

On Hatred*

Tilman Klumpp[†]
Emory University

Hugo M. Mialon[‡]
Emory University

December 2011

Abstract

This paper investigates the effects of hatred in two-player games. We model hate as “reverse-altruism” or a preference for low opponent payoffs, and derive implications for behavior in conflicts where players are motivated by hate. We use these results to illuminate several policy issues, both historical and contemporary: the strategy of non-violent resistance during the American civil rights era, shifts in U.S. national security strategy following 9/11, and the justification for penalty enhancements for hate crimes.

Keywords: Hate; conflict; (non)violence; (counter)terrorism; hate crime.

JEL codes: D74, H11, K14, K42.

*We thank Andrew Francis, Sue Mialon, Paul Rubin, and Xuejuan Su for many fruitful conversations.

[†]Department of Economics, Rich Building 316, 1602 Fishburne Dr., Atlanta, GA 30322. E-mail: tklumpp@emory.edu.

[‡]Department of Economics, Rich Building 317, 1602 Fishburne Dr., Atlanta, GA 30322. E-mail: hmialon@emory.edu.

“The price of hating other human beings is loving oneself less.”

— Eldrige Cleaver, *Soul on Ice*

1 Introduction

The central point we wish to make in this essay is that hate can be understood in much the same way as altruism, namely as an interdependent preference, and that this view can inform our thinking about important policy issues.

What is hate? Like altruism, hate is a concern for someone else’s well-being. Unlike an altruistic person, however, a hateful individual experiences a higher utility the *lower* another individual’s utility is. If individual 1 hates individual 2, then 1’s overall utility can be written as

$$v_1 = (1 - \lambda)u_1 + \lambda(-u_2), \tag{1}$$

where u_1 and u_2 are “material” payoffs and v_1 is individual 1’s “emotional” utility, which includes utility generated by hate. The weight $\lambda \in [0, 1]$ is the degree to which 1 hates 2. Notice that as λ increases, the relative weight placed by individual 1 on her own fundamental utility decreases. That is, “the price of hating other human beings is loving oneself less” (Cleaver 1968).

References to hatred as an other-regarding preference appear in the economics literature as early as in Smith’s *Theory of Moral Sentiments* and Bentham’s *Principles of Morals and Legislation*, where it is discussed alongside altruism.¹ Despite these early appearances, economic investigations of the causes and effects of hatred have been far less frequent than investigations of altruism. This, perhaps, is because benevolence appears to be a more puzzling phenomenon than malice in a world of scarcity and competition. In many human relationships, however, hate matters as much as love. This paper investigates such interactions.

Consider, for example, the phenomenon of suicide attacks. The fact that an individual is willing to sacrifice his own life in order to harm others may seem difficult to reconcile with rational behavior. This becomes less contradictory if one assumes that suicide attackers are motivated by hatred of their intended victims. However, we do not claim to truly know what motivates such individuals—hateful preferences toward their victims is only one possibility. Alternatively, suicide killers may feel altruistic toward their own families, and if their families are financially compensated for the killers’ sacrifices then love, rather than hate, might provide motivation for their choices. Similarly, a belief in heavenly rewards may be the motivating factor. What we claim is that, if these financial or spiritual rewards are correlated with the harm inflicted on others, the observed behavior of suicide killers will be seen to be consistent with a model of hate as in (1). Therefore, our simple model can go a long way to understanding and predicting the behavior of these individuals.

¹See Smith (1759), ch. 1.2.3 and Bentham (1783), ch. 5.27, 6.27. Bentham uses the term “antipathy” for hatred, and “sympathy” for altruism. Smith uses the terms “resentments” and “sympathy.”

We focus, in particular, on three implications of hatred. First, the presence of hatred can impart a strong zero-sum element to arbitrary games, even if only one player hates the other. The way in which a player should react to increased hatred by an opponent depends on the properties of the game’s material payoffs. To illustrate this, we consider a stylized model of asymmetric conflict. For this game, we show that the strong side should adopt a more aggressive strategy when facing a hateful opponent, while the weaker side should adopt a less aggressive strategy. Second, the effectiveness of penalties in deterring certain behaviors can be severely limited when applied to a hating player. We show that the promise of compensating the victims of hatred will sometimes be a stronger deterrent than the threat of punishing the perpetrators of hate. Third, whether hatred is beneficial to a player depends, again, on the material payoffs of the game. In particular, we show that hatred can help the stronger side of an asymmetric conflict, but not the weaker side.

We then examine the meaning of these implications in the context of a historical application and two contemporary applications:

- The first concerns the efficacy of the tactics of non-violent resistance that were employed by the African American Civil Rights Movement of the 1950s and 1960s. We argue that a strategy of non-violence—even in the face of hatred and violence—is consistent with our model’s implications for asymmetric conflict. Non-violent tactics that generate media attention and thereby elicit sympathy from others are especially effective in light of our “rewards as deterrence” argument. Furthermore, Martin Luther King’s call to “love one’s enemy” is consistent with our model’s implications concerning whether or not one should hate.
- The second application concerns shifts in U.S. national security strategy following the terrorist attacks of September 11, 2001. We argue that the United States’ reliance on proactive terrorism prevention, instead of reactive deterrence, can be explained by the difficulties of deterring hateful opponents with the threat of penalties. In particular, the use of preemptive force in the “War on Terror” is in accord with a shift toward a zero-sum view of conflict, which, we argue, can be appropriate in conflicts against hateful adversaries. Language used by the George W. Bush administration to describe the “War on Terror” affirms the zero-sum perspective.
- The third application concerns criminal justice, and specifically the punishment of hate crime. Conventional justifications for penalty enhancements for hate crime are based on the argument that hate crimes cause greater social harm than do equivalent non-hate crimes and are more difficult to avoid. Interpreting “hate crime” literally—a crime committed with the predominant aim of reducing the victim’s well-being—suggests a much more direct justification. To achieve a desired level of deterrence, a hateful individual must be threatened with a more severe penalty than otherwise necessary.

The aim of this paper is not to provide a realistic economic model of the complex,

social-psychological phenomenon of hatred. In real life, hateful emotions are likely to be the product of an individual’s personality, experience, and information (or lack thereof), and the social influence of others. For the purpose of this article, we are content to deal with a simple, “reduced form” model of hatred. Neither is it our goal to make a technical contribution to the theory of interdependent preferences or to the theory of conflict—the mathematical results of this paper are straightforward and require little in the way of formal proofs. Our aim, instead, is to show that even a simple model of hatred can inform our understanding of conflict in new, and sometimes unexpected, ways. By addressing the implications of hatred in three very different but equally important applications, we hope to convince the reader of the usefulness of our approach for thinking about human struggle in general.

The remainder of the paper proceeds as follows. In Section 2 we review the related economic literature on hatred and conflict. In Section 3 we define a general two-player game with hatred, and in Section 4 we derive several general implications within this framework. In Section 5 we discuss these implication in the context of the African American Civil Rights Movement (Section 5.1), U.S. national security strategy (Section 5.2), and the punishment of hate crime (Section 5.3). Section 6 concludes.

2 Literature Review

Starting with Becker (1974), an extensive economic literature has formally modeled altruism as positively interdependent utilities and explored its implications for behavior in various strategic environments. For a review of this literature, see Kolm (2006). In contrast, much fewer papers have analyzed “unpleasant emotions”—such as hate, spite, and envy—as negatively interdependent utilities. We review this literature below.

Possajennikov (2000), Bolle (2000), Konrad (2004), Koçkesen *et al.* (2000a,b), and Heifetz *et al.* (2007) explore the evolutionary stability of these interdependent preferences. For example, Koçkesen *et al.* (2000a,b) analyze games in which players are symmetric in their material payoffs, but a fraction of them also care about their payoff relative to the average payoff of their opponents. For a subset of these games, which includes public goods contribution games, the authors find that the players who care also about their relative payoff do better than those who care only about their material payoff. Heifetz *et al.* (2007) consider a general form of payoff interdependence in two-player games. In their model, a player’s perceived utility is his material utility plus a potentially non-zero “disposition” that depends positively or negatively on the other player’s material utility. They show that there always exists a non-zero disposition that would give a player a higher material payoff in equilibrium. Thus, dispositions will not disappear through evolutionary selection. Their specification of interdependent utilities encompasses ours; however, Heifetz *et al.* (2007) do not specifically focus on the implications of hatred for strategy in conflict, as we do here.

Our paper is also related to work on the political economy of hatred. Glaeser (2005) analyzes conditions under which political leaders foster hatred to increase their probability of reelection. There is an “in-group” and an “out-group,” and the political leader of the in-group can send a message that creates hatred towards the out-group. The leader then faces an election, in which the voting decisions of the in-group members depend on whether they hate the out-group. For example, if the leader is opposed to redistribution and the out-group is poor, the leader has a higher probability of reelection if the in-group members hate the out-group. Fomenting hatred is particularly likely when out-groups are politically relevant but socially segregated. Like Glaeser’s (2005) model, ours has implications for the creation, as well as attenuation, of hatred by group leaders (see Section 4.3). However, our theoretical approach is much different. We model hatred explicitly in terms of preferences, instead of only its effects on behavior. Furthermore, we analyze a conflict between groups (rather than a within-group election contest), and examine the effects of hatred on the hated group’s strategy and the hating group’s payoff.

Baliga and Sjöström (2011) analyze conditions under which extremists can manipulate conflict between two groups. In their model, an extremist within one group can send a public message that conveys information on the group’s cost of conflict. This information can alter the other group’s strategy: When actions are strategic complements, a hawkish extremist can send a provocative message that induces the other group to become more aggressive. This in turn induces the extremist’s own group to become more aggressive. Similarly, with strategic substitutes, a dovish extremist can send a peaceful message that induces the other group to become more aggressive, which in turn makes the extremist’s own group less aggressive. Our model also has implications for the manipulation of conflict. As in Baliga and Sjöström (2011), these depend on the strategic substitutability or complementarity of actions. However, the mechanism through which conflict can be manipulated in our model is very different. Most importantly, it works directly on the level of players’ interdependent preferences, instead of their information.²

Lastly, our paper is related to the law and economics literature on hate crime. Dharmapala and Garoupa (2004) and Gan *et al.* (2010) both argue that a hate crime causes greater social harm than an equivalent non-hate crime, and that, as a result, it should be punished more severely. The explanation derived by Dharmapala and Garoupa (2004) is that a hate crime is an attack on a network of individuals, and thus induces a larger number of costly avoidance activities. Gan *et al.* (2010) argue that hate crime is more difficult to avoid because a person cannot change the characteristic targeted by the crime (e.g., skin color) and the pool of potential targets is smaller. Our model provides a different, and more direct, justification for penalty enhancements for hate crime: If a criminal is motivated by hate, he cares more about reducing his victim’s utility, and thus cares relatively less about his own utility. It therefore requires a more severe punishment to deter this individual, compared to one who is not motivated by hate.

²In reality, of course, hatred is often created by the careful manipulation of information. Our model abstracts from this, and simply examines the effects of changes in a player’s hate parameter.

3 Two-Player Games with Hateful Preferences

In this section we will lay out a general model of hatred in two-player games, and provide an example that demonstrates how the framework can be applied to the study of conflict.

3.1 Definitions

We call the players 1 and 2. The opponent of player $i \in \{1, 2\}$ is denoted by $-i$. Let S be player 1's set of actions, and let T be player 2's set of actions. We assume that S and T are intervals; representative elements from these sets are denoted $s \in S$ and $t \in T$. Together with a payoff function $u : S \times T \rightarrow \mathbb{R}^2$, these sets give rise to a normal form game $G = (S \times T, u)$, called the underlying game. The payoff function u , called the material payoff function, is twice continuously differentiable, and satisfies $\partial u_1 / \partial t < 0$ and $\partial u_2 / \partial s < 0$. The players' actions are hence ordered by their level of aggressiveness, in the sense that increasing one player's action lowers the other player's payoff (*ceteris paribus*).

The following definitions will play an important role in the analysis later. Actions are *strategic substitutes (SS)* or *strategic complements (SC)* for player i in G if

$$\underbrace{\frac{\partial^2 u_i}{\partial s \partial t} < 0}_{\text{SS}}, \quad \underbrace{\frac{\partial^2 u_i}{\partial s \partial t} > 0}_{\text{SC}}.$$

With strategic substitutes, an increase in player $-i$'s action decreases the marginal payoff of player i 's own action. With strategic complements, an increase in player $-i$'s action decreases the marginal payoff of player i 's action.

To introduce hatred, let $\lambda = (\lambda_1, \lambda_2) \in [0, 1]^2$ and define a new payoff function $v : S \times T \rightarrow \mathbb{R}^2$ as follows:

$$v_i(s, t) \equiv (1 - \lambda_i)u_i(s, t) + \lambda_i(-u_{-i}(s, t)). \quad (2)$$

The payoff v_i reflects player i 's preferences if he harbors hateful emotions against player $-i$. The resulting normal form game $G(\lambda) = (S \times T, v)$ is the game with hatred.³

We assume that $G(\lambda)$ has a unique interior Nash equilibrium in pure strategies, characterized by the first-order conditions

$$\frac{\partial v_1}{\partial s} = (1 - \lambda_1) \frac{\partial u_1}{\partial s} - \lambda_1 \frac{\partial u_2}{\partial s} = 0, \quad (3)$$

$$\frac{\partial v_2}{\partial t} = (1 - \lambda_2) \frac{\partial u_2}{\partial t} - \lambda_2 \frac{\partial u_1}{\partial t} = 0. \quad (4)$$

³The underlying game is therefore $G = G(0, 0)$. Note that we could have defined hateful preferences also recursively, i.e. $v_i \equiv (1 - \gamma_i)u_i + \gamma_i(-v_{-i})$ for some $\gamma_i \in [0, 1]$. This is equivalent to the non-recursive formulation in (2) with $\lambda_i = (1 - \gamma_i)/(1 - \gamma_i \gamma_{-i})$.

We further assume that the equilibrium is locally strictly stable. That is, small changes in the equilibrium strategies would result in best reply dynamics converging back to the equilibrium. Stability implies that $\partial^2 v_1 / \partial s^2 < 0$ and $\partial^2 v_2 / \partial t^2 < 0$; moreover, the Hessian

$$H \equiv \begin{bmatrix} \frac{\partial^2 v_1}{\partial s^2} & \frac{\partial^2 v_1}{\partial s \partial t} \\ \frac{\partial^2 v_2}{\partial s \partial t} & \frac{\partial^2 v_2}{\partial t^2} \end{bmatrix}$$

is negative definite at the equilibrium (see Bulow *et al.* 1985), so $|H| > 0$.

We will be especially concerned with how hatred alters the strategic choices of the players. Differentiating (3)–(4) implicitly with respect to λ , we obtain

$$\begin{bmatrix} \frac{ds}{d\lambda_1} & \frac{ds}{d\lambda_2} \\ \frac{dt}{d\lambda_1} & \frac{dt}{d\lambda_2} \end{bmatrix} = \frac{1}{|H|} \begin{bmatrix} \frac{\partial u_2}{\partial s} \frac{\partial^2 v_2}{\partial t^2} & -\frac{\partial u_1}{\partial t} \frac{\partial^2 v_1}{\partial s \partial t} \\ -\frac{\partial u_2}{\partial s} \frac{\partial^2 v_2}{\partial s \partial t} & \frac{\partial u_1}{\partial t} \frac{\partial^2 v_1}{\partial s^2} \end{bmatrix}. \quad (5)$$

The assumptions that $\partial u_1 / \partial t < 0$ and $\partial u_2 / \partial s < 0$ imply a positive sign for $ds/d\lambda_1$ and $dt/d\lambda_2$. In other words, the players increase their actions (i.e., they become more aggressive) as their own hate parameters increase.

3.2 An example: Asymmetric conflict

Hatred plays an important role in conflicts. We now provide a simple model of conflict to be used as our underlying game. While stylized, the model has the important property that parties in conflict are asymmetric in their relative strengths—a notion we will frequently refer to in the remainder of the paper.

To model conflict, we employ a contest game in which the players compete over a fixed and indivisible prize, such as winning a war, by investing costly non-recoverable efforts. We will focus on a particular functional form for simplicity, the **Tullock contest** (Tullock 1980).⁴ Denoting by s and t the efforts invested by the players, the probability that player 1 wins the contest is

$$f(s, t) = \frac{s}{s + t},$$

and the probability that player 2 wins is $1 - f(s, t) = t/(s + t) = f(t, s)$. Thus, the probability of success for each player is proportional to the players' efforts.⁵ Assume now that effort has a per unit cost of 1 for player 1, and $k > 1$ for player 2. We say

⁴The Tullock contest is one of several “workhorse models” used for the game-theoretic study of conflict. An excellent overview of this and other contest models is given in Konrad (2009). For an application of the Tullock contest to conflict resolution, see Garfinkel and Skaperdas (2000).

⁵Alternatively, these probabilities might be viewed as the fractions of a limited but divisible resource that the players obtain in the conflict; e.g., the percentage of territory that the players control after the conflict.

that player 1 is the stronger party and player 2 is the weaker party. Normalizing the value of winning to one, the expected payoffs in the contest are then given by

$$u_1(s, t) = f(s, t) - s, \quad u_2(s, t) = f(t, s) - kt. \quad (6)$$

Notice that (6) satisfies our assumption that an increase in one player's effort decreases the opponent's payoff. Observe also that

$$\frac{\partial^2 u_1(s, t)}{\partial s \partial t} = \frac{s - t}{(s + t)^3}, \quad \frac{\partial^2 u_2(s, t)}{\partial s \partial t} = \frac{t - s}{(s + t)^3}.$$

Thus, at any effort profile (s, t) with $s \neq t$, the player who spends the larger amount of effort regards efforts as strategic complements, while the player who spends the smaller amount regards efforts as strategic substitutes.

When we introduce hatred to the Tullock contest, we generate a new game with the following emotional payoffs:

$$\begin{aligned} v_1(s, t) &= (1 - \lambda_1)[f(s, t) - s] - \lambda_1[f(t, s) - kt], \\ v_2(s, t) &= (1 - \lambda_2)[f(t, s) - kt] - \lambda_2[f(s, t) - s]. \end{aligned}$$

The Nash equilibrium of this game can be shown to be

$$s^*(\lambda) = \frac{k(1 - \lambda_2)}{((1 - \lambda_1) + k(1 - \lambda_2))^2}, \quad t^*(\lambda) = \frac{1 - \lambda_1}{((1 - \lambda_1) + k(1 - \lambda_2))^2}. \quad (7)$$

As expected, each player's effort increases in this player's own hate parameter. Furthermore, the stronger player invests more effort than the weaker player if and only if

$$k > \Lambda \equiv \frac{1 - \lambda_1}{1 - \lambda_2}. \quad (8)$$

Thus, as long as $k > \Lambda$, efforts are strategic complements for the stronger player and strategic substitutes for the weaker player. In particular, (8) will be satisfied if the weaker party does not hate. However, if λ_2 is large enough so that $k < \Lambda$, player 1's cost advantage will be outweighed by player 2's "hate advantage," in which case player 2 outspends player 1.

4 Implications

In the previous section we introduced a game $G(\lambda)$, obtained from an underlying two-player game G by introducing a preference for low opponent payoffs into each player's objective. In the present section, we develop a number of themes and implications which arise within this framework. We specifically will argue the following points:

First, the presence of hatred can eliminate any common interest among the players, and this can be true even when only one player hates the other. In general, the way a player should react to opponent hatred depends on whether actions are strategic substitutes or complements, and in our conflict example this will depend on a player's

relative strength. Second, the effectiveness of penalties in deterring certain actions can be severely limited when applied to a hateful player. Third, whether hate can help a player achieve higher materials payoffs again is determined by whether actions are strategic substitutes or complements. In our conflict example, this depends on the player's relative strength.

4.1 Strategic responses to hatred

As we have shown, hate causes a player to become more aggressive than he would otherwise be. By the same token, facing a hateful opponent may cause a player to adopt different strategies than he would have otherwise chosen.

Consider for a moment the special case where $\lambda_1 + \lambda_2 = 1$. In this case, $G(\lambda)$ becomes a two-person, zero-sum game:

$$\begin{aligned} v_1(s, t) + v_2(s, t) &= [(1-\lambda_1)u_1(s, t) - \lambda_1 u_2(s, t)] + [(1-\lambda_2)u_2(s, t) - \lambda_2 u_1(s, t)] \\ &= (1 - [\lambda_1 + \lambda_2])u_1(s, t) + (1 - [\lambda_2 + \lambda_1])u_2(s, t) \\ &= 0. \end{aligned}$$

In zero-sum games, the players' interests are diametrically opposed and equilibrium reasoning must result in a pair of *max-min strategies*: The players expect the worst possible outcome from each strategy, and select the strategy with the best such worst-case outcome (von Neumann and Morgenstern 1944). Now observe that, in particular, the games $G(0, 1)$ and $G(1, 0)$ are zero-sum games. This is remarkable in so far as it takes exactly one completely hateful player to turn any two-player game into one in which both players' interests are opposed, and this will be true despite the fact that the players may have some common interest in the underlying game. For example, if $\lambda_2 = 1$ while $\lambda_1 = 0$, then player 1 maximizes $v_1 = u_1$ and player 2 maximizes $v_2 = -u_1$. In this case, player 1 must treat the situation as one in which he must adopt a max-min strategy,

$$s^* = \arg \max_{s \in S} \min_{t \in T} u_1(s, t).$$

His strategic approach to the situation will therefore be precisely the same as that of his hateful opponent, who chooses $t^* = \arg \max_t \min_s -u_1(s, t)$.

We can say more for the general, variable-sum case. From (5), observe that $ds/d\lambda_2$ has the same sign as

$$\frac{\partial^2 v_1}{\partial s \partial t} = (1-\lambda_1) \frac{\partial^2 u_1}{\partial s \partial t} - \lambda_1 \frac{\partial^2 u_2}{\partial s \partial t}.$$

Thus, if actions are strategic complements for player 1 in the underlying game, and strategic substitutes for player 2 in G , then player 1 must increase s as player 2 becomes more hateful. Similarly, if actions are strategic substitutes for 1 and complements for 2, then player 1 must decrease s . A similar statement can be made for player 2, of course.

Thus, we have identified two ways in which a player might respond to an opponent's hatred. The first is what we call the *stooping-down effect*: In response

to an increase in the opponent's hatred a player becomes more aggressive himself. The second is what we call the *turn-the-other-cheek effect*: In response to an increase in the opponent's hatred, a player becomes less aggressive himself. Both the stooping-down response and a turn-the-other-cheek response can arise in equilibrium, depending on whether actions are viewed as strategic substitutes or complements.⁶

In Section 3.2 we demonstrated that, in a simple model of conflict, strategic substitutability and complementarity were tied to the relative strength of players. Provided the condition (8) is satisfied, the weaker party in a conflict will “turn the other cheek” in response to increased hatred by the stronger party. On the other hand, the stronger party will “stoop down” in response to increased hatred by the weaker party.

4.2 The (im)possibility of deterrence

Society may have an interest in regulating the players' actions. Suppose the goal is to limit the activities represented by s and t to \bar{s} and \bar{t} , respectively. The conventional way to enforce such limits is to impose costs, say $c_1(s) > 0$ and $c_2(t) > 0$, on actions $s > \bar{s}$ and $t > \bar{t}$. With these penalties in place, player 1's material payoff becomes $\hat{u}_1(s, t) = u_1(s, t) - c_1(s)$ and player 2's material payoff becomes $\hat{u}_2(s, t) = u_2(s, t) - c_2(t)$. The players' emotional payoffs become

$$\hat{v}_1(s, t) = (1 - \lambda_1) [u_1(s, t) - c_1(s)] - \lambda_1 [u_2(s, t) - c_2(t)]$$

and

$$\hat{v}_2(s, t) = (1 - \lambda_2) [u_2(s, t) - c_2(t)] - \lambda_2 [u_1(s, t) - c_1(s)].$$

Observe that the higher λ_i is, the less weight player i places on the penalty applied to his own actions. Thus, the deterrence effect of a given penalty is diminished in the presence of hatred. It is therefore questionable whether conventional punishments work in the presence of hatred.

If player 1 chooses his maximal permitted action \bar{t} , then for player 1's action not to exceed the limit \bar{s} we need $\hat{v}_1(s, \bar{t}) \leq \hat{v}_1(\bar{s}, \bar{t})$, or

$$\begin{aligned} (1 - \lambda_1) [u_1(s, \bar{t}) - c_1(s)] - \lambda_1 [u_2(s, \bar{t}) - c_2(\bar{t})] \\ \leq (1 - \lambda_1) [u_1(\bar{s}, \bar{t}) - c_1(\bar{s})] - \lambda_1 [u_2(\bar{s}, \bar{t}) - c_2(\bar{t})]. \end{aligned}$$

Setting $c_1(\bar{s}) = 0$, the penalty that deters player 1 from taking action $s > \bar{s}$ satisfies the condition

$$c_1(s) \geq [u_1(s, \bar{t}) - u_1(\bar{s}, \bar{t})] + \frac{\lambda_1}{1 - \lambda_1} [u_2(\bar{s}, \bar{t}) - u_2(s, \bar{t})]. \quad (9)$$

(A similar expression can be derived for $c_2(t)$.) The term $[u_1(s, \bar{t}) - u_1(\bar{s}, \bar{t})]$ in (9) is the private gain of player 1 from playing s instead of \bar{s} . The term $[u_2(\bar{s}, \bar{t}) - u_2(s, \bar{t})]$

⁶These effects are straightforward and appear in different form elsewhere in the literature. For example, our “turn the other cheek” effect is reminiscent of Fudenberg and Tirole's (1984) “puppy dog ploy.” What is new here is our application to games with hateful players.

is the harm inflicted on player 2 by player 1's choice of s instead of \bar{s} . This term must be included in the calculation of the penalty applied to player 1, because player 1 is (in part) motivated by the desire to inflict harm on others. If λ_1 increases, the penalty required to deter s increases as well and grows to infinity as $\lambda_1 \rightarrow 1$. This is so because a player motivated by pure hatred puts a zero weight on his own payoffs, and thus cannot be deterred by *any* penalty applied to his actions.

An alternative deterrence mechanism is to reward player i 's opponent instead. Specifically, assume that, instead of reducing 1's material payoff by $c_1(s)$ if action $s > \bar{s}$ is taken, player 2's payoff is increased by $b_2(s)$. Similarly, player 1's payoff is increased by $b_1(t)$ whenever $t > \bar{t}$. Player 1's emotional payoff then becomes

$$\hat{v}_1(s, t) = (1 - \lambda_1) [u_1(s, t) + b_1(t)] - \lambda_1 [u_2(s, t) + b_2(s)].$$

Setting $b_2(\bar{s}) = 0$, requirement $\hat{v}_1(s, \bar{t}) \leq \hat{v}_1(\bar{s}, \bar{t})$ now boils down to

$$b_2(s) \geq \frac{1 - \lambda_1}{\lambda_1} [u_1(s, \bar{t}) - u_1(\bar{s}, \bar{t})] + [u_2(\bar{s}, \bar{t}) - u_2(s, \bar{t})]. \quad (10)$$

The minimum reward (10) is again a weighted sum of 1's gain and 2's loss resulting from 1's choice of s . As player 1's hatred increases, the weight on private gains approaches zero, as a player motivated by pure hatred does not care about his own payoffs. On the other hand, the weight attached to the social harm component stays constant at one. That is, in the limit as $\lambda_1 \rightarrow 1$, it is sufficient to simply compensate player 2 for his loss resulting from 1's actions.

4.3 Fomenting and dissuading hatred

While our basic model does not explain where hatred comes from, it still allows us to examine who benefits from hatred and who does not. Specifically, we now ask the question: In equilibrium of the game $G(\lambda)$, how are the players' material payoffs affected as their hate parameters change?

To examine the effect of player 1's hate parameter on 1's payoff, differentiate the emotional payoff v_1 with respect to λ_1 , and rearrange, to get

$$\begin{aligned} \frac{du_1}{d\lambda_1} &= \frac{1}{1 - \lambda_1} \left[\frac{dv_1}{d\lambda_1} + \lambda_1 \frac{du_2}{d\lambda_1} \right] \\ &= \frac{1}{1 - \lambda_1} \left[\left(\frac{\partial v_1}{\partial s} \frac{ds}{d\lambda_1} + \frac{\partial v_1}{\partial t} \frac{dt}{d\lambda_1} \right) + \lambda_1 \left(\frac{\partial u_2}{\partial s} \frac{ds}{d\lambda_1} + \frac{\partial u_2}{\partial t} \frac{dt}{d\lambda_1} \right) \right] \\ &= \frac{\partial u_1}{\partial t} \frac{dt}{d\lambda_1} + \frac{\lambda_1}{1 - \lambda_1} \frac{\partial u_2}{\partial s} \frac{ds}{d\lambda_1}. \end{aligned} \quad (11)$$

(We used the fact that $\partial v_1 / \partial s = 0$ in equilibrium of $G(\lambda)$.) Recall that $\partial u_1 / \partial t < 0$, $\partial u_2 / \partial s < 0$, and $ds / d\lambda_1 > 0$. Thus, the second summand in (11) is negative.

If actions are strategic complements for player 2 and strategic substitutes for player 1, then player 2 "stoops down" ($dt / d\lambda_1 > 0$), as shown earlier. In this case, we have $du_1 / d\lambda_1 < 0$, so player 1's hatred unambiguously reduces his material payoff. In the

context of our conflict model, this implies that a weaker player cannot benefit from increasing his hatred toward his opponent (provided condition (8) is satisfied). In particular, a weaker player who does not hate the stronger player will see his material payoff decrease as he starts to hate.

If, on the other hand, actions are strategic substitutes for player 2 and strategic complements for 1, then player 2 “turns the other cheek” ($dt/d\lambda_1 < 0$). In this case, if $\lambda_1 = 0$ the second term on the right-hand side of (11) vanishes, and the first term will be positive. An increase of λ_1 from zero to a small positive value will therefore result in a higher material utility for player 1. At least a small amount of hatred is hence desirable from the perspective of player 1: If the opponent regards actions as strategic substitutes, he responds to hatred with a less aggressive strategy. In the Tullock contest of Section 3.2, this must be the case for the weaker player, provided this player does not hate himself. An already strong contestant can therefore benefit from developing at least a small amount of hatred toward his weaker opponent.⁷

To the extent that hate is a malleable emotion that can be influenced by others, these results suggest that military or political leaders in conflicts may attempt to manipulate their followers’ preferences strategically. Note that leaders who simply want to induce their followers to pursue more aggressive actions can do so by fomenting hatred toward their adversaries, for example through the use of propaganda and indoctrination.⁸ Whether this is in the ultimate interest of the manipulated depends on their relative strength. For the stronger side in a conflict, “a little bit of hate” may help their cause. On the other hand, responsible leaders of the weaker side want to do the opposite, and dissuade hatred in their followers.

5 Applications

The themes developed in the previous section can illuminate a number of policy issues, historical and contemporary, from novel, and perhaps surprising, angles. In the following, we discuss the implications of our theory of hatred in three different contexts: The African American Civil Rights Movement, U.S. national security strategy following 9/11, and the legal debate surrounding the punishment of hate crime.

⁷Note that hatred makes a player adopt a more aggressive strategy than is optimal. However, since the player was optimizing in equilibrium, a small increase in aggressiveness has only a second-order effect on that player’s payoff. On the other hand, it typically has a first-order effect on the opponent’s strategy, and thus on payoffs, and it is this strategic effect which can help the hating player. Recent results on the evolution of other-regarding preferences are based on similar observations; see, for example, Heifetz *et al.* (2007) and related work discussed in Section 2.

⁸The famous shock experiments by Milgram (1965) and the Stanford Prison experiment by Zimbardo *et al.* (1974) suggest that even ordinary people may obey directives to commit hateful acts if the directives come from an authority figure. Hatred by authority figures may then be sufficient for hateful acts to be carried out on a large scale (Harrington 2004).

5.1 Non-violence and the African American Civil Rights Movement

In a famous act of defiance, African American seamstress Rosa Parks refused to give up her seat to a white passenger in a segregated bus in Montgomery, Alabama, on December 1, 1955. Park was subsequently arrested, convicted of disorderly conduct, and ordered to pay \$14 in fines and court fees. The day of Park’s trial marked the beginning of what would turn into a 381 day boycott of Montgomery’s public transit system by the city’s black residents.⁹

Park’s refusal to give up her seat, her arrest, and the ensuing bus boycott, are examples of a strategy of non-violent resistance, which was pursued by members of the African American Civil Rights Movement of the 1950s and 1960s. In his Nobel Prize lecture, Martin Luther King Jr. (1964) describes the strategy of non-violence as follows:

“Broadly speaking, nonviolence in the civil rights struggle has meant not relying on arms and weapons of struggle. Nonviolence has meant noncooperation with customs and laws which are institutional aspects of a regime of discrimination and enslavement. [...] Nonviolence has also meant that my people in the agonizing struggles of recent years have taken suffering upon themselves instead of inflicting it on others. It has meant [...] that we are no longer afraid and cowed. But in some substantial degree it has meant that we do not want to instill fear in others or into the society of which we are a part. The movement does not seek to liberate Negroes at the expense of the humiliation and enslavement of whites. It seeks to liberate American society and to share in the self-liberation of all the people.”

Non-violent acts in resisting the laws and customs of apartheid—examples of “no longer being afraid and cowed”—generated strong resentments among white southerners. These resentments sometimes resulted in outright violence against blacks, including the bombing of churches and residences. Should blacks have reacted to such attacks with violence of their own? As whites held a clear advantage in power, our turn-the-other-cheek result actually implies the opposite. King (1964) offers the following explanation for preferring non-violent tactics in fighting racial injustices:

“I am only too well aware of [...] the doubts about the efficacy of nonviolence, and the open advocacy of violence by some. But I am still convinced that nonviolence is both the most practically sound and morally excellent way to grapple with the age-old problem of racial injustice. [...] Violence is impractical because it is a descending spiral ending in destruction for all. It is immoral because it seeks to humiliate the opponent rather than win his understanding: it seeks to annihilate rather than convert. Violence is immoral because it thrives on hatred rather than love. Violence ends up defeating itself. It creates bitterness in the survivors and brutality in the destroyers.”

⁹The events are described in detail in King (1958).

It is not difficult to translate King’s arguments into the logic of strategic substitutes and complements. If the oppressed minority took to violence, it would provoke further violence by the oppressive majority, which views actions as strategic complements. Such a reaction might then initiate a downward cycle of violence that would only result in further oppression of the minority. Thus, as King concluded, non-violence may indeed be the only practical strategy for an oppressed minority.

This strategic argument for non-violence does not necessarily require an assumption of hatred, on the part of either the majority or the minority group. Interestingly, however, our theory of hatred does shed light on some of the specific tactics of non-violent resistance and civil disobedience that were employed and advocated by the Civil Rights Movement. In describing his non-violence strategy, King (1964) emphasizes the importance of mass participation in peaceful protest, instead of actions behind the scenes by fragmented groups:

“[Nonviolence] has meant direct participation of masses in protest, rather than reliance on indirect methods which frequently do not involve masses in action at all.”

An important characteristic of peaceful, large-scale, and coordinated protests is that they tend to attract media attention. This may directly benefit an oppressed minority by arousing a collective conscience. Moreover, violent escalation by the oppressive side only increases the attention that is given to the oppressed side’s cause and thereby can generate further sympathy from outsiders. Outside sympathy can function as a “reward” that is bestowed on the oppressed side in response to suffering inflicted by their oppressors. Our arguments of Section 4.2 suggest that such rewards can provide a disincentive for the resentful side to further escalate violence.¹⁰

Martin Luther King Jr. also went beyond simply calling for non-violent and civil disobedience, and actively attempted to attenuate hatred by his followers. For example, on the evening of December 5, 1955, King announced that the Montgomery bus boycott would follow the slogan “Thou shall not requite violence with violence.” King further invoked a biblical passage, Matthew 5:44, to discourage hatred against whites despite the outrage of Montgomery’s black community because of Park’s arrest and a number of retaliatory acts by whites against blacks:

“Our method will be that of persuasion not coercion. We will only say to the people, ‘Let your conscience be your guide.’ Our actions must be

¹⁰In *The Theory of Moral Sentiments*, Adam Smith puts forward the idea that sympathy toward the oppressed side is most likely forthcoming if it refrains from any angry displays of its own (Smith 1759, ch. 1.2.3). The African American Civil Rights Movement was extremely successful in securing outside sympathy by remaining peaceful in the face of violence. For example, in May of 1961 a group of black and white “Freedom Riders” travelled by bus from Washington D.C. to the deep south in support of racial equality. The brutal response by local police and mobs in several southern cities attracted the attention of the national media. The ensuing nationwide outrage at the events compelled John F. Kennedy’s administration to negotiate the safety of the Freedom Riders, and ultimately to support the struggle for civil rights in the South. (A detailed account of the Freedom Riders’ journey through the American south is given in Arsenault 2007.)

guided by the deepest principles of our Christian faith. [...] Once again we must hear the words of Jesus echoing across the centuries: ‘Love your enemies, bless them that curse you, and pray for them that spitefully use you.’ (Jahn 1964)

In his call for love instead of hate, King was heavily influenced by Mohandas K. Gandhi, who previously had employed a similar strategy to fight the exploitation of his people by the British. In Section 4.3, we showed that hatred reduces the payoffs of the weaker side in an asymmetric conflict. Therefore, as leaders of the arguably weaker sides in their respective struggles, Gandhi and King chose to attenuate their followers’ hate and even advocated love for their oppressors.¹¹

5.2 U.S. National security strategy following 9/11

On September 17, 2002—almost one year exactly following the terrorist attacks of 9/11—George W. Bush’s administration announced its new National Security Strategy (2002 NSS hereafter). The 2002 NSS constituted a significant departure from the hitherto stated approach to U.S. security, which was to a large part based on the Cold War strategy of deterrence. Before we begin to discuss the strategy in the light of our theoretical results, let us review some key passages of the 2002 NSS that are contained in Section V of the document (National Security Council 2002).

First, it is argued that the United States’ terrorist adversaries are motivated by hate toward the United States:

“These [rogue] states [...] reject basic human values and hate the United States and everything for which it stands. [...] As was demonstrated by the losses on September 11, 2001, mass civilian casualties is the specific objective of terrorists.”

One of the more controversial aspects of the 2002 NSS is the explicit authorization of preemptive military force. Section V makes a case for a wide range of proactive efforts to prevent attacks before they happen, including counter-proliferation efforts, the use of diplomacy, and the formation of strategic alliances, but also the use of preemptive force:

“We must be prepared to stop rogue states and their terrorist clients before they are able to threaten or use weapons of mass destruction against the United States and our allies and friends. [...] We must deter and defend against the threat before it is unleashed.”

Finally, at several points an explicit link is proposed between the goals of terrorist adversaries, their choices, and the prescribed proactive strategy:

¹¹Ghandi, however, advocated the strategy of “loving the enemy” not only in India’s struggle against oppression by the British, but also later when Nazi Germany appeared poised and ready to invade Britain (Gandhi 1972). This might have been a mistake, since the Allies were perhaps not the weaker side in World War II.

“Given the goals of rogue states and terrorists, the United States can no longer solely rely on a reactive posture as we have in the past. The inability to deter a potential attacker, the immediacy of today’s threats, and the magnitude of potential harm that could be caused by our adversaries’ choice of weapons, do not permit that option. We cannot let our enemies strike first. [...] In the Cold War, weapons of mass destruction were considered weapons of last resort whose use risked the destruction of those who used them. Today, our enemies see weapons of mass destruction as weapons of choice. [...] Traditional concepts of deterrence will not work against a terrorist enemy whose avowed tactics are wanton destruction and the targeting of innocents; whose so-called soldiers seek martyrdom in death and whose most potent protection is statelessness. [...] We must adapt the concept of imminent threat to the capabilities and objectives of today’s adversaries.”

In 2006, a revised version of the document was released. Its language was somewhat less aggressive, shifting emphasis from preemption to cooperation with allies (National Security Council 2006). Presumably these changes were made in reaction to the U.S. military’s failure to find weapons of mass destruction in Iraq (a state against which the new preemptive doctrine was used). In practice, however, the United States’ preemptive approach to the “War on Terror” seems to have changed merely in scale and scope. Preemptive efforts have shifted notably during Barack Obama’s presidency from a small number of large-scale operations against entire regimes, to a larger number of small-scale operations against individual militants.¹²

We therefore argue that 9/11 brought about a major paradigm shift in U.S. national security strategy, and that any subsequent changes in strategy were of the second order. We will now interpret this paradigm shift in light of our formal arguments made earlier, by elaborating on several of the themes that we developed in the previous sections.

Recall that actors who are motivated by hatred cannot be easily deterred from taking certain harmful actions by imposing action-dependent penalties. Interestingly, the 2002 NSS explicitly refers to the difficulties in relying on a “reactive posture” and implementing “traditional concepts of deterrence” against an enemy who is motivated by hatred. (Our theoretical results suggest that, perhaps, a better deterrent would be to reward oneself in case of an enemy attack. However, in the context of national security, a strategy of “living well is the best revenge” hardly seems to be a feasible option.) Consequently, and consistent with the United States’ own stated strategy, the focus of U.S. national security strategy following 9/11 shifted from a reactive approach that was based on the logic of deterrence to a proactive approach of terrorism prevention.

¹²Most of these operations appear to be assassination attempts that used unmanned aircraft. For example, on October 10, 2011, PBS NewsHour reported that “Under Obama’s watch [...] the number of drone strikes has increased exponentially from 13 during years 2004–2007 to 122 in 2010.” (www.pbs.org/newshour/run-down/2011/10/drone-strikes-1.html.)

One such preventive measure is the use of preemptive force: Instead of threatening to retaliate against certain enemy actions, the forceful removal of these options from the enemy’s strategy set became a choice that the U.S. was prepared to make in the “War on Terror.” The logic of preemption is most forceful when one’s adversary is motivated by pure hatred. In this case, strategic interaction becomes zero-sum. The key observation is that, if a game is zero-sum for one player, it is also zero-sum for the other—even if this second player is not motivated by hate. Thus, if hatred of “the United States and everything for which it stands” is in fact what motivates its adversaries, one must acknowledge max-minimization as a logical approach to U.S. national security. In practical terms, playing a max-min strategy entails making decisions based on the opponent’s capabilities, without guessing their intentions (Luce and Raiffa 1957, pp. 64–65). The use of preemptive force to eliminate enemy capabilities before they are used is consistent with this approach.

In this regard, it is interesting to note that for a very long time, the United States’ military decision doctrine was in fact a capabilities-oriented doctrine (Haywood 1954, pp. 365–366). Such a perspective is problematic, however, once conflict is a variable-sum game. Here, the Cold War stands out as a conflict with immensely variable-sum payoffs, as both the U.S. and the Soviet Union shared the common goal of avoiding nuclear war. Thus, the fact that large nuclear capabilities had been amassed on both sides was a less important determinant of strategy than whether these capabilities would be used. Under these circumstances, deterrence became the central approach to U.S. national security strategy (Schelling 1960). Because the presence of hatred has the potential to turn variable-sum conflicts into zero-sum conflicts, the “War on Terror” represents, in important aspects, a reversal of the nature of conflict to a pre-Cold War state, and a number of strategic insights can be gained from pre-Cold War game theoretic reasoning. The 2002 NSS, in returning to a more capabilities-oriented doctrine, reflects this view.

In addition to its willingness to employ preemptive military force, the United States government engaged in a number of aggressive anti-terror practices in its response to 9/11. These include the use of warrantless domestic wiretapping, indefinite detention, so-called “enhanced interrogation techniques” such as waterboarding, and rendition of terror suspects to other countries.¹³ The adoption of such practices can be regarded as the U.S. “stooping down” to the level of its enemies, and has been characterized as such by domestic and international media commentators. Accepting the premise of a hateful adversary, however, our theoretical arguments suggest that a sufficiently strong player should indeed react by adopting a more aggressive stance. To the extent that preemptive wars are costly, and aggressive anti-terror measures tarnish American reputation abroad, a hateful adversary perversely *benefits* from these actions. However, if the same measures are indeed equilibrium strategies, it is clear that one cannot do better by adopting a less aggressive approach. The troubling observation is that equilibrium in a game against a hateful opponent is bound to be

¹³For a summary of the U.S. government’s drive to allow the use of enhanced interrogation techniques in counterterrorism following 9/11 and an analysis of the potential effects of legalizing torture in counterterrorism on national security and welfare, see Mialon *et al.* (2012).

an unpleasant outcome.

5.3 The punishment of hate crime

On October 7, 1989, in Kenosha, Wisconsin, Gregory Reddick, a white, fourteen-year-old boy, while walking home from a pizza parlor, was beaten into a coma by a group of African American individuals. The group was led by Todd Mitchell who had incited the beating after the group had watched the movie *Mississippi Burning*, in which a scene depicts a white man beating a young black boy. On June 7, 1998, in Jasper, Texas, James Byrd, an African American man, was beaten, tied to a pickup truck, dragged nearly three miles, and decapitated by three members of a white supremacist group. On October 6, 1998, in Laramie, Wyoming, Matthew Shephard, a student who was perceived to be homosexual, was pistol-whipped, tortured, tied to a fence, and left to die by Russell Henderson and Aaron McKinney. On September 15, 2001, in Dallas, Texas, Waqar Hasan, a Pakistani Muslim, was shot in the face by Mark Stroman, who later confessed on a radio show that he had killed Hasan as a revenge for the terrorist attacks.

Between 1995 and 2009, 147,099 individuals were victims of reported hate crimes in the U.S. (Federal Bureau of Investigation 1995–2009), including the high-profile cases of Gregory Reddick, James Byrd, Matthew Shepard, and Waqar Hasan. Our theoretical results have implications for the punishment of such crimes. Before discussing these implications, we will review the prevailing definitions of hate crime, as well as existing justifications for enhanced punishments for hate crime.

Most states have enacted laws that specifically concern hate crime.¹⁴ Several of these laws define hate crimes simply as crimes involving *discriminatory selection* of victims, regardless of the reason for the selection. Other statutes define hate crime as crimes in which the reason for discriminatory selection is *racial animus*. For example, Massachusetts’s hate crime statute states:

“Hate Crime [is] any criminal act coupled with overt actions motivated by bigotry and bias [...]” (Mass. Gen. Laws ch. 22c, §32, 1997)

Furthermore, several statutes provide penalty enhancements for hate crime under the racial animus definition. For example, Florida’s statute stipulates that:

“The penalty for any felony or misdemeanor shall be [enhanced] if the commission of such felony or misdemeanor evidences prejudice based on

¹⁴For a classification and analysis of these laws, see Lawrence (2002). A few states, including Arizona and Georgia, currently have no hate crime laws, and several states have only recently enacted hate crime statutes. Texas, for example, had no hate crime law at the time of the James Byrd case and few prosecutions have occurred under the hate crime statute that Texas enacted in 2001. Wyoming’s hate crime law did not include crimes motivated by a victim’s sexual orientation at the time of the Matthew Shepard case. In 2009, Congress passed the Matthew Shepard and James Byrd, Jr. Hate Crimes Prevention Act (Public Law No. 111–84, §§4701–4713, 123 Stat. 2835, 2009), which expanded federal hate crime law to include crimes that are motivated by sexual orientation and provides additional funding for state agencies to enforce hate crime laws.

the race, color, ancestry, ethnicity, religion, sexual orientation, or national origin of the victim.” (Fla. Stat. Ann. §775.085, 1995)

Enactment and enforcement of hate crime laws in the U.S. have been met by a great deal of political and legal opposition. Two main arguments against penalty enhancements in particular have been proposed. The first argument is that penalty enhancements punish defendants for their beliefs or thoughts, in violation of the free speech principles of the First Amendment (see Dillof 1997; Jacobs and Potter 1998; Harel and Parchomovsky 1999; Nearpass 2003). In *Wisconsin v. Mitchell*, the Supreme Court ruled that a defendant’s beliefs, no matter how reprehensible, cannot in and of themselves be the grounds for an enhanced sentence (508 U.S. 476, 1993). However, the same ruling also states that sentence enhancement can be justified on the separate grounds that a hate crime produces greater individual or societal harm than a parallel non-hate crime. For example, a hate crime may cause greater emotional distress to the victim, or may provoke a more violent retaliatory response, if the social network to which the victim belongs feels threatened by the crime. The second argument against penalty enhancement for hate crime challenges the “greater harms” argument as being empirically unsupported (see, e.g., Jacobs and Potter 1998, pp. 81–88; Harel and Parchomovsky 1999, pp. 514–515). Moreover, Kahan (2001) has argued that, even if hate crimes produce greater harms, these additional harms are emotional and may simply be the product of the hateful emotions that motivated the hate crimes.¹⁵

Using the theoretical framework developed in this paper, we can provide a different and presently overlooked justification for hate crime penalty enhancements. It relies not on a hate crime producing a greater level of emotional harm, but on it being more difficult to deter. Interpreting the actors in our model as a potential criminal who spends effort to commit a crime, and a potential victim (who potentially spends effort to avoid being victimized), our arguments in Section 4.2 imply that any given penalty is less likely to deter a crime if the crime is motivated by hate than if hate is not the motivation. Thus, applying the same penalty to both a hate crime and an equivalent non-hate crime would result in less deterrence of the hate crime. A sentencing scheme can only achieve the same level of deterrence for a hate crime if it involves a greater penalty for the hate crime than for an (equivalent) non-hate crime.¹⁶

We note that this argument requires courts to be able to detect whether a defendant’s actions were motivated by hatred toward his or her victim—that is, that the perpetrator acted out of (racial) animus. To some extent, at least, this condition is satisfied. In many instances, the role of animus in the crime can be proven di-

¹⁵In other words, victims may feel greater emotional harm precisely because of their aversion to their attackers’ animus. Similarly, hate crimes may provoke a greater retaliatory response because the groups that feel threatened by these crimes judge the motives behind them to be more reprehensible.

¹⁶Curry and Klumpp (2009) show that, in order to achieve a constant degree of deterrence across individuals, penalties must depend on individual characteristics such as wealth or income. Some European countries, for example, penalize non-violent offenses with day fines, which are penalty enhancements for richer defendants. Penalty enhancements for hate crime are similar, except that the penalty is a function of hate instead of income.

rectly. For example, in the case of James Byrd, the three attackers were known white supremacists, and in the case of Waqar Hasan, the attacker publicly admitted that his motive was animus toward people of Arabic descent. More generally, the role of animus in a crime can be assessed indirectly by asking whether the crime would have been committed *but for* the victim’s race or sexual orientation. For example, if the victim had little or no acquaintance with the attacker, and the attacker only beat the victim but did not rob him, then these facts suggest animus as the principal motive.

We emphasize that our justification for enhanced hate crime penalties holds even if the individual and social harms of a hate crime are the same as those of an equivalent non-hate crime. Our argument is based on the simple idea that, in order to provide a desired degree of deterrence, more severe punishments are necessary when potential perpetrators are motivated by hatred of their victims. Interestingly, this argument suggests that constitutional support for penalty enhancements perhaps could be derived from the Fourteenth Amendment: To the extent that penalty enhancements do indeed have the potential to equalize the chances of being the victim of a hate crime versus an (equivalent) non-hate crime, the equal protection clause seems to not only permit but actually require such enhancements.

6 Conclusion

We have developed a simple model of the effects of hate in two-player games with the aim of acquiring an understanding of human conflict and policy issues that are related to it. The model implies that one player’s hatred toward another has a strategic effect on the hated. If actions are strategic complements, the hated player will “stoop down,” whereas if actions are strategic substitutes, the hated player will “turn the other cheek.” We showed that it only takes one completely hateful player to turn a game into one in which both players’ interests are opposed. Furthermore, we found that a greater penalty is required to deter a player from acting aggressively if the player is motivated by hatred. In application of the model to asymmetric conflict, the weaker side was shown to be better off if it harbors no hate than if it harbors a small amount of hate. This implies that a responsible leader of the weaker side would benefit from attenuating hate. On the other hand, a leader of any side who simply wants to induce his side to become more aggressive can do so by fomenting hate.

We argued that the “turn the other cheek” effect and the incentive for a responsible leader of the weaker side to attenuate hate of the stronger side can rationalize Martin Luther King’s strategy of non-violent resistance and his message of “love your enemy” in the African American Civil Rights Movement. We argued that the “stooping down” effect and the greater difficulty of deterring an enemy who is motivated by hate can help explain the more aggressive military stance of the U.S. and its shift in national security strategy from one of deterrence to one of preemption following the 9/11 attacks. Lastly, we argued that the greater difficulty of deterring an individual who is motivated by hatred also provides a rationale for enhancement of the penalties for hate crimes.

Future work could further explore the formation of hatred in conflict. While ha-

tred may be strategically fomented or attenuated by political leaders, it also grows and subsides through non-strategic processes. Hate is, at least to some extent, an instinctive emotional response to being the target of violence. One can imagine specifying a repeated Hawk-Dove game in which one side's hate parameter would increase or decrease depending on whether the other side previously chose to act like a hawk or a dove. More hawkishness or violence would spark more hatred, which might fuel further violence. In this context, one could explore the conditions under which cycles of violence are likely to be sustained. For example, is a cycle of violence more likely if the two sides are more equal in strength?

It would also be intriguing to explore the revelation of hatred in conflict. Hate is an emotion that is private information and revealed through actions. One can imagine a sequential model in which one side conceals any hate that it may feel for the other side, and moves first, deciding whether to attack. If it attacks, the second side infers whether the first side harbors hate and decides whether to respond aggressively. An aggressive response may prompt a further attack by the instigating side, especially if the instigator is motivated by hate. In this context, one could explore the conditions under which hate can be inferred to provide a motivation for an attack. For example, an attack may be a stronger sign of hate if it is carried out by a side that faces lower material payoffs or by a weaker side against a stronger side. A great deal of interesting work lies ahead in exploring the formation and revelation of hate in games.

References

- [1] Arsenault, Raymond. *Freedom Riders: 1961 and the Struggle for Racial Justice*. Oxford, UK: Oxford University Press (2007).
- [2] Baliga, Sandeep and Tomas Sjöström. "The Strategy of Manipulating Conflict." *American Economic Review*, forthcoming (2011).
- [3] Becker, Gary. "A Theory of Social Interaction." *Journal of Political Economy* 82, 1063–1094 (1974).
- [4] Bentham, Jeremy. *An Introduction to the Principles of Morals and Legislation*. Oxford, UK: Oxford University Press (1789).
- [5] Bolle, Friedel. "Is Altruism Evolutionarily Stable? And Envy and Malevolence? Remarks on Bester and Güth." *Journal of Economic Behavior and Organization* 42, 131–133 (2000).
- [6] Bulow, Jeremy, John Geanakoplos, and Paul Klemperer. "Multimarket Oligopoly: Strategic Substitutes and Complements." *Journal of Political Economy* 93, 488–511 (1984).
- [7] Cleaver, Eldrige. *Soul on Ice*. New York: Dell Books (1968).
- [8] Curry, Philip and Tilman Klumpp. "Crime, Punishment, and Prejudice." *Journal of Public Economics* 93, 73–84 (2009).

- [9] Dharmapala, Dhammika and Nuno Garoupa. “Penalty Enhancement for Hate Crimes: An Economic Analysis.” *American Law and Economics Review* 6, 185–207 (2004).
- [10] Dillof, Anthony. “Punishing Bias: An Examination of the Theoretical Foundations of Bias Crime Statutes.” *Nothwestern Law Review* 91, 1015–1081 (1997).
- [11] Federal Bureau of Investigation. *Uniform Crime Reports* (1995–2009). Available online at www.fbi.gov/about-us/cjis/ucr/ucr.
- [12] Fudenberg, Drew and Jean Tirole. “The Fat-Cat Effect, the Puppy-Dog Ploy, and the Lean and Hungry Look.” *American Economic Review* 74, 361–366 (1984).
- [13] Gan, Li, Robertson Williams, and Thomas Wiseman. “A Simple Model of Optimal Hate Crime Legislation.” *Economic Inquiry* 49, 674–684 (2007).
- [14] Gandhi, Mohandas. *Non-Violence in Peace and War, 1942–1949*. New York: Garland Publishing (1972).
- [15] Garfinkel, Michelle and Stergios Skaperdas. “Conflict Without Misperceptions or Incomplete Information: How the Future Matters.” *Journal of Conflict Resolution* 44, 793–807 (2000).
- [16] Glaeser, Edward. “The Political Economy of Hatred.” *Quarterly Journal of Economics* 120, 45–86 (2005).
- [17] Haywood, Oliver. “Military Decision and Game Theory.” *Journal of the Operations Research Society of America* 2, 365–385 (1954).
- [18] Harel, Alon and Gideon Parchomovsky. “On Hate and Equality.” *Yale Law Journal* 109, 507–539 (1999).
- [19] Harrington, Evan. “The Social Psychology of Hatred.” *Journal of Hate Studies* 3, 49–82 (2004).
- [20] Heifetz, Aviad, Chris Shannon, and Yossi Spiegel. “What to Maximize if You Must.” *Journal of Economic Theory* 133, 31–57 (2007).
- [21] Jacobs, James and Kimberley Potter. *Hate Crimes: Criminal Law and Identity Politics*. New York: Oxford University Press (1998).
- [22] Jahn, Gunnar. Award Ceremony Speech, Nobel Peace Prize (1964). Available online at www.nobelprize.org/nobel_prizes/peace/laureates/1964/press.html.
- [23] Kahan, Dan. “Two Liberal Fallacies in the Hate Crime Debate.” *Law and Philosophy* 20, 175–193 (2001).
- [24] King Jr., Martin Luther. *Nobel Lecture*, Nobel Peace Prize (1964). Available online at www.nobelprize.org/nobel_prizes/peace/laureates/1964/king-lecture.html.

- [25] King Jr., Martin Luther. *Stride Toward Freedom: The Montgomery Story*. New York: Harper (1954).
- [26] Koçkesen, Levent, Efe Ok, and Rajiv Sethi. “Evolution of Interdependent Preferences in Aggregative Games.” *Games and Economic Behavior* 31, 303–310 (2000a).
- [27] Koçkesen, Levent, Efe Ok, and Rajiv Sethi. “The Strategic Advantage of Negatively Interdependent Preferences.” *Journal of Economic Theory* 92, 274–299 (2000b).
- [28] Kolm, Serge-Christophe. “Introduction to the Economics of Giving, Altruism, and Reciprocity.” In: S.C. Kolm and J.M. Ythier (eds.), *Handbook of the Economics of Giving, Altruism, and Reciprocity Vol. 1*. Elsevier: North-Holland (2006).
- [29] Konrad, Kai. “Altruism and Envy in Contests: An Evolutionarily Stable Symbiosis.” *Social Choice and Welfare* 22, 479–490 (2004).
- [30] Konrad, Kai. *Strategy and Dynamics in Contests*. Oxford, UK: Oxford University Press (2009).
- [31] Lawrence, Frederick. *Punishing Hate: Bias Crimes Under American Law*. Cambridge, MA: Harvard University Press (2002).
- [32] Luce, Duncan and Howard Raiffa. *Games and Decisions*. New York: Dover (1957).
- [33] Mialon, Hugo, Sue Mialon, and Maxwell Stinchcombe. “Torture in Counterterrorism: Agency Incentives and Slippery Slopes.” *Journal of Public Economics* 96, 33–41 (2012).
- [34] Milgram, Stanley. “Some Conditions of Obedience and Disobedience to Authority.” *Human Relations* 18, 57–75 (1965).
- [35] National Security Council of the United States. *National Security Strategy* (2002). Available online at georgewbush-whitehouse.archives.gov/nsc/nss/2002.
- [36] National Security Council of the United States. *National Security Strategy* (2006). Available online at georgewbush-whitehouse.archives.gov/nsc/nss/2006.
- [37] Nearpass, Gregory. “The Overlooked Constitutional Objection and Practical Concerns to Penalty-Enhancement Provisions of Hate Crime Legislation.” *Albany Law Review* 66, 547–573 (2003).
- [38] Possajennikov, Alex. “On The Evolutionary Stability of Altruistic and Spiteful Preferences.” *Journal of Economic Behavior and Organization* 42, 125–129 (2000).

- [39] Schelling, Thomas. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press (1960).
- [40] Smith, Adam. *The Theory of Moral Sentiments*. London: A. Millar, in the Strand (1759).
- [41] Tullock, Gordon. “Efficient Rent Seeking.” In: J.M. Buchanan, R.D. Tollison and G. Tullock (eds.), *Toward a Theory of the Rent Seeking Society*. College Station: Texas A&M University Press (1980).
- [42] von Neumann, John and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press (1944).
- [43] Zimbardo, Philip, Craig Haney, Curtis Banks, and David Jaffe. “The Psychology of Imprisonment: Privation, Power and Pathology.” In Zick Rubin (ed.) *Doing Unto Others: Explorations in Social Behavior*, pp. 61–73. Englewood Cliffs, NJ: Prentice-Hall (1974).