Explaining Preferences from Behavior: A Cognitive Dissonance Approach

Avidit Acharya, Stanford University
Matthew Blackwell, Harvard University
Maya Sen, Harvard University

The standard approach in positive political theory posits that action choices are the consequences of preferences. Social psychology—in particular, cognitive dissonance theory—suggests the opposite: preferences may themselves be affected by action choices. We present a framework that applies this idea to three models of political choice: (1) one in which partisanship emerges naturally in a two-party system despite policy being multidimensional, (2) one in which interactions with people who express different views can lead to empathetic changes in political positions, and (3) one in which ethnic or racial hostility increases after acts of violence. These examples demonstrate how incorporating the insights of social psychology can expand the scope of formalization in political science.

What are the origins of interethnic hostility? How do young people become lifelong Republicans or Democrats? What causes people to change deeply held political preferences? These questions are the bedrock of many inquiries within political science. Numerous articles and books study the determinants of racism, partisanship, and preference change. Throughout, a theme linking these seemingly disparate literatures is the formation and evolution of political and social preferences as an object of study.

Although the empirical literature in these areas is well developed, formal theories of preference change have been substantially more scarce in political science.[1] 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 article, we adopt the perspective that preferences 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 preferences of individuals; they instead may be chosen for a variety of reasons, including imitation, experimentation, and habit. Preferences then adjust to justify the behaviors that were adopted.

Our framework builds on an insight originating in social psychology with the work of Festinger (1957) that suggests that actions could affect preferences through *cognitive dissonance*. One key aspect of cognitive dissonance theory is that individuals experience a mental discomfort after taking actions that appear to be in conflict with their starting preferences. To minimize or avoid this discomfort, they change their preferences to more closely align with their actions.

We show via three examples that the cognitive dissonance approach can be applied to settings in politics in which individuals make choices and then later change their intrinsic preferences to be consistent with those choices. Because the theory views preferences as the consequences of actions, the approach is well suited to applications where actions are the main independent variables and preference parameters are the dependent variables. Indeed, a vast subfield of political science—political behavior—is concerned with the origins of partisanship, ideology, ethnic identification, and so on. Our examples show how the traditional rational choice approach can be extended to provide a better understanding of the sources of these preferences by incorporating ideas from cognitive dissonance theory.

Avidit Acharya (avidit@stanford.edu) is an assistant professor at Stanford University, Stanford, CA 94305. Matthew Blackwell (mblackwell@gov.harvard.edu) is an assistant professor at Harvard University, Cambridge, MA 02138. Maya Sen (maya_sen@hks.harvard.edu) is an associate professor at Harvard University, Cambridge, MA 02138.

Data and supporting materials necessary to reproduce the numerical results in the article are available in the *JOP* Dataverse (https://dataverse.harvard.edu/dataverse/jop). An online appendix with supplementary material is available at http://dx.doi.org/10.1086/694541.

1. There are some exceptions, however. We discuss these below.

The Journal of Politics, volume 80, number 2. Published online March 1, 2018. http://dx.doi.org/10.1086/694541

We proceed as follows. We begin by providing a conceptual overview of our approach and by developing the basic framework. We then develop the three applications. The first demonstrates how the cognitive dissonance approach can explain the development of partisan affiliation. The second demonstrates how individuals with differing political preferences—but who feel empathy or kinship toward one another—find compromise by adjusting their policy positions. The third shows how cognitive dissonance can explain the emergence and persistence of ethnic or racial hostility from acts of violence. We conclude with a discussion of other areas of politics in which these ideas may be applied.

ACTIONS CAN AFFECT PREFERENCES

Studies by social psychologists have documented the possibility that action choices affect preferences. For example, Davis and Jones (1960) and Glass (1964) demonstrated that individuals are likely to lower their opinions of others whom they are made to speak ill of or harm. They interpreted these lowered opinions as consequences of the choice to harm. Several other experiments (e.g., Brehm 1956; Festinger 1957; Festinger and Carlsmith 1959) provide similar evidence that making a choice or undertaking an action—oftentimes blindly or forcibly—can lead to an increased preference over time for the chosen alternative. The theory has been tested in experiments involving young children, animals, and amnesiacs (Lieberman et al. 2001), suggesting that the idea that preferences follow actions may be innate across species. Egan, Bloom, and Santos (2010) and Egan, Santos, and Bloom (2007), for example, showed how children and monkeys that chose a certain kind of toy or candy would then, in the next round of experimentation, devalue other toys or candies, even when the initial choice was made blindly (cf. Chen and Risen 2010). In addition, neurologists have documented physiological changes consistent with subjects forming stronger commitments to their choices after the choice has been made (Sharot, De Martino, and Dolan 2009).

These findings and their interpretations contrast with the traditional rational-choice approach. When an action that an individual chooses, or might choose, is in conflict with the individual's preference, 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 approach is that preferences dictate actions, not vice versa (cf. Dietrich and List 2011, 2013). Nevertheless, our work demonstrates how the views of social psychology can be consistent with a broader interpretation of the rational choice approach and may even be considered a part of it. We develop a framework for how a decision maker chooses preference parameters to maximize an objective function, which can be interpreted as a utility. The decision maker seeks to minimize certain costs, which happen to be psychological rather than material. Our model uses the language of the rational choice approach—"maximize utility given costs"—to explain preference change. The result is that individuals bring their preferences into alignment with their actions.

Framework

We develop our main theoretical framework in this section. We consider a person with a starting preference parameter x^o, which is fixed. There is an action a that is taken and a new preference parameter x^n that is chosen by the individual. These choices influence two terms that we refer to as "action dissonance" and "preference change dissonance." Action dissonance is given by the function $d_A(a, x^n)$ that is increasing in some measure of the discrepancy between the action a and the new preference parameter x^n. Preference change dissonance is a function $d_P(x^n, x^o)$ that is increasing in some measure of the discrepancy between the new and old preference parameters, x^n and x^o. "Total dissonance" is the sum of action and preference change dissonance,

$$d(a, x^n, x^o) = d_A(a, x^n) + d_P(x^n, x^o). \quad (1)$$

We can think of the decision maker as seeking to maximize $-d(a, x^n, x^o)$, that is, to minimize total dissonance. In this case, we can consider $u = -d(a, x^n, x^o)$ to be the decision maker's utility and both a and x^n to be choice variables. Alternatively, the decision maker may choose the action a according to some behavioral rule (e.g., to maximize a different objective function) and choose x^n to maximize u. In yet another alternative, the action may be chosen by someone other than the decision maker or forced on the decision maker by a third party. Or, some components of a may be chosen by the decision maker while other components are chosen by others. In all of these cases, the decision maker chooses at least x^n to maximize u, and in this sense maximizing u is an objective of the decision maker.

Our first example, on partisanship, considers a simple decision-theoretic problem for a voter choosing a and x^n to minimize total dissonance $d(a, x^n, x^o)$ absent any strategic considerations. The next example, on socialization and empathy, considers two individuals who each choose a component of a two-dimensional $a = (a_1, a_2)$ and a new political viewpoint x^n. This application considers a strategic interaction between two individuals. The third example, on attitudes shaped by violence, considers a behavioral model in which the action a is not optimized but rather imitated from others, and agents change their preference parameter x^n to cope with the dissonance created by the mismatch between the initial preference parameter x^o and the nonoptimal a.

Other approaches

The discussion above clarifies how the ideas of cognitive dissonance theory can be consistent with a broad interpretation of rational choice, but it also makes clear the important caveat that our approach is not to formalize cognitive dissonance theory; rather, it is to develop a formal theory of preference change that is inspired by some of the ideas that were developed in the cognitive dissonance literature. In short, we are exploring the consequences, not the causes, of cognitive dissonance.

Our focus in this article is on how actions can induce changes in preferences, but there are other studies that use cognitive dissonance to explain preferences without appealing to any action. In one alternative approach, Jost et al. (2003) argue that political ideology is a form of motivated cognition, under which individuals develop ideology in response to deep-seated motivations to reduce uncertainty and perceived threat. In this framework, cognitive dissonance, along with other fundamental motivations, helps shape a person's basic political ideology, which in turn forms the basis of preferences over policies and candidates. Changes to the perception of threat or uncertainty can lead to changes in ideology. This theory provides an explanation of how preferences might develop and change in the absence of any concrete actions, which is an important consideration but one that we do not model here. Nevertheless, if taking an action changes a person's beliefs about threat or uncertainty, actions would lead to ideological shifts due to cognitive dissonance in both our model and that of Jost et al. (2003). In this case, motivated cognition would be a force that shapes the initial preferences, x^o, which would, in turn, affect future preferences.

In addition, early work by Festinger, Riecken, and Schachter (1956) presents evidence that individuals can reinforce their existing beliefs despite learning information that appears inconsistent with these beliefs (see also Jost and Banaji 1994; Nyhan and Reifler 2010).[2] Our model does not speak directly to this possibility, although some work in behavioral economics does address the fact that cognitive dissonance may arise from the conflict between an individual's existing beliefs and new information, or existing beliefs and known facts (Benabou and Tirole 2006). Our work complements this work by maintaining focus on the discrepancies between preferences and actions, rather than the discrepancies between beliefs and information.[3]

Our work also differs from other models of preference change. For example, it differs from evolutionary approaches (e.g., Dekel, Ely, and Yilankaya 2007; Güth and Yaari 1992; Little and Zeitzoff 2017) in that preferences are chosen optimally rather than being the outcome of a natural selection process. It differs also from models of endogenous belief formation that rely on anticipatory effects of uncertainty (e.g., Benabou 2008; Minozzi 2013). Instead, it is most closely related to the models of Akerlof and Dickens (1982) and Rabin (1994), who apply the cognitive dissonance concept to study applications in which individuals rationalize the choice of "immoral" actions, and to a recent model by Penn (2017), who applies the concept to study the endogenous adoption of economically productive skills in understanding economic inequality. Our article differs from these contributions in that the outcomes of interest in our applications are political preferences, formalized as preference parameters (such as ideal points). We now turn to these applications.

PARTISANSHIP

In this section, we develop 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. Political competition between two policy-motivated parties endogenously produces an electorate that is ideologically unidimensional in the sense that voter preferences become perfectly correlated across dimensions. This occurs because voters wanting to minimize cognitive dissonance will adjust their policy preferences toward the platform of the party that they support. Partisanship emerges as a natural outcome of this process.

Model

The policy space, $X = [0, 1] \times [0, 1]$, is two-dimensional with generic policy denoted (x_1, x_2). For concreteness, one can think of the first dimension as economic policy and the second dimension as social policy. A left party L runs on policy $(x_1^L, x_2^L) = (0, 0)$, and a right party R runs on policy $(x_1^R, x_2^R) = (1, 1)$.[4] A voter with initial ideal point $x^o = (x_1^o, x_2^o) \in X$ decides both which party to support and what to choose as her

2. Some authors, however, have provided alternative theories to account for such evidence. See, e.g., Bem's (1967) theory of self-perception and Cooper and Fazio's (1984) theory of aversive consequences.

3. In yet another application of cognitive dissonance theory, Festinger and Carlsmith (1959) suggest that the theory explains why monetary rewards could crowd out intrinsic motivation. Again, our model does not directly address this kind of application, although some aspects of this theory have also been formalized and developed further by Benabou and Tirole (2003).

4. Here, we assume that parties have fixed party platforms, but in app. B, available online, we present a version of this model that allows the parties to choose their positions strategically. Much of the intuition of the more simple approach here carries over to that setting.

new ideal point $x^n = (x_1^n, x_2^n) \in X$. If the voter supports party j, then her choice $a = (a_1, a_2)$ equals j's platform (x_1^j, x_2^j). The voter has action dissonance and preference change dissonance given by

$$\begin{aligned} d_A(a, x^n) &:= |a_1 - x_1^n| + \gamma|a_2 - x_2^n| \\ d_P(x^n, x^o) &:= \kappa(|x_1^n - x_1^o| + \gamma|x_2^n - x_2^o|), \end{aligned} \tag{2}$$

where $\gamma > 0$ is the salience of the second issue with respect to the first and $\kappa > 0$ represents the salience of preference change dissonance with respect to action dissonance. The voter chooses a and x^n to minimize total dissonance, so the voter's preferences are represented by $u(a, x^n|x^o) := -d(a, x^n, x^o)$, where $d(a, x^n, x^o)$, given by (1), is the sum of action and preference change dissonance.

Proposition 1. A voter with initial ideal position (x_1^o, x_2^o) supports party L if x_2^o is smaller than

$$\ell(x_1^o) := \frac{1}{2}\left(\frac{1+\gamma}{\gamma}\right) - \frac{1}{\gamma}x_1^o$$

and supports party R if x_2^o is greater than $\ell(x_1^o)$. If $\kappa < 1$, then she changes her ideal point to the platform of the party she supports (i.e., $(x_1^n, x_2^n) = (x_1^j, x_2^j)$, where $j = L, R$ is her party), while if $\kappa > 1$, she keeps her initial ideal point (i.e., $(x_1^n, x_2^n) = (x_1^o, x_2^o)$).

Line $\ell(x_1^o)$ is negatively sloped in (x_1, x_2)-space and passes through the point (1/2, 1/2). A voter with an initial ideal point below this line supports the left party, while a voter with an initial ideal point above the line supports the right party. The line ℓ gets steeper as γ, the importance of the second issue, falls. This has the natural implication that voters who are right wing on the first issue but left wing on the second issue shift away from the right party and move to the left party as the second issue becomes more important. If $\kappa < 1$, they sort into being right partisans rather than left partisans. If voter ideal points are distributed across the policy space and γ and $\kappa < 1$ are shared across individuals, then ℓ is the "cutting line" that partitions the electorate into left and right partisans. Preferences become one-dimensional as a result of partisanship.

Discussion

The above example shows that while the two parties adopt their own preferred positions, voters whose initial preferences can lie anywhere in the two-dimensional policy space may change their ideal point to match the positions taken by the party they support. Partisanship, in this example, emerges naturally from voters wanting to minimize the psychological cost associated with supporting a party that takes a position different from their own ideal position.[5]

The example provides some support for empirical findings that document how earlier political actions have downstream effects on preferences toward parties or candidates. For example, McCann (1997) argues that citizens changed their core values to match the values of their preferred candidate in a previous presidential election, conjecturing that cognitive dissonance may explain the changes. Similarly, Lenz (2012) shows that voters in the United States first choose a politician to back and then shift their positions to adopt that leader's policy views, and Levendusky (2009) shows that elite polarization leads to mass opinion sorting along partisan lines.[6]

Finally, the model can be extended to highlight the possibility that variation in political knowledge could affect 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 party. Low-information voters may have less cognitive dissonance simply because they are less likely to have knowledge of the parties' political platforms.[7] This assumption could help explain why political knowledge predicts the consistency of mass political preferences with party elites (Zaller 1992). This is also in line with Layman and Carsey (2002), who show that only high-information voters have polarized along with the parties in recent decades.

SOCIALIZATION AND EMPATHY

When two individuals socialize, it is possible that their preferences 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). One channel for this is *empathy*. 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

5. In this sense, the above example speaks to ideological scaling efforts documenting that policy preferences of political elites in the United States can be scaled onto no more than two dimensions and usually just one (Poole and Rosenthal 1991).

6. The findings are also consistent with the literature showing persistence in the turnout decision (e.g., Bølstad, Dinas, and Riera 2013; Meredith 2009; Mullainathan and Washington 2009).

7. If a voter does not know these positions (or does not know any one of the components of a party's position), then it is natural to assume that the voter does not experience any cognitive dissonance rather than to assume that voters have beliefs about the positions of parties and experience the "expected level of cognitive dissonance" from being out of step with respect to these beliefs.

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 preferences to become closer to one another's.

Model

Two individuals, $i = 1, 2$, 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, which is common knowledge to both individuals. Each individual simultaneously decides what her new ideal point x_i^n will be and which ideal position a_i to express. Let $a = (a_1, a_2)$ denote the pair of actions chosen. In this application, we assume that action dissonance and preference change dissonance are given by

$$\begin{aligned} d_{A,i}(a, x_i^n) &:= (x_i^n - a_i)^2 + e_i(x_i^n - a_{-i})^2 \\ d_{P,i}(x_i^n, x_i^o) &:= \kappa_i(x_i^n - x_i^o)^2, \end{aligned} \tag{3}$$

respectively, where $-i$ is the usual notation for the other individual and e_i and κ_i are positive parameters. As in the previous example, both individuals have preferences represented by the negative of total dissonance. We write the utility of voter i as

$$u_i(a, x_i^n | x_i^o) = -d_i(a, x_i^n, x_i^o) := -d_{A,i}(a, x_i^n) - d_{P,i}(x_i^n, x_i^o)$$

and posit that i chooses (a_i, x_i^n) to maximize this utility. Thus, each individual desires to express a position a_i that matches her new ideal position x_i^n. Each individual, however, also experiences some discomfort when her new ideal position x_i^n is different from the ideal position expressed by the other individual a_{-i}. This discomfort is weighted by $e_i > 0$, which we interpret as the level of empathy that individual i has toward $-i$. Finally, there is a cost to changing one's ideal position from x_i^o to x_i^n. This cost is weighted by the salience of preference change dissonance, κ_i.

The first-order conditions for the maximization of $u_i(a, x_i^n | x_i^o)$ with respect to a_i and x_i^n imply that

$$\begin{aligned} a_i &= x_i^n \\ x_i^n &= \frac{e_i}{e_i + \kappa_i} a_{-i} + \frac{\kappa_i}{e_i + \kappa_i} x_i^o. \end{aligned} \tag{4}$$

This means that individual i's new preference parameter x_i^n is a weighted average of the old preference parameter x_i^o and the other individual's expressed preference a_{-i}, where the weights are determined by the level of empathy e_i and the salience of preference change dissonance κ_i. When empathy is high, the new preference parameter is closer to the other individual's expressed preference, and when the cost of preference change is high, the new preference parameter is instead closer to the old preference parameter.

Finally, in a game in which each of the two individuals simultaneously best responds to the choices made by the other, their choices solve the system of equations implied by (4) for $i = 1, 2$. We report the unique Nash equilibrium of this game as follows.

Proposition 2. In the unique Nash equilibrium, each individual $i = 1, 2$ chooses (a_i, x_i^n) given by

$$\begin{aligned} a_i &= x_i^n = \alpha_i x_i^o + (1 - \alpha_i) x_{-i}^o, \text{ where } \alpha_i \\ &= \frac{e_{-i}\kappa_i + \kappa_{-i}\kappa_i}{e_{-i}\kappa_i + e_i\kappa_{-i} + \kappa_{-i}\kappa_i}. \end{aligned}$$

To summarize, in equilibrium, individual i expresses a position a_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 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. The weight α_i is increasing in the degree of empathy e_{-i} that the other individual $-i$ feels toward i and decreasing in the difficulty κ_{-i} that $-i$ experiences in changing his position. In the relationship, if one individual does not feel very much empathy toward the other, or if he finds it difficult to change his views, then the other individual ends up compromising her position more.

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 standard dynamic adjustment (i.e., *tâtonnement*) argument for how the players might arrive at equilibrium through socialization.[8]

In our setup, 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. The "reaction functions" (i.e., best response functions) for each player are

8. An alternative approach, which we do not pursue here, would be to have the players interact repeatedly, taking the pair of new ideal points (x_1^n, x_2^n) from the last period interaction as the current period state variables, and then characterize the limit of the sequence of ideal points under a Markov perfect equilibrium. However, our dynamic adjustment approach can be interpreted as a dynamic game in which myopic players have the objective of best responding to the other player's last period announcement but in which dissonance with the old preference parameters (x_1^o, x_2^o) is persistent.

$$r_i(x^n_{-i}) = \frac{e_i}{e_i + \kappa_i} x^n_{-i} + \frac{\kappa_i}{e_i + \kappa_i} x^o_i, \qquad i = 1, 2. \tag{5}$$

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:

$$\begin{aligned} x_1[0] &= x^o_1 \\ x_2[0] &= x^o_2 \\ x_1[t] &= \frac{e_1}{e_1 + \kappa_1} x_2[t-1] + \frac{\kappa_1}{e_1 + \kappa_1} x^o_1, \qquad t > 0 \\ x_2[t] &= \frac{e_2}{e_2 + \kappa_2} x_1[t] + \frac{\kappa_i}{e_2 + \kappa_2} x^o_2, \qquad t > 0, \end{aligned} \tag{6}$$

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^o_1 and x^o_2 the sequence of positions for each player converges to the equilibrium positions given in proposition 2 above.

Proposition 3. For all x^o_1 and x^o_2, the sequences of $\{x_1[t]\}_t$ and $\{x_2[t]\}_t$ converge to the equilibrium values of x^n_1 and x^n_2 given in proposition 2 above.

Figure 1 illustrates the socialization process described above. It shows how the dynamic adjustment process leads to the players eventually reaching the equilibrium values (x^n_1, x^n_2) from the starting point (x^o_1, x^o_2). The two oblique lines 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^o_1, x^o_2). What the figure does not reveal is that each individual's final position lies between his original position and the original position of the other individual. This follows from the fact that α_i in proposition 2 lies between 0 and 1.

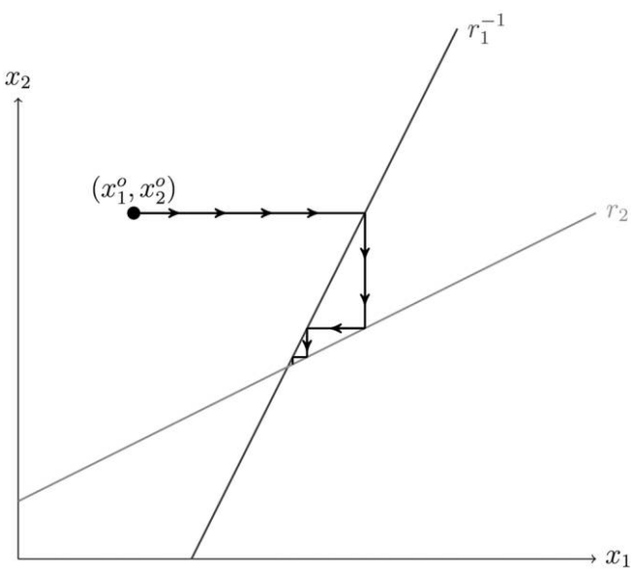

Figure 1. Socialization as a dynamic adjustment process

In addition, figure 1 shows 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.

Discussion

This application provides theoretical support to two related literatures. The first documents the stability of partisanship over time along with its ability to change as a result of major life events, including marriage and divorce (Green, Palmquist, and Schickler 2002) or emigration (Brown 1981). For example, Green et al. (2002) observe that partisanship operates similarly to religious affiliation in the sense that close, empathetic relationships have the potential to change it. They write that an "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" (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, the model sheds light on how empathy can lead to changes in specific policy positions. Several studies have documented that close relationships have the capacity to affect decision making on certain issues. For example, leveraging a natural experiment, Washington (2008) finds that male members of Congress who have daughters tend to vote

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

in more liberal directions on issues having a gender component. This finding was replicated in the judicial context by Glynn and Sen (2015). The application also speaks to a broader literature on political persuasion, which examines campaign tactics in the United States and documents that sending demographically similar campaign workers is more effective than sending dissimilar workers, perhaps because similarity activates empathy (Enos and Hersh 2015; Leighley 2001; Shaw, de la Garza, and Lee 2000).

One question that our application leaves unanswered is: What determines the level of empathy to begin with? That is, what determines the values of the empathy parameters e_i ($i = 1, 2$)? Whether empathy leads to substantial convergence in preferences between the two individuals depends on how large these parameters are. If they are small, then socialization will not lead to much convergence in preferences. And, if they are negative—meaning that the individuals feel antipathy rather than empathy toward one another—then socialization will lead to further preference divergence. Since there is considerable variation in the success of interventions designed to increase empathy (e.g., Gubler 2013), it would be valuable to empirically investigate the determinants of the model's parameters.[10]

ATTITUDES SHAPED BY VIOLENCE

The conventional view is that violence is the outcome of prejudice: individuals engage in violence against those they hate. Holmes, in his introduction to *Behemoth*, however, attributes to Hobbes another equally plausible view: "In 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, 32).

In this section, we develop an application in which ethnic or racial animosity 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.[11] The application supports Hobbes's conjecture and provides a formal theoretical 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 coevolve with violence, tracking the amount of violence over time. We explain how ethnic hostility may in fact persist even after violence disappears, a result that has many applications that we discuss below.[12]

Model

Consider a dynasty r of one-period-lived individuals. The individual that is alive in each period $t = 0, 1, 2 \ldots$ decides whether to engage in an aggressive action $a_t(r) \in \{0, 1\}$ against a member of another group, which we will refer to as the "target group" ($a_t(r) = 1$ means that the individual from dynasty r alive in period t chooses the aggressive action; $a_t(r) = 0$ means that he does not). The individual alive in period t starts the period with attitude $x_t^o(r) \in [0, 1]$ toward members of the target group, where high values of $x_t^o(r)$ indicate more hostile attitudes. At the end of the period, the individual forms a new attitude $x_t^n(r) \in [0, 1]$ and then passes down this attitude to the next generation so that $x_{t+1}^o(r) = x_t^n(r)$. The individual from dynasty r alive in period t has action and preference change dissonances given by, respectively,

$$\begin{aligned} d_A(a_t(r), x_t^n(r)) &= |x_t^n(r) - a_t(r)| \\ d_P(x_t^n(r), x_t^o(r)) &= \frac{1}{2\kappa}[x_t^n(r) - x_t^o(r)]^2, \end{aligned} \tag{7}$$

where $\kappa > 0$ is a parameter that determines the salience of preference change dissonance. The generation t individual chooses $a_t(r)$ according to a behavior rule that we specify below and chooses $x_t^n(r)$ to minimize total dissonance (i.e., the sum of action and preference change dissonances) given the choice of $a_t(r)$. That is, after the individual at r chooses $a_t(r)$ in period t, she chooses $x_t^n(r) \in [0, 1]$ to minimize

$$d_A(a_t(r), x_t^n(r)) + d_P(x_t^n(r), x_t^o(r)).$$

The following lemma characterizes intergenerational attitude change as a function of actions and inherited attitudes.

Lemma 1. Given the choice of $a_t(r)$ and the inherited attitude $x_t^o(r)$, an individual who chooses $x_t^n(r)$ to minimize total dissonance chooses

$$x_t^n(r) = \begin{cases} \min\{x_t^o(r) + \kappa, 1\} & \text{if } a_t(r) = 1 \\ \max\{0, x_t^o(r) - \kappa\} & \text{if } a_t(r) = 0. \end{cases} \tag{8}$$

10. This would enable us to address other related questions, including the role of social networks in changing policy preferences and the impact of close contact between people of different ethnic groups (including both "contact theory" and the "racial threat" hypothesis).

11. Although we use the term "violence" here, this framework can apply to instances involving any kind of negative action that requires costly effort but has diffuse benefits, including (but not limited to) verbal exchanges, the policing of racial roles, etc.

12. We do not address the question of how exposure to violence affects the preferences or attitudes of the target group. Past work on this, e.g., Shayo and Zussman (2011) and Voors et al. (2012), has emphasized the importance of threat perception and trade-offs in social-identity choice to explain the relationship between violence and attitudes for the target group. Whether cognitive dissonance theory can provide alternative explanations for the attitude development of the target group remains an open and interesting question.

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 also assume that the parameter κ is small so that attitudes move incrementally within the interval [0, 1].

Violence decisions

We now study violence decisions under the assumption that agents are connected to each other in a network and choose the action on the basis of imitation of others in their network. Each dynasty r is identified with a real number; thus, $r \in (-\infty, +\infty)$. We refer to the interval $\mathcal{B}(r) = [r - (\mu/2), r + (\mu/2)]$ as the "local community" of dynasty r. The assumption that the local community of a dynasty does not vary over generations is implicit and serves only to simplify the analysis. The model can be extended without much complication to the case in which communities change over time.

Let $\rho_t(r)$ denote the fraction of individuals in r's local community that engage in violence against the target group. We assume then that the "material payoff" to an individual who lives at r is

$$u_t(r) = w_t\rho_t(r) - va_t(r), \tag{9}$$

where $w_t \geq 0$ is a time-varying parameter, and $v > 0$ is the material cost of violence. Since the gains from violence, $w_t\rho_t(r)$, are proportional to the total amount of violence produced in r's local community, our assumption is that violence can influence individual payoffs only socially.

The dynamic linkage across periods in our model arises from intergenerational socialization: each individual observes the material payoffs of the members of his parents' generation that lived in his local community and then decides whether to engage in violence by "imitating" the individual from the previous generation who received the highest material payoff. More formally, define the sets of members of the tth generation individual in dynasty r's local community that respectively do not engage, and engage, in violence to be

$$\begin{aligned} \mathcal{A}_t^0(r) &= \{\tilde{r} \in \mathcal{B}(r) : a_t(r) = 0\}, \\ \mathcal{A}_t^1(r) &= \{\tilde{r} \in \mathcal{B}(r) : a_t(r) = 1\}. \end{aligned} \tag{10}$$

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

$$a_{t+1}(r) = \begin{cases} 0 & \text{if } \sup\, u_t(\mathcal{A}_t^1(r)) < \sup\, u_t(\mathcal{A}_t^0(r)) \\ 1 & \text{if } \sup\, u_t(\mathcal{A}_t^1(r)) \geq \sup\, u_t(\mathcal{A}_t^0(r)), \end{cases} \tag{11}$$

and if $\mathcal{A}_t^0(r) = \varnothing$, then $a_{t+1}(r) = 1$, while if $\mathcal{A}_t^1(r) = \varnothing$, 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 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.

The dynamic evolution of attitudes and violence

Since the path of violence is generated by recursive imitation, characterizing this path requires making assumptions about the initial conditions. 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 we focus on how violence may spread or decline after this point. Formally, our assumptions about the initial conditions are as follows:

i) $\lambda_0 \geq \mu$.

ii)

$$(a_0(r), x_0^a(r)) = \begin{cases} (1, \kappa) & \text{if } r \in \left[-\dfrac{\lambda_0}{2}, \dfrac{\lambda_0}{2}\right] \\ (0, 0) & \text{otherwise.} \end{cases}$$

Given assumption ii, assumption i guarantees that there is at least one individual whose entire local community engages in violence in the first period. Assumption ii states that the small community of individuals who 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 (although, in fact, they are the first generation of individuals in the model).

Our main result characterizes the recursive paths of violence and attitudes under these assumptions about the initial conditions. To state the result, we divide the set of periods into two subsets, $T_0 = \{t : v < w_t/2\}$ and $T_1 = \{t : v > w_t/2\}$. In what follows we identify the degenerate interval [0, 0] with the empty set $\varnothing$. The following proposition characterizes the coevolution of violence and attitudes in the population over time.

Proposition 4. Given $\lambda_t \geq 0$ and the value of w_t in period t, let

$$\lambda_{t+1} = \begin{cases} \lambda_t + \mu\left(1 - \dfrac{2v}{w_t}\right) & \text{if } t \in T_0 \text{ and } \lambda_t \geq \mu\left(\dfrac{1}{2} + \dfrac{v}{w_t}\right) \\ \max\{0, \lambda_t - \mu\} & \text{otherwise.} \end{cases}$$

Then, the paths of violence and attitudes are recursively given by

$$(a_{t+1}(r), x^n_{t+1}(r)) = \begin{cases} (1, \min\{x^n_t(r) + \kappa, 1\}) \text{ for all } r \in \left[-\frac{\lambda_{t+1}}{2}, \frac{\lambda_{t+1}}{2}\right] \\ (0, \max\{x^n_t(r) - \kappa, 0\}) \text{ for all } r \notin \left[-\frac{\lambda_{t+1}}{2}, \frac{\lambda_{t+1}}{2}\right]. \end{cases}$$

Proposition 4 implies that the mass of individuals who adopt violence grows in any period $t \in T_0$, provided that a large enough interval of individuals adopted violence in the previous period. The proposition also implies that as individuals adopt violence in successive periods, their attitudes toward the target group become increasing hostile.

It is also worth noting that the assumption that agents imitate members of their local community (of the previous generation), rather than optimally decide whether 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. Our assumption that a small concentrated mass λ_0 of individuals choose violence in the first period is also important for this result.

We conclude with an example that will inform our discussion of the related empirical findings below. Suppose that $w_t \in \{w^H, w^L\}$ for all t with $w^L/2 < v < w^H/2$, and $t \in T_0$ if and only if $t \leq t^\star$ for some $t^\star > 0$. Then violence grows up to period $t^\star$ after which it declines. If the critical period $t^\star$ is sufficiently large (and $2v/w^H$ is sufficiently small), then a large mass of individuals continue to develop increasingly hostile attitudes even after violence begins to decline. Consequently, average hostility—that is, the population average of $a_t(r)$—may peak in a period $t^{\star\star} > t^\star$, after which it begins 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. Figure 2 presents the results of a simulation. The figure shows that aggregate hostility goes to 0 at some time $\bar{t}$ after the time $\underline{t}$ at which all violence disappears.

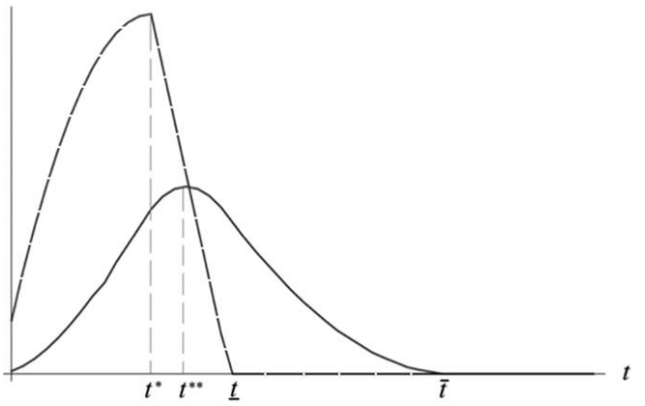

Figure 2. Evolution of aggregate violence (i.e., the integral of $a_t(r)$ over individuals r; *dashed dark gray line*) and aggregate attitudes (i.e., the integral of $x^n_t(r)$ over individuals r; *solid gray line*).

Discussion

The example above says that individuals committing violence against members of another group will develop hostile attitudes toward their victims as a way of minimizing cognitive dissonance. These attitudes may persist longer than the acts of violence that created them.

The model contributes to our understanding of how group-based prejudices might originate and develop. It provides a plausible theoretical framework with which to integrate instrumentalist (strategic) and constructivist approaches in the study of ethnicity and violence.[13] 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, 856) write, individual "actions may … result in the construction of new or altered identities, which themselves change cultural boundaries." Moving from violence to other kinds of hostile actions (e.g., segregation, discrimination) accommodates other theories of how cognitive dissonance may contribute to the propagation of racial/ethnic attitudes or their formation—including how perceptions of threat could lead to racist attitudes. Finally, the mechanism posited by our framework can also operate alongside other mechanisms, including recurring economic incentives or exogenous shocks.

Furthermore, the results provide a theoretical foundation for recent empirical studies documenting the historical persistence of ethnic or racial prejudices that originate in violence. For example, Voigtländer and Voth (2012) document 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 twentieth century. Similarly,

13. It also supports the many empirical studies that have found that violence can be—and has historically been—used by elites as a mechanism of fostering in-group solidarity and furthering anti-outgroup attitudes (Brass 1997; Gagnon 1994). Although we do not directly model such elite strategies in this example, our first application might provide one possible link between elite and racial/ethnic attitudes among the public.

Acharya, Blackwell, and Sen (2016) explore the legacy of American slavery, finding that those parts of the US 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 postbellum racial violence, which was used to terrorize newly freed slaves, solidifying antiblack attitudes.

CONCLUDING REMARKS

One of the main contributions of this article is to demonstrate how ideas from social psychology can help expand the scope of formalization in political science. We developed a framework for how individuals adjust their political and social preferences to minimize cognitive dissonance—the discomfort that arises when choices come into conflict with preexisting preferences. With its roots in social psychology, this simple intuition explains why people often change their preferences to bring them into closer alignment with their actions. It therefore provides a conceptual basis for a model of preference formation.

We conclude by noting that our approach is amenable to introducing other concepts from social psychology that are closely related to cognitive dissonance, such as confirmation bias and motivated reasoning (Lodge and Taber 2013). Confirmation biases are instances when individuals refuse to absorb or engage with potentially conflicting information, choosing instead to update on the basis of information that conforms with preexisting attitudes. Motivated reasoning is the tendency for people to explicitly view new evidence as entirely consistent with their preexisting views (Druckman and Bolsen 2011; Taber and Lodge 2006). Both of these concepts are, at their core, instances when individuals seek to avoid cognitive dissonance. With confirmation bias, cognitive dissonance is minimized by avoiding potentially challenging information that could create mental discomfort; with motivated reasoning, objective information is actively ignored also to reduce such discomfort. Both can be considered special cases of the broader framework that we suggest here, meaning that our approach can be used to formalize these increasingly important concepts and explore their consequences.

APPENDIX A

Proof of proposition 1

The voter's optimization problem is piecewise linear, so the solution lies at a corner: the voter either keeps her initial ideal position on an issue or adopts the position of the party she supports. Suppose the voter supports the left party. If the voter adopts the left party's platform as her ideal point, that is, $(x_1^n, x_2^n) = (0, 0)$, then the voter's utility is $-\kappa(x_1^o + \gamma x_2^o)$. If she keeps her initial ideal point, that is, $(x_1^n, x_2^n) = (x_1^o, x_2^o)$, then her utility is $-(x_1^o + \gamma x_2^o)$. (It cannot be optimal for her to adopt the left party's position on one issue and maintain her initial position on the other since doing so would result in a utility of either $-x_1^o - \kappa\gamma x_2^o$ or $-\kappa x_1^o - \gamma x_2^o$, so the payoff is guaranteed to be lower than the payoff from keeping her initial ideal point or changing her ideal point to the platform of the left party, whichever is greater.) Keeping her initial ideal point is preferable if $\kappa > 1$, while adopting the left's platform as the new ideal point is preferable if $\kappa < 1$. The symmetric argument holds for the case in which the voter supports the right party. The voter then supports the left party if

$$-x_1^o - \gamma x_2^o > -(1 - x_1^o) - \gamma(1 - x_2^o),$$

which rearranges to $x_2^o < \ell(x_1^o)$, where $\ell(x_1^o)$ is defined in the proposition. She supports the right party when the reverse inequality holds. QED

Proof of proposition 2

Follows from solving the best response system of equations defined by (4) for $i = 1, 2$. QED

Proof of proposition 3

Solving the system of recursive equations in (6) yields

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

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 2, for each $i = 1, 2$. QED

Proof of lemma 1

If $a_t(r) = 1$, then total dissonance is $1 - x_t^n(r) + (1/2\kappa)[x_t^n(r) - x_t^o(r)]^2$, so the value of $x_t^n(r) \in [0, 1]$ that minimizes this is $\min\{x_t^o(r) + \kappa, 1\}$. If $a_t(r) = 0$, then total dissonance is $x_t^n(r) + (1/2\kappa)[x_t^n(r) - x_t^o(r)]^2$, so the value of $x_t^n(r) \in [0, 1]$ that minimizes this is $\max\{x_t^o(r) - \kappa, 0\}$. QED

Proof of proposition 4

The proof is by induction. Since the set of individuals that engage in violence in the first period is an interval $[-\lambda_0/2, \lambda_0/2]$, the proposition can be proven by showing that if the set of individuals that engage in violence in period t is an interval $[-\lambda_t/2, \lambda_t/2]$, then the set that engages in violence in period $t + 1$ is $[-\lambda_{t+1}/2, \lambda_{t+1}/2]$, where λ_{t+1} is given in the

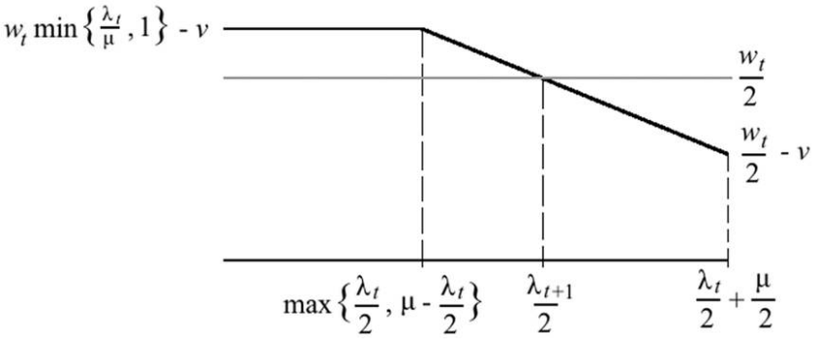

Figure A1. If $(w_t/2) < w_t \min\{(\lambda_t/\mu), 1\} - v$, then the mass of individuals choosing violence grows from λ_t to λ_{t+1}, as depicted. But if the reverse of this inequality holds, then this mass of individuals declines from λ_t to $\max\{0, \lambda_t - \mu\}$.

statement of the proposition. The path of attitudes $x_t^n(r)$ described in the proposition is then an immediate implication of lemma 1. Let us assume that in period t, the set of individuals that choose violence is $[-\lambda_t/2, \lambda_t/2]$.

We focus on values of $r \geq 0$, since the analysis for values of $r < 0$ will be symmetric: the individual at $-r$ makes the same choices as the individual at r. Note that individuals at $r > (\lambda_t/2) + (\mu/2)$ do not choose violence since nobody in their local community commits violence, while if $\lambda_t \geq \mu$, individuals $r \in [0, (\lambda_t/2) - (\mu/2))$ all commit violence since everyone in their local community does. Therefore, all that is required is to characterize the violence decisions of individuals $r \in [\max\{0, (\lambda_t/2) - (\mu/2)\}, (\lambda_t/2) + (\mu/2)] =: \mathcal{R}$. So in what follows, we will assume that r lies in this interval.

If $\lambda_t < \mu/2$, then $\sup u_t(\mathcal{A}_t^1(r)) \leq (w_t\lambda_t/\mu) - v$ and $\sup u_t(\mathcal{A}_t^0(r)) = w_t\lambda_t/\mu$ for all $r \in \mathcal{R}$. Therefore, all individuals choose nonviolence.

If $\mu/2 \leq \lambda_t < \mu$, then $\sup u_t(\mathcal{A}_t^1(r)) \leq (w_t\lambda_t/\mu) - v$, but now $\sup u_t(\mathcal{A}_t^0(r)) = w_t/2$ for all $r \in \mathcal{R}$. Now there are two cases to consider. The first is $\lambda_t < \mu[(1/2) + (v/w_t)]$. In this case, $(w_t\lambda_t/\mu) - v < w_t/2$, so all individuals again choose nonviolence. The second case is $\lambda_t \geq \mu[(1/2) + (v/w_t)]$, which requires $t \in T_0$ by the hypothesis that $\lambda_t < \mu$. Here, $\sup u_t(\mathcal{A}_t^1(r))$ equals $(w_t\lambda_t/\mu) - v$ for $r \leq (\mu/2) + [(\mu/2) - (\lambda_t/2)]$ and is linearly decreasing from $(w_t\lambda_t/\mu) - v$ to $(w_t/2) - v$ on the interval $[(\mu/2) + [(\mu/2) - (\lambda_t/2)], (\mu/2) + (\lambda_t/2)]$, as shown in figure A1. Therefore, $\sup u_t(\mathcal{A}_t^1(r)) \geq \sup u_t(\mathcal{A}_t^0(r))$ if and only if

$$r \leq \frac{\lambda_t}{2} + \frac{\mu}{2}\left(1 - \frac{2v}{w_t}\right).$$

Then λ_{t+1} is defined so that this threshold on r equals $\lambda_{t+1}/2$.

Finally, suppose that $\lambda_t \geq \mu$. In this case, $\sup u_t(\mathcal{A}_t^1(r))$ equals $w_t - v$ for all $r \leq \lambda_t/2$ and is linearly decreasing from $w_t - v$ to $(w_t/2) - v$ on the interval $[\lambda_t/2, (\lambda_t/2) + (\mu/2)]$ (again see fig. A1). But, $\sup u_t(\mathcal{A}_t^0(r)) = w_t/2$ for all $r \in \mathcal{R}$. So if $t \in T_0$, then $\sup u_t(\mathcal{A}_t^1(r)) \geq \sup u_t(\mathcal{A}_t^0(r))$ if and only if $r \leq (\lambda_t/2) + (\mu/2)[1 - (2v/w_t)]$, as before, and λ_{t+1} is again defined so that this threshold on r equals $\lambda_{t+1}/2$. If, however, $t \in T_1$, then $\sup u_t(\mathcal{A}_t^1(r)) < \sup u_t(\mathcal{A}_t^0(r))$ for all $r \in \mathcal{R}$, so $\lambda_{t+1} = \lambda_t - \mu$. QED

ACKNOWLEDGMENTS

Some of the results from our previous working paper titled "Attitudes Shaped by Violence" have been incorporated into this article. We thank Ryan Enos, Jeffry Frieden, Alice Hsiaw, Josh Kertzer, David Laitin, Ken Shepsle, Paul Sniderman, and Dustin Tingley for helpful feedback. Ruxi Zhang provided outstanding research assistance.

REFERENCES

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "The Political Legacy of American Slavery." *Journal of Politics* 78 (3): 621–41.

Akerlof, George A., and William T. Dickens. 1982. "The Economic Consequences of Cognitive Dissonance." *American Economic Review* 72 (3): 307–19.

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

Benabou, Roland. 2008. "Ideology." *Journal of the European Economic Association* 6 (2–3): 321–52.

Benabou, Roland, and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70 (3): 489–520.

Benabou, Roland, and Jean Tirole. 2006. "Belief in a Just World and Redistributive Politics." *Quarterly Journal of Economics* 121 (2): 699–746.

Bølstad, Jørgen, Elias Dinas, and Pedro Riera. 2013. "Tactical Voting and Party Preferences: A Test of Cognitive Dissonance Theory." *Political Behavior* 35 (3): 429–52.

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

Brehm, Jack W. 1956. "Postdecision Changes in the Desirability of Alternatives." *Journal of Abnormal and Social Psychology* 52 (3): 384–89.

Brown, Thad A. 1981. "On Contextual Change and Partisan Attributes." *British Journal of Political Science* 11 (4): 427–47.

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): 573–94.

Cooper, Joel, and Russell H. Fazio. 1984. "A New Look at Dissonance Theory." *Advances in Experimental Social Psychology* 17:229–66.

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): 402–10.

Dekel, Eddie, Jeffrey C. Ely, and Okan Yilankaya. 2007. "Evolution of Preferences." *Review of Economic Studies* 74 (3): 685–704.

Dietrich, Franz, and Christian List. 2011. "A Model of Non-informational Preference Change." *Journal of Theoretical Politics* 23 (2): 145–64.

Dietrich, Franz, and Christian List. 2013. "Where Do Preferences Come From?" *International Journal of Game Theory* 42 (3): 613–37.

Druckman, James N., and Toby Bolsen. 2011. "Framing, Motivated Reasoning, and Opinions about Emergent Technologies." *Journal of Communication* 61 (4): 659–88.

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): 204–7.

Egan, Louisa C., Laurie R. Santos, and Paul Bloom. 2007. "The Origins of Cognitive Dissonance: Evidence from Children and Monkeys." *Psychological Science* 18 (11): 978–83.

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): 252–78.

Fearon, James D., and David D. Laitin. 2000. "Violence and the Social Construction of Ethnic Identity." *International Organization* 54 (4): 845–77.

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

Festinger, Leon, and James M. Carlsmith. 1959. "Cognitive Consequences of Forced Compliance." *Journal of Abnormal and Social Psychology* 58 (2): 203–10.

Festinger, Leon, Henry W. Riecken, and Stanley Schachter. 1956. *When Prophecy Fails*. Minneapolis: University of Minnesota Press.

Gagnon, Valere Philip. 1994. "Ethnic Nationalism and International Conflict: The Case of Serbia." *International Security* 19 (3): 130–66.

Glass, David C. 1964. "Changes in Liking as a Means of Reducing Cognitive Discrepancies between Self-Esteem and Aggression." *Journal of Personality* 32 (4): 531–49.

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. 2002. *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters*. New Haven, CT: Yale University Press.

Gubler, Joshua R. 2013. "When Humanizing the Enemy Fails: The Role of Dissonance and Justification in Intergroup Conflict." Working paper.

Güth, Werner, and Menahem Yaari. 1992. "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game." In Ulrich Witt, ed., *Explaining Process and Change: Approaches to Evolutionary Economics*. Ann Arbor: University of Michigan Press, 23–34.

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

Jost, John T., and Mahzarin R. Banaji. 1994. "The Role of Stereotyping in System-Justification and the Production of False Consciousness." *British Journal of Social Psychology* 33 (1): 1–27.

Jost, John T., Jack Glaser, Arie W. Kruglanski, and Frank J. Sulloway. 2003. "Political Conservatism as Motivated Social Cognition." *Psychological Bulletin* 129 (3): 339–75.

Layman, Geoffrey C., and Thomas M. Carsey. 2002. "Party Polarization and 'Conflict Extension' in the American Electorate." *American Journal of Political Science* 46 (4): 786–802.

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

Lenz, Gabriel S. 2012. *Follow the Leader? How Voters Respond to Politicians' Policies and Performance*. Chicago: University of Chicago Press.

Levendusky, Matthew. 2009. *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans*. Chicago: University of Chicago 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): 135–40.

Little, Andrew T., and Thomas Zeitzoff. 2017. "A Bargaining Theory of Conflict with Evolutionary Preferences." *International Organization* 71 (3): 523–57.

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

McCann, James A. 1997. "Electoral Choices and Core Value Change: The 1992 Presidential Campaign." *American Journal of Political Science* 41 (2): 564–83.

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

Minozzi, William. 2013. "Endogenous Beliefs in Models of Politics." *American Journal of Political Science* 57 (3): 566–81.

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

Nyhan, Brendan, and Jason Reifler. 2010. "When Corrections Fail: The Persistence of Political Misperceptions." *Political Behavior* 32 (2): 303–30.

Penn, Elizabeth Maggie. 2017. "Inequality, Social Context, and Value Divergence." *Journal of Politics* 79 (1): 153–65.

Poole, Keith T., and Howard Rosenthal. 1991. "Patterns of Congressional Voting." *American Journal of Political Science* 35 (1): 228–78.

Rabin, Matthew. 1994. "Cognitive Dissonance and Social Change." *Journal of Economic Behavior and Organization* 23 (2): 177–94.

Sharot, Tali, Benedetto De Martino, and Raymond J. Dolan. 2009. "How Choice Reveals and Shapes Expected Hedonic Outcome." *Journal of Neuroscience* 29 (12): 3760–65.

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): 338–46.

Shayo, Moses, and Asaf Zussman. 2011. "Judicial Ingroup Bias in the Shadow of Terrorism." *Quarterly Journal of Economics* 126 (3): 1447–84.

Taber, Charles S., and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50 (3): 755–69.

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): 1339–92.

Voors, Maarten J., Eleonora E. M. Nillesen, Philip Verwimp, Erwin H. Bulte, Robert Lensink, and Daan P. Van Soest. 2012. "Violent Conflict and Behavior: A Field Experiment in Burundi." *American Economic Review* 102 (2): 941–64.

Washington, Ebonya L. 2008. "Female Socialization: How Daughters Affect Their Legislator Fathers." *American Economic Review* 98 (1): 311–32.

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