

University of Illinois College of Law
Law and Economics Working Papers

Year 2006

Paper 60

The Just World Bias and Hate Crime Statutes

Dhammika Dharmapala*

Nuno Garoupa[†]

Richard H. McAdams[‡]

*University of Connecticut and University of Michigan, dharmap@illinois.edu

[†]Universidade Nova de Lisboa, ngaroupa@illinois.edu

[‡]University of Illinois College of Law, rmcadams@uchicago.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://law.bepress.com/uiuclwps/art60>

Copyright ©2006 by the authors.

The Just World Bias and Hate Crime Statutes

Dhammika Dharmapala, Nuno Garoupa , and Richard H. McAdams

Abstract

The issue of whether and how to justify penalty enhancements for hate crimes against members of disfavored groups has attracted widespread attention. Harel and Parchomovsky (1999) justify penalty enhancements on egalitarian grounds, arguing that such crimes lead to the disproportionate victimization of minorities. However, within an economic framework, no distinctive harm is caused by disparate victimization *per se*. This paper addresses this issue by extending the standard economic model of crime in two ways. First, it introduces potential offenders' beliefs about the characteristics of potential victims as a factor that may affect the net benefits from crime (or, more generally, from hostile acts). Second, based on psychological evidence, it assumes that individuals are subject to a just world bias in inference - i.e., a tendency to attribute disproportionate victimization to negative characteristics of the victimized group, rather than to the hate-motivated preferences of offenders. In a simple two-period setting, we show that disproportionate victimization of the disfavored group in the first period can lead to additional crime against that group in the second period. The reason is that potential offenders subject to the just world bias infer that the cause of disproportionate victimization is not hate motivation but the victims' negative characteristics, and this inference raises the net benefits of crime against that group. Our main result is that penalty enhancements can reduce the social harm due to these extra crimes (or, more generally, to socially costly acts of discrimination). We also consider the implications of the just world bias for a more general welfare analysis of optimal enforcement policy.

The Just World Bias And Hate Crime Statutes*

Dhammika Dharmapala[†]
University of Connecticut
University of Michigan

Nuno Garoupa[‡]
Universidade Nova de Lisboa, Portugal
FEDEA, Madrid
CEPR, London

Richard McAdams[§]
University of Illinois at Urbana-Champaign
University of Chicago

September 20, 2006

*We thank Dominique Demougin, Lee Fennell, Louis Kaplow, Friedrich Kubler, Henrik Lando, William Landes, Anup Malani, Thomas Miles, Janice Nadler, Eric Posner, Jeff Rachlinski, Jennifer Robbenolt, Chris Sanchirico, Katherine Strandburg, Lu-in Wang, seminar participants at the University of Chicago and participants at the 2005 American Law and Economics Association meetings and the 2006 Conference on “Uncertainty, Risk and Regulation: The Behavioral Law and Economics Perspective,” at Technische Universität, Berlin for valuable comments. The usual disclaimer applies. Garoupa acknowledges the financial support of the FCT, POCI/JUR/55752/2004, Portugal.

[†]Department of Economics, University of Connecticut, 341 Mansfield Rd., U-1063, Storrs, CT 06269-1063, USA, and Ford School of Public Policy, University of Michigan, Weill Hall 5219, 735 S. State St., Ann Arbor, MI 48109. Email: dhammika@umich.edu

[‡]Faculdade de Economia, Universidade Nova de Lisboa, Campus de Campolide, P-1099-032 Lisboa, Portugal. Email: ngaroupa@fe.unl.pt

[§]University of Illinois at Urbana-Champaign College of Law, 504 E. Pennsylvania Avenue, Champaign, IL 61820, USA, and University of Chicago Law School, 1111 East 60th

Abstract

The issue of whether and how to justify penalty enhancements for hate crimes against members of disfavored groups has attracted widespread attention. Penalty enhancements have been defended on egalitarian grounds, as such crimes lead to the disproportionate victimization of minorities. However, within an economic framework, no distinctive harm is caused by disparate victimization *per se*. This paper extends the standard economic model of crime in two ways. First, it introduces potential offenders' beliefs about the characteristics of potential victims as a factor that may affect the net benefits from crime. Second, based on psychological evidence, it assumes that individuals are subject to a "just world bias" in inference – i.e. a tendency to attribute disproportionate victimization to negative characteristics of the victimized group, rather than to the hate-motivated preferences of offenders. In a simple two-period setting, we show that disproportionate victimization of the disfavored group in the first period can lead to additional crime against that group in the second period. The reason is that potential offenders subject to the just world bias infer that the cause of disproportionate victimization is not hate motivation but the victims' negative characteristics, and this inference raises the net benefits of crime against that group. Our main result is that penalty enhancements can reduce the social harm due to these extra crimes.

JEL Classification: K4.

Keywords: hate crimes, behavioral economics.

1 Introduction

In recent years, the Federal Bureau of Investigation (2005) has tabulated reports of “hate crimes” involving 9,000-12,000 victims per year. Most of these crimes are based on race. Some (such as the murderous rampages described in Dharmapala and McAdams (2005)) are severe, but most are less serious, involving crimes such as vandalism, intimidation, or simple assault (Federal Bureau of Investigation, 2005). Because many crimes go unreported, victim surveys report higher numbers. The National Crime Victimization Survey shows an annual average of 210,000 hate crime victims in the United States from 2000 to 2003 (Harlow, 2005).

Most US jurisdictions define a crime as a hate crime if it is committed because of the perpetrator’s animus or hatred toward a racial or other specified group, or, more broadly, if the victim is selected because of membership in a specified group.¹ For example, Federal sentencing guidelines provide for a “penalty enhancement” if the defendant “intentionally selected any victim or any property as the object of the offence of conviction because of the actual or perceived race, color, religion, national origin, ethnicity, gender, disability, or sexual orientation of any person.”² Most states have similar laws for state offences, though the list of possible groups varies widely (Grattet, Jenness and Curry, 1998). As a result, a hate crime may be punished more harshly than the same crime absent the hatred or discriminatory motive the perpetrator had for committing the offence.³

Bias-motivated crimes have attracted the scholarly attention of a large legal and philosophical literature (e.g. Lawrence, 1999; Hurd and Moore, 2004) and a small but growing economic literature (Dharmapala and Garoupa, 2004; Dharmapala and McAdams, 2005; Gan, Williams and Wiseman, 2004). The central issue in this analysis is whether and how to justify penalty enhancements for hate crimes. In an important contribution to this debate,

¹The latter type of statute has been upheld by the US Supreme Court (*Wisconsin v. Mitchell*, 508 U.S. 476 (1993)).

²28 U.S.C. § 994 (1994).

³Similarly, in the UK, although “hate crime” has no specific legal meaning, hate motivation is an aggravating factor in sentencing since the Criminal and Disorder Act 1998 (CDA 1998) and incitement to racial hatred has been extended to include religious grounds by the Anti-Terrorism, Crime and Security Act 2001 (ATCSA 2001).

Harel and Parchomovsky (1999) argue that bias-motivated crimes lead to the disproportionate victimization of minorities. They adopt a nonconsequentialist “fair protection paradigm” in which the inequality inherent in disproportionate victimization is in itself sufficient to justify government action, such as hate crime penalty enhancements, to equalize rates of victimization.

Dharmapala and Garoupa (2004) formalize the notion of disproportionate victimization by extending the economic model of optimal enforcement to the case where the population is divided into dominant and disfavored groups. (defined e.g. by ethnicity or religion). A subset of potential offenders from the dominant group is assumed to have discriminatory or hateful preferences, so that they derive greater benefits from committing an otherwise identical crime against members of the disfavored group than from committing the same crime against a member of their own group. Assuming that expected sanctions are independent of the victim’s group, this hate motivation gives rise to an equilibrium in which members of the disfavored group face a disproportionately high probability of victimization.⁴ A central lesson of this analysis, however, is that penalty enhancements may or may not be optimal on efficiency grounds, depending on factors such as the distribution of benefits from crime and the social costs of sanctions. It appears that, within a consequentialist (economic or utilitarian) framework, there is no harm caused by disparate victimization *per se*.

In this paper, we revisit this issue by analyzing the possibility that the observation of disparate victimization may influence inferences about group characteristics. We extend the model of Dharmapala and Garoupa (2004) by introducing a new variable representing the moralistic beliefs that a potential offender holds about the characteristics of a potential victim. Specifically, these beliefs concern the perceived intrinsic value or “moral worth” of the individual one may victimize. We assume the beliefs affect the potential offender’s costs (net benefits) of crime.⁵ In particular, members of the dom-

⁴Note that victimization is disproportionate here in the sense that it occurs at a higher rate than in the absence of hate motivation. This does not necessarily correspond to *statistically* disproportionate victimization (where members of a group constitute a larger percentage of crime victims than appears warranted by their percentage in the total population).

⁵The motivation for this assumption is that potential offenders sometimes expect to

inant group may hold negative beliefs about the characteristics of members of the disfavored group, e.g., their loyalty to a foreign nation, their honesty, or their propensity for participation in welfare programs. The more negative these characteristics are believed to be, the less are the expected costs from committing a crime against a member of the disfavored group.⁶

We use this framework to show that rational Bayesian inferences about group characteristics cannot, in general, be influenced by penalty enhancements. The basic intuition is that the penalty enhancements themselves will convey as much information about such characteristics as will disproportionate victimization. Then, we analyze how penalty enhancements may influence inferences about victims' characteristics in a setting where potential offenders are subject to a particular inferential bias that has been found to be empirically relevant in studies of attitudes towards crime victims. A series of psychological experiments (discussed in more detail in Section 2 below) demonstrates that, while they do not believe the world is perfectly just or that every outcome is fully deserved, individuals tend to systematically overestimate the degree to which people deserve the outcomes they receive. Moreover, subjects are found to distort other judgments in order to maintain their prior belief in the basic justness of the world. For example, when subjects observe a person suffering from a bad outcome that might be unjust and the subjects are unable to correct or ameliorate the situation, they tend to evaluate the person's intrinsic qualities more negatively than they otherwise would. In other words, when people observe others suffering through no fault of their own, they negatively revise their beliefs about the victims to a greater extent than Bayesian rationality warrants, so that victimization appears to be more "deserved" or, at least, less undeserved. This psychological phenomenon (generally known as "belief in a just world") has not attracted

incur guilt or shame from committing a crime, where the precise amount depends on the perceived characteristics of the victim. The offender expects to incur less psychological aversion (guilt) or social disapproval (shame) – i.e. less cost – to commit an offense against a person perceived to have negative characteristics (low moral worth) than against a person perceived to have positive characteristics (high moral worth). For example, defrauding a "liar" and assaulting a "bully" are less costly than committing the same crimes against a person without those negative characteristics. It would be still more costly to commit the crimes against one perceived to be the moral "pillar of the community."

⁶This general notion extends beyond the context of crime, and could also be applied to a variety of acts that manifest hostility but are not necessarily criminal.

much attention in economics,⁷ although it is closely related to other biases that have become well-known in the behavioral economics literature, such as cognitive dissonance (e.g. Akerlof and Dickens, 1982), optimism bias (e.g. Jolls, Sunstein and Thaler, 1998) and the fundamental attribution error (e.g. Dharmapala and McAdams, 2005).⁸

The basic intuition underlying the model in this paper can be summarized as follows: in a setting with uniform sanctions (i.e. no penalty enhancements), both haters and nonhaters within the dominant group commit crimes against the disfavored group, which suffers from disproportionate victimization. Observers (in particular, a new cohort of nonhaters within the dominant group) observe the rate of victimization, and know that crimes committed by haters are attributable to the offenders' preferences, while crimes committed by nonhaters are attributable to the offenders' beliefs about the victimized group's negative characteristics. Not all crimes are solved, so there are at least some crimes for which there is uncertainty about the offenders' motivations (i.e. about whether the crimes were hate-motivated or not). These crimes could be attributed either to nonhaters (whose motivation is related to the disfavored group's negative characteristics) or to haters (who are motivated by discriminatory preferences).

Psychological evidence suggests that most people view crime victimization caused by hate motivation as more unjust than victimization caused by negative characteristics of the victims (Rayburn, Mendoza and Davison, 2003).⁹ Thus, attributing the unsolved crimes to haters (which entails that the victims were targeted through no fault of their own) potentially conflicts with a belief in a just world (and hence involves some utility cost to the observer). If observers are subject to the inferential bias outlined above, they will trade off this disutility against the informational benefits of unbiased inference, and will tend to attribute fewer of the unsolved crimes to haters than would

⁷Benabou and Tirole (2006) is an exception. Note that in their model, individuals' belief in a just world is correct in equilibrium, and so is not a bias, as we characterize it here.

⁸For example, a likely mechanism for being optimistic about one's future is the belief that the world tends to reward deserving people like oneself, which is to say that outcomes are basically just. In contrast, it is harder to be optimistic if the world is arbitrary.

⁹Note that this is not a normative claim about the actual justice or injustice of disproportionate victimization (as made by Harel and Parchomovsky (1999)), but rather a *positive* claim about what people tend to view as just.

a rational Bayesian. However, because nonhaters' crimes are motivated by the victimized group's negative characteristics, the only way to reconcile this attribution with the observed rate of victimization is to revise (negatively) beliefs about the victimized group's characteristics. By underestimating the role of discriminatory preferences in the face of uncertainty, observers preserve (to some degree) their belief in the basic justness of the world. The revised beliefs about the victimized group's negative characteristics in turn raise the net benefits from crimes against that group. This can lead to additional crimes being committed against that group. Our central result is that penalty enhancements for hate crimes can reduce the social harms associated with these additional crimes. This result points to a social harm from disproportionate victimization that is potentially relevant to any consequentialist analysis of penalty enhancements.

It should be emphasized that the mechanism outlined above is driven by uncertainty about the motivations of offenders. In particular, we are *not* claiming that crimes that are known to be hate-motivated will lead to negative inferences about victims, e.g. through an inference that the victims must have done something to deserve such hatred. To the contrary, if one observes what is known to be a hate-motivated crime, there is no reason to change one's views about the characteristics of the victim; the discriminatory preferences of haters constitute a sufficient explanation for their crimes. Rather, the bias we explore arises when crimes are unsolved and the motive is ambiguous. The bias consists of attributing a larger proportion of *unsolved* crimes to nonhaters than would an unbiased Bayesian.

The crucial role of uncertainty leads to some additional insights (beyond our central point about penalty enhancements). First, any reduction in the level of uncertainty (such as an increase in the fraction of crimes that are solved) will reduce the scope of the inferential bias. When crimes are known to have been committed because of hate motivation, they cannot be attributed to nonhating offenders (and hence "explained" by the supposed negative characteristics of the victimized group). This highlights the potential importance of laws that force the revelation of hate motives (through e.g. inquiries related to whether penalty enhancements should be applied). Second, the provision of information about the role of hate motivation in the victimization of minority groups can ameliorate the inferential bias. Consistent with this notion is the observation that human rights organizations

opposed to hate crimes often reveal and disseminate information about the role of hate motivation in crimes against the victimized group. This information is intended to attribute victimization to hate motivation (as opposed, for example, to the negative characteristics of the victimized group). Thus, such publicity tends to counteract the inferential bias, by providing information that makes observers less likely to underestimate the extent of discriminatory preferences.

Another insight that emerges from our model is that the beneficial effect of penalty enhancements noted above applies only to penalties for hate-motivated crimes, rather than for all crimes against the disfavored group (see Proposition 3 below). This highlights an important difference between hate crime penalty enhancements and what are termed “vulnerability” penalty enhancements that are aimed at protecting certain particularly vulnerable groups. A common example is the enhancement of penalties for crimes against the elderly (e.g. Moskowitz and DeBoer, 1999: 41-42). These kinds of enhancements do not depend on the offenders’ motivation (e.g. do not require that the perpetrator is motivated by hatred of the elderly or selected a victim because of his or her advanced age) and are not reciprocal (in the sense that they do not apply if the victim is from any age group other than the elderly). Vulnerability enhancements can be straightforwardly understood within the standard economic model of crime: attacks against particularly vulnerable victims are less costly to perpetrators (e.g. in terms of possible resistance or retaliation by the victims) and may cause victims greater direct harm (e.g. if the elderly are more likely to be injured by an attack of given severity). On either of these grounds, enhanced penalties can be justified on optimal deterrence grounds.

On the other hand, hate crime penalty enhancements are typically characterized by a concern for offenders’ motivation, and by reciprocity: for any given statutory criteria, the enhancement attaches to a crime committed against *anyone* on the basis of that criteria. When race is the criterion, for example, anyone in the population, not just members of a particular race, can be the victim of a hate crime justifying the enhancement. These features cannot be readily explained within the standard economic model; however (as argued below) they are consistent with the framework developed in this paper. For example, reciprocity arises in our model because it is possible for multiple groups to be disproportionately victimized (in the sense that

victimization occurs at a higher rate than in the absence of hate motivation) simultaneously.

The paper proceeds as follows. The just world bias and some caveats are outlined in more detail in Section 2. Then, the basic model is presented in Section 3. A more general welfare analysis is discussed in Section 4, while Section 5 concludes.

2 Belief in a Just World as a Bias

Lerner (1965) and Lerner and Simmons (1966) first proposed the idea of belief in a just world (BJW) to explain the effect that the victimization of an innocent individual has on an observer. The resulting research literature is extensive, drawing on a number of different methodologies of psychology. However, to date, none of these experiments uses the distinctive methods of experimental economics, involving tests of behavior in settings where subjects are provided with monetary incentives. One may, for example, imagine subjects being exposed to an individual's victimization and then placed in a situation in which they have the opportunity to trade with the victim. We hope that such economic experiments will be carried out in the future (and that this paper highlights some of the important policy applications that would be at issue in such an exercise). In the current absence of such an approach, we provide a brief summary of the existing psychological evidence.¹⁰

In the pioneering experiment of Lerner and Simmons (1966), subjects viewed on a television what appeared to be a contemporaneous experiment on learning, in which a subject (actually a confederate of the experimenters) received extremely painful electric shocks for giving incorrect answers. After ten minutes, the experimenters asked subjects to evaluate this "victim." Before making their evaluation, however, the experimenters told the subjects either (1) that they would thereafter watch the same person in another ten minute session of the same experiment (the midpoint condition) or (2) that they would thereafter anonymously vote on whether the person would continue with the negative reinforcement experiment with electric shocks or be

¹⁰For more extensive reviews, see Lerner and Miller (1978), Lerner (1980), Maes (1998), Ross and Miller (2003), and Dharmapala, Garoupa and McAdams (2006).

moved to a positive reinforcement experiment with monetary rewards (the reward condition). In the latter reward condition, the result of the vote - which was always to move the victim into the reward scenario - was announced before the subjects evaluated her. The main result was that subjects evaluated the victim significantly more negatively in the midpoint condition than the reward condition. Lerner and Simmons inferred that the midpoint condition was more threatening to the subjects' sense of justice than the reward condition, because only in the latter could the subjects restore justice by ending the suffering and rewarding the victim for past suffering. Without that power to correct injustice, the subjects adjusted their views of the victim negatively to make her seem more deserving of her bad outcome.¹¹

Psychologists have found similar effects using a variety of more typical victims, from those suffering poverty in developing nations (Reichle and Schmitt, 1998), to cancer victims (Maes, 1994), and, most relevant for our purposes, crime victims (Wyer et al., 1985; see also Lincoln and Levinger (1972)). Moreover, the evidence does not rely entirely on self-reports. Hafer (2000a) uses a "modified Stroop task"¹² to measure the effects of exposure to perceived injustice on individuals, and found unique anxiety or stress over the challenge to one's BJW. Hafer (2000a) also found that those subjects' suffering greater anxiety from a challenge to their JWB appear to relieve that anxiety by derogating the victim, thus reducing the perceived injustice of her victimization. This tendency appears more pronounced among those with a stronger BJW (Hafer, 2002).

Belief in a just world might seem excessively naïve, but the point is not that people consciously believe that the world is perfectly just, but that they

¹¹One implication of this experiment is that, if observers are able to intervene in ways that help the victim, the negative inferences about the victim's characteristics may be ameliorated. However, in the application to hate crimes in this paper, it appears reasonable to assume that most individuals will not be in a position to help crime victims they do not personally know or encounter. It is worth noting, nonetheless, that the existence of organizations that, for instance, accept donations to help crime victims may work to counter negative inferences about victims.

¹²A Stroop effect is a delay in performing an identification task; other research shows that the delay reveals psychological stress. Typically, subjects are asked to identify the color of a word flashed briefly on a screen. "[P]eople take longer to identify the color of words that are associated with whatever is of emotional concern (i.e., whatever is threatening) than the color of neutral words." (Hafer, 2000a: 166).

subconsciously strive to interpret it as being as just as possible. Put differently, causal attributions for bad outcomes are complex and difficult. We would expect even rational Bayesians to make errors, either overestimating or underestimating the degree to which an individual is responsible for bad things that happen to him. But instead of these errors being randomly distributed around a mean that represents the correct causal attribution, the BJW suggests that, at least when individuals cannot act to correct the bad outcomes, errors are skewed towards over-attributing bad outcomes to the negative characteristics of the individuals who suffer them.

There are two possible explanations for the persistence of this bias (e.g. Lerner, 1998: 255). The first is that the just world belief is a rule of thumb for boundedly rational individuals. The heuristic is particularly strong in children (Piaget 1965: 260; Rubin and Peplau, 1975: 72).¹³ Adults outgrow the crudest form of the heuristic (Lerner, 1998: 267), but the rule of thumb continues to influence the process of inference. Indeed, the bias may even be functional in adults by increasing the perceived value of long-term planning (Hafer, 2000b), and so may have advantages that offset (or render imperceptible) its disadvantages. Lerner (1998) also argues that individuals derive utility from believing in a just world, or suffer increasing anxiety and stress from believing the world is increasingly unjust. People are therefore willing to trade off the utility of BJW against the informational benefits from unbiased inferences about the world. There is some evidence supporting the benefit of some trade-off, for instance in the case of bereavement (e.g. Bonanno *et al.*, 2002). Given either cause of the just world belief - as a rule of thumb for making complex attributions or as a way of coping with anxiety - the result in our terms is a just world bias.¹⁴

¹³Indeed, some evidence links children’s ability to delay gratification with their expectation of fairness or justice (see Long and Lerner (1974) and Mischel (1974)).

¹⁴An alternative strand of psychological research stresses individual heterogeneity in the extent of the belief in a just world, and constructs survey-based measures of this belief (Rubin and Peplau, 1973, 1975; see also Lerner, 1998). While there are differences between this approach and the “fundamental delusion” view (Lerner, 1980), they are not crucial for the points we make. Our model assumes that all non-hate-motivated individuals are subject to the just world bias to the same extent. However, even if the incidence of this bias varies across these individuals, the results would be essentially unaffected (unless those individuals who are most likely to commit crimes are immune to the just world bias, which appears unlikely). Thus, whether the just world belief is conceptualized as a “fundamental delusion” or as an individual-specific factor, we are justified in treating it

In one important respect, our model goes beyond the JWB as it is usually discussed in psychology. We focus on inferences about groups, although the JWB is typically discussed in terms of inferences about individuals. However, we believe that our extension is reasonable. In explaining how people reconcile the emotion driven bias with rationality, Lerner (1998) says that “the person who derogates a victim will generate a culturally plausible basis for that condemnation.” The perceived negative characteristics of a disfavored group to which the victim belongs naturally provides such a basis. Indeed, there is evidence suggesting that people are less likely to make negative inferences about individuals who are similar to them in some salient respect (e.g. Lerner and Agar, 1972). If differences between the individual and victim are sufficient to heighten the individual’s bias, then we believe it is a reasonable extension to posit that the individual will use such differences to “explain” his or her derogation of the victim.¹⁵ As an example, surveys of Americans during World War II suggests that they became more likely over this time to view Jews as wielding too much power in the United States. “Far from evoking sympathy, the Nazi persecutions apparently sparked a rise in anti-Semitism in this country.” (Selznick and Steinberg, 1969: 63).

The relevance of the just world belief and related psychological phenomena for the study of hate crimes has been recognized previously by Wang (1997; 1999). Her argument, however, is primarily that victims of hate crimes suffer greater psychological harms than victims of parallel non-hate crimes, because the former victims’ belief in a just world is more seriously impaired. She also argues (1997:129), as we do, that hate crimes will “promote[] prejudice against the [victim’s] group” even among those who do not hate the group. However, we specify a precise mechanism through which this effect

as a bias and our basic results hold.

¹⁵We also believe we are justified in ignoring the JWB in modeling the decision to offend outside the group context, even though the negative inferences apply to all crime victims. If an individual is randomly victimized, others may make negative inferences about her characteristics, and this may raise the benefits to future crimes against her. However, the probability that those who may commit those future crimes will (a) know of this particular individual’s victimization and (b) encounter this individual again in a setting where a crime may be committed, is negligible. In contrast, if negative inferences are made about the characteristics of an entire disfavored group, the probability of future criminals encountering a member of that group is relatively high. Thus, the increased crime effect could be argued to apply only to the latter case.

operates, and link the increased prejudice to increased future crime against the victimized group. In examining the effects of disproportionate victimization on the beliefs and behavior of potential offenders, our approach here is closely related to the approach of Dharmapala and McAdams (2005), which identifies the costs of hate speech by focusing on the behavior of potential offenders, rather than on the psychic harms suffered by victims.

3 Model

3.1 Basic Assumptions

Our model is based on the standard economic theory of law enforcement (as reviewed in Garoupa, 1997 and Polinsky and Shavell, 2000). We extend this setup to a two-period framework, where the set of potential offenders in period one is replaced by a new “generation” or cohort of potential offenders who enter the model in period two. Each of these (risk-neutral) potential offenders has the opportunity to commit up to one crime against each of two groups of equal size¹⁶ – the X ’s and the Y ’s. For simplicity, we assume that all potential offenders in both periods are X ’s, the same approach as in Dharmapala and Garoupa (2004).¹⁷ The illegal gains from committing the crime are represented by the variable b (as is conventional in the economic theory of enforcement, each individual’s b is unobservable, although the distribution is known). The distribution of offenders (X ’s) over b is given by the cumulative distribution function (cdf) $F(b)$ (note that $F(b)$ is the fraction of X ’s who receive benefits less than b from committing the crime).

Let the probability of detection of a crime against a member of group j be denoted by p_j , where $j = X, Y$. Assume also that this probability does not change over time. The sanction imposed in period i for a crime against a member of group j is denoted s_i^j . The complete set of sanctions is denoted by $\{s_1^X, s_1^Y, s_2^X, s_2^Y\}$.

¹⁶Different group sizes would have a scale effect, but would make no qualitative difference to the results of the model.

¹⁷Introducing crimes committed by Y ’s would complicate the model, but would not affect the basic results.

3.2 Rational Learning

In period 1, suppose that p_X and p_Y are both known to all potential offenders. Then, (risk-neutral) potential offenders commit the crime against X 's if $b \geq p_X s_1^X$. Bearing in mind that the groups are of equal size, the rate of victimization of X 's is:

$$r_{X1} = 1 - F(p_X s_1^X) \quad (1)$$

Potential offenders commit the crime against Y 's if $b \geq p_Y s_1^Y$. The fraction of X 's who commit the crime against Y 's is thus $(1 - F(p_Y s_1^Y))$ and the rate of victimization of Y 's is:

$$r_{Y1} = 1 - F(p_Y s_1^Y) \quad (2)$$

Suppose that there is no penalty enhancement, and sanctions are uniform (i.e. $s_1^X = s_1^Y$), but that Y 's are more "vulnerable," in the sense that $p_Y < p_X$ (for example, the police are less likely to investigate crimes against Y 's, or Y 's are less likely to report crimes to the authorities). Then, it follows straightforwardly that disparate victimization occurs (i.e. $r_{Y1} > r_{X1}$).

In period 2, suppose that a new cohort of potential offenders knows p_X but not p_Y . Potential offenders commit the crime against X 's if $b \geq p_X s_2^X$, so that the rate of victimization of X 's is:

$$r_{X2} = 1 - F(p_X s_2^X) \quad (3)$$

Let \widehat{p}_Y be the probability of detection inferred by the new cohort of potential offenders (who can observe the sanctions and rates of victimization that occurred in period 1). The crime level against Y 's depends on the inferred \widehat{p}_Y ; from Equation (2), this inference problem can be implicitly characterized by:

$$r_{Y1} = 1 - F(\widehat{p}_Y s_1^Y) \quad (4)$$

This straightforwardly implies that $\widehat{p}_Y = p_Y$. Thus, potential offenders commit the crime against Y 's if $b \geq p_Y s_2^Y$, and the rate of victimization is:

$$r_{Y2} = 1 - F(p_Y s_2^Y) \quad (5)$$

In this setup, rational learning implies that a higher rate of victimization of Y 's in period 1 will lead *ceteris paribus* to a lower inferred \widehat{p}_Y and hence to a

higher period-2 rate of victimization of Y 's. That is, crime in period 1 generates additional crime in period 2. However, penalty enhancement in period 1 (i.e. setting $s_1^X < s_1^Y$) cannot address this problem, as long as period-1 sanctions are observable to the new period-2 cohort.¹⁸ For instance, suppose s_1^Y were increased, holding everything else fixed. Then, r_{Y1} would fall, but the inference problem (Eq. 4 above) would lead to the same inference $\widehat{p}_Y = p_Y$. In essence, period-2 offenders realize that the reason the victimization rate for Y 's is low is because of the penalty enhancement, and they incorporate this into their inference. It follows that penalty enhancements for hate crimes cannot be explained by a dynamic rational learning setup of this kind (although penalty enhancements for vulnerable groups may still be optimal for purely static reasons - i.e. to reduce the harms from crimes in period 1).

3.3 Introducing Hate Motivation

Now, assume that members of group X can be partitioned into two (exhaustive and mutually exclusive) subgroups: those with hate motivation (in a sense defined more precisely below), denoted by X^H , and those without hate motivation, denoted by X^N .¹⁹ In each period, the former constitutes a fraction $\alpha \in (0, 1)$ of all X 's, and the latter a fraction $(1 - \alpha)$ (these proportions are assumed to be common knowledge). The illegal gains from committing the crime, represented by the variable b , differ across the two subsets of X . The distribution of non-hate-motivated offenders (X^N 's) over b is given by the cumulative distribution function (cdf) $F_N(b)$, independently of whether the victim of the crime is an X or a Y (note that $F_N(b)$ is the fraction of X^N 's who receive benefits less than b from committing the crime). The distribution of hate-motivated offenders (X^H 's) over the benefits from crimes against other X 's is also given by $F_N(b)$. However, the distribution of X^H 's over the benefits from crimes against Y 's is given by the cdf $F_H(b)$.

¹⁸This can be generalized to the case where sanctions are not observable, as long as no systematic errors are made by potential offenders in inferring the sanctions.

¹⁹The assumption that not all X 's are haters is intended to reflect reality. For the sake of additional realism, we introduce the assumption that haters do not think that hate crimes are unjust; however, our results would only be strengthened if we assumed that haters themselves subject to the just world bias.

The notion of hate motivation is captured by assuming that X^H 's derive greater benefits, *ceteris paribus*, from crimes against Y 's. Specifically, following Dharmapala and Garoupa (2004: 190), we assume that for a given b in the relevant range (in particular, high enough values of b such that the crime may be committed):

Assumption 1a: $F_N(b) > F_H(b)$,

i.e., the fraction of X^H 's who receive benefits less than b from crimes against X 's is greater than the corresponding fraction for crimes against Y 's. We also make the following assumptions about the probability density functions (pdf's):

Assumption 1b: For low values of b , $f_N(b) > f_H(b)$; for high values of b , $f_N(b) < f_H(b)$.

To ensure that the cdf's $F_H(b)$ and $F_N(b)$ are strictly monotonically increasing, it is also assumed that:

Assumption 2: For all b in the relevant range, $f_j(b) > 0$ for $j = N, H$.

Assumption 1a implies that (in the absence of penalty enhancements) Y 's will suffer disproportionate victimization as a result of the discriminatory preferences of X^H 's. This paper introduces another distinct source of disparate victimization - negative beliefs held by X^N 's about the characteristics of Y 's. These may induce X^N 's to disproportionately target Y 's, even in the absence of any intrinsic hate motivation. We extend the standard framework by assuming that the net benefit from a crime committed by a member of X^N depends not only on b , but also on group Y 's perceived "characteristics" (denoted by c).²⁰ These characteristics are assumed to bear on the (perceived) moral inappropriateness of the individual being criminally victimized, and correspond to the amount of guilt and shame the perpetrator expects to feel. The higher is c , the worse the group's characteristics are perceived to be, so the *net* benefit from the crime to a potential offender can be characterized as $(b + c)$ (i.e. a larger c for a particular group implies a greater net benefit of attacking a member of that group).²¹

²⁰For simplicity, it is assumed that X^N 's only have beliefs about group Y 's characteristics.

²¹It is intuitive to think of c as the variation in the guilt or shame from committing the offense. However, since there is no formal cost term in the model, it is easier, and analytically equivalent, to place c on the benefit side.

For the sake of simplicity, this variable c is assumed to only influence the behavior of X^N 's, and not that of X^H 's. The latter's hatred of Y 's is presumed to be independent of these perceived characteristics (or alternatively, the perception of Y 's characteristics among X^H 's can be viewed as being implicitly incorporated into the specification of the $F_H(b)$ function, and to be fixed across periods).

Without loss of generality, we normalize c in period 1 to zero, so that X^N 's in period 1 make their decisions about whether to commit the crime on the basis that $c = 0$.²² It is assumed that this first-period value of the characteristics is not observed by the new cohort of X^N 's who enter the model in period 2. In order to make their crime decisions, these individuals must infer the value of c (more generally, they must update some prior about c). Our focus is on inferences about the victimized group's intrinsic characteristics. In practice, it is possible that the inference from disparate victimization may instead be that members of the targeted group take greater risks (such as carrying large amounts of cash). This amounts, in effect, to an inference that the probability of detection (and hence the expected sanction) is lower when attacking such groups. In our analysis, we assume that the probability of detection and the expected sanction are common knowledge, so no updating of beliefs about these variables takes place. Even if the inference from disparate victimization is that members of the disfavored group take greater risks, this will lead to a higher level of crime, and so operate in the same direction as the effect we identify. We denote this inferred value by \hat{c} .²³

²²The normalization $c = 0$ is not crucial for the results. If observers in period 2 start with a prior that $c < 0$ (i.e. a belief that group Y has positive characteristics), then the updating following observation of the actual crime rate will be very similar to that when $c = 0$ (with the only difference being that the inferred \hat{c} may now be negative, albeit less favorable than the prior). Now consider the opposite case, where the prior is $c > 0$ (i.e. observers start with a prior belief that the outgroup has negative characteristics). Here, it is possible that observers expect more crime against Y 's than actually occurs, and so revise their priors in a favorable direction (i.e. they end up with more favorable beliefs about the Y 's characteristics). Even in this case, however, the just world bias (JWB) will reduce the extent of this favorable updating (i.e. \hat{c} will be more positive in the presence of the JWB, when all crime is attributed to nonhaters, than in the case of Bayesian inference, where only some of the crime were attributed to nonhaters). Thus, the JWB works in the same direction, regardless of the assumption about the prior about c .

²³This type of updating could also take place within a cohort. For simplicity, however, we assume that all members of a given cohort are symmetrically informed, and that updating

Define $p = p_X = p_Y < 1$ to be the probability of detection, which is assumed to be time invariant and independent of the victim's group. The sanction for a crime against a member of group j in period t is denoted by s_t^j where $j = X, Y$ and $t = 1, 2$. In addition, it is assumed that hate motivation is perfectly observable *ex post* by courts, so the sanction for a crime against a Y can differ depending on whether or not the perpetrator was motivated by hate. Given the assumptions made above, all X^H 's who have sufficiently large b that they commit the crime in equilibrium derive greater benefits from targeting Y 's. In this sense, all crimes by X^H 's against Y 's that are detected are revealed to be hate crimes (and subject to penalty enhancements, if they exist). Specifically, if an X^H commits a crime against a Y in period 1 and it is detected, then the sanction is s_1^{YH} . The complete set of sanctions is denoted by $\{s_1^X, s_1^{YH}, s_1^{YN}, s_2^X, s_2^{YH}, s_2^{YN}\}$. Note that the penalty enhancement for a hate crime in period 1 (the policy variable on which the analysis focuses) is $(s_1^{YH} - s_1^{YN})$.²⁴

The harm to an individual victim of a crime is denoted by $h > 0$ (and is independent of the victim's group and the perpetrator's motivation). This assumption, consistent with Dharmapala and Garoupa (2004), entails that the private harm to an individual victim from a hate crime is identical to that from an equivalent non-hate crime, so that we can *endogenously* derive disparities in the *social* harms that result from a pattern of discriminatory selection of victims.

3.4 Outcomes with Bayesian Inference

In this subsection, we briefly characterize the outcomes when individuals engage in Bayesian inference. In period 1, (risk-neutral) potential offenders commit the crime against X 's if $b \geq ps_1^X$. Bearing in mind that the groups are of equal size, the rate of victimization of X 's is:

$$r_{X1} = 1 - F_N(ps_1^X) \tag{6}$$

only occurs across cohorts.

²⁴This formulation is sufficiently general to include the case where penalty enhancements do not exist. The analysis below begins with the assumption of uniform sanctions ($s_1^X = s_1^{YH} = s_1^{YN}$ in period 1), before proceeding to analyze the consequences of penalty enhancements.

X^H 's commit the crime against Y 's if $b \geq ps_1^{YH}$, and X^N 's commit the crime against Y 's if $b \geq ps_1^{YN}$. The fraction of X 's who commit the crime against Y 's is thus $\alpha(1 - F_N(ps_1^{YH})) + (1 - \alpha)(1 - F_N(ps_1^{YN}))$ and the rate of victimization of Y 's is:

$$r_{Y1} = 1 - \alpha F_H(ps_1^{YH}) - (1 - \alpha)(1 - F_N(ps_1^{YN})) \quad (7)$$

Suppose that there is no penalty enhancement, and sanctions are uniform (i.e. $s_1^X = s_1^{YH} = s_1^{YN}$). Then, by Assumption 1a, it follows that disparate victimization occurs (i.e. $r_{Y1} > r_{X1}$). Disparate victimization, under these assumptions, involves a higher rate of victimization being suffered by Y 's than by X 's. However, a more general (and more analytically relevant) interpretation of the disparity is that Y 's suffer a higher rate of victimization than they would in the absence of hate motivation among those who commit offenses against them. In this setup, the two interpretations are equivalent (because of the assumption that, but for hate motivation, X 's and Y 's would be victimized at the same rates, given the same penalties), but the latter generalizes more readily to cases where multiple groups may experience hate-motivated victimization.²⁵

In period 2, potential offenders commit the crime against X 's if $b \geq ps_2^X$, while X^H 's commit the crime against Y 's if $b \geq ps_2^{YH}$. The crime level of X^N 's depends on their inferred \hat{c} ; from Equation (2), this inference problem can be implicitly characterized by:

$$1 - r_{Y1} = \alpha F_H(ps_1^{YH}) + (1 - \alpha)F_N(ps_1^{YN} - \hat{c}) \quad (8)$$

This straightforwardly implies that $\hat{c} = 0$. Thus, X^N 's commit the crime against Y 's if $b \geq ps_2^{YN}$. The rates of victimization in period 2 are $r_{X2} = 1 - F_N(ps_2^X)$ and $r_{Y2} = 1 - \alpha F_H(ps_2^{YH}) - (1 - \alpha)(1 - F_N(ps_2^{YN}))$.

Note that Equation (8) implies that the new cohort of X^N 's in period 2 infers \hat{c} only on the basis of the observed behavior of the previous cohort of X^N 's, not on that of X^H 's. This assumption captures the intuition that, in

²⁵Also, the empirical evidence cited above (e.g. Rayburn, Mendoza and Davison, 2003) seems to suggest that hate-motivated victimization, rather than different victimization rates across different groups *per se* are seen as unjust; thus, the former would be more likely than the latter to trigger the just world bias.

learning about the characteristics of the outgroup, non-hate-motivated members of the dominant group will not be influenced by the behavior of those who are known to hate the outgroup. Since X^H 's are known to hate Y 's, their attacks are correctly interpreted as stemming from discriminatory preferences, and so do not convey any new information about the characteristics of Y 's.

3.5 Introducing the Just World Bias

In the previous subsection, it was assumed that the new cohort of X^N 's in period 2 observes the first-period outcomes (in particular, the rates of victimization) and policies (in particular, the expected sanctions), and knows the preferences (i.e. the distributions $F_N(b)$ and $F_H(b)$). That is, these individuals are aware that X^H 's in period 1 have discriminatory preferences, and correctly attribute the higher victimization rate of Y 's to this hate motivation. However, this extra victimization of Y 's through no fault of their own may come into conflict with the desire of X^N 's to believe in a just world.

Rayburn, Mendoza and Davison (2003) offer empirical support for this claim. They find (p. 1063) that subjects exposed to a hate crime scenario view the perpetrator as being more culpable than the perpetrator in an otherwise identical non-hate crime. Similarly, even though subjects blame all crime victims to some degree (p. 1069), they rate the victim of a hate crime as less culpable than the victim of a non-hate crime (pp. 1062 – 63). These results support the assumption that most people view hate-motivated victimization as particularly unjust.²⁶ Moreover, the study supports the importance of the inferences made in the face of crime: if observers knew for certain that a crime was the product of discriminatory preferences, as the scenarios in the experiment made clear, they would blame the victim less. But because individuals do not know for certain the cause of most crimes (particularly if the law does not force the revelation of hate motives), the

²⁶Even so, the study does not prove that the subset of the dominant group that commits hate crimes shares this conception of justice. Thus, our model assumes that only those individuals in the dominant group who are not hate-motivated view disparate victimization as unjust (and are therefore subject to the just world bias). This assumption does not mean that haters are more “rational” in making inferences, but simply emphasizes that the results do not depend on haters viewing disparate victimization as unjust.

bias has scope to operate, allowing them to attribute the crime to the victim's negative characteristics and to blame the victim. By underestimating the role of discriminatory preferences in the face of uncertainty, observers preserve their belief in the basic justness of the world.

In this subsection, we introduce the assumption that X^N 's are subject to the just world bias (JWB) discussed in Section 2 above. In the first period, the outcomes are identical to those characterized in Section 3.4 above. In the second period, however, the just world bias will affect the inferences, and hence behavior, of the new cohort of X^N 's. Specifically, we formulate the just world bias as follows: X^N 's underestimate the extent of hate motivation on the part of X^H 's in period 1. An extreme way to capture this idea is to assume that the hate motivation is completely ignored, so that F_N is used instead of F_H in inferring \hat{c} .²⁷ Intuitively, all crime against Y 's in period 1 is attributed to nonhaters. Then, the inference problem is:

$$1 - r_{Y1} = \alpha F_N(ps_1^{YH}) + (1 - \alpha)F_N(ps_1^{YN} - \hat{c}) \quad (9)$$

i.e.

$$\alpha F_H(ps_1^{YH}) + (1 - \alpha)F_N(ps_1^{YN}) = \alpha F_N(ps_1^{YH}) + (1 - \alpha)F_N(ps_1^{YN} - \hat{c}) \quad (10)$$

The solution to this inference problem, and the main results that follow, are presented in the following subsection.

3.6 Results

Given Equation (10) above, it follows that:

Remark 1: The just world bias leads to biased inference; i.e. $\hat{c} > 0$.

Proof: Rearranging Equation (10):

$$F_N(ps_1^{YN} - \hat{c}) - F_N(ps_1^{YN}) = \frac{\alpha}{1 - \alpha}(F_H(ps_1^{YH}) - F_N(ps_1^{YH})) \quad (11)$$

Suppose that $\hat{c} \leq 0$. Then, $F_N(ps_1^{YN} - \hat{c}) \geq F_N(ps_1^{YN})$ (as F_N is strictly monotonically increasing), so that:

$$\frac{\alpha}{1 - \alpha}(F_H(ps_1^{YH}) - F_N(ps_1^{YH})) \geq 0 \quad (12)$$

²⁷Similar results would hold even if the underestimation of the extent of hatred were only partial.

But, this contradicts Assumption 1a (that $F_N(b) > F_H(b)$). So, $\hat{c} > 0$.

Thus, beliefs about group Y 's characteristics are more negative in period 2 than in period 1, as a result of the just world bias and the disproportionate victimization suffered by Y 's in period 1.²⁸ As the Bayesian inference is $\hat{c} = 0$, the extent of this bias can be measured straightforwardly by the magnitude of \hat{c} . Importantly, this measure of the extent of the bias depends on the period-1 sanction imposed on hate crimes against Y 's. An important caveat concerns the observability of the crimes that occur in period 1. The analysis assumes that the crime rates suffered by each group (r_{X1} and r_{Y1}) are observable, and that courts can perfectly observe *ex post* whether crimes are hate-motivated or not (i.e. whether they are committed by X^H 's or X^N 's). Thus, if the outcomes of all trials were observable to period 2 X^N 's, they would be able to directly observe the number of hate crimes (committed by X^H 's) and the number of non-hate-motivated crimes (committed by X^N 's) that occurred in period 1. Under such circumstances, it would appear that the underestimation of hate motivation involved in the just world bias requires not just biased inference, but also that observers ignore available information that contradicts their inferences. This point would be reinforced if courts could observe not merely whether a perpetrator is an X^H or X^N , but also her b . This would enable courts to publicize these b 's (perhaps by calibrating the penalty enhancement to the degree of hate motivation). If all crimes were solved, this would make the true distribution $F_H(b)$ directly observable. However, as long as there are some unsolved crimes (i.e. $p < 1$, as seems likely in reality), the result in Proposition 1 will continue to hold, even when the perpetrators of the solved crimes are known. For unsolved crimes, the identity of the perpetrator is unknown, and an observer subject

²⁸These revised beliefs about the disfavored group's characteristics may potentially be costly to those who engage in biased inference. For instance, these beliefs may induce suboptimally low interaction or trade with the disfavored group. In the experiments cited above, this was not an issue (the person who was evaluated by the subjects was a stranger whom they would never meet again), but it may be important in real-world settings. We do not explicitly model these costs. The implicit underlying assumption is that the desire to believe in a just world is traded off against these costs. As long as some utility is derived from the belief in a just world, then the biased inference and increased crime that we identify would continue to hold. Note, however, that because beliefs represent a direct source of utility in this setup (as e.g. in Akerlof and Dickens, 1982), we cannot impose the usual equilibrium condition that beliefs are correct in equilibrium. These wider conceptual issues are not addressed here, but would be an interesting subject for future research.

to the just world bias can underattribute these crimes to X^H 's (and hence infer a positive \hat{c}) without contradicting any directly observable information. Thus, if the motivations behind at least some subset of crimes is unknown, then there is scope for the just world bias to operate (i.e. for observers to underestimate the role of hate motivation and to overestimate the role of the victimized group's negative characteristics).

Proposition 1 For marginal changes in s_1^{YH} , \hat{c} is decreasing in s_1^{YH} .

Proof: Bearing in mind that the inferred \hat{c} adjusts to maintain the equality in Equation (10), it can be expressed as an identity:

$$\alpha F_H(ps_1^{YH}) + (1-\alpha)F_N(ps_1^{YN}) - \alpha F_N(ps_1^{YH}) - (1-\alpha)F_N(ps_1^{YN} - \hat{c}) \equiv 0 \quad (13)$$

Using the implicit-function rule:

$$\frac{d\hat{c}}{ds_1^{YH}} = -\frac{\alpha p(f_H(ps_1^{YH}) - f_N(ps_1^{YH}))}{(1-\alpha)f_N(ps_1^{YN} - \hat{c})} < 0 \quad (14)$$

by Assumption 1b.

Thus, the larger the sanction imposed on hate crimes against Y 's in period 1, the less pronounced is the tendency to attribute negative characteristics to Y 's. Note that as the argument in Proposition 1 involves holding all other variables fixed, a marginal increase in s_1^{YH} is equivalent to a marginal increase in the penalty enhancement ($s_1^{YH} - s_1^{YN}$).

Equation (10) suggests a straightforward intuition for why this result holds. Consider a given \hat{c} , and suppose that s_1^{YH} (or equivalently the penalty enhancement ($s_1^{YH} - s_1^{YN}$)) is increased marginally. The observer (a period-2 member of X^N who is subject to the just world bias) expects this change to lead to a relatively small decrease in crimes against Y 's (specifically, a fall of $f_N(ps_1^{YH})$). However, because the observer has underestimated the extent of hatred among X^H 's, she is "surprised" by the size of the fall in crime against Y 's (given by $f_H(ps_1^{YH})$, where $f_H(ps_1^{YH}) > f_N(ps_1^{YH})$ by Assumption 1b). Intuitively, this surprise occurs because the upper tail of the distribution of X^H 's is thicker than the observer believes. The observed change in the level of crime in response to the increase in the penalty enhancement is thus inconsistent with the given value of \hat{c} . This forces the observer to revise her estimate of \hat{c} downwards (i.e. the characteristics of the outgroup are viewed as being less negative).

Proposition 1 establishes that the bias due to the belief in a just world can be reduced by increases in the penalty enhancement for hate crimes. The importance of this bias stems from its role in influencing behavior, and in particular the crime rates in period 2. It can be shown that the just world bias generates additional crime in period 2 (i.e. crimes that would not have been committed under the Bayesian inference that $\hat{c} = 0$). The additional harm (or decrease in social welfare) from these extra crimes is denoted by A , and can be characterized as:

$$A = h(1 - \alpha)(F_N(ps_2^{YN}) - F_N(ps_2^{YN} - \hat{c})) \quad (15)$$

It follows that:

Proposition 2: A is strictly positive when the just world bias exists (i.e. when $\hat{c} > 0$), and is strictly increasing in \hat{c} (the extent of the just world bias).

Proof: Straightforwardly, $\hat{c} > 0$ implies $A > 0$, and

$$\frac{\partial A}{\partial \hat{c}} = h(1 - \alpha)f_N(ps_2^{YN} - \hat{c}) > 0 \quad (16)$$

Consequently, there is some amount of crime against Y 's in period 2 that is directly attributable to the just world bias. In this sense, the just world bias leads to disproportionate victimization *per se* being associated with harms (extra crimes) that even a consequentialist would recognize. Moreover, combining Propositions 1 and 2, it follows that the social harm from these extra crimes is increasing in the extent of the bias, and hence decreasing in the penalty enhancement for hate crimes. Thus, increasing the sanction imposed on hate-motivated crimes in period 1 reduces the extent to which observers in period 2 draw negative inferences about the disfavored group's characteristics. This, in turn, reduces the level of crime against the disfavored group in period 2.²⁹

There are a number of factors that may reinforce or counteract the basic effects we identify in Propositions 1 and 2. For example, recall the assumption

²⁹If the model were extended to multiple periods, the just world bias would potentially lead to extra crime in each period, and perhaps eventually to a situation where all X^N 's commit crime against Y 's. This is possible for some distributions of preferences. However, for more realistic distributions, a large fraction of X^N 's would be expected to have values of b sufficiently small that they would never commit the crime, even for the maximum possible \hat{c} . Then, there will be some limit to the escalation of crime.

above that X^H 's do not view disproportionate victimization of Y 's as unjust, and hence are not subject to the just world bias. While the focus has been entirely on the inferences made by X^N 's, the basic results would only be strengthened if X^H 's also committed additional crimes in period 2 due to the just world bias. Similarly, the results would be reinforced if the period-2 X^N 's underestimate the proportion of haters in the population, as well as the extent of their hatred.

The formulation in Equation (9) assumes that the new generation of X^N 's in period 2 correctly observes the sanctions that were in place in period 1. Thus, the observer recognizes that X^H 's faced a sanction s_1^{YH} (which potentially includes a penalty enhancement), even while underestimating the hate motivation of X^H 's. If period 2 X^N 's underestimate the sanction imposed on hate crimes in period 1, then this may reduce the degree of surprise occasioned by a marginal increase in the penalty enhancement. It is possible that this could reduce (or even reverse) the revision of \hat{c} described above. In addition, it may be the case that (while underestimating the extent of hate motivation among X^H 's, as assumed above) period 2 X^N 's attribute to X^H 's some private information that Y 's have negative characteristics. This would potentially "explain" the thickness in the tail of the distribution of X^H 's without necessarily revising \hat{c} downwards, and hence may reduce or reverse the effect in Proposition 1. However, all of these scenarios involve some misperception or cognitive bias in addition to the just world bias. The aim of the analysis above is to derive the effects attributable solely to the just world bias *per se* by abstracting from all other potential information asymmetries and cognitive biases.

Proposition 1 shows that marginal increases in s_1^{YH} reduce \hat{c} . On the other hand, increases in s_1^{YN} (the sanction for non-hate-motivated crimes against the disfavored group) do not have an unambiguous effect on the extent of the just world bias. Intuitively, an increase in s_1^{YN} will not generate a surprise for the observer, who has an accurate expectation of the consequences (a fall of $f_N(ps_1^{YN})$ in crimes against Y 's). Instead, the effect (if any) of changes in s_1^{YN} depends on the slope of the pdf f_N , and is not readily susceptible to an intuitive interpretation. Specifically:

Proposition 3 For marginal changes in s_1^{YN} , $\frac{d\hat{c}}{ds_1^{YN}} \begin{matrix} \geq \\ \leq \end{matrix} 0$ if $f_N(ps_1^{YN}) \begin{matrix} \geq \\ \leq \end{matrix} f_N(ps_1^{YN} - \hat{c})$.

Proof: Using the implicit-function rule (as in the proof of Proposition 1):

$$\frac{d\hat{c}}{ds_1^{YN}} = -\frac{p(f_N(ps_1^{YN}) - f_N(ps_1^{YN} - \hat{c}))}{f_N(ps_1^{YN} - \hat{c})} \quad (17)$$

from which the result follows straightforwardly.

This suggests that it is the penalty enhancement for hate-motivated crimes, rather than the sanction for crimes against Y 's *per se*, that has a clearer impact on the perception that the outgroup has negative characteristics. Proposition 3 can thus explain an important and distinctive feature of hate crime penalty enhancements, namely, that they are typically characterized by a concern for offenders' motivation (either in terms of animus against Y 's, or the discriminatory selection of Y 's as victims). This contrasts with "vulnerability" penalty enhancements that are aimed at protecting certain particularly vulnerable groups, such as the elderly (e.g. Moskowitz and De-Boer, 1999: 41-42)). The latter kinds of enhancements are imposed without inquiring into offenders' motivation. They are also not reciprocal (in the sense that they do not apply if the victim is from any age group other than the elderly).

Vulnerability enhancements can be straightforwardly understood within the standard economic model of crime: attacks against particularly vulnerable victims are less costly to perpetrators (e.g. in terms of possible resistance or retaliation by the victims) and may cause victims greater direct harm (e.g. if the elderly are more likely to be injured by an attack of given severity). On either of these grounds, enhanced penalties can be justified on optimal deterrence grounds.

On the other hand, hate crime penalty enhancements are typically characterized by a concern for offenders' motivation, and by reciprocity: for any given statutory criteria, the enhancement attaches to a crime committed against *anyone* on the basis of that criteria. When race is the criterion, for example, anyone in the population, not just members of a particular race, can be the victim of a hate crime justifying the enhancement. These features cannot be readily explained within the standard economic model; however they are consistent with the framework developed here. Proposition 3 shows that this framework gives rise to a concern with offenders' motivation (rather than simply with the group identity of the victim). Reciprocity arises in our

model because it is possible for multiple groups to simultaneously be disproportionately victimized. In particular, the model can be readily extended to the case where there is hate motivation against the other group among subsets of both X 's and Y 's, and where the just world bias operates to increase future crime against both X 's and Y 's. In such a setting, both X 's and Y 's suffer more crime than they would in the absence of hate motivation among offenders of the other group, and so each is subject to disparate victimization. Consequently, penalty enhancements can reduce the social harm from increased future crime against both groups. This argument also extends straightforwardly to cases where there are more than two groups.

Finally, it should be noted that in a general welfare analysis, fully equalizing victimization rates (as advocated by Harel and Parchomovsky (1999)) may not be optimal because of the costs of imposing sanctions. The results from a comprehensive welfare analysis are not qualitatively different from those in Dharmapala and Garoupa (2004), where the optimality of penalty enhancements depends crucially on enforcement costs and the probability density function. With no enforcement costs, maximal penalties are optimal, and the question of penalty enhancement is moot. Hence, enforcement costs determine the structure of efficient sanctions. The probability distribution function determines the exact direction and magnitude of penalty enhancements, where a higher marginal benefit in terms of deterrence caused by a penalty enhancement must be traded off against the higher marginal cost of imposing it.

The main novelties in the welfare analysis caused by the explicit consideration of the just world bias are the following. On the one hand, the just world bias increases social welfare, because criminal actions are vindicated by the belief that the victims deserve their victimization. The reduction in the perceived wrongfulness of the criminal act makes those who commit crimes better off. While this positive effect on social welfare may be viewed by some as ethically dubious, it is nonetheless worth noting from a purely utilitarian standpoint. On the other hand, the just world bias dilutes deterrence by raising the net benefits from crime; thus, it leads to a higher level of crime for any given expected sanction. In the solution to the social planner's problem, the "extra crime" effect derived in Propositions 1 and 2 leads to a greater willingness to impose penalty enhancements for hate crimes, other things equal. Therefore, if we take the view that this second effect is pri-

mary, the model justifies penalty enhancements for hate crimes for reasons quite similar to Dharmapala and Garoupa (2004), that is, not because hate crimes are exogenously more harmful, but because they generate more harm endogenously by increasing the number of committed crimes.

4 Discussion and Conclusion

The argument of this paper can be summarized as follows. The existence of animus by some members of one group against members of a second group leads to (at least some) potential offenders from the first group deriving greater benefits from crimes against members of the second group. This results in greater victimization of the second group (under a system of uniform sanctions), which conflicts with notions of justice that are widely held (at least among non-hate-motivated members of the first group). In order to maintain (to some degree) their belief in a just world, some members of the first group ascribe negative characteristics to the second group, making the latter's greater victimization appear more deserved (or less undeserved). This, in turn, raises their net benefits from committing crimes against the second group, and thus results in additional crimes against the latter. These additional crimes (or other manifestations of hostility) constitute a distinctive social harm associated with disproportionate victimization. Our model shows that this social harm can be reduced by imposing penalty enhancements for hate crimes against the second group in order to ameliorate the bias. Our analysis generalizes to reciprocal bias and bias among more than two groups. In particular, because the potential for bias is reciprocal, so that members of the second group can also commit hate-motivated crimes against members of the first group, our analysis suggests that the penalty enhancement should also be reciprocal.

One of the central challenges facing any theory of penalty enhancements (e.g. Blake, 2001) is whether and why they should apply to certain kinds of groups (e.g. racial or religious minorities) but not others (e.g. young males aged 18-25, the homeless or the poor). Posner (2001: 233) argues that “. . . advocates of enhanced punishment for hate crimes mean by the term [only] . . . crimes against members of groups for which they have a particular solicitude, such as blacks, Jews, and homosexuals.” We

do not claim to have dispelled Posner’s concern; interest group politics may indeed explain how legislatures define hate crimes in the real world. However, our analysis suggests some new, hitherto neglected, factors that are relevant to the issue. Most fundamentally, our model suggests the importance of determining empirically, not only when disproportionate victimization exists (because some offenders are motivated by hatred of the group), but when such victimization is perceived as unjust. If there is no hatred or no perception of injustice, then there is no scope for the bias to operate.³⁰ The existing psychological literature does not directly address many of these questions, so it is difficult to reach any firm conclusions. Nonetheless, we hope that our analysis clarifies the issues involved, and suggests significant areas for future inquiry.

Although we have used our model to reveal the benefit of hate crime penalty enhancements, one could derive alternative policy implications from our model. Most obviously, one could seek to offset the effect of the just world bias on crime not only by greater penalties but also by an enhanced probability of detection. There is some evidence that hate crime statutes may actually work this way in jurisdictions where the police department creates a special detective unit for investigating hate crimes that would not be investigated as seriously or at all were there no hate motivation (Bell, 2002). We have focused on penalty enhancements because that is more directly relevant to existing policy debates.

Less obviously, one could seek to offset the effect of the just world bias by policies suppressing the dissemination of information about crime. If members of the majority do not learn of crime, then the bias cannot cause them to believe such victimization is deserved, and thereby increase crime. We have not explored this avenue, however, because there are some significant costs to suppressing true information about crimes (e.g. Dharmapala and McAdams, 2005), for instance, the reduced ability of potential victims to take precautions against crime.³¹ In view of these costs, and because of

³⁰These issues are also important for the question of how hate crime statutes are framed. There are two types of these statutes (e.g. Wang, 1999). One is based on an “animus” model, focusing on the perpetrator’s hostility to the victim’s group. The other is the “discriminatory selection” model (upheld by the US Supreme Court in *Wisconsin v. Mitchell*, *supra* note 1), which punishes the selection of victims on the basis of group membership, independently of motivation.

³¹In addition, if it were known that information about crime is being suppressed, ob-

the difficulty of suppressing information about crimes (even if this were a desirable course of action), human rights and advocacy groups seeking to combat hate crimes typically do not pursue this approach. Rather, they seek to publicize information that makes clear the motivations of offenders - in particular, they emphasize the role of hate motivation (rather than victims' negative characteristics) in these crimes. This reduces the propensity of observers to attribute the crimes to nonhaters (and thus to revise their beliefs about the victimized group's characteristics).

A more general implication of the discussion above is that the discovery and provision of information about the motivations of offenders can limit the scope of the just world bias. The detection of crimes that occur in period 1, and the determination by the courts of whether the perpetrators were hate-motivated or not, will tend to publicize information about the extent of hate motivation among offenders. Thus, a larger probability of detection p will tend to reduce the extent of the just world bias, as there is less scope for inferences about victims' characteristics when more information is available concerning offenders' motives. Greater accuracy in courts' determination of offenders' motivations will also operate in the same direction. This suggests a novel expressive benefit of law enforcement, which is independent of whether or not penalty enhancements are imposed for hate-motivated crimes.³²

There are also a number of other possible informational effects of hate crime statutes that may be relevant to our argument. For example, individuals who are unaware that hate crimes against group X occur in their community may infer from the passage of a hate crime statute that such crimes in fact occur, and that members of group X suffer disproportionate victimization. This, in turn, may lead to negative beliefs about group X through the just world bias, an effect that may partially counteract the deterrent effect from the hate crime statute. This would not be relevant, of course, in situations where there has already been considerable publicity

servers would infer some expected level of crime, based on their prior beliefs. This would reintroduce the just world bias. Indeed, it may even exacerbate it, as there would be no directly observable information about offenders' motivation to constrain the inferential bias.

³²However, courts may have more incentive to accurately determine whether crimes are hate-motivated if the sanctions they administer include penalty enhancements for hate crimes.

about hate crimes. However, where there has been no such publicity (in particular, because there have been no hate crimes), purely symbolic legislation might be counterproductive.³³

As we have noted above, the effects of disproportionate victimization and the just world bias may not necessarily be manifested in the form of increased crime. While our model focuses on crime, the basic idea could equally well be illustrated by other manifestations of hostility, such as social or economic discrimination. Those individuals subject to the just world bias who are not willing to engage in violence against the disfavored group may become more likely to engage in acts of discrimination that (while less extreme than violent crimes) are socially harmful. When our results are viewed in this broader perspective, the social harm from hate crimes (and the concomitant social benefit from penalty enhancements) that we identify can occur even in the absence of an increase in crime.

Similarly, analogous social harms may be caused by other types of disproportionate impacts. For example, the just world bias may cause similar social harms to arise from racial profiling. Suppose that profiling by law enforcement personnel is due (at least partly) to animus towards the disfavored group. Then, observers would ascribe negative characteristics (such as a high probability of guilt) to the victimized group in order to reconcile these practices with belief in a just world. This could lead to an increased level of discrimination (or indeed to increased hate crimes) against the disfavored group.

Finally, our analysis may also be related to one of the most famous early scholarly studies of discrimination. Myrdal (1944) proposes the existence of the following “vicious circle” in ethnic relations. An exogenous increase in discrimination against the disfavored group leads to worse outcomes for that group. These outcomes are then viewed by members of the dominant group as evidence of the disfavored group’s intrinsic negative characteristics, leading

³³The passage of legislation can potentially convey a variety of types of information, beyond the hate crime context. New statutes may, for instance, reduce uncertainty about the content of the law, or reduce uncertainty concerning the preferences of others (in particular, concerning the extent to which people disapprove of criminal behavior, and the extent to which criminals benefit from crime). The overall effects are quite ambiguous. For a discussion of the informational consequences of legislative enactments, see Dharmapala and McAdams (2003).

to more discrimination, which causes even worse outcomes for the disfavored group, and to ever more negative inferences about the disfavored group's characteristics. This process (and the corresponding "virtuous circle" that can lead to reductions in discrimination) is difficult to reconcile with Bayesian rationality, as it requires that members of the dominant group naively ignore the effects of discrimination on the disfavored group's outcomes. However, the "vicious circle" is consistent with the existence of a just world bias, where the (presumably unjust) outcomes experienced by the disfavored group are attributed to the negative characteristics of that group, rather than to the discriminatory preferences of the dominant group.

5 References

Akerlof, G. and W. T. Dickens (1982) "The Economic Consequences of Cognitive Dissonance" *American Economic Review*, 72, 307-319.

Bell, J. (2002) *Policing Hatred: Law Enforcement, Civil Rights, and Hate Crime* New York: New York University Press.

Benabou, R. and J. Tirole (2006) "Belief in a Just World and Redistributive Politics" *Quarterly Journal of Economics*, forthcoming.

Blake, M. (2001) "Geeks and Monsters: Bias Crime and Social Identity" *Law and Philosophy*, 20, 121-139.

Bonanno, G., C. Wortman, D. Lehman, R. Tweed, M. Haring, J. Sonnega, D. Carr, and R. Nesse (2002) "Resilience to Loss and Chronic Grief: A Prospective Study from Preloss to 18-Months Postloss" *Journal of Personality and Social Psychology*, 83, 1150-1164.

Dharmapala, D. and Garoupa, N. (2004) "Penalty Enhancement for Hate Crimes: An Economic Analysis" *American Law and Economics Review*, 6, 185-207.

Dharmapala, D., N. Garoupa and R. McAdams (2006) "The Just World Bias and Hate Crime Statutes" University of Illinois Law and Economics Research Working Paper 06-11.

Dharmapala, D. and R. McAdams (2003) "The Condorcet Jury Theorem and the Expressive Function of Law: A Theory of Informative Law" *Ameri-*

can Law and Economics Review, 5, 2003, 1-31.

Dharmapala, D. and R. McAdams (2005) "Words that Kill? An Economic Model of the Influence of Speech on Behavior (with Particular Reference to Hate Speech)" *Journal of Legal Studies*, 34, 93-136.

Federal Bureau of Investigation (2005) *Hate Crime Statistics 2004*, Washington, DC: US Department of Justice.

Furnham, A. (1993) "The Just World Belief in Twelve Cultures" *Journal of Social Psychology* 133, 317-29.

Gan, L., R. C. Williams, III, and T. Wiseman (2004) "A Simple Model of Optimal Hate Crime Legislation" NBER Working Paper #10463.

Garoupa, N. (1997) "The Theory of Optimal Law Enforcement" *Journal of Economic Surveys*, 11, 267-295.

Grattet, R., V. Jenness, and T. R. Curry (1998) "The Homogenization and Differentiation of Hate Crime Law in the United States, 1978 to 1995: Innovation and Diffusion in the Criminalization of Bigotry" *American Sociological Review*, 63, 286-307.

Hafer C. (2000a) "Do Innocent Victims Threaten the Belief in a Just World? Evidence From a Modified Stroop Task" *Journal of Personality and Social Psychology*, 79, 165-173.

Hafer C. (2000b) "Investment in Long-Term Goals and Commitment to Just Means Drive the Need to Believe in a Just World" *Personality and Social Psychology Bulletin*, 26, 1059-1073.

Hafer C. (2002) "Why We Reject Innocent Victims" in M. Ross and D. Miller (eds.) *The Justice Motive in Everyday Life*, New York: Cambridge University Press, pp.109-126.

Harel, A. and G. Parchomovsky (1999) "On Hate and Equality" *Yale Law Journal*, 109, 507-539.

Harlow, C.W. (2005) "Hate Crime Reported by Victims and Police" *Bureau of Justice Statistics Special Report*, Washington, DC: US Department of Justice.

Hurd, H. and M. Moore (2004) "Punishing Hate and Prejudice" *Stanford Law Review*, 56, 1081-1146.

Jolls, C., C. R. Sunstein and R. H. Thaler (1998) "A Behavioral Approach to Law and Economics" *Stanford Law Review*, 50, 1471-1550.

Lawrence, F. (1999) *Punishing Hate: Bias Crime under American Law* Cambridge, MA: Harvard University Press.

Lerner, M. (1965) "Evaluation of Performance as a Function of Performer's Reward and Attractiveness" *Journal of Personality and Social Psychology*, 1, 355-360.

Lerner, M. (1980) *The Belief in a Just World: A Fundamental Delusion* New York, NY: Plenum Press.

Lerner, M. (1998) "The Two Forms of Belief in a Just World: Some Thoughts on Why and How People Care about Justice" in L. Montada and M. Lerner (eds.) *Responses to Victimization and Belief in a Just World* New York, NY: Plenum Press, pp. 247-269.

Lerner, M. and E. Agar (1972) "The Consequences of Perceived Similarity: Attraction and Rejection, Approach and Avoidance" *Journal of Experimental Research in Personality*, 6, 69-75.

Lerner, M. and D. Miller (1978) "Just World Research and the Attribution Process: Looking Back and Ahead" *Psychological Bulletin*, 85, 1030-1051.

Lerner, M. and C. H. Simmons (1966) "The Observer's Reaction to the 'Innocent Victim': Compassion or Rejection?" *Journal of Personality and Social Psychology*, 4, 203-210.

Lincoln, A. and G. Levinger (1972) "Observers' Evaluations of the Victim and the Attacker in an Aggressive Incident" *Journal of Personality and Social Psychology*, 22, 202-210.

Long, G.T. and M.J. Lerner (1974) "Deserving, the 'Personal Contract,' and Altruistic Behavior by Children" *Journal of Personality and Social Psychology*, 29, 551-56.

Maes, J. (1994) "Blaming the Victim: Belief in Control or Belief in Justice?" *Social Justice Research*, 7, 69-90.

Maes, J. (1998) "Eight Stages in the Development of Research on the Construct of a Belief in a Just World" in L. Montada and M. Lerner (eds.) *Responses to Victimization and Belief in a Just World* New York, NY: Plenum Press, pp.163-185.

- McAdams, R. (1995) "Cooperation and Conflict: The Economics of Group Status Production and Race Discrimination" *Harvard Law Review*, 108, 1003-1084.
- Mischel, W. (1974) "Processes in Delay of Gratification" in L. Berkowitz (ed.), *Advances in Experimental Social Psychology*, New York: Academic Press.
- Moskowitz, S. and M. J. DeBoer (1999) "When Silence Resounds: Clergy and the Requirement to Report Elder Abuse and Neglect" *DePaul Law Review*, 49, 1-83.
- Myrdal, G. (1944) *An American Dilemma: The Negro Problem and Modern Democracy* New York, NY: Harper and Row.
- Piaget, J. (1965) *The Moral Judgment of the Child*, New York: Free Press.
- Polinsky, A. M. and S. Shavell (2000) "The Economic Theory of Public Enforcement of Law" *Journal of Economic Literature*, 38, 45-76.
- Posner, R. (2001) *Frontiers of Legal Theory* Cambridge, MA: Harvard University Press.
- Rayburn, N. R., M. Mendoza and G. C. Davison (2003) "Bystanders' Perceptions of Perpetrators and Victims of Hate Crime" *Journal of Interpersonal Violence*, 18, 1055-1074.
- Reichle, B. and M. Schmitt (2002) "Helping and Rationalization as Alternative Strategies for Restoring the Belief in a Just World: Evidence from Longitudinal Change Analyses" in M. Ross and D. Miller (eds.) *The Justice Motive in Everyday Life*, New York: Cambridge University Press, pp.127-148.
- Ross, M. and D. Miller (eds.)(2002) *The Justice Motive in Everyday Life*, New York: Cambridge University Press.
- Rubin, Z. and L. A. Peplau (1973) "Belief in a Just World and Reactions to Another's Lot: A Study of Participants in the National Draft Lottery" *Journal of Social Issues*, 29, 73-93.
- Rubin, Z. and L. A. Peplau (1975) "Who Believes in a Just World?" *Journal of Social Issues*, 31, 65-89.
- Selznick, G.J. and S. Steinberg (1969) *The Tenacity of Prejudice*, New

York: Harper and Row.

Wang, L. (1997) "The Transforming Power of "Hate": Social Cognition Theory and the Harms of Bias-Related Crime" *Southern California Law Review*, 71, 47-135.

Wang, L. (1999) "The Complexities of 'Hate'" *Ohio State Law Journal*, 60, 799-900.

Wyer, R., G. Bodenhausen, and T. Gorman (1985) "Cognitive Mediators of Reactions to Rape" *Journal of Personality and Social Psychology*, 48, 324-338.