0)$ *while those with parameter* $\gamma > 1$ *change their ideal points to* $(1,1)$. *Voters with initial ideal point* $(1,0)$ *and preference parameter* $\gamma < 1$ *change their ideal point to* $(1,1)$ *while those with parameter* $\gamma > 1$ *change their ideal point to* $(0,0)$.

Proof. Suppose that $(x^L, y^L) = (0,0)$ and $(x^R, y^R) = (1,1)$ so that both parties announce their ideal policies as their positions. To show that this is an equilibrium, we first start by proving the claim that voters do as the proposition describes. Clearly, voters with initial ideal point $(x^o, y^o) = (0,0)$ will support party L and not change their ideal point. Voters with initial ideal point $(x^o, y^o) = (1,1)$ will support party R and not change their ideal point.

For voters with initial ideal points $(x^o, y^o) = (0,1)$ the payoff from not changing their preference and supporting party L is $-\gamma$. The payoff from not changing their preference and supporting party R is -1. If these voters are to support party R and change their ideal point, then the most profitable ideal point to choose is $(x^n, y^n) = (1,1)$ in which case their payoff is $-\kappa$. If they are to support party L and change their ideal point, then the most profitable ideal point to choose is $(x^n, y^n) = (0,0)$. The payoff from this is $-\kappa\gamma$. Now, because of our assumption that $\kappa < 1$ we know that $-\kappa\gamma > -\gamma > -1$ and $-\kappa\gamma > -\kappa$ whenever $\gamma < 1$. This means that voters with initial ideal point $(x^o, y^o) = (0,1)$ and parameter $\gamma < 1$ support party L and change their ideal point to $(x^n, y^n) = (0,0)$. On the other hand, if $\gamma > 1$ then $-\kappa > -1 > -\gamma$ and

$-\kappa > -\kappa\gamma$. This means that voters with initial ideal point $(x^o, y^o) = (0,1)$ change their ideal points to $(x^n, y^n) = (1,1)$ and support party R.

Therefore, to summarize, among voters with initial ideal point $(x^o, y^o) = (0,1)$ those with $\gamma < 1$ change their ideal point to $(x^n, y^n) = (0,0)$ and support party L while those with $\gamma > 1$ change their ideal point to $(x^n, y^n) = (1,1)$ and support party R. Those with $\gamma = 1$ are a measure zero set so to compute the parties' vote shares from this group of voters we do not need to specify what they do.

Analogously we can show that among voters with initial ideal point $(x^o, y^o) = (1,0)$ those with $\gamma < 1$ change their ideal point to $(x^n, y^n) = (1,1)$ and support party R while those with $\gamma > 1$ change their ideal point to $(x^n, y^n) = (0,0)$ and support party L. Again, voters with $\gamma = 1$ are a measure zero set so we ignore them.

Party L's vote share given state θ is then $\frac{1}{4}\left(1 + \phi^-_{(0,1)}[\theta] + \phi^+_{(1,0)}[\theta]\right)$ while party R's vote share is one minus this quantity. By assumptions (i) and (ii) party L's vote share is larger than $1/2$ at state $\theta = \theta_\ell$ and smaller than $1/2$ at state $\theta = \theta_r$. Because the two states are equally likely, this means that parties L and R both have equal probability of winning the election.

We now use these facts to show that it is an equilibrium for the two parties to run on their ideal policies. If party R announces $(x^R, y^R) = (1,1)$ and party L stays at $(x^L, y^L) = (0,0)$ then L's expected payoff is $-\frac{1}{2}(0) - \frac{1}{2}(1 + \lambda^L)$. If party L deviates to $(1,1)$ then its expected payoff is $-(1+\lambda^L)$ so the deviation would not be profitable. If L deviates to $(0,1)$ then its vote share would be $\frac{1}{4}\left(1 + \phi^-_{(0,0)}[\theta] + \phi^+_{(1,1)}[\theta]\right)$, which follows from an argument analogous to the one above in which we derived the (purported) equilibrium vote share. This vote share is larger than $1/2$ when $\theta = \theta_r$ and smaller than $1/2$ when $\theta = \theta_\ell$, by assumption (ii). Therefore, party L's probability of winning is $1/2$. This means that L's expected payoff from the deviation is $-\frac{1}{2}(\lambda^L) - \frac{1}{2}(1 + \lambda^L)$, so again the deviation would not be profitable. Finally, suppose that party L were to deviate to $(1,0)$. In this case, its vote share would be $\frac{1}{4}\left(1 + \phi^+_{(0,0)}[\theta] + \phi^-_{(1,1)}[\theta]\right)$. This quantity is larger than $1/2$ when $\theta = \theta_\ell$ and smaller than $1/2$ when $\theta = \theta_r$, by assumption (i). Therefore, the probability that L wins is again $1/2$. Consequently, the expected payoff from the deviation is $-\frac{1}{2}(1) - \frac{1}{2}(1 + \lambda^L)$. Therefore this deviation is also not profitable.

We have left to show that the equilibrium is unique. We show this by ruling out the remaining cases in Appendix A. □

3.3 Discussion

The main conclusion of Proposition 1 is that while the two parties adopt their own preferred positions, voters who are initially evenly distributed across four ideal points in a two-dimensional policy space change their ideal policies to match the positions taken by the two parties. Partisanship in a two-party system emerges naturally in our model from voters wanting to minimize the psychological cost associated with supporting a party that takes a position that is different from their own ideal positions on a particular issue.

This analysis makes several contributions to our theoretical understanding of partisan politics. First, the results speak to ideological scaling efforts documenting that voters' policy preferences can be scaled onto no more than two dimensions, and usually just one dimension (Poole and Rosenthal, 2000, 1991).[6] Although scaling has become more sophisticated in recent years and papers have extended scaling to members of the general public (e.g., Bonica, 2014; Tausanovitch and Warshaw, 2013), there are few theoretical explanations for the relatively low dimensionality of American politics despite the fact that there are so many issue areas. Implicitly, the explanation for this result is usually grounded in the existence of two major political parties. For example, Poole and Rosenthal (1991, p. 230) note that "[e]xcept for very brief periods, the United States has always had a two-party political system. It is not surprising, therefore, that one dimension strings from strong loyalty to one party (Democrat-

[6] In times of "stability," for example, Poole and Rosenthal (1991, p. 230), note that "a one dimensional-model accounts for most voting in Congress."

Republican or Democrat) to weak loyalty to either party to strong loyalty to a second, competing, party (Federalist, Whig, or Republican)." Recent findings by Tomz and Sniderman (2005) also support this view. They show that attaching a party label to a policy position enhances the correlation of voters' positions on different issues. Our model supports both of these findings.

Second, the results provide a broader theoretical framework for a number of papers documenting that earlier political actions have downstream effects on the attitudes of voters. For example, Beasley and Joslyn (2001) find that voters and non-voters differed in their eventual assessments of potential candidates, thus lending credence to the idea that the act of casting a vote is an action that then changes preferences downstream. Likewise, Mullainathan and Washington (2009) use the sharp discontinuity in American voting age restrictions (e.g., the fact that those 18 and over can vote but those under 18 cannot) to examine voting in future elections. They find that American voters narrowly eligible to vote in a certain election show greater levels of polarization later on than those who were narrowly ineligible to vote. The authors conclude that voting early on for a candidate, but then seeing that candidate struggle triggers a kind of cognitive dissonance, which in turn makes those who cast a vote increase their support for the candidate once elected—in large part to justify their earlier decision. This is consistent with the subsequent analyses of Bølstad, Dinas and Riera (2013), who find that the act of voting in favor of a party in the U.K. leads those voters to hold more favorable opinions of that party. Other papers have examined the issue in the context of voter turnout. For example, Meredith (2009) also uses the sharp discontinuity in American voting age requirements to examine voting in future elections; he finds that voting in one election increases the probability of voting in the next election by about 5 percentage points and also increases the rates of partisan affiliations.

Finally, we note that our model perhaps has its greatest relevance in situations where voters have relatively high levels of political information, implying that variation in political information could impact the extent to which cognitive dissonance shapes partisanship. In particular, voters must know the political positions of the parties in order to incur the psychological cost of being "out of step" with their supported party. Low-information voters, then, will have lower cognitive dissonance costs simply because they are less likely to have knowledge on the policy positions of the parties. In fact, a straightforward extension of the model could explain why levels of political knowledge predict the consistency of mass political attitudes with party elites (Zaller, 1992).

4 Attitudes Shaped by Violence

It is often argued that violence is the outcome of prejudice: individuals engage in violence against those whom they hate. However, an alternative view was, according to Stephen Holmes, articulated by Thomas Hobbes. In his introduction to Hobbes' *Behemoth*, Holmes wrote that

> [i]n his abridged "translation" of [Aristotle's] Rhetoric, Hobbes departed from Aristotle's original by adding intriguingly that individuals have a tendency 'to hate' anyone "whom they have hurt," simply because they have hurt him (Holmes, 1990).

Holmes speculates in a footnote also on the "contrary proclivity," which is "the irrational tendency to love those whom we have helped, simply because we have helped them" (Holmes, 1990).

In this section, we develop an application of cognitive dissonance theory in which ethnic or racial hatred increases when an individual commits an act of violence toward someone from a different ethnic or racial group and decreases when the individual does not commit any such act of violence. The application provides a social psychological basis for the constructivist viewpoint that ethnic and racial divisions can be socially or individually constructed, possibly from acts of violence (Fearon and Laitin, 2000). The model also demonstrates how ethnic animosities can be passed down across generations and how they may co-evolve with violence, tracking the amount of violence over time. We show that ethnic hostility may in fact persist even after violence disappears, a result that has applications both in comparative and American politics.

4.1 Model Setup

Consider a society consisting of two groups, A and B, that are geographically structured on the real line $\mathbb{R}$. Located at each point $r \in \mathbb{R}$ at any given period of time, $t = 0, 1, 2, ..., \infty$, is exactly one member of group A and one member of group B. We assume that each individual from each group lives for exactly one period, after which he is replaced by exactly one offspring who inherits both his location on the interval and his group identity. Since we take group A to be the dominant group, while group B is a passive group, we will identify members of group A with their location on the interval. The assumption that society is geographically structured means that we can define the notion of a community: we will refer to the interval $\mathcal{B}(r) = \left[r - \frac{\mu}{2}, r + \frac{\mu}{2}\right]$ as the "local community" of r.

In each period, members of group A must decide whether or not to engage in violence against group B. Denote the choice by $a_t(r) \in \{0, 1\}$ ($a_t(r) = 1$ means that the member of group A located at $r \in \mathbb{R}$ chooses violence in period t; $a_t(r) = 0$ means that he does not). Members of group A must also choose the kind of attitude $x_t^n(r) \in [0, 1]$ to have towards group B. We interpret higher values of $x_t^n(r)$ as reflecting more hostile attitudes. If $\rho_t(r)$ is the fraction of individuals in r's local community that engage in violence against group B, then the "material payoff" received by the group A individual who lives at r is

$$u_t(r) = \pi_t(\rho_t(r)) - v \cdot a_t(r) \tag{2}$$

where $\pi_t(\rho_t(r))$ is a part of the material payoff that depends only on the aggregate violence against members of group B in r's local community and $v > 0$ is the material cost of violence. By assuming that $\pi_t(\cdot)$ depends only on the total amount of violence produced in a local community, we are implicitly assuming that violence can produce benefits to group A only socially.

In addition to the material payoffs described above, we assume that the payoff to each member of group A also contains a psychological part; in particular, the psychological payoff to the individual located at r is

$$\psi(r) = -\gamma \left| x_t^n(r) - a_t(r) \right| - \frac{1}{2\kappa} \left[x_t^n(r) - x_t^o(r) \right]^2 \tag{3}$$

where $\gamma > 0$ and $\kappa > 0$ are parameters, and the quantity $x_t^o(r) \in [0, 1]$ is the attitude of r's parent; thus, $x_t^o(r) = x_{t-1}^n(r)$ if $t > 0$ and we take some fixed $x_0^o(r)$ given by a function $x_0^o : \mathbb{R} \to [0, 1]$ as the distribution of attitudes in the initial period. The interpretation of this is that, in early life, children socialize with their parents, acquiring information and perspectives from this interaction before developing their own social beliefs. They form their own opinions in later life, taking $x_t^o(r)$ as a benchmark. One can therefore think of the term $-\frac{1}{2\kappa}\left[x_t^n(r) - x_t^o(r)\right]^2$ as being the psychological cost of changing one's inherited attitude, which reflects the idea that larger changes incur greater costs. Analogously, the term $-\gamma \left| x_t^n(r) - a_t(r) \right|$ can be interpreted as a cognitive dissonance cost, which may arise from engaging in violence toward group B but not holding sufficiently hostile views towards the group, or holding hostile views but not engaging in violence. This implies that the total payoff for a member of group A located at r is the sum of material and psychological payoffs, $u_t(r) + \psi(r)$.

Behavioral Assumptions. We assume that the dynamic linkage across periods arises from intergenerational socialization: each group A member r observes the material payoffs of group A members from his parents' generation that lived in his local community, and then decides whether or not to engage in violence by "imitating" the individual from the previous generation that received the highest material payoff. More formally, define the sets of group A members in r's local community that respectively do not engage, and engage, in violence in period t to be

$$\begin{aligned} \mathcal{A}_t^0(r) &= \{\tilde{r} \in \mathcal{B}(r) : a_t(\tilde{r}) = 0\}, \\ \mathcal{A}_t^1(r) &= \{\tilde{r} \in \mathcal{B}(r) : a_t(\tilde{r}) = 1\}. \end{aligned} \tag{4}$$

We assume that the individual who lives at r in period $t+1$ engages in violence if and only if the highest material payoff among individuals in his local community that commit violence in period t is larger than the highest material payoff among individuals who choose not to commit violence; in other words, if $\mathcal{A}_t^0(r)$ and $\mathcal{A}_t^1(r)$ are both nonempty, then

$$a_{t+1}(r) = \begin{cases} 0 & \text{if sup } u_t(\mathcal{A}_t^1(r)) < \text{sup } u_t(\mathcal{A}_t^0(r)) \\ 1 & \text{if sup } u_t(\mathcal{A}_t^1(r)) \geq \text{sup } u_t(\mathcal{A}_t^0(r)) \end{cases} \tag{5}$$

and if $\mathcal{A}_t^0(r) = \emptyset$, then $a_{t+1}(r) = 1$, while if $\mathcal{A}_t^1(r) = \emptyset$, then $a_{t+1}(r) = 0$. The latter part of this assumption says that if every member of group A in r's local community took the same action in the previous period, then r takes that action in the current period. This imitation rule is an "optimistic" imitation rule in the sense that r aspires to the highest material payoff received by his parents' neighbors and then imitates the individual who received the highest material payoff.

Finally, note that given the choice of $a_t(r)$, the optimal attitude is simply a function of $x_t^o(r)$, and is

$$x_t^n(r) = \begin{cases} \min\{x_t^o(r) + \kappa, 1\} & \text{if } a_t(r) = 1 \\ \max\{0, x_t^o(r) - \kappa\} & \text{if } a_t(r) = 0 \end{cases} \tag{6}$$

This implies that an individual always pays a cost of at most $\kappa/2$ for changing his attitude, which he pays when the attitude rises or falls by the maximum optimal change of κ. We will assume that after agents choose $a_t(r)$ by imitation, they choose $x_t^n(r)$ optimally, so (6) characterizes how attitudes change. We also take the perspective that the parameter κ is small so that attitudes move incrementally within the interval $[0, 1]$.

4.2 Analysis

Our main result characterizes the recursive paths of violence and attitudes in the society described in the previous section. To focus on a substantively interesting case where violence is initially increasing and then decreasing (see our discussion in Section 4.3) we add two assumptions to the model as follows.

First, recall that we assumed that violence produces benefits to group A only socially, since $\pi_t(\cdot)$ depends only on the fraction $\rho_t(r)$ of individuals in r's local community that engage in violence. To this, we add the following assumptions.

(i) $\pi_t(\rho)$ is continuous and strictly increasing in ρ for all t

(ii) $\exists t^* > 0$ s.t. $v < \pi_t(1) - \pi_t(\frac{1}{2})\ \forall t < t^*$ and $v > \pi_t(1) - \pi_t(\frac{1}{2})\ \forall t \geq t^*$.

Assumption (i) implies that the return to violence is increasing in the amount of violence produced in the local community. Assumption (ii) states that in early periods the cost of violence is relatively low, but in later periods it is high.

Second, note that if no individual engages in violence in the first period, then by our imitation rule no individual will ever engage in violence. So, we will assume that a concentrated mass, λ_0, of individuals adopt violence in the first period, and focus on how violence may spread or decline after this point. Formally, our assumptions about the initial conditions of the model are as follows:

(iii) $\lambda_0 \geq \mu$

(iv) $(\alpha_0(r), a_0(r)) = \begin{cases} (1, \kappa) & \text{if } r \in \left[-\frac{\lambda_0}{2}, \frac{\lambda_0}{2}\right] \\ (0, 0) & \text{otherwise} \end{cases}$

Given the assumption that a concentrated mass λ_0 of group A individuals adopt violence in the first period, assumption (iii) guarantees that there is at least one individual whose entire local community engages in violence in the first period. (This assumption is stronger than necessary for our purposes, as we explain following the statement of

Proposition 2 below.) Assumption (iv) states that the small community of individuals that adopt violence in the first period is centered at 0, and that these individuals have the same attitudes that they would have chosen if their parents' attitudes were 0 (but, in fact, they are the first generation of group A individuals in the model).

Before stating our main result, note that for all periods $t < t^*$ there is a unique number $\rho_t^* \in (\frac{1}{2}, 1)$ satisfying $\pi_t(\rho_t^*) - v = \pi_t(\frac{1}{2})$, which follows from assumptions (i) and (ii). In addition, in all that follows we will identify the "degenerate interval" $[0, 0]$ with the empty set $\emptyset$. Our main result below recursively characterizes the spread and decline of violence and attitudes in the population over time.

Proposition 2 *The paths of violence and attitudes are recursively given by*

$$\big(a_{t+1}(r), x_{t+1}^n(r)\big) = \begin{cases} (1, \min\{x_t^n(r) + \kappa, 1\}) & \textit{for all } r \in [-\frac{\lambda_{t+1}}{2}, \frac{\lambda_{t+1}}{2}] \\ (0, \max\{x_t^n(r) - \kappa, 0\}) & \textit{for all } r \notin [-\frac{\lambda_{t+1}}{2}, \frac{\lambda_{t+1}}{2}] \end{cases} \tag{$*$}$$

$$\textit{where } \lambda_{t+1} = \begin{cases} \lambda_t + 2\mu(1 - \rho_t^*) & \textit{if } t < t^* \\ \max\{0, \lambda_t - \mu\} & \textit{if } t \geq t^* \end{cases} \tag{$\dagger$}$$

Proof. See Appendix B. □

Proposition 2 implies that the mass of individuals that adopt violence grows up to period t^* after which it declines. The proposition also implies that as group A individuals adopt violence toward group B, they also develop increasingly hostile attitudes towards that group. If the critical period t^* is sufficiently large (and ρ_t^* is sufficiently lower than 1 in all of these periods), then a large mass of group A individuals continue to develop increasingly hostile attitudes toward group B, even after violence begins to decline. Consequently, average attitudes may peak in a period $t^{**} > t^*$ after which they begin to decline. In particular, it will take longer for average attitudes to decline all the way to 0 than it will for the mass of individuals adopting violence to go to 0.

Finally, note that our assumption that agents imitate members of their local community (of the previous generation), rather than optimally decide whether or not to engage in violence, is important for the result that violence can spread in the population. Because violence produces benefits only socially, whereas its costs are private, optimizing agents would succumb to the free-rider problem and choose not to contribute to violence. That said, our assumption that a small concentrated mass λ_0 of individuals choose violence in the first period is also important for this result. Nevertheless, as we mentioned before, the assumption that $\lambda_0 \geq \mu$ is stronger than necessary for Proposition 2 to hold. If $v < \pi_0(1) - \pi_0(\frac{1}{2})$, then there exists $\lambda^* \in \left(\frac{\mu}{2}, \mu\right)$ such that $v = \pi\left(\frac{\lambda^*}{\mu}\right) - \pi\left(\frac{1}{2}\right)$. Proposition 2 holds when we replace the assumption that $\lambda_0 \geq \mu$ with the weaker assumption that $\lambda_0 \geq \lambda^*$. On the other hand, if $\lambda_0 < \lambda^*$ then the mass of group A individuals who adopt violence in the first period is too small for violence to survive, let alone spread, and it disappears completely from society.

4.3 Discussion

Proposition 2 shows that individuals committing violence against members of another group will develop hostile attitudes towards their victims as a way of minimizing cognitive dissonance. Importantly, the hostile attitudes may persist even after the violence itself declines.

The model contributes to our understanding of how group-based prejudices originate and develop. First, it provides a theoretical framework with which to integrate instrumentalist (or strategic) and constructivist approaches in the study of ethnicity.[7] While the traditional constructivist approaches focus on the social construction of ethnic identity, the strategic and constructivist approaches move toward each other when individual agency is considered—that is, individuals can become agents of constructing ethnic identities and furthering hostility against other groups (Fearon and Laitin, 2000). Our framework speaks to this possibility by showing how actions by

[7] Indeed, our analysis rejects primordialist arguments, which focus on conflict as being the result of deep-rooted or immutable antipathies that have biological or innate origins (Fearon and Laitin, 2000). This is in contrast to the traditional rational choice approach, which seems to be more in line with the primordialist view of the world in that it takes attitudes as given and immutable.

individuals (violence) can affect individual attitudes (ethnic or racial animosity). Although this is somewhat in contrast with the standard ethnic violence literature, which starts from the proposition that violence is usually the result of ethnic cleavages (rather than vice versa), our argument supports many empirical studies that have found that violence can be—and has historically been—used by elites as a way of fostering in-group solidarity and furthering anti-outgroup group attitudes (e.g., studies like Brass, 1997).[8] In addition, our findings engage the broader possibility that individuals have a significant role to play in the development or propagation of ethnic or racial prejudice. As Fearon and Laitin (2000, p. 856) write, individual "actions may ... result in the construction of new or altered identities, which themselves change cultural boundaries."

Second, the results provide theoretical justification for an increasing number of empirical studies documenting the historical persistence of ethnic or racial prejudices that originate in violence. For example, Voigtländer and Voth (2012) document a remarkable persistence in anti-Semitic attitudes in Germany. They show that regions that had medieval anti-Jewish pogroms during periods of the Black Death are also those places that had the most intense anti-Semitism in the 1920s and greater support for the Nazi Party. The link in their work is violence: violence against the Jews over 500 years ago led to a persistently anti-Semitic climate well into the 20th century. Similarly, Acharya, Blackwell and Sen (2015) explore the legacy of U.S. Southern slavery, finding that those parts of the U.S. South where slavery was highly prevalent are also those areas where whites today are the most conservative and racially hostile. The reason, they posit, lies in the postbellum economic and political incentives for

[8] Our theory can be expanded to explain group or racial solidarity among the *oppressing* group. For example, in a famous study, Aronson and Mills (1959) assigned individuals undergoing an initiation ritual to mild or embarrassing rituals. Those who had the more embarrassing initiation were the most enthusiastic about membership later on. That is, the more costly the initiation, the stronger the group attachment. Such a component of cognitive dissonance could help explain the use of shocking forms of ethnic or racial violence to cultivate membership in ethnic or religious nationalist organizations, for example white supremacist groups in the United States (particularly in the early 20th century) or even ISIS in the Middle East.

whites to expand racially targeted violence (e.g. rapes and lynchings) which was used to terrorize newly freed slaves against exercising their basic economic and political rights. They document the persistence of these attitudes over time and argue that intergenerational socialization plays an important role in explaining the persistence.

5 Socialization and Empathy

It is often argued that when two individuals socialize, their attitudes converge to each other's even when they do not exchange information or evidence, and even on issues on which there may be no evidence to exchange (such as religion). By *empathizing* with another individual—that is, by internalizing the other person's preferences and action choices—an individual may experience some level of cognitive dissonance arising from the fact that her initial preferences are in conflict with the preferences or actions of the individual with whom she shares this connection.

In this section, we develop a model in which individuals seek to minimize such cognitive dissonance by changing their initial attitudes. This example speaks to a growing literature showing empathy's role in shaping individual partisanship and elite decision making (e.g., Glynn and Sen, 2015; Washington, 2008) and in aiding in political persuasion and inter-group relationships (e.g., Enos and Hersh, 2015; Pettigrew and Tropp, 2008). It also speaks to the important idea that partisanship is very stable but might change as a result of close, personal relationships, such as those with husbands, wives, or other family members (Green, Palmquist and Schickler, 2004).

5.1 Model Setup

Consider two individuals, $i = 1, 2$ who have preferences on a one-dimensional issue space represented by the real line $\mathbb{R}$. Each individual i has an initial ideal point x_i^o, both of which are common knowledge. Each individual simultaneously decides what her new ideal point x_i^n will be, and which attitude x_i to express. Both have quadratic loss preferences represented by

$$u_i(x_i^n, x_i \mid x_i^o) = -(x_i^n - x_i)^2 - e_i(x_i^n - x_{-i})^2 - \kappa_i(x_i^n - x_i^o)^2$$

where $-i$ is the usual notation for the other individual and e_i and κ_i are positive parameters. The first term in these preferences reflects the individual's desire to express a position x_i that matches her new ideal position x_i^n. We can interpret this also as cognitive dissonance. But, more interestingly, the second term of this expression reflects the cognitive dissonance that the individual experiences when she has ideal position x_i^n and the *other* individual expresses a position x_{-i} different from that. This dissonance is weighted by $e_i > 0$ which we interpret as the level of *empathy* that individual i has towards $-i$. Finally, there is cost to changing one's ideal position from x_i^o to x_i^n that is reflected in the third term.

Proposition 3 *In the unique Nash equilibrium each individual* $i = 1, 2$ *chooses*

$$x_i = x_i^n = \alpha_i x_i^o + (1 - \alpha_i)x_{-i}^o, \textit{ where } \alpha_i = \frac{e_{-i}\kappa_i + \kappa_{-i}\kappa_i}{e_{-i}\kappa_i + e_i\kappa_{-i} + \kappa_{-i}}.$$

Proof. The first order conditions for the maximization of $u_i(x_i^n, x_i \mid x_i^o)$ with respect to x_i^n and x_i are given by

$$0 = x_i^n - x_i \tag{7}$$

$$0 = (x_i^n - x_i) + e_i(x_i^n - x_{-i}) + \kappa_i(x_i^n - x_i^o), \qquad i = 1, 2 \tag{8}$$

The equilibrium choices of (x_1^n, x_1) and (x_2^n, x_2) solve this system of four equations in four unknowns. In fact, because $x_i^n = x_i$ the system reduces to two equations in two unknowns. Solving these best response equations yields the result. □

In equilibrium, an individual expresses a position x_i equal to her new ideal position x_i^n and her new ideal position is a convex combination of her starting position x_i^o and the starting position of the other individual x_{-i}^o. The weight α_i that individual i puts

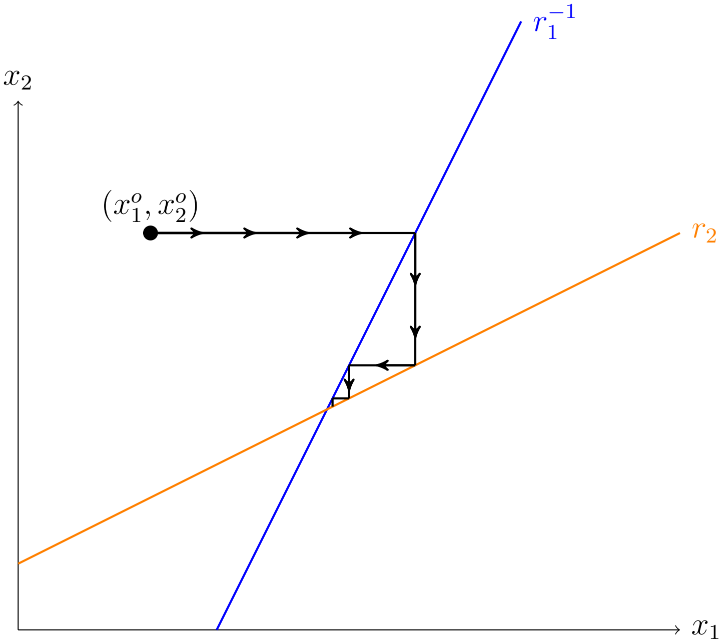

Figure 1: Socialization as a Dynamic Adjustment Process

on her own starting position x_i^o is decreasing in the degree of empathy e_i that she feels toward the other individual, and increasing in the difficulty κ_i in changing her own position. Interestingly, the weight α_i is increasing in the degree of empathy e_{-i} that the other individual $-i$ feels towards i and decreasing in the difficulty κ_{-i} that the other individual experiences in changing his position. ("If you don't feel very much empathy toward me, or if you find it hard to change your views, then I end up compromising my position more.")

5.2 Socialization as a Dynamic Adjustment Process

If equilibrium is instantly achieved, then the model above does not fully capture the process of socialization, which takes time. In this section, we provide a dynamic adjustment (or *tâtonnement*) argument for how the players might arrive at equilibrium through socialization.

In our set-up, players take turns reacting to changes in each other's positions by iteratively choosing best responses before they settle on their final position. Player 1 first reacts to player 2's initial position; player 2 then reacts to player 1's new position; player 1 then reacts to player 2's new position, and so on. To write the "reaction function" for each player, we substitute (5) into (6) and solve for x_i^n. The reaction function for individual i is then:

$$r_i(x_{-i}^n) = \frac{e_i}{e_i + \kappa_i} x_{-i}^n + \frac{\kappa_i}{e_i + \kappa_i} x_i^o, \qquad i = 1, 2 \tag{9}$$

These are simply best response functions. The sequence of positions that the players take when they take turns reacting to each other is then given by the following initial conditions and recursive relationships:

$$\begin{aligned}
x_1[0] &= x_1^o \\
x_2[0] &= x_2^o \\
x_1[t] &= \frac{e_1}{e_1 + \kappa_1} x_2[t-1] + \frac{\kappa_1}{e_1 + \kappa_1} x_1^o, \qquad t > 0 \\
x_2[t] &= \frac{e_2}{e_2 + \kappa_2} x_1[t] + \frac{\kappa_i}{e_2 + \kappa_2} x_2^o, \qquad t > 0
\end{aligned} \tag{10}$$

where $x_i[t]$ denotes player i's position after he has reacted t times. The following result states that for all starting values of x_1^o and x_2^o the sequence of positions for each player converges to the equilibrium positions given in Proposition 3 above.

Proposition 4 *For all x_1^o and x_2^o the sequences of $\{x_1[t]\}_t$ and $\{x_2[t]\}_t$ converge to the equilibrium values of x_1^n and x_2^n given in Proposition 3 above.*

Proof. Solving the system of recursive equations in (10) yields the following

$$x_i[t] = (\tau_1 \tau_2)^t x_2^o + \left[\tau_i(1 - \tau_{-i}) x_{-i}^o + (1 - \tau_i) x_i^o\right] \left[\frac{1 - (\tau_1 \tau_2)^t}{1 - \tau_1 \tau_2}\right] \qquad i = 1, 2$$

where $\tau_i = e_i/(e_i + \kappa_i) \in (0, 1)$, $i = 1, 2$. This implies that

$$\lim_{t \to \infty} x_i[t] = \frac{\tau_i(1 - \tau_{-i}) x_{-i}^o + (1 - \tau_i) x_i^o}{1 - \tau_1 \tau_2}, \qquad i = 1, 2$$

Substituting in τ_i, $i = 1, 2$, and simplifying, we find that this limit equals the equilibrium value of x_i^n given in Proposition 3, for each $i = 1, 2$. $\square$

Figure 1 illustrates the socialization process described above. The figure shows how the dynamic adjustment process leads to the players eventually reaching the equilibrium values (x_1^n, x_2^n) starting from value of (x_1^o, x_2^o). The two oblique lines depicted are the reaction functions, or best responses. The vertical and horizontal lines with arrows depict the socialization path, which starts from the original positions (x_1^o, x_2^o). What the figure does not reveal, but is true (from the fact that α_i in Proposition 3 lies between 0 and 1), is that each individual's final position lies between his original position and the original position of the other individual.

In addition, Figure 1 reveals the nuanced prediction that the convergence of $x_i[t]$ to the equilibrium position need not be monotonic in the beginning. Early in the socialization process, player 1 may entertain a very different perspective than his own as he makes an effort to put himself in player 2's shoes. As player 2 reveals that she is doing the same, player 1 may decide to take a step back. It is then player 2 who takes successive steps closer to player 1's position, and player 1 who takes small steps back, as the players figure out where they each will stand. In this process, player 1 makes too large a compromise in the beginning and spends the rest of the socialization process taking small steps back. Player 2, however, always moves in the direction of her final position.[9] Such a process may be quite natural for two empathetic individuals working together to understand each other's perspectives and develop their own new positions.

5.3 Discussion

This example provides theoretical support to two related literatures. The first is a literature documenting the stability of partisanship over time in tandem with its ability

[9] If we had reversed the order of moves—assuming that player 2 reacts first—then, the reverse would be true. Player 2 would initially take too large a step, and then spend the rest of the interaction taking small steps back. Player 1 would consistently move towards his final position.

to change as a result of major life events, including major interpersonal events, such as marriage, re-marriage or divorce (Green, Palmquist and Schickler, 2004), or the act of moving or emigrating to a new place (Brown, 1981). For example, Green, Palmquist and Schickler (2004) note that partisanship operates similarly to religious affiliation in the sense that close, empathetic relationships have the potential to change it over time. They note that a

> common avenue for shifting religious affiliation is a changing small-group environment, in particular, marriage to a person of another faith. In such instances, people...may alter their perception of the new religion as they come to see it through their spouse's eyes. Parallel observations may be made about partisan identities, which also change as regional and occupational mobility put adults into contact with new friends and social groups (Green, Palmquist and Schickler, 2004, p. 6).

Our analysis provides a theoretical foundation for how exactly these sorts of major life events could lead to the transformation of political preferences over time.

Second, and relatedly, our analysis sheds light not only on how close relationships can lead to changes in *partisanship* but also how empathy can lead to changes in specific *policy positions*. For example, several studies have documented that close personal relationships have the capacity to affect decision making on issues having a substantive connection. For example, leveraging a natural experiment, Washington (2008) finds that male Congressional representatives who have daughters tend to vote in more liberal directions on issues having a gender component; this finding is replicated in the judicial context using a similar methodology by Glynn and Sen (2015), who show that federal judges with daughters are more likely to vote in a progressive direction on cases involving reproductive rights and gender-based employment discrimination. Our findings also speak to a broader literature on political persuasion. Here, a growing literature on campaign tactics in the American context has documented sending demographically similar campaign workers (e.g., empathy) is a more winning strategy than sending campaign workers who look very different from voters (e.g., Enos and Hersh, 2015; Leighley, 2001; Shaw, de la Garza and Lee, 2000).

6 Concluding Remarks

Our contribution in this paper is to develop a framework for understanding the origins of political preferences and attitudes and how they evolve. We developed a formal theory of how individuals adjust their political and social preferences to minimize cognitive dissonance—the discomfort that arises when choices or actions come into conflict with pre-existing preferences. With strong roots in social psychology, this simple intuition shows how people often change their preferences to bring them into closer alignment with their actions. The concept is a powerful predictor of preference formation and change, and, as we argued, it provides a theoretical foundation for a variety of political and social phenomena.

We applied our modeling approach to three examples: partisanship, ethnic hatred, and empathy or socialization. But there are other applications. For example, our approach may fruitfully be applied to social-psychology concepts that are closely related to cognitive dissonance, such as confirmation bias and motivated reasoning (Lodge and Taber, 2013). Confirmation biases are instances where individuals refuse to absorb or engage with potentially conflicting information, choosing instead to update on the basis of information that conforms with pre-existing attitudes, while motivated reasoning is the tendency for people to explicitly view new evidence as entirely consistent with their pre-existing views.[10] Both confirmation bias and motivated reasoning are,

[10] Thus, as explained by Taber and Lodge (2006), confirmation bias is "when given a chance to pick and choose what information to look at—rather than when presented with pro and con arguments—people will actively seek out sympathetic, nonthreatening sources." As for motivated reasoning (sometimes called "disconfirmation bias"), we look to the explanation of Druckman and Bolsen (2011), who define it as "the tendency to seek out and/or view new evidence as consistent with one's prior views, even if this is not objectively accurate."

at their core, instances where individuals seek to avoid cognitive dissonance. With confirmation bias, cognitive dissonance is minimized by avoiding potentially challenging information; with motivated reasoning, objective information is actively ignored. Both are examples of the broader framework that we suggest here, meaning that our approach can be used to explore and define these increasingly important concepts.

Our approach can also be used to understand broader findings in political science, including the important role of social networks in changing policy preferences, the impact of close contact between people of different ethnic or racial groups (both in terms of "contact theory" or "racial threat"), and the way that polarization can be exacerbated by media and social environments. We leave the detailed formal analysis of these concepts to future research.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2015. "The Political Legacy of American Slavery." Working Paper.

Akerlof, George A. and William T. Dickens. 1982. "The Economic Consequences of Cognitive Dissonance." *American Economic Review* 72(3):307–319.

Aronson, Elliot and Judson Mills. 1959. "The Effect of Severity of Initiation on Liking for a Group." *The Journal of Abnormal and Social Psychology* 59(2):177.

Beasley, Ryan K. and Mark R. Joslyn. 2001. "Cognitive Dissonance and Post-Decision Attitude Change in Six Presidential Elections." *Political Psychology* 22(3):521–540.

Bem, Daryl J. 1967. "Self-Perception: An Alternative Interpretation of Cognitive Dissonance Phenomena." *Psychological Review* 74(3):183–200.

Bølstad, Jørgen, Elias Dinas and Pedro Riera. 2013. "Tactical Voting and Party Preferences: A Test of Cognitive Dissonance Theory." *Political Behavior* 35(3):429–452.

Bonica, Adam. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science* 58(2):367–386.

Brass, Paul R. 1997. *Theft of an Idol: Text and Context in the Representation of Collective Violence*. Princeton University Press.

Brehm, Jack W. 1956. "Postdecision Changes in the Desirability of Alternatives." *Journal of Abnormal and Social Psychology* 52(3):384.

Brown, Thad A. 1981. "On Contextual Change and Partisan Attributes." *British Journal of Political Science* 11(04):427–447.

Chen, M. Keith and Jane L. Risen. 2010. "How Choice Affects and Reflects Preferences: Revisiting the Free-Choice Paradigm." *Journal of Personality and Social Psychology* 99(4):573–594.

Cooper, Joel and Russell H. Fazio. 1984. "A New Look at Dissonance Theory." *Advances in Experimental Social Psychology* 17:229–266.

Davis, Keith E. and Edward E. Jones. 1960. "Changes in Interpersonal Perception as a Means of Reducing Cognitive Dissonance." *Journal of Abnormal and Social Psychology* 61(3):402–410.

Dietrich, Franz and Christian List. 2011. "A Model of Non-Informational Preference Change." *Journal of Theoretical Politics* 23(2):145–164.

Dietrich, Franz and Christian List. 2013. "Where Do Preferences Come From?" *International Journal of Game Theory* 42(3):613–637.

Druckman, James N. and Toby Bolsen. 2011. "Framing, Motivated Reasoning, and Opinions About Emergent Technologies." *Journal of Communication* 61(4):659–688.

Egan, Louisa C., Laurie R. Santos and Paul Bloom. 2007. "The Origins of Cognitive Dissonance: Evidence from Children and Monkeys." *Psychological Science* 18(11):978–983.

Egan, Louisa C., Paul Bloom and Laurie R. Santos. 2010. "Choice-induced Preferences in the Absence of Choice: Evidence from a Blind Two Choice Paradigm with Young Children and Capuchin Monkeys." *Journal of Experimental Social Psychology* 46(1):204–207.

Enos, Ryan D. and Eitan D. Hersh. 2015. "Party Activists as Campaign Advertisers: The Ground Campaign as a Principal-Agent Problem." *American Political Science Review* 109(2):252–278.

Fearon, James D. and David D. Laitin. 2000. "Violence and the Social Construction of Ethnic Identity." *International Organization* 54(4):845–877.

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance*. Stanford University Press.

Festinger, Leon, Henry W. Riecken and Stanley Schachter. 1956. *When Prophecy Fails*. Pinter & Martin Publishers.

Festinger, Leon and James M. Carlsmith. 1959. "Cognitive Consequences of Forced Compliance." *Journal of Abnormal and Social Psychology* 58(2):203.

Glass, David C. 1964. "Changes in Liking as a Means of Reducing Cognitive Discrepancies Between Self-Esteem and Aggression." *Journal of Personality* 32(4):531–549.

Glynn, Adam N. and Maya Sen. 2015. "Identifying Judicial Empathy: Does Having Daughters Cause Judges to Rule for Women's Issues?" *American Journal of Political Science* 59(1):37–54.

Green, Donald P., Bradley Palmquist and Eric Schickler. 2004. *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters*. Yale University Press.

Harsanyi, John C. 1967. "Games with Incomplete Information played by 'Bayesian' Players, Part I: The Basic Model." *Management Science* 14(3):159–182.

Harsanyi, John C. 1968*a*. "Games with Incomplete Information played by 'Bayesian' Players, Part II: Bayesian Equilibrium Points." *Management Science* 14(5):320–334.

Harsanyi, John C. 1968*b*. "Games with Incomplete Information played by 'Bayesian' Players, Part III: Basic Probability Distribution of the Game." *Management Science* 14(7):486–502.

Holmes, Stephen. 1990. Introduction to *Behemoth; Or, The Long Parliament*. University of Chicago Press.

Konow, James. 2000. "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions." *American Economic Review* 90(4):1072–1091.

Leighley, Jan E. 2001. *Strength in Numbers? The Political Mobilization of Racial and Ethnic Minorities*. Princeton University Press.

Lieberman, Matthew D., Kevin N. Ochsner, Daniel T. Gilbert and Daniel L. Schacter. 2001. "Do Amnesiacs Exhibit Cognitive Dissonance Reduction? The Role of Explicit Memory and Attention in Attitude Change." *Psychological Science* 12(2):135–140.

Lodge, Milton and Charles S. Taber. 2013. *The Rationalizing Voter*. Cambridge University Press.

McKelvey, Richard D. 1976. "Intransitivities in Multidimensional Voting Models and Some Implications for Agenda Control." *Journal of Economic theory* 12(3):472–482.

Meltzer, Allan H. and Scott F. Richard. 1981. "A Rational Theory of the Size of Government." *Journal of Political Economy* 89(5):914–927.

Meredith, Marc. 2009. "Persistence in Political Participation." *Quarterly Journal of Political Science* 4(3):187–209.

Mullainathan, Sendhil and Ebonya L. Washington. 2009. "Sticking with Your Vote: Cognitive Dissonance and Voting." *American Economic Journal: Applied Economics* 1(1):86–111.

Pettigrew, Thomas F. and Linda R. Tropp. 2008. "How Does Intergroup Contact Reduce Prejudice? Meta-analytic Tests of Three Mediators." *European Journal of Social Psychology* 38(6):922–934.

Plott, Charles R. 1967. "A Notion of Equilibrium and its Possibility under Majority Rule." *American Economic Review* 57(4):787–806.

Poole, Keith T. and Howard Rosenthal. 1991. "Patterns of Congressional Voting." *American Journal of Political Science* 35(1):228–278.

Poole, Keith T. and Howard Rosenthal. 2000. *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press.

Roemer, John E. 2000. *Political competition: Theory and Applications*. Harvard University Press.

Sharot, Tali, Benedetto De Martino and Raymond J. Dolan. 2009. "How Choice Reveals and Shapes Expected Hedonic Outcome." *Journal of Neuroscience* 29(12):3760–3765.

Shaw, Daron, Rodolfo O. de la Garza and Jongho Lee. 2000. "Examining Latino Turnout in 1996: A Three-state, Validated Survey Approach." *American Journal of Political Science* 44(2):338–346.

Staw, Barry M. 1976. "Knee-Deep in the Big Muddy: A Study of Escalating Commitment to a Chosen Course of Action." *Organizational Behavior and Human Performance* 16(1):27–44.

Taber, Charles S. and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50(3):755–769.

Tausanovitch, Chris and Christopher Warshaw. 2013. "Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities." *Journal of Politics* 75(02):330–342.

Tomz, Michael and Paul M. Sniderman. 2005. "Brand Names and the Organization of Mass Belief Systems." Unpublished Manuscript.

Voigtländer, Nico and Hans-Joachim Voth. 2012. "Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi Germany." *Quarterly Journal of Economics* 127(3):1339–1392.

Washington, Ebonya L. 2008. "Female Socialization: How Daughters Affect Their Legislator Fathers." *American Economic Review* 98(1):311–332.

Zaller, John. 1992. *The Nature and Origins of Mass Opinion*. Cambridge University Press.

Appendix

A. Proof of the Uniqueness Claim in Proposition 1

We prove the uniqueness claim of Proposition 1 by ruling out all other possible equilibria, one by one.

If party R announces $(x^R, y^R) = (1, 1)$ and party L announces $(x^L, y^L) = (0, 1)$, then L's probability of winning is $\frac{1}{2}$, which means that L's expected payoff is $-\frac{1}{2} - \lambda^L$. If L deviates to $(0, 0)$, its expected payoff is $-\frac{1}{2} - \frac{\lambda^L}{2}$. Therefore, the deviation is profitable.

If party R announces $(x^R, y^R) = (1, 1)$ and party L announces $(x^L, y^L) = (1, 0)$, then L's probability of winning is $\frac{1}{2}$, which means that L's expected payoff is $-1 - \frac{\lambda^L}{2}$. If L deviates to $(0, 0)$, its expected payoff is $-\frac{1}{2} - \frac{\lambda^L}{2}$. Therefore, the deviation is profitable.

If party L announces $(x^L, y^L) = (0, 0)$ and party R announces at $(x^R, y^R) = (1, 0)$, then voters with initial ideal point $(0, 0)$ vote for L and those with initial ideal point $(1, 0)$ vote for R. For voters with initial ideal point $(0, 1)$ the payoff from not changing their preference and supporting party L is $-\gamma$, the payoff from not changing their preference and supporting party R is $-1-\gamma$. If they are to support party L and change their ideal point, then the most profitable ideal point is $(x^n, y^n) = (0, 0)$, giving them a payoff of $-\gamma\kappa$. If they are to support party R and change their ideal point, then the most profitable ideal point would be $(x^n, y^n) = (1, 0)$, giving them a payoff of $-\kappa - \gamma\kappa$. Since $-1 - \gamma < -\gamma < -\gamma\kappa$ and $-\kappa - \gamma\kappa < -\gamma\kappa$, voters with initial ideal point $(0, 1)$ will change to $(0, 0)$ and support party L. Following an analogous argument we can show that voters with initial ideal point $(1, 1)$ will change to $(1, 0)$ and support party R. Thus party R's probability of winning is $\frac{1}{2}$, which means that R's expected payoff is $-\frac{1}{2} - \lambda^R$. If R deviates to $(1, 1)$, its probability of winning is $\frac{1}{2}$, which means its expected payoff is $-\frac{1}{2} - \frac{\lambda^R}{2}$. Therefore, the deviation is profitable.

If party L announces $(x^L, y^L) = (0, 0)$ and party R announces $(x^R, y^R) = (0, 1)$, then voters with initial ideal point $(0, 0)$ vote for L and voters with initial ideal point $(0, 1)$ vote for R. For voters with initial ideal point $(1, 0)$, the highest payoff comes from changing to $(0,0)$ and supporting L. For voters with initial ideal point $(1,1)$, the highest payoff comes from changing to $(0,1)$ and supporting R. Thus, R's probability of winning is $\frac{1}{2}$, which means its expected payoff is $-1-\frac{\lambda^R}{2}$. If R deviates to $(1,1)$, its probability of winning is $\frac{1}{2}$, which means its expected payoff is $-\frac{1}{2}-\frac{\lambda^R}{2}$. Therefore, the deviation is profitable.

If party R announces $(x^R, y^R) = (0,1)$ and party L announces $(x^L, y^L) = (1,0)$, then voters with initial ideal point $(1,0)$ will support L and voters with initial ideal point $(0,1)$ will support R. For voters with initial ideal point $(0,0)$, if they change the ideal point and support L, then the most profitable ideal point to choose is $(1,0)$, which gives a payoff of $-\kappa$; if they change the ideal point and support R, then the most profitable ideal point to choose is $(0,1)$, which gives a payoff of $-\gamma\kappa$. When $\gamma > 1$, voters with initial ideal point $(0,0)$ change to $(1,0)$ and support L; when $\gamma > 1$, they change to $(0,1)$ and support R. Using an analogous argument we can show that when $\gamma > 1$, voters with initial ideal point $(1,1)$ change to $(0,1)$ and support R; when $\gamma > 1$, they change to $(1,0)$ and support L. Party L's vote share given state θ is then $\frac{1}{4}\left(1+\phi^{+}_{(0,0)}[\theta]+\phi^{-}_{(1,1)}[\theta]\right)$, which is bigger than $\frac{1}{2}$ when $\theta = \theta_r$ and smaller than $\frac{1}{2}$ when $\theta = \theta_\ell$ by assumption (ii). Therefore, party L's probability of winning is $\frac{1}{2}$, which means that L's expected payoff is $-\frac{1}{2}-\frac{\lambda^L}{2}$. If L deviates to $(0,0)$, then voters with initial ideal point $(0,0)$ support L and voters with initial ideal point $(0,1)$ support R. For voters with initial ideal point $(1,1)$, the highest payoff comes from changing to $(0,1)$ and supporting R. For voters with initial ideal point $(1,0)$, the highest payoff comes from changing to $(0,0)$ and support L. Thus L's probability of winning is $\frac{1}{2}$, which means it expected payoff is $-\frac{\lambda^L}{2}$. Therefore, the deviation is profitable.

If party R announces $(x^R, y^R) = (1,0)$ and party L announces $(x^L, y^L) = (0,1)$, then voters with initial ideal point $(0,1)$ will support L and voters with initial ideal point $(1,0)$ will support R. Voters with initial ideal point $(0,0)$ get a payoff of $-\gamma$ by not changing their ideal point and supporting L, and -1 by not changing their ideal point and supporting R. If they change the ideal point and support L, then the most profitable ideal point is $(0,1)$, which gives a payoff of $-\gamma\kappa$; if they change the ideal point and support R, then the most profitable ideal point is $(1,0)$, which gives a payoff of $-\kappa$. When $\gamma > 1$, $-\gamma < -1 < -\kappa$ and $-\gamma\kappa < -\kappa$, which means voters with ideal point $(0,0)$ support R and when $\gamma < 1$ they support L. Using an analogous argument we can show that when $\gamma > 1$, voters with ideal point $(1,1)$ support L and when $\gamma < 1$ they support R. Party L's vote share given state θ is then $\frac{1}{4}\left(1 + \phi^{-}_{(0,0)}[\theta] + \phi^{+}_{(1,1)}[\theta]\right)$, which is bigger than $\frac{1}{2}$ when $\theta = \theta_\ell$ and smaller than $\frac{1}{2}$ when $\theta = \theta_r$ by assumption (ii). Therefore, party L's probability of winning is $\frac{1}{2}$, which means that L's expected payoff is $-\frac{1}{2} - \frac{\lambda^L}{2}$. If L deviates to $(0,0)$, then its probability of winning is $\frac{1}{2}$, which means its expected payoff is $-\frac{1}{2}$. Therefore, the deviation is profitable.

If party L announces $(x^L, y^L) = (1,0)$ and R stays $(x^R, y^R) = (1,0)$, then voters with any initial ideal point are indifferent between choosing either party. Thus each party's probability of winning is $\frac{1}{2}$, which means party L's expected payoff is $-\lambda^L$. If L deviates to $(0,0)$, voters with initial ideal point $(1,0)$ will support L and voters with initial ideal point $(1,1)$ will support R. Voters with initial ideal point $(0,0)$ get the highest payoff by changing to $(1,0)$ and supporting L. Voters with initial ideal point $(0,1)$ get the highest payoff by changing to $(1,1)$ and supporting R. Thus party R's probability of winning is $\frac{1}{2}$, which means its expected payoff is $-\frac{\lambda^R}{2}$. Therefore, the deviation is profitable.

If party R announces $(x^R, y^R) = (0,1)$ and party L announces $(x^L, y^L) = (0,1)$, then voters with any initial ideal point are indifferent between choosing either party. Thus each party's probability of winning is $\frac{1}{2}$, which means party L's expected payoff is $-\lambda^L$. If L deviates to $(0,0)$, then L's probability of winning is $\frac{1}{2}$, which means its expected payoff is $-\frac{\lambda^L}{2}$. Therefore, the deviation is profitable.

If party R announces $(x^R, y^R) = (1,1)$ and party L announces $(x^L, y^L) = (1,1)$, then voters with any initial ideal point are indifferent between choosing either party. Thus each party's probability of winning is $\frac{1}{2}$, which means party L's expected payoff is $-1 - \lambda^L$. If L deviates to $(0,1)$, then its expected payoff is $-\frac{1}{2} - \lambda^L$. Therefore, the deviation is profitable.

If party L announces $(x^L, y^L) = (0,0)$ and party R announces $(x^R, y^R) = (0,0)$, then voters with any initial ideal point are indifferent between choosing either party. Thus each party's probability of winning is $\frac{1}{2}$, which means party R's expected payoff is

$-1-\lambda^R$. If R deviates to $(0,1)$, then voters with initial ideal point $(0,1)$ will support R and voters with initial ideal point $(0,0)$ will support L. Voters with initial ideal point $(1,1)$ get the highest payoff by changing to $(0,1)$ and supporting R. Voters with initial ideal point $(1,0)$ get the highest payoff by changing to $(0,0)$ and supporting L. Thus party R's probability of winning is $\frac{1}{2}$, which means its expected payoff is $-1-\frac{1}{2}\lambda^R$. Therefore, the deviation is profitable.

If party R announces $(x^R, y^R) = (0,0)$ and party L announces $(x^L, y^L) = (1,1)$, then voters with initial ideal point $(0,0)$ will support R and voters with initial ideal point $(1,1)$ will support L. Voters with initial ideal point $(1,0)$ get $-\gamma$ by not changing their ideal point and supporting L and -1 by not changing their ideal point and supporting R. If they change their ideal point and support L, the most profitable ideal point is $(x^n, y^n) = (1,1)$, which gives them a payoff of $-\gamma\kappa$. If they change their ideal point and support R, the most profitable ideal point is $(0,0)$, which gives them a payoff of $-\kappa$. When $\gamma > 1$, $-\gamma < -1 < -\kappa$ and $-\gamma\kappa < -\kappa$, so voters with initial ideal point $(1,0)$ will change their ideal point to $(0,0)$ and support R. When $\gamma < 1$, $-1 < -\gamma < -\gamma\kappa$ and $-\kappa < -\gamma\kappa$, they will change their ideal point to $(1,1)$ and support L. Using an analogous argument, we can show that when $\gamma > 1$, voters with ideal point $(0,1)$ will change their ideal point to $(1,1)$ and support L; when $\gamma < 1$, they will change their ideal point to $(0,0)$ and support R. Thus party L's vote share is $\frac{1}{4}\left(1+\phi^{+}_{(0,1)}[\theta]+\phi^{-}_{(1,0)}[\theta]\right)$, which is bigger than $\frac{1}{2}$ when $\theta = \theta_r$ and smaller than $\frac{1}{2}$ when $\theta = \theta_\ell$ by assumption (i). It means that party L's expected payoff is $-\frac{1}{2}-\frac{\lambda^L}{2}$. If L deviates to $(0,1)$, voters with initial ideal point $(0,1)$ will support L and voters with initial ideal point $(0,0)$ will support R. Voters with initial ideal point $(1,1)$ get the highest payoff by changing to $(0,1)$ and supporting L; voters with initial ideal point $(1,0)$ get the highest payoff by changing to $(0,0)$ and supporting R. Thus party L's probability of winning is $\frac{1}{2}$, which means it expected payoff is $-\frac{\lambda^L}{2}$. Therefore, the deviation is profitable.

If party R announces $(x^R, y^R) = (0,1)$ and party L announces $(x^L, y^L) = (1,1)$, then voters with initial ideal point $(0,1)$ will support R and voters with initial ideal point $(1,1)$ will support L. Voters with initial ideal point $(1,0)$ get the highest payoff by changing to $(1,1)$ and supporting L; voters with initial ideal point $(0,0)$ get the highest payoff by changing to $(0,1)$ and supporting R. Thus party L's probability of winning is $\frac{1}{2}$, which means its expected payoff is $-\frac{1}{2} - \lambda^L$. If L deviates to $(0,1)$, its probability of winning is $\frac{1}{2}$, which means its expected payoff is $-\lambda^L$. Therefore, the deviation is profitable.

If party R announces $(x^R, y^R) = (1,0)$ and party L announces $(x^L, y^L) = (1,1)$, then voters with initial ideal point $(1,0)$ will support R and voters with initial ideal point $(1,1)$ will support L. Voters with initial ideal point $(0,1)$ get the highest payoff by changing to $(1,1)$ and supporting L; voters with initial ideal point $(0,0)$ get the highest payoff by changing to $(1,0)$ and supporting R. Thus party L's probability of winning is $\frac{1}{2}$, which means its expected payoff is $-1 - \frac{1}{2}\lambda^L$. If L deviates to $(0,1)$, its probability of winning is $\frac{1}{2}$, which means that L's expected payoff is $-\frac{1}{2} - \frac{\lambda^L}{2}$. Therefore, the deviation is profitable.

If party L announces at $(x^L, y^L) = (0,1)$ and party R announces $(x^R, y^R) = (0,0)$, then voters with initial ideal point $(0,0)$ will support R and voters with initial ideal point $(0,1)$ will support L. Voters with initial ideal point $(1,1)$ get the highest payoff by changing to $(0,1)$ and supporting L; voters with initial ideal point $(1,0)$ get the highest payoff by changing to $(0,0)$ and supporting R. Thus party R's probability of winning is $\frac{1}{2}$, which means its expected payoff is $-1 - \frac{1}{2}\lambda^R$. If R deviates to $(0,1)$, then its probability of winning is $\frac{1}{2}$, which means its expected payoff is -1. Therefore, the deviation is profitable.

If party R announces $(x^R, y^R) = (0,0)$ and party L announces $(x^L, y^L) = (1,0)$, then voters with initial ideal point $(0,0)$ will support R and voters with initial ideal point $(1,0)$ will support L. Voters with initial ideal point $(0,1)$ get the highest payoff by changing to $(0,0)$ and supporting R; voters with initial ideal point $(1,1)$ get the highest payoff by changing to $(1,0)$ and supporting L. Thus party L's probability of winning is $\frac{1}{2}$, which means its expected payoff is $-\frac{1}{2}$. If L deviates to $(0,0)$, its probability of winning is $\frac{1}{2}$, which means it expected payoff is 0. Therefore, the deviation is profitable.

B. Proof of Proposition 2

The proof is by induction. Since the set of individuals that engage in violence in the first period is an interval $[-\frac{\lambda_0}{2}, \frac{\lambda_0}{2}]$ the proposition can be proven by showing that if the set of individuals that engage in violence in period t is an interval $[-\frac{\lambda_t}{2}, \frac{\lambda_t}{2}]$ then the set that engage in violence in period $t+1$ is $[-\frac{\lambda_{t+1}}{2}, \frac{\lambda_{t+1}}{2}]$, where λ_{t+1} is given by (†) in the statement of the proposition. The path of attitudes $x_t^n(r)$ described in $(*)$ is then a immediate implication of this result. For any set $\mathcal{R} \subseteq \mathbb{R}$ we use $u_t(\mathcal{R})$ denotes the image of u_t on $\mathcal{R}$.

Note that in periods $t < t^*$, we have $\lambda_t \geq \mu$ along the conjectured path. We focus on values of $r \geq 0$, since the argument for values of $r < 0$ will be symmetric. For every individual located at $r \in \left[0, \frac{\lambda_t}{2} - \frac{\mu}{2}\right]$, $\mathcal{A}_t^0(r) = \emptyset$, so $a_{t+1}(r) = 1$. For individuals located at $r > \frac{\lambda_t}{2} + \frac{\mu}{2}$, $\mathcal{A}_t^1(r) = \emptyset$, so $a_{t+1}(r) = 0$. For individuals $r \in (\frac{\lambda_t}{2} - \frac{\mu}{2}, \frac{\lambda_t}{2} + \frac{\mu}{2}]$, we have

$$\sup u_t(\mathcal{A}_t^1(r)) = u_t\left(\max\{0, r - \tfrac{\mu}{2}\}\right) = \begin{cases} \pi_t\left(1 - \frac{r - \lambda_t/2}{\mu}\right) - v & \text{if } r > \lambda_t/2 \\ \pi_t(1) - v & \text{if } r \leq \lambda_t/2 \end{cases} \tag{11}$$

The first equality follows from the fact that $\pi_t(\rho)$ is strictly increasing in ρ by Assumption (i), so $u_t(\tilde{r})$ is highest for $\tilde{r} = \max\{0, r - \frac{\mu}{2}\}$ in the set $\mathcal{B}(r)$. The second follows from noting that $\rho_t\left(\max\{0, r - \frac{\mu}{2}\}\right)$ equals 1 for $r \leq \lambda_t/2$, and equals $1 - \frac{r - \lambda_t/2}{\mu}$ for $r > \lambda_t/2$; and then substituting $u_t(r)$ from (2).

On the other hand, for these individuals we also have

$$\begin{aligned} \sup u_t(\mathcal{A}_t^0(r)) &= \lim_{\varepsilon \to 0^+} \pi_t\left(\rho_t\left(\tfrac{\lambda_t}{2} + \varepsilon\right)\right) \\ &= \lim_{\varepsilon \to 0^+} \pi_t\left(\tfrac{1}{\mu}\left((\tfrac{\lambda_t}{2} + \varepsilon) - (\tfrac{\lambda_t}{2} - \tfrac{\mu}{2})\right)\right) = \pi_t\left(\tfrac{1}{2}\right) \end{aligned} \tag{12}$$

which follows from the continuity of $\pi_t(\cdot)$ by assumption (i). Since $v < \pi_t(1) - \pi_t(\frac{1}{2})$ for all periods $t < t^*$, these results imply that $\sup u_t\left(\mathcal{A}_t^1(r)\right) \geq \sup u_t\left(\mathcal{A}_t^0(r)\right)$ for all $r \leq \lambda_t/2$. For $r > \lambda_t/2$ we have $\sup u_t\left(\mathcal{A}_t^1(r)\right) \geq \sup u_t\left(\mathcal{A}_t^0(r)\right)$ if and only if

$$\pi_t\left(1 - \frac{r - \lambda_t/2}{\mu}\right) - v \geq \pi_t\left(\frac{1}{2}\right) \tag{13}$$

At $r = \lambda_t/2$, the left side is bigger than the right side by assumption (i), and at $r = (\lambda_t + \mu)/2$ it is smaller (since $v > 0$). Since $\pi_t(\rho)$ is increasing in ρ, there is a critical r_t^* such that the two sides are equal. By definition, we have

$$1 - \frac{r_t^* - \lambda_t/2}{\mu} = \rho_t^* \qquad \Longrightarrow \qquad r_t^* = \frac{\lambda_t}{2} + \mu(1 - \rho_t^*) \tag{14}$$

Since we need $r_t^* = \lambda_{t+1}/2$, the result obtains for periods $t < t^*$.

Now consider periods $t \geq t^*$, and suppose that $\lambda_t \geq \mu$. Again, for all $r \in \left[0, \frac{\lambda_t}{2} - \frac{\mu}{2}\right]$, $\mathcal{A}_t^0(r) = \emptyset$, so $a_{t+1}(r) = 1$; and for all $r > \frac{\lambda_t}{2} + \frac{\mu}{2}$, $\mathcal{A}_t^1(r) = \emptyset$, so $a_{t+1}(r) = 0$. For all $r \in \left(\frac{\lambda_t}{2} - \frac{\mu}{2}, \frac{\lambda_t}{2} + \frac{\mu}{2}\right]$, the expressions for $\sup u_t\left(\mathcal{A}_t^1(r)\right)$ and $\sup u_t\left(\mathcal{A}_t^0(r)\right)$ are given by (11) and (12) above, so $v > \pi_t(1) - \pi_t(\frac{1}{2})$ and the fact that $\pi_t(\rho)$ is strictly increasing in ρ imply that $\sup u_t\left(\mathcal{A}_t^1(r)\right) < \sup u_t\left(\mathcal{A}_t^0(r)\right)$ for all $r \in \left(\frac{\lambda_t}{2} - \frac{\mu}{2}, \frac{\lambda_t}{2} + \frac{\mu}{2}\right]$. Thus, the set of group A individuals who engage in violence in period $t+1$ is $\left[-\frac{\lambda_t}{2} + \frac{\mu}{2}, \frac{\lambda_t}{2} - \frac{\mu}{2}\right]$.

Now suppose $0 < \lambda_t < \mu$. Again, since $\mathcal{A}_t^1(r) = \emptyset$ for all $r > \frac{\lambda_t}{2} + \frac{\mu}{2}$ we know that $a_{t+1}(r) = 0$ for all $r > \frac{\lambda_t}{2} + \frac{\mu}{2}$. On the other hand, for all $r \in \left[0, \frac{\lambda_t}{2} + \frac{\mu}{2}\right]$, we have

$$\sup u_t(\mathcal{A}_t^0(r)) = \lim_{\varepsilon \to 0^+} \pi_t\left(\tfrac{\lambda_t}{2} + \varepsilon\right) = \begin{cases} \pi_t\left(1/2\right) & \text{if } \lambda_t \geq \mu/2 \\ \pi_t\left(\lambda_t/\mu\right) & \text{if } \lambda_t < \mu/2 \end{cases} \tag{15}$$

For all $r \geq 0$, $\sup u_t(\mathcal{A}_t^1(r))$ is bounded above by $u_t(0) = \pi_t\left(\lambda_t/\mu\right) - v$. So if $\lambda_t < \mu/2$, then for all $r \geq 0$ we have $a_{t+1}(r) = 0$, since $v > 0$ implies $\sup u_t(\mathcal{A}_t^1(r)) < \sup u_t(\mathcal{A}_t^0(r))$ in this case. If, instead, $\lambda_t \geq \mu/2$, then we also have $\sup u_t(\mathcal{A}_t^1(r)) < \sup u_t(\mathcal{A}_t^0(r))$, since

$$\pi_t\left(1/2\right) > \pi_t(1) - v > \pi_t\left(\lambda_t/\mu\right) - v \tag{16}$$

where the first inequality follows by assumption (ii), and the second follows by our hypothesis that $\lambda_t < \mu$ and the fact that $\pi_t(\rho)$ is strictly increasing in ρ. So we have $a_{t+1}(r) = 0$ for all $r \geq 0$. Finally, if $\lambda_t = 0$ then $\mathcal{A}_t^1(r) = \emptyset$, so $a_{t+1}(r) = 0$ for all r. □