On Hatred*

Tilman Klumpp† Hugo M. Mialon‡

January 2013

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 criminal and civil penalty enhancements for hate crime.

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

JEL codes: D74, H11, K14, K42.

*We thank Phil Curry, Andrew Francis, Fabio Mendez, Sue Mialon, Paul Pecorino, Paul Rubin, Xuejuan Su, Joshua Teitelbaum, an anonymous referee, and participants at the 2012 Meetings of the American Law and Economics Association and the 2012 Southern Economic Association Conference for many helpful comments.

†University of Alberta, Department of Economics. 9-20 Tory Building, Edmonton, AB, T6G 2H4, Canada. E-mail: klumpp@ualberta.ca.

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

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

— Eldridge Cleaver, *Soul on Ice*

1 Introduction

The central point we wish to make in this essay is that understanding hate as an interdependent preference 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 individual 1's and individual 2's "material" utilities, respectively, and $-u_2$ is individual 1's "emotional" utility generated by hate for individual 2. 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.[2]

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

[1] See Smith (1759), ch. 1.2.3 and Bentham (1789), ch. 5.27, 6.27. Bentham uses the terms "antipathy" for hatred and "sympathy" for altruism. Smith uses the terms "resentments" and "sympathy."

[2] 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).

should adopt a more aggressive strategy when facing a hateful opponent, while the weaker side should adopt a less aggressive strategy. Second, 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. Third, 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.

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

1. 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. Moreover, Martin Luther King's call to "love one's enemy" is consistent with our model's implications concerning whether or not one should hate.

2. 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.

3. The third application concerns 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. Our analysis provides a more direct justification. To achieve a desired level of deterrence, a hateful individual must be threatened with a more severe penalty than otherwise necessary. Our analysis also provides a separate justification for additional civil penalties for hate crime. Compensation or damages awards for victims provide a greater deterrent in the presence of hatred.

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 develop a simple,

"reduced form" model of hatred. 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 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.

Bester and Güth (1998), Bolle (2000), Dufwenberg and Güth (2000), Possajennikov (2000), Koçkesen *et al.* (2000a,b), Güth and Peleg (2001), Konrad (2004), and Heifetz *et al.* (2007) explore the evolutionary stability of these interdependent preferences. Bester and Güth (1998) show that altruism may be evolutionarily stable in games where actions are strategic complements, while Bolle (2000), Dufwenberg and Güth (2000), and Possajennikov (2000) show that spite may be evolutionarily stable in games where actions are strategic substitutes. These papers analyze symmetric two-player games where actions are either strategic substitutes for both players or strategic complements for both players. We also investigate when a player benefits from hate in asymmetric games where actions may be strategic complements for one player but strategic substitutes for the other. In particular, we examine a Tullock model of contests (Tullock 1980), where the stronger side views actions as strategic complements while the weaker side views them as strategic substitutes as long as the weaker side does not hate too much.[3]

Konrad (2004) shows that altruists and envious players who meet in contests are symbiots, in the sense that the benefit of behaving altruistically is larger if the share of envious contestants is higher and the benefit of behaving enviously is larger if the share of altruists is higher. 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

[3]The 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).

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. Güth and Peleg (2001) and Heifetz *et al.* (2007) analyze conditions for material payoff maximization to be evolutionarily stable in symmetric and in more general two-player games, respectively. In their models, 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. The specification of interdependent utilities in Heifetz *et al.* (2007) encompasses ours. However, none of the aforementioned papers 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 4.2). However, our theoretical approach is very 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 (2012) 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 (2012), 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.[4]

Lastly, our paper is related to the law and economics literature on hate crime. Dharmapala and Garoupa (2004) and Gan *et al.* (2011) 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 more avoidance activities, which are costly. Gan *et al.* (2011) 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 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. Also, our model provides a novel justification for tort remedies in hate crime cases. If a perpetrator is motivated mainly by hate, he will be deterred less by the threat of punishment and more by the promise that his victims will be compensated.

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 The underlying game

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 $u : S \times T \to \mathbb{R}^2$, 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).

[4] In 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.

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

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

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 increases the marginal payoff of player i's action.

The notions of strategic substitutability and strategic complementarity allow the classification of many two-player models of conflict into three broad categories. The first is the class of games in which actions are strategic substitutes for both players. A simple discrete example of such a game is the classic hawk-dove game (Maynard Smith and Price 1973): For each player, the best response to "dove" is "hawk," and the best response to "hawk" is "dove." Thus, a more aggressive choice on the part of one player makes the other player want to adopt a less aggressive choice. The second class consists of games in which actions are strategic complements for both players. An example is the arms race game (Schelling 1960), where the best response to an opponent acquiring new arms is to acquire new arms oneself, and the best response to an opponent not acquiring new arms is to not acquire new arms. A more aggressive choice on the part of one player therefore makes the other player want to adopt a more aggressive choice.

The third class, finally, consists of games in which actions are strategic substitutes for one player and strategic complements for the other player. Many of our results have particularly interesting implications for such games. In Section 3.3, we will therefore introduce a simple model of a asymmetric conflict, where the "weak side" regards actions as strategic substitutes and the "strong side" regards actions as strategic complements.

3.2 The game with hatred

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

$$v_i(s,t) \equiv (1-\lambda_i)u_i(s,t) + \lambda_i(-u\_i(s,t)). \tag{2}$$

That is, v_i is a convex combination of player i's material payoff and the negative of i's opponent's material payoff. Thus, the payoff v_i reflects player i's preferences if he

harbors hateful emotions against player $-i$, with the weight λ_i measuring the strength of i's hatred. The resulting normal form game $G(\lambda) = (S \times T, v)$ is the game with hatred.[5]

We now characterize an interior Nash equilibrium of $G(\lambda)$, assuming that it exists. Note that, since $\partial u_1/\partial t < 0$ and $\partial u_2/\partial s < 0$, $\lambda_i < 1$ for $i = 1, 2$ is necessary for existence of an interior equilibrium. (We will look at the case $\lambda_i = 1$ in Section 4.1.) The equilibrium is characterized by the first-order conditions

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

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

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

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

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

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

Note that

$$\frac{\partial(u_1+u_2)}{\partial s} < 0 \quad \text{and} \quad \frac{\partial(u_1+u_2)}{\partial t} < 0 \tag{6}$$

[5]The underlying game is therefore $G = G(0,0)$. Note that we could have defined hateful preferences recursively, i.e.,

$$v_i \equiv (1-\gamma_i)u_i + \gamma_i(-v_{-i}) = (1-\gamma_i)u_i - \gamma_i\big[(1-\gamma_{-i})u_{-i} - \gamma_{-i}v_i\big]$$

for some $\gamma_i \in [0,1]$. Solving for v_i, one obtains

$$v_i = \frac{1-\gamma_i}{1-\gamma_i\gamma_{-i}}u_i - \frac{\gamma_i - \gamma_i\gamma_{-i}}{1-\gamma_i\gamma_{-i}}u_{-i},$$

which is the non-recursive formulation in (2) with $\lambda_i = (\gamma_i - \gamma_i\gamma_{-i})/(1-\gamma_i\gamma_{-i})$.

in equilibrium.[6] This implies 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.

3.3 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***. 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.[7] Assume now that effort has a per unit cost of 1 for player 1, and $k > 1$ for player 2. We say 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

$$u_1(s,t) = f(s,t) - s, \quad u_2(s,t) = f(t,s) - kt. \tag{7}$$

Notice that (7) 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.

[6] To see this, note that (3) implies that $(1-\lambda_1)\partial u_1/\partial s = \lambda_1 \partial u_2/\partial s$. Because $\partial u_2/\partial s < 0$ by assumption, the right-hand side is non-positive. To make the left-hand side non-positive, we need $\partial u_1/\partial s \leq 0$ (recall that $\lambda_1 < 1$ is necessary for an interior equilibrium, as noted above). Thus, $\partial u_1/\partial s \leq 0$ and $\partial u_2/\partial s < 0$, and it follows that $\partial(u_1+u_2)/\partial s < 0$ in an interior equilibrium. A similar argument can be made to show that $\partial(u_1+u_2)/\partial t < 0$.

[7] Alternatively, 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.

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

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

The Nash equilibrium of this game can be shown to be

$$s^*(\lambda) = \frac{k(1-\lambda_2)}{\big((1-\lambda_1)+k(1-\lambda_2)\big)^2}\ , \quad t^*(\lambda) = \frac{1-\lambda_1}{\big((1-\lambda_1)+k(1-\lambda_2)\big)^2}\ . \tag{8}$$

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

$$k \ > \ \Lambda \equiv \frac{1-\lambda_1}{1-\lambda_2}. \tag{9}$$

Thus, as long as $k > \Lambda$, efforts are strategic complements for the stronger player and strategic substitutes for the weaker player. In particular, (9) 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 the opponent's hatred depends on whether actions are strategic substitutes or complements; in our conflict example this will depend on a player's relative strength. Second, whether hate can help a player achieve higher materials payoffs again is determined by whether actions are strategic substitutes or complements. Third, penalties for a player who hates may be a less effective deterrent than rewards for the player who is the target of the hate.

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 will cause a player to

adopt different strategies than he would have otherwise chosen. We now examine these strategic responses to an opponent's hatred.

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:

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

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 $t^* = \arg\max_t \min_s -u_1(s,t)$.[8]

We can say more for the general, variable-sum case. From (5)–(6), observe that $ds/d\lambda_2$ has the same sign as

$$\frac{\partial^2 v_1}{\partial s \partial t} \;=\; (1-\lambda_1)\frac{\partial^2 u_1}{\partial s \partial t} \;-\; \lambda_1 \frac{\partial^2 u_2}{\partial s \partial t}. \tag{10}$$

Thus, if actions are strategic complements for player 1 and strategic substitutes for player 2 in the underlying game 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. If actions are strategic substitutes or complements for both players, the sign of (10) is generally ambiguous. Whether player 1 increases or decreases s in response to an increase in λ_2 then depends on player 1's own hate parameter, λ_1. In particular, for sufficiently small values of λ_1 (i.e., λ_1 close to zero), (10) will be positive/negative if actions are strategic complements/substitutes. The opposite holds

[8]Since u_1 is strictly decreasing in t, for max-min strategy t^* to exist the strategy set T must be compact. For example, in the Tullock contest with $\lambda_2 = 1$, we would have to assume a maximal strategy $\bar{t} < \infty$, such that player 2 can only choose efforts $t \leq \bar{t}$.

for sufficiently large values of λ_1 (i.e., λ_1 close to one). Similar statements can be made for player 2's strategy, t.

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.[9] Both the stooping-down response and the turn-the-other-cheek response can arise in equilibrium, depending on whether actions are viewed as strategic substitutes or complements. In Section 3.3 we demonstrated that, in a Tullock model of conflict, strategic substitutability and complementarity are tied to the relative strength of players. Provided condition (9) 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 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. In particular, we now examine whether developing hateful emotions can be in the ultimate interest of a player, in the sense that it such emotions can increase the player's material payoffs.

To examine the effect of player 1's hate parameter on 1's payoff, differentiate 1's overall payoff v_1 with respect to λ_1:

$$\frac{du_1}{d\lambda_1} = \frac{\partial u_1}{\partial s}\frac{ds}{d\lambda_1} + \frac{\partial u_1}{\partial t}\frac{dt}{d\lambda_1}. \tag{11}$$

Note that player 1's first-order condition (3) implies that, in equilibrium,

$$\frac{\partial u_1}{\partial s} = \frac{\lambda_1}{1-\lambda_1}\frac{\partial u_2}{\partial s}. \tag{12}$$

Substituting (12) into (11), we get

$$\frac{du_1}{d\lambda_1} = \frac{\lambda_1}{1-\lambda_1}\frac{\partial u_2}{\partial s}\frac{ds}{d\lambda_1} + \frac{\partial u_1}{\partial t}\frac{dt}{d\lambda_1}. \tag{13}$$

Recall now that $\partial u_2/\partial s < 0$, $\partial u_1/\partial t < 0$, and $ds/d\lambda_1 > 0$. Thus, the first summand in (13) is non-positive, while the sign of the second summand depends on the sign of $dt/d\lambda_1$.

[9] These effects 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.

If actions are strategic complements for player 2 and strategic substitutes for player 1 in the underlying game G, 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 all other cases, the sign of (13) is ambiguous. However, for the case $\lambda_1 = \lambda_2 = 0$, the first term on the right-hand side of (13) vanishes and the sign of the second term depends only on the sign of $dt/d\lambda_1$. This, in turn, depends on whether player 2 regards actions as strategic complements or substitutes in G (see Section 4.1). In the first case, $dt/d\lambda_1 > 0$, so that an increase in λ_1 will reduce player 1's material payoff. In the second case, $dt/d\lambda_1 < 0$, so that an increase of λ_1 from zero to a small positive value will increase 1's material payoff. A moderate amount of hatred is hence desirable from the perspective of player 1, if his opponent regards actions as strategic substitutes in G.[10]

The literature on the evolution of other-regarding preferences, reviewed in Section 2, relies on the same mechanism to show that a hateful disposition can emerge through evolutionary selection if actions are strategic substitutes. Observe, however, that hate is also a malleable emotion that can be influenced by others. Military or political leaders in conflicts can thus commit their followers to more aggressive actions by fomenting hatred toward their adversaries. In the context of the asymmetric Tullock contest, our results imply that leaders of the stronger side can advance the material interest of their group by inducing a modest amount of hate. This strategic manipulation of emotions and preferences can be achieved through propaganda and indoctrination, but also through more subtle psychological tactics.[11]

Responsible leaders of the weaker side, on the other hand, should do the opposite and *dissuade* hatred in their followers. If the weaker side is the target of aggressive actions by the stronger side, this may be a challenging task—hence the reason why those who preach a message of peace in the face of oppression are often celebrated. Conversely, leaders who preach a message of hate to the weaker side can be suspected of merely instrumentalizing their followers for their own personal gain. For example, a general who commands an violent rebel group may reap large political rents in negotiations with his opponents (a related argument is made in Glaeser (2005). The fomentation of hatred on the weaker side in an asymmetric conflict is thus an indication that ulterior motives are at play.

[10]The reason is the following. If actions are strategic substitutes for 2, then 2 responds to an increase in λ_1 with a less aggressive strategy, which helps player 1. At the same time, the increased λ_1 makes player 1 adopt a more aggressive strategy than would be optimal for maximization of u_1. But since 1 was optimizing the (differentiable) function u_1 in equilibrium of $G(0,0)$, this has only a second-order effect on the value of u_1. The adjustment in the 2's strategy, on the other hand, has a first-order effect on u_1.

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

4.3 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 $\overline{s}$ and $\overline{t}$, respectively. The conventional way to enforce such limits is to impose costs, say $c_1(s) > 0$ and $c_2(t) > 0$, on actions $s > \overline{s}$ and $t > \overline{t}$. 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' overall payoffs become

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

and

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

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

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

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

$$c_1(s) \;\geq\; \Big[u_1(s,\overline{t}) - u_1(\overline{s},\overline{t})\Big] + \frac{\lambda_1}{1-\lambda_1}\Big[u_2(\overline{s},\overline{t}) - u_2(s,\overline{t})\Big]. \tag{14}$$

(A similar expression can be derived for $c_2(t)$.) The term $[u_1(s,\overline{t}) - u_1(\overline{s},\overline{t})]$ in (14) is the private gain of player 1 from playing s instead of $\overline{s}$. The term $[u_2(\overline{s},\overline{t}) - u_2(s,\overline{t})]$ is the harm inflicted on player 2 by player 1's choice of s instead of $\overline{s}$. 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 $\lambda_1 \to 1$. 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 $s > \overline{s}$ 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 overall payoff then becomes

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

Observe that the higher λ_1 is, the more weight player 1 places on player 2's reward. Thus, the deterrence effect of a given reward is increased in the presence of hatred.

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

$$b_2(s) \;\geq\; \frac{1-\lambda_1}{\lambda_1}\Big[u_1(s,\bar{t})-u_1(\overline{s},\bar{t})\Big]+\Big[u_2(\overline{s},\bar{t})-u_2(s,\bar{t})\Big]. \tag{15}$$

The minimum reward (15) 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 $\lambda_1 \to 1$, it is sufficient to simply compensate player 2 for his loss resulting from 1's actions.[12]

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 after 9/11, and the legal debate surrounding the punishment of hate crime.

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

[12]This argument suggests an interesting possibility to achieve deterrence, namely to transfer some utility from player i to player i's opponent. Assume that it is possible to reduce player 1's material payoff by $c_1(s)$, and simultaneously increase player 2's material payoff by the same amount $c_1(s)$, if action $s > \overline{s}$ is taken. (For example, if $c_1(s)$ represents a monetary transfer and players' material utilities are quasilinear in wealth with identical coefficients, this assumption holds.) Similarly, player 2's material payoff is reduced by $c_2(t)$, and player 1's material payoff is increased by the same amount $c_2(t)$, if action $t > \bar{t}$ is taken. Player 1's overall payoff then becomes

$$\begin{aligned}\hat{v}_1(s,t) &= (1-\lambda_1)\left[u_1(s,t)-c_1(s)+c_2(t)\right]-\lambda_1\left[u_2(s,t)+c_1(s)-c_2(t)\right]\\ &= (1-\lambda_1)u_1(s,t)-\lambda_1 u_2(s,t)-c_1(s)+c_2(t)\\ &= v_1(s,t)-c_1(s)+c_2(t).\end{aligned} \tag{16}$$

The deterrence effect of the transfer from player 1 to player 2 is now unaffected by λ_1. While the deterrence effect of a pure penalty is diminished, and that of a pure reward is increased, by the presence of hatred, the deterrence effect of a transfer remains the same regardless of the extent of hatred.

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.[13]

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."

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

[13] The events are described in detail in King (1958).

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, especially if the oppressed side remains non-violent in the face of the violent escalation by the oppressive side. Outside sympathy can function as a "reward" that is bestowed on the oppressed side in response to bearing the fsuffering inflicted by their oppressors. Our arguments of Section 4.3 suggest that such rewards can provide a disincentive for the resentful side to further escalate violence.[14]

[14]In *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).)

More generally, the presence of an external constituency of potential sympathizers may be a predictor of the use of non-violent resistance. In the American civil rights movement, Southern blacks could appeal to Northern whites to pass legislation, and these appeals were more likely to succeed if the movement was non-violent. Similarly, in the Indian independence movement, people in India could appeal to a democratic electorate in Britain. While people in Britain might have preferred to keep India as a colony, they would not feel directly threatened by its loss, and a peaceful independence movement could gain sympathizers. In contrast, in the fight against apartheid in South Africa, the African National Congress did not have a similar audience to which it could appeal with non-violent methods because white rulers lived in South Africa and felt threatened by black rule. This may partly explain why the fight against apartheid in South Africa was more often characterized by armed resistance and less often by non-violent resistance than were the American civil rights movement under Martin Luther King Jr. and the Indian independence movement under Mohandas K. Gandhi.

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 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 despitefully 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 4.2, 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.[15]

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:

[15] It is worth noting, however, that Gandhi advocated the strategy of "loving the enemy" not only in India's struggle against oppression by the British, but also during World War II when Nazi Germany appeared poised and ready to invade Britain (Gandhi 1972)—a recommendation that can be regarded as misguided in hindsight. In light of our model, there are two types of misclassifications that could lead one to wrongly advocate appeasement. First, if the conflict can be described by a Tullock contest, the Allies were perhaps not the weaker side in World War II, and Nazi Germany not the stronger side. On the other hand, it is not obvious that Nazi Germany actually regarded actions as strategic complements. In a hawk-dove game, for instance, where each side views actions as strategic substitutes, a pacifying policy would have resulted in even more aggressiveness from the Nazis, as was demonstrated by the Munich Agreement.

> "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:

> "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. Under Barack Obama's presidency, finally, preemptive efforts have shifted notably from a small number of large-scale operations against entire regimes, to a larger number of

small-scale operations against individual militants, as well as "light footprint" operations against individual programs belonging to adversarial regimes.[16]

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 now interpret this paradigm shift in light of our formal arguments made earlier, by elaborating on several of the themes developed in 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 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.

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.

[16] Most of these operations appear to be targeted killings that used unmanned aircraft. Sanger (2012, p. 244) counts roughly 265 confirmed drone strikes during the first three years of Obama's presidency, compared to about 40 during George W. Bush's entire administration. In addition, the Obama administration has embraced the use of cyberweapons to limit the military capabilities of its adversaries, and in particular Iran's nuclear capabilities (Sanger 2012).

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.[17] 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 an unpleasant outcome.

5.3 The punishment of hate crime

Between 1995 and 2009, 147,099 individuals were victims of reported hate crimes in the U.S. (Federal Bureau of Investigation 1995–2009). 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.[18] 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)

[17] For 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).

[18] For a classification and analysis of these laws, see Lawrence (2002). A few states, including Arkansas, Georgia, Kansas, Indiana, and South Carolina currently have no hate crime laws (Borys 2012).

Furthermore, several statutes provide enhancements to criminal penalties and additional civil penalties for hate crime under the racial animus definition.[19] 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 the race, color, ancestry, ethnicity, religion, sexual orientation, or national origin of the victim. [...] A person or organization which establishes by clear and convincing evidence that it has been coerced, intimidated, or threatened in violation of this section shall have a civil cause of action for treble damages, an injunction, or any other appropriate relief in law or in equity." (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 enhanced punishments in particular have been proposed. The first argument is such 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 simply the product of the emotions that motivated the hate crimes.[20]

Using the theoretical framework developed in this paper, we can provide a presently overlooked justification for enhanced punishments for hate crime. 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),

[19] Among the states that have hate crime statutes, those that do not have explicit provisions for penalty enhancements include Colorado, Idaho, Massachussetts, North Dakota, Oklahoma, Oregon, South Dakota, Washington, West Virginia, and Wyoming (Borys 2012).

[20] In 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.

our arguments in Section 4.3 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.[21]

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 directly. For example, in the case of James Byrd, an African American man who was tied to a pickup truck, dragged nearly three miles, and decapitated in Jasper, Texas in 1988, the three perpetrators were known members of a white supremacist group. In the case of Waqar Hasan, a Pakistani Muslim who was shot in the face in Dallas, Texas, on September 15, 2001, the perpetrator went on a radio show immediately after committing the crime to confess that he had killed Hasan as a revenge for the terrorist attacks. 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 ethnicity. 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 the aforementioned justification for enhanced hate crime punishments 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.[22]

Our analysis also provides a separate justification for civil penalties or tort remedies (such as treble damages) for hate crimes. Our arguments of Section 4.3 suggest that compensation or damages awards to victims provide a deterrent to hate crime. Moreover, a civil remedy in the form of a payment of damages by offenders to their victims provides a deterrent regardless of the extent to which potential offenders are motivated by hate.

[21] Curry and Klumpp (2009) show that, to achieve a constant degree of deterrence across individuals, penalties must depend on individual characteristics such as wealth or income. Penalty enhancements for hate crime are similar, except that the penalty is a function of hate instead of income.

[22] Harel and Parchomosvsky (1999) make a similar argument: "under the fair protection paradigm, victims who are particularly vulnerable to crime may have a legitimate claim on fairness grounds to greater protection against crime" (p. 509).

If they are not motivated by hate, their loss provides a deterrent, while if they are purely motivated by hate, their victim's compensation or damages award provides a deterrent.

A few papers analyze and advocate the use of civil suits in combatting hate crime (Braithwaite 1996 and Rustad and Koenig 2007), but these papers have mainly emphasized the deterrent effects of imposing financial hardship on hate organizations and not the deterrent effects of providing financial awards to the victims of violence by members of these organizations. In the U.S., a number of successful civil suits have been carried out against hate groups. For example, in 1998, a faction of the Ku Klux Klan was ordered to pay $37.8 million in damages for conspiring to burn down an African American church in South Carolina (*Macedonia Baptist Church v. Christian Knights of the Klu Klux Klan*, No. 96-CP-14-217, SC Ct. Com. P1. 1998). Assets of the Klan faction and its high-ranking members were seized to help rebuild the church (Knickerbocker 2000). Our model suggests that such penalties could be effective in deterring hate organizations—especially if they are used to help the communities that these organizations target.

Lastly, our result that rewarding those who are targets of attacks has a greater deterrent effect if the attacks are motivated by hatred may have implications for government programs to compensate victims of crime. In the U.S., all states have established crime victim compensation programs with funding from the federal government through the Department of Justice's Office for Victims of Crime.[23] Most programs reimburse victims for crime-related expenses, such as medical costs and lost wages. Maximum awards per claim range widely by state, from $\$10,000$ to $\$150,000$ (Eddy 2003). Our analysis suggests that such programs may have the additional benefit of being a deterrent in the case of crimes motivated by hate.

6 Conclusion

We 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. We emphasized three key messages: First, one player's hatred toward another has a strategic effect on the hated—the latter may either "stoop down" or "turn the other cheek" in response to being hated, and we delineated conditions under which each response occurs. Second, the deterrence effect of punishing a player is diminished if the player is more hateful, and rewarding the target of hatred may be a stronger deterrent than conventional punishment of the perpetrator. Third, hatred can both benefit and hurt a player's material interests, and we derived conditions under which each effect occurs.

We then made these messages concrete by discussing their implications in three specific contexts: The strategy of non-violent resistance in the African American Civil Rights Movement, U.S. national security strategy post 9/11, and the debate surrounding the punishment of hate crime. While much has been written about each of these areas, we

[23] For details on these programs, see www.ojp.usdoj.gov/ovc/.

believe that our approach of examining a game theoretic model that explicitly incorporates hatred as a preference informs our understanding of the pertinent policy issues in new and important ways.

Our work leaves ample room for further investigations. Below we briefly discuss two avenues for future research that seem particularly promising.

Hatred in dynamic games. First, while hatred may be strategically fomented or attenuated by political leaders—a mechanism examined in this paper—it also grows and subsides through non-strategic processes. Hate is, at least to some extent, an instinctive response to being the target of violence. One can imagine specifying a repeated game in which one side's hate parameter increases in the aggressiveness of the other sides's past actions. More aggressive actions spark more hatred, which then fuels further aggressiveness. Specifying such a law of motion for hatred explicitly would allow one to ask many interesting questions. One could explore conditions under which cycles of violence are likely to be sustained (e.g., is a cycle of violence more likely if the two sides are more equal in strength?). One could also examine whether there is an optimal length of memory with which past actions should be remembered. In repeated games, memory can foster mutually beneficial cooperation via the folk theorem. However, if hateful preferences evolve with past actions, remembering the past too precisely may counteract this effect and lead to non-cooperation. The fact that some ethnic conflicts seem to be fueled by centuries-old grudges indeed suggests that, when hatred is an instinctive emotional response to previous opponent actions, too much knowledge of history may inhibit cooperation.

Revelation of hatred. It would also be intriguing to explore the revelation of hatred in conflict. Hate is an emotion whose strength 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. 2007. *Freedom Riders: 1961 and the Struggle for Racial Justice*. Oxford, UK: Oxford University Press.

[2] Baliga, Sandeep and Tomas Sjöström. 2012. The Strategy of Manipulating Conflict. *American Economic Review* 102: 2897–2922.

[3] Becker, Gary. 1974. A Theory of Social Interaction. *Journal of Political Economy* 82: 1063–1094.

[4] Bentham, Jeremy. 1789. *An Introduction to the Principles of Morals and Legislation*. Oxford, UK: Oxford University Press.

[5] Bester, Helmut and Werner Güth. 1998. Is Altruism Evolutionarily Stable? *Journal of Economic Behavior and Organization* 34: 193–209.

[6] Bolle, Friedel. 2000. Is Altruism Evolutionarily Stable? And Envy and Malevolence? Remarks on Bester and Güth. *Journal of Economic Behavior and Organization* 42: 131–133.

[7] Borys, Gregory. 2012. Hate Crimes: An Empirical Analysis of the Impact of Legislation. Emory University Honors Thesis. Available online at etd.library.emory.edu/view/record/pid/emory:bqbrv.

[8] Braithwaite, David. 1996. Combatting Hate Crimes: The Use of Civil Alternatives to Criminal Prosecution. *Public Interest Law Journal* 6: 243–265.

[9] Bulow, Jeremy, John Geanakoplos, and Paul Klemperer. 1985. Multimarket Oligopoly: Strategic Substitutes and Complements. *Journal of Political Economy* 93: 488–511.

[10] Cleaver, Eldridge. 1968. *Soul on Ice*. New York, NY: Dell Books.

[11] Curry, Philip and Tilman Klumpp. 2009. Crime, Punishment, and Prejudice. *Journal of Public Economics* 93: 73–84.

[12] Dharmapala, Dhammika and Nuno Garoupa. 2004. Penalty Enhancement for Hate Crimes: An Economic Analysis. *American Law and Economics Review* 6: 185–207.

[13] Dillof, Anthony. 1997. Punishing Bias: An Examination of the Theoretical Foundations of Bias Crime Statutes. *Northwestern Law Review* 91: 1015–1081.

[14] Dufwenberg, Martin and Werner Güth. 2000. Why Do You Hate Me? On the Survival of Spite. *Economics Letters* 67: 147–152.

[15] Eddy, Dan. 2003. State Crime Victim Compensation Programs: Nature and Scope. National Center for Victims of Crime, Discussion Paper.

[16] Federal Bureau of Investigation. *Uniform Crime Reports* (1995–2009). Available online at www.fbi.gov/about-us/cjis/ucr/ucr.

[17] Fudenberg, Drew and Jean Tirole. 1984. The Fat-Cat Effect, the Puppy-Dog Ploy, and the Lean and Hungry Look. *American Economic Review* 74: 361–366.

[18] Gan, Li, Roberton Williams, and Thomas Wiseman. 2011. A Simple Model of Optimal Hate Crime Legislation. *Economic Inquiry* 49: 674–684.

[19] Gandhi, Mohandas. 1972. *Non-Violence in Peace and War, 1942–1949*. New York, NY: Garland Publishing.

[20] Garfinkel, Michelle and Stergios Skaperdas. 2000. Conflict Without Misperceptions or Incomplete Information: How the Future Matters. *Journal of Conflict Resolution* 44: 793–807.

[21] Glaeser, Edward. 2005. The Political Economy of Hatred. *Quarterly Journal of Economics* 120: 45–86.

[22] Güth, Werner and Bezalel Peleg. 2001. When Will Payoff Maximization Survive? An Indirect Evolutionary Analysis. *Journal of Evolutionary Economics* 11: 479–499.

[23] Haywood, Oliver. 1954. Military Decision and Game Theory. *Journal of the Operations Research Society of America* 2: 365–385.

[24] Harel, Alon and Gideon Parchomovsky. 1999. On Hate and Equality. *Yale Law Journal* 109: 507–539.

[25] Harrington, Evan. 2004. The Social Psychology of Hatred. *Journal of Hate Studies* 3: 49–82.

[26] Heifetz, Aviad, Chris Shannon, and Yossi Spiegel. 2007. What to Maximize if You Must. *Journal of Economic Theory* 133: 31–57.

[27] Jacobs, James and Kimberley Potter. 1998. *Hate Crimes: Criminal Law and Identity Politics*. New York, NY: Oxford University Press.

[28] Jahn, Gunnar. 1964. Award Ceremony Speech, Nobel Peace Prize. Available online at www.nobelprize.org/nobel_prizes/peace/laureates/1964/press.html.

[29] Kahan, Dan. 2001. Two Liberal Fallacies in the Hate Crime Debate. *Law and Philosophy* 20: 175–193.

[30] King Jr., Martin Luther. 1958. *Stride Toward Freedom: The Montgomery Story.* New York, NY: Harper.

[31] King Jr., Martin Luther. 1964. *Nobel Lecture*, Nobel Peace Prize. Available online at www.nobelprize.org/nobel_prizes/peace/laureates/1964/king-lecture.html.

[32] Knickerbocker, Brad. 2000. Latest Tactic Against Hate Groups: Bankruptcy. *Christian Science Monitor*, August 25.

[33] Koçkesen, Levent, Efe Ok, and Rajiv Sethi. 2000a. Evolution of Interdependent Preferences in Aggregative Games. *Games and Economic Behavior* 31: 303–310.

[34] Koçkesen, Levent, Efe Ok, and Rajiv Sethi: 2000b. The Strategic Advantage of Negatively Interdependent Preferences. *Journal of Economic Theory* 92: 274–299.

[35] Kolm, Serge-Christophe. 2006. Introduction to the Economics of Giving, Altruism, and Reciprocity. In *Handbook of the Economics of Giving, Altruism, and Reciprocity Vol. 1*, ed. S.C. Kolm and J.M. Ythier. Amsterdam, The Netherlands: North-Holland.

[36] Konrad, Kai. 2004. Altruism and Envy in Contests: An Evolutionarily Stable Symbiosis. *Social Choice and Welfare* 22: 479–490.

[37] Konrad, Kai. 2009. *Strategy and Dynamics in Contests.* Oxford, UK: Oxford University Press.

[38] Lawrence, Frederick. 2002. *Punishing Hate: Bias Crimes Under American Law.* Cambridge, MA: Harvard University Press.

[39] Luce, Duncan and Howard Raiffa. 1957. *Games and Decisions.* New York, NY: Dover.

[40] Maynard Smith, John and George Price. 1973. The Logic of Animal Conflict. *Nature* 246: 15–18.

[41] Mialon, Hugo, Sue Mialon, and Maxwell Stinchcombe. 2012. Torture in Counterterrorism: Agency Incentives and Slippery Slopes. *Journal of Public Economics* 96: 33–41.

[42] Milgram, Stanley. 1965. Some Conditions of Obedience and Disobedience to Authority. *Human Relations* 18: 57–75.

[43] National Security Council of the United States. 2002. *National Security Strategy.* Available online at georgewbush-whitehouse.archives.gov/nsc/nss/2002.

[44] National Security Council of the United States. 2006. *National Security Strategy.* Available online at georgewbush-whitehouse.archives.gov/nsc/nss/2006.

[45] Nearpass, Gregory. 2003. The Overlooked Constitutional Objection and Practical Concerns to Penalty-Enhancement Provisions of Hate Crime Legislation. *Albany Law Review* 66: 547–573.

[46] Possajennikov, Alex. 2000. On The Evolutionary Stability of Altruistic and Spiteful Preferences. *Journal of Economic Behavior and Organization* 42: 125–129.

[47] Rustad, Michael and Thomas Koenig. 2007. 'Hate Torts' to Fight Hate Crimes: Punishing the Organizational Roots of Evil. *American Behavioral Scientist* 51: 302–318.

[48] Sanger, David. 2012. *Confront and Conceal: Obama's Secret Wars and Surprising Use of American Power*. New York, NY: Crown.

[49] Schelling, Thomas. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

[50] Smith, Adam. 1759. *The Theory of Moral Sentiments.* London, UK: A. Millar, in the Strand.

[51] Tullock, Gordon. 1980. Efficient Rent Seeking. In *Toward a Theory of the Rent Seeking Society*, ed. J.M. Buchanan, R.D. Tollison, and G. Tullock. College Station, TX: Texas A&M University Press.

[52] von Neumann, John and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

[53] Zimbardo, Philip, Craig Haney, Curtis Banks, and David Jaffe. 1974. The Psychology of Imprisonment: Privation, Power and Pathology. In *Doing Unto Others: Explorations in Social Behavior*, ed. Zick Rubin. Englewood Cliffs, NJ: Prentice-Hall.