The Just World Bias and Hate Crime Statutes

Dhammika Dharmapala∗ Nuno Garoupa†
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, which may affect the net benefits from crime (or, more generally, from other hostile acts). Second, based on psychological evidence, it assumes that individuals may be subject to the “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 illustrate this argument by showing that disproportionate victimization of the disfavored group in the first period can lead to additional crime against that group in the second period (as offenders’ perceptions about the disfavored group’s characteristics become more negative due to the just world effect, and their net benefits from crime thereby increase). Our main result is that penalty enhancements can reduce the social harm due to these extra crimes (or, more generally, to other 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.

JEL Classification: K4.

Keywords: hate crimes, behavioral economics.
1 Introduction

Bias-motivated crimes against members of disfavored groups have attracted widespread attention in recent years. The issue has been the subject of a large legal 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 the penalty enhancements for hate crimes that have been enacted, for example, in the hate crime statutes of many US states (and upheld by the US Supreme Court in Wisconsin v. Mitchell).

In an important contribution to this debate, Harel and Parchomovsky (1999) argue that bias-motivated crimes lead to the disproportionate victimization of minorities. They adopt a nonconsequentialist victim-centered “fair protection paradigm.” Within their perspective, the inequality inherent in disproportionate victimization is in itself sufficient to justify government action (through, for instance, penalty enhancements for hate crimes) to equalize rates of victimization.

Dharmapala and Garoupa (2004) formalize this notion of disproportionate victimization by extending the economic model of optimal enforcement to the case where potential offenders have a bias motivation (i.e. derive greater benefits from crimes against members of a disfavored group outgroup than against members of their own group). 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 criminal benefits and the social costs of sanctions. Essentially, within a consequentialist (economic or utilitarian) framework, there is no distinctive harm that is caused by disparate victimization of a group, per se.

The aim of this paper is to revisit this issue of the distinctive harms associated with disparate victimization. While our starting point is the economic model of crime, we extend this framework by introducing the assumption that individuals (and, in particular, potential offenders) are subject to an inferential bias of a type that has been widely discussed in the psychological literature. A series of psychological experiments initiated by Lerner (1965)

\footnote{508 U.S. 476 (1993).}
and Lerner and Simmons (1966) suggests that people tend to hold the belief that the world is “just” (in the sense that people deserve the outcomes that they experience). More importantly, they appear to distort other judgments in order to maintain this belief (see also Lerner, 1980, 1998). These experiments involve subjects in the treatment group observing a person who is (for no particular reason) subjected to what appear to be electric shocks. These subjects tend to evaluate the person’s intrinsic qualities more negatively than do subjects in a control group that observed no shocks. Thus, when people observe others being victimized through no fault of their own, they do not revise their prior belief that the world is a just place; rather, they negatively revise their beliefs about the characteristics of the victims (to a greater extent than warranted by Bayesian inference), so that the latter’s victimizations appears more “deserved” (or at least less undeserved). It is argued in the psychology literature that this distortion occurs in order to maintain the belief in a just world. This “fundamental delusion” (Lerner, 1980) is widely discussed in psychology, but has not been analyzed extensively in economics.²

This paper builds on the model of Dharmapala and Garoupa (2004), where the population is divided into a dominant group and a disfavored group. Potential offenders from the dominant group are assumed to 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 the same, regardless of the group to which the victim belongs, this bias motivation will give rise to an equilibrium in which members of the disfavored group face a higher probability of victimization.

In this paper, we extend this framework by introducing a new variable: the beliefs that a potential offender holds about a potential crime victim that affect the potential offender’s benefits of committing a crime against him or her. These beliefs concern the perceived intrinsic value or “moral worth” of the individual one may victimize. The idea is that potential offenders expect to suffer some amount of costly guilt or shame from committing a crime, but that the precise amount depends on the perceived characteristics of the victim. The offender expects to incur less psychological aversion (guilt) or

²Benabou and Tirole (2005) is a recent exception.
social disapproval (shame) - i.e. less cost - to commit an offense against a person perceived to have negative characteristics (and hence low moral worth) than against a person perceived to have positive characteristics (and hence 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; more costly still would be to commit the crimes against one perceived to be the moral “pillar of the community.”

With this new variable, the disfavored group status of a victim affects crime not only because of the majority’s preferences but also because of their beliefs. Members of the dominant group may have negative beliefs about the characteristics of the disfavored group (e.g. their loyalty to a foreign nation, their acquisitiveness, or their propensity for participation in welfare programs). The more negative these characteristics are believed to be, the greater the net benefits from committing a crime against a member of the disfavored group. Of course, this 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. For example, it may be less costly to engage in discrimination - by e.g. avoiding social contact or economic interaction - against members of a group that is perceived to have negative characteristics. However, the paper focuses on the example of crime, both because the social harms from crime are incontrovertible, and because the effects of beliefs on crime can be conveniently modeled within the same formal setup that is used to analyze hate crimes.

The paper then illustrates its central claim by showing how differential victimization due to the dominant group’s preferences can influence individuals’ beliefs in a way that produces more crime (or other manifestations of hostility). We use a simple two-stage framework. In the first period, potential offenders choose whether or not to commit crimes, based on their net benefits from the crime (taking as given the perceived characteristics of the disfavored group) and on the expected sanction. As noted above, this will entail the disproportionate victimization of the disfavored group (in the absence of sufficiently large penalty enhancements). In the second period, a new generation of potential offenders enters the model. Members of this new generation observe the rates of victimization in period one, and update their priors about the disfavored group’s characteristics based on these observations. If members of this new generation were rational Bayesians, they would
correctly infer the same characteristics as in the first period. The level of crime in period two will thus depend on the same net benefits as in the first period, and on the expected sanction in period two. The disproportionate victimization that occurs in period one will not, in itself, affect the level of crime in period two.

Suppose, however, that members of the new cohort are subject to the just world bias in inference. We assume that most individuals will view differential victimization caused by a bias motivation - different preferences - as arbitrary and unjust, while differential victimization caused by negative characteristics of the disfavored group is just (or at least not unjust).\(^3\) As a result, a just world bias will cause members of the new cohort to over-attribute the observed differential victimization to the disfavored group’s negative characteristics and to underestimate the role of discriminatory preferences. In other words, to preserve their view that the world is basically just, they will infer, to a greater extent than is warranted by Bayesian rationality, that discriminatory victimization is deserved, being produced by the disfavored group’s negative characteristics. Thus, they will revise their beliefs about the disfavored group’s characteristics in a negative direction. This, in turn, will affect the net benefits from crimes against the disfavored group, and lead to a higher level of crime against it than in the first period, holding the expected sanction fixed. Thus, the disproportionate victimization in the first period can lead to additional crime (or other manifestations of hostility) in period two. This, we argue, is a potential social harm associated with disproportionate victimization that even a consequentialist must recognize.

The relevance of the just world bias and related psychological phenomena for the study of hate crimes has been recognized previously by Wang (1997). Her argument, however, is primarily that victims of hate crimes suffer greater psychological harms than victims of parallel crimes, because the former’s belief in a just world is more seriously impaired. She also discusses the effects of hate crimes on members of the nontargeted group (who do not themselves

\(^3\)Note that we are not making a normative claim about the actual justice or injustice of disproportionate victimization (as Harel and Parchomovsky (1999) do), but rather a positive claim about what people tend to view as just. This assumption may raise the question of whether those who commit hate crimes against a disfavored group would necessarily view the disproportionate victimization of that group as unjust. This is a valid concern, which is addressed in Section 2 below.
commit these crimes), and argues that they may derive psychological benefits from not being targets (e.g. through greater opportunities for “downward comparison”). In contrast, this paper analyzes the effects of disproportionate victimization on the beliefs and behavior of potential offenders. In spirit, it is closely related to the approach in Dharmapala and McAdams (2005), which also focuses on the behavior of potential offenders, rather than on the psychic harms suffered by victims.

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 The Just World Bias

As described above, the just world bias was proposed by Lerner (1965) and Lerner and Simmons (1966), based on a series of experiments that sought to isolate the impact on observers’ perceptions of witnessing the victimization of an innocent individual. The basic finding is that observers tend to derogate the victim, even when it is clear that the victim is not at fault. On the basis of these and subsequent results, Lerner (1980) attributes this derogation to observers’ belief in a just world, and describes this phenomenon as a “fundamental delusion” in human cognition. In terms of the language of economics, the essential idea is that individuals derive utility from maintaining their belief in a just world. They are willing to trade off this utility against the informational benefits from unbiased inferences about the characteristics of others. Thus, they make biased inferences about the characteristics of innocent victims, in order to preserve to some degree their belief in a just world.

An 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 individuals are subject to the just world bias to the same extent. However, even if this biased inferential process only applies to a
subset of individuals, the results would be essentially unaffected, unless those individuals who are most likely to commit hate crimes are systematically less likely to be subject to the just world bias. It appears that this latter condition does not hold - survey-based measures of the extent of belief in a just world are positively correlated with measures of right-wing authoritarianism (Lambert, Burroughs and Chasteen, 1998, p. 112), which is often thought to be associated with hostility to ethnic and other minorities. Thus, whether the just world bias is conceptualized as a “fundamental delusion” or as an individual-specific factor, our basic results continue to hold.

In one important respect, our model goes beyond the just world bias as it is usually discussed in psychology. We focus on inferences about groups, although the just world bias is typically discussed in terms of inferences about individuals. However, we believe that our extension is a reasonable one. First, there is some 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). Second, even if the negative inferences apply to all crime victims, and not just to victims of hate crimes, there are some important differences between the two contexts. 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 are relatively high. Thus, the increased crime effect could be argued to apply only to the latter case.

Our focus is on inferences about the victimized group’s intrinsic characteristics. We contend that individuals will infer from greater victimization against a group that the group has negative characteristics that make its criminal victimization more deserved, or at least less undeserved. 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 de-
tection 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. Note, however, that the classic experiments on the just world bias were specifically designed to preclude extraneous inferences, to the extent possible. Thus, the experimental evidence cited above seems to suggest that there will be some revision of beliefs about the targeted group’s intrinsic worth.

The biased inference assumed by the model requires that the observed behavior (in this case, the disproportionate victimization of the disfavored group) is viewed as being unjust. Rayburn, Mendoza and Davison (2003) find that subjects exposed to a hate crime scenario tend to view the perpetrator as being more culpable than the perpetrator in a similar non-hate crime scenario. This would be consistent with the existence of a widespread perception that hate-motivated victimization is unjust. Even so, however, this does not necessarily imply that the subset of the dominant group that commits hate crimes (or is on the margin of doing so) shares this view. Our model assumes implicitly that all individuals in the dominant group share a common conception of justice, and revise their beliefs in an identical way. If those who are most likely to commit hate crimes do not view disproportionate victimization as unjust, however, they will not revise their beliefs in a negative direction, and hence may not commit additional hate crimes. In other words, the biased negative inference may be restricted to those sections of the dominant group that are least likely to commit hate crimes, and their changed beliefs may not be sufficient to tip them over into criminal behavior. Even if they are not willing to engage in violence against the disfavored group, however, it may nonetheless be the case that they become more likely to engage in discrimination, or other acts that (while less extreme than violent crimes) are socially harmful. As was noted above, our model focuses on the effect of the just world bias in increasing hate crime, but the basic idea could equally well be illustrated by other manifestations of hostility, such as discrimination. When our results are viewed in this broader perspective, the social harm from disproportionate victimization that we identify can occur even in the absence of an increase in hate crime.

Finally, the revised beliefs about the disfavored group’s characteristics may
potentially be costly to the individuals who engage in biased inference. For instance, they 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), one 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.

3 Model

3.1 Basic Assumptions

Our model is based on the standard economic model of law enforcement (as reviewed in Garoupa (1997) and Polinsky and Shavell (2000)). We extend this setup to a two-period framework, where there is a set of potential offenders in period one, with a new “generation” or cohort of potential offenders entering in period two. These (risk-neutral) potential offenders have the opportunity to each commit up to one crime against each of two groups of equal size – the Xs and the Ys (for simplicity, we assume all potential offenders are Xs in both periods, the same approach as in Dharmapala and Garoupa, 2004).\footnote{Different group sizes would have a scale effect, but would make no qualitative difference to the results of the model.} Define $p$ to be the probability of detection, which is assumed to be time

\footnote{More realistically, we can assume that the Xs are a partition or a subset of a larger group of individuals who do not commit hate groups against Ys; therefore, they can be ignored in the same way that we ignore crimes committed by Xs against Xs. As long as hate crimes are observable (that is, crimes committed with a hate motivation as distinct from crimes committed without a hate motivation), the analysis is unaffected.}
invariant and independent of the victim’s group, and \( s_{jt} \) the sanction enforced for crimes committed against victims of group \( j = X,Y \) at time \( t = 1, 2 \). The set of sanctions is described by \{ \( s_{X1}, s_{Y1}, s_{X2}, s_{Y2} \) \}. As in Dharmapala and Garoupa (2004), we will treat the probability as an exogenous variable determined in the context of general law enforcement, and will concentrate on the sanctions for bias crimes.

We use \( h \) to designate the social damage caused by the underlying offense, committed against the \( X \)'s and the \( Y \)'s. 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. That is, our results show that even if a particular hate crime causes the same private harm to its victim as an identical non-hate crime, a pattern of crimes disproportionately targeting an identifiable group causes greater social harm than does a pattern of crimes where targets are randomly chosen. This endogenous disparity in social harms is generated rather than imposed by our modeling technique.

Let \( F_X \) be the cumulative distribution function (cdf) of gains to potential offenders from crimes against \( X \)'s, and \( F_Y \) be the distribution of gains to potential offenders from crimes against \( Y \)'s. Note that \( F_j(b) \), for \( j = X,Y \), is the fraction of potential offenders who receive benefits less than \( b \) from a crime against \( j = X,Y \). The notion of bias motivation is captured by assuming that potential offenders (who are by assumption all members of group \( X \)) derive greater benefits, ceteris paribus, from crimes against \( Y \)'s. Following Dharmapala and Garoupa (2004, p. 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:** \( F_X(b) > F_Y(b) \)

i.e., the fraction of potential offenders 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 \( b \), \( f_X(b) > f_Y(b) \); for high values of \( b \), \( f_X(b) < f_Y(b) \)
In this formulation, all $X$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 $X$s against $Y$s in our model could be characterized as hate crimes (and subject to penalty enhancements, if they exist). Of course, this may seem to be an excessively broad definition of hate crimes. However, this assumption is made only for analytical convenience. In a more realistic model, we could divide potential offenders from group $X$ into two types, one of which derives greater benefits from targeting $Y$s, while the other receives the same benefits from the crime, regardless of the victim’s group. Then, even though both types may commit crimes against $Y$s in equilibrium, only the crimes of the first type would be classified as hate crimes.\footnote{Of course, courts would have to be able to distinguish between the two types \textit{ex post} in order to impose any penalty enhancements.}

It is also assumed that:

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

This ensures that the cdf’s $F_X(b)$ and $F_Y(b)$ are invertible. The variable $b$ represents the illegal gain that is conventional in the economic theory of enforcement.

In the introduction we emphasized the distinction between preferences and beliefs, in the sense of disproportionate victimization being the product of the dominant group members’ (a) discriminatory preferences or (b) negative beliefs about the characteristics of the minority. So far, we have characterized (a); we now introduce (b). We extend the framework by assuming that the net benefit from a crime also depends on the group’s “characteristics” (denoted $c$). These characteristics are assumed to bear on the (perceived) moral inappropriateness of the individual being criminally victimized, which corresponds 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 $(b + c)$.$^7$

Hence, a larger $c$ for a particular group implies a greater net benefit of attacking a member of that group. Because potential offenders are all members...
of group X, the characteristics of group X are irrelevant, and assumed to be 0; thus, the focus here is on potential offenders’ (or more generally, group X’s) beliefs about group Y’s characteristics. Without loss of generality, we normalize c in period 1 to zero, so that Xs in period 1 make their decisions about whether to commit the crime on the basis that c = 0. It is assumed that this first-period value of the characteristics is not observed by the new cohort of X’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); we denote this inferred value by \( \hat{c} \).

3.2 Outcomes with Bayesian Inference

In this subsection, we briefly characterize the outcomes when all individuals engage in Bayesian inference. In period 1, (risk-neutral) potential offenders will commit the crime against Xs if \( b \geq ps_{X1} \), while they commit the crime against Ys if \( b \geq ps_{Y1} \). That is, bearing in mind that the groups are of equal size, a fraction \( r_{X1} = 1 - F_X(ps_{X1}) \) of Xs are victimized in period one, while a fraction \( r_{Y1} = 1 - F_Y(ps_{Y1}) \) of Ys are victimized in period one. Thus, when \( s_{X1} = s_{Y1} \) (same penalty regardless of victimization), disproportionate victimization occurs (i.e. \( r_{Y1} > r_{X1} \)) because benefits are higher from crimes against Ys (by Assumption 1a).

In period 2, the new generation of Xs must infer c. Given that all the other variables in the model are observed, and that \( F_Y \) is invertible and unique (Assumption 2), then rational inference is straightforward. Recalling that \( r_{Y1} = 1 - F_Y(ps_{Y1}) \), we have \( \hat{c} = ps_{Y1} - F_Y^{-1}(1 - r_{Y1}) = 0 \). Thus, in period 2, potential offenders will commit the crime against Xs if \( b \geq ps_{X2} \), while they commit the crime against Ys if \( b \geq ps_{Y2} \). The rates of victimization in period 2 are \( r_{X2} = 1 - F_X(ps_{X2}) \) and \( r_{Y2} = 1 - F_Y(ps_{Y2}) \).

As noted above, Harel and Parchomovsky (1999) argue that sentencing policy should be used to equalize rates of victimization across groups. Thus we define:

**Definition:** Harel-Parchomovsky sentencing is defined as a set of sanctions \( \langle s_{X1}^{HP}, s_{Y1}^{HP}, s_{X2}^{HP}, s_{Y2}^{HP} \rangle \) such that \( F_X(ps_{X1}^{HP}) = F_Y(ps_{Y1}^{HP}) \) and \( F_X(ps_{X2}^{HP}) = F_Y(ps_{Y2}^{HP}) \).

That is, Harel-Parchomovsky sentencing leads to identical rates of victim-
ization across the two groups in any given period. Note that, given Assumption 1a, \( (s^{HP}_{X1}, s^{HP}_{Y1}, s^{HP}_{X2}, s^{HP}_{Y2}) \) would entail that \( s_{X1} < s_{Y1} \) and \( s_{X2} < s_{Y2} \) (i.e. penalty enhancement for crimes against victims belonging to group \( Y \)).

### 3.3 Introducing the Just World Bias

In the previous subsection, it was assumed that the new cohort of \( X \)s observes the first-period outcomes (in particular, the rates of victimization), policies (in particular, the expected sanctions), and preferences (i.e. the distributions \( F_X(b) \) and \( F_Y(b) \)). That is, these individuals are aware that potential offenders in period 1 were biased against \( Y \)s, and correctly attribute the higher victimization rate of \( Y \)s to this bias motivation. However, this extra victimization of \( Y \)s through no fault of their own may come into conflict with the desire of \( X \)s to believe in a just world. This creates the possibility for biased inference of the type discussed in Section 2 above. In this subsection, we analyze the outcomes under the assumption that individuals are subject to the just world bias.

In the first period, the outcomes are identical to those characterized in Section 3.2 above. In the second period, however, the just world bias will affect the inferences, and hence behavior, of the new cohort. There are many possible ways to translate the psychological notion of the just world bias into an economic model. Specifically, we formulate the just world bias as follows, in a way that is natural and straightforward in this particular context. The new generation of \( X \)s infrs \( \hat{c} \) without taking into account the period-one offenders’ bias against \( Y \)s. That is, while the appropriate distribution of benefits to use in the inference problem is \( F_Y \) (which incorporates the bias motivation), they instead use \( F_X \) (which ignores the bias motivation). Then, the inference is:

\[
\hat{c} = p s_{Y1} - F^{-1}_X (1 - r_{Y1})
\]  

(1)

Note that this formulation assumes that the new generation observes the sanctions that were in place in period 1. An alternative would be to assume that the sanctions are unobserved, and that the new generation naively assumes that the sanction for crimes against group \( Y \) is identical. Of course, this would make no difference in the case where there are no penalty enhancements. If penalty enhancements existed in period 1 (i.e. \( s_{Y1} > s_{X1} \)),...
then the new cohort would underestimate the expected sanction for crimes against Ys. This would reduce the extent of the just world bias, and make penalty enhancements a more powerful tool for combatting hate crimes than in our model (indeed, the just world bias would be eliminated by Harel-Parchomovsky sentencing). Our assumptions, on the other hand, focus solely on the just world bias per se by abstracting from all other information asymmetries and cognitive biases.

Given the assumptions above, it follows that:

**Remark 1**  The inference satisfies \( \hat{c} > 0 \).

**Proof** By Assumption 1a, \( F_X(b) > F_Y(b) \) in the relevant range. Therefore \( F_X^{-1}(b) < F_Y^{-1}(b) \) in the relevant range. Rearranging, we can write \( \hat{c} = ps_{Y1} - F_X^{-1}(F_Y(ps_{Y1})) \). We know that \( F_Y^{-1}(F_Y(ps_{Y1})) = ps_{Y1} \); as \( F_X^{-1}(b) < F_Y^{-1}(b) \), \( F_X^{-1}(F_Y(ps_{Y1})) < F_Y^{-1}(F_Y(ps_{Y1})) \) i.e. \( F_X^{-1}(F_Y(ps_{Y1})) < ps_{Y1} \) Therefore, it must be true that \( \hat{c} > 0 \).

Thus, the just world bias, as we have formulated it, always leads to biased inference. In particular, beliefs about group Y’s characteristics are more negative in period 2 than in period 1, as a result of the disproportionate victimization suffered by Ys in period 1. Importantly, the extent of this bias depends on the period-1 sanction on crimes against Ys:

**Proposition 1** The estimated \( \hat{c} \) under the just world bias is decreasing in \( s_{Y1} \).

**Proof:** Notice that if \( F_X(b) > F_Y(b) \) in the relevant range, then it must be the case that \( f_X(b) > f_Y(b) \) for low values of \( b \) but it should be the case that \( f_X(b) < f_Y(b) \) for the highest values of \( b \), and there will be a switching point such that \( f_X(b) = f_Y(b) \). In the particular case that \( F_Y(ps_{Y1}) = F_X(ps_{Y1} - \hat{c}) \), we should have \( f_Y(ps_{Y1}) > f_X(ps_{Y1} - \hat{c}) \) (because if the cumulative distribution function takes the same value at that point, it must be the case we must be to the right of the switching point in the distribution). Hence, we can easily show:

\[
\frac{\partial \hat{c}}{\partial s_{Y1}} = p - \frac{f_Y(ps_{Y1})}{f_X(ps_{Y1} - \hat{c})} \tag{2}
\]

As \( f_Y(ps_{Y1}) > f_X(ps_{Y1} - \hat{c}) \) in the relevant range, \( \frac{f_Y(ps_{Y1})}{f_X(ps_{Y1} - \hat{c})} > 1 \) and so \( \frac{\partial \hat{c}}{\partial s_{Y1}} < 0 \). QED

The immediate consequence of Proposition 1 is that, in period two, of-
fenders will commit the crime against Ys if $b + \hat{c} \geq ps_{Y2}$, and $r_{Y2} = (1 - F_Y(ps_{Y2} - \hat{c}))$ of Ys are victimized in period two. Consequently, there is some amount of crime against Ys 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 have to recognize.\(^8\)

The harm from these extra crimes, denoted $H$, can be characterized as follows:

**Proposition 2** (i) The social harm $H$ from the additional crimes attributable to the just world bias is: $h[F_Y(ps_{Y2}) - F_Y(ps_{Y2} - \hat{c})] > 0$ when $\hat{c} > 0$.

(ii) $H$ is increasing in $\hat{c}$.

**Proof** (i) The social harm is the harm $h$ from a crime, aggregated over all the extra crimes that are attributable to the biased inference - i.e.

$$H = h \int_{ps_{Y2} - \hat{c}}^{ps_{Y2}} f_Y(b)db = h[F_Y(ps_{Y2}) - F_Y(ps_{Y2} - \hat{c})]$$

(3)

Straightforwardly, $H > 0$ when $\hat{c} > 0$.

(ii) Consider:

$$\frac{\partial H}{\partial \hat{c}} = h[0 + f_Y(ps_{Y2} - \hat{c})] > 0$$

(4)

(by Assumption 2). Thus, $H$ is increasing in $\hat{c}$.

Thus, the social harm from disproportionate victimization is larger the greater the bias in inference due to a belief in a just world. Moreover, combining Propositions 1 and 2, it follows that an increase in $s_{Y1}$ (the sanction for crimes against Ys in period 1) will reduce $\hat{c}$, and thus reduce $H$: i.e. enhancing the penalty for bias-motivated crimes infinitesimally will reduce the prevalence of negative beliefs about the disfavored group, and thus reduce the social harm from the just world bias and disproportionate victimization.

\(^8\)Our model only considers two periods. However, if it were extended to multiple periods, this effect would potentially lead to extra crime in each period, and perhaps eventually to a situation where all Xs commit crime against Ys. This is possible for some distributions of preferences. However, for more realistic distributions, a large fraction of Xs 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.
Identical victimization would require $s_{X1} = s_{X2}$ (crimes against $X$s should be punished in the same way in both periods), $s_{X2} < s_{Y2}$ (penalty enhancement for victims $Y$), in particular $F_X(p_{sX2}) = F_Y(p_{sY2} - \hat{c})$ and $s_{Y1} < s_{Y2}$ (penalty enhancement for victims $Y$ more substantial in period-two than in period-one), in particular. If the government also aims at identical victimization across time as well as across groups, a third restriction must be imposed, so that $F_Y(p_{sY1}) = F_Y(p_{sY2} - \hat{c})$. As a consequence, identical victimization requires the following enforcement policy: $s_{X1} = s_{X2} < s_{Y1} < s_{Y2}$.

Recall that we defined above a set of Harel-Parchomovsky sanctions that ensure identical victimization across groups in each time period, assuming Bayesian inference. Given the just world bias, this definition has to be modified for period 2, with period 2 sanctions set so that $F_X(p_{sX2}^{HP}) = F_Y(p_{sY2}^{HP} - \hat{c})$. While it is possible to achieve identical victimization using these sanctions, it should be noted that we have so far said nothing about the social costs of sanctions. Thus, it is not being claimed that Harel-Parchomovsky sanctions will necessarily be socially optimal from an efficiency standpoint (see the discussion in Section 4 below).

Furthermore, even if Harel-Parchomovsky sanctions are used to equate rates of victimization in period 1, this does not imply that the just world bias in inference will be eliminated:

**Remark 2** Under a regime of Harel-Parchomovsky sanctions, $\hat{c} = p[s_{Y1}^{HP} - s_{X1}^{HP}] > 0$.

**Proof** By definition, $F_X(p_{sX1}^{HP}) = F_Y(p_{sY1}^{HP})$. Thus, $\hat{c} = p s_{Y1} - F_X^{-1}(1 - r_{Y1}) = p s_{Y1} - F_X^{-1}(F_X(p_{sX1}^{HP})) = p[s_{Y1}^{HP} - s_{X1}^{HP}]$. This is strictly positive because Assumption 1a implies that $s_{Y1}^{HP} > s_{X1}^{HP}$.

The intuition here is that, even though victimization rates are identical, observers know that sanctions are larger for crimes against $Y$s. Thus, they can infer that disproportionate victimization would occur, in the event that sanctions were uniform. This leads to some degree of just world bias. However, it should be recalled from Propositions 1 and 2 above that, even though the just world bias cannot be eliminated, a small increase in $s_{Y1}$ will always reduce the bias in inference, and lower the associated social harm.
4 A More General Welfare Analysis

It is important to note that in a comprehensive welfare analysis, fully equalizing victimization rates may not be optimal, because of the costs of sanctions. The results from a comprehensive welfare analysis are not qualitatively different from those in Dharmapala and Garoupa (2004), where penalty enhancements crucially rely 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 shape 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. Here we sketch the main novelties in the welfare analysis caused by the explicit consideration of the just world bias, referring the reader to Dharmapala and Garoupa (2004) for the analytical derivations.

In this framework, we should start by noting the following important points. 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.

These effects can be easily distinguished if we express aggregate social welfare in the form that is conventional in the literature:

\[
W = \int_{\bar{Y}_X}^{\bar{b}_X} (b - h - pt(s_{X1}))dF_X(b) + \int_{\bar{Y}_X}^{\bar{b}_X} (b - h - pt(s_{X2}))dF_X(b) \\
+ \int_{\bar{Y}_Y}^{\bar{b}_Y} (b - h - pt(s_{Y1}))dF_Y(b) + \int_{\bar{Y}_Y}^{\bar{b}_Y} (b + c - h - pt(s_{Y2}))dF_Y(b)
\]

where \(t(.)' > 0\) and \(t(.)'' > 0\), \(\bar{b}_X\) and \(\bar{b}_Y\) are the superior limits of the distribution of illegal gains, and the discount factor
from the first to the second period is assumed to be one. The two effects of the bias can be immediately seen in the last integral. It appears with a positive sign inside the integrand due to the utility effect on criminals, and with a negative sign on the lower limit of integration because it reduces the expected opportunity cost of crime (and induces more crime).

Whereas an intertemporal effect is absent from all the other three sanctions (those for crimes against $X$s because there is no *just world bias* and with respect to $s_{Y2}$ because it is the last period), we should note that the determination of $s_{Y1}$ is subject to the effect on the third and fourth integrals in the social welfare expression. An increase in $s_{Y1}$ leads to a reduction of crimes against $Y$s in period one. By Proposition 1, this leads to a less substantial overestimation of $c$, thereby reducing crime against $Y$s in period two. This intertemporal effect should imply that a penalty enhancement for crimes against $Y$s in the first period could be justified as a mechanism to deter future crime.

## 5 Discussion and Conclusion

The argument of this paper can be summarized as follows. The existence of animus against certain disfavored groups leads to (at least some subset of) potential offenders from the dominant group deriving greater benefits from crimes against members of the disfavored group. This results in disproportionate victimization of the disfavored group (under a system of uniform sanctions), which conflicts with widely-held notions of justice. In order to maintain (to at least some degree) their belief in a just world, members of the dominant group ascribe negative characteristics to the disfavored group in order to make the latter’s disproportionate victimization appear more deserved (or less undeserved). This, in turn, raises the net benefits of committing crimes against the disfavored group, and thus results in additional crimes against its members. 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 crimes against the disfavored group.

One of the central challenges facing any theory of penalty enhancements 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).\(^9\) We do not claim to solve this difficult problem. However, our analysis suggests some new, hitherto neglected, factors that may be relevant to this debate. Most fundamentally, our model suggests that a series of empirical questions about perceptions about the justice of disproportionate victimization are important, because these perceptions may affect inferences about the characteristics of the affected group.\(^10\) For example, in determining whether penalty enhancements should be applied to crimes against young males, it would be relevant to consider not simply whether this group is disproportionately victimized, but also whether people tend to view this situation as unjust (for example, because the disproportionate victimization of 18-25 year old males is caused by the fact that these victims choose to associate with other 18-25 year old males who commit a lot of crime). The discussion at the end of Section 3 also suggests that the extent to which inferences about group characteristics adjust for the existence of penalty enhancements may be a relevant factor. If this adjustment is made, then not all disproportionate victimization would lead to biased inference; only disproportionate victimization that is attributable to animus against the disfavored group would do so. In such circumstances, it could be argued that hate crime statutes should apply only to groups against which animus exists, rather than to all groups that may face disproportionate victimization. Of course, the answers to many of these questions do not appear to have been directly addressed in the existing literature, and 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

---

\(^9\)See e.g. Blake (2001). Posner (2001, p. 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.”

\(^10\)These 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.
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 hate crimes. If members of the majority do not learn of hate crimes, 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, as explained in Dharmapala and McAdams (2005).

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 serious publicity of 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.11

As we have emphasized above, the effects of disproportionate victimization and the just world bias may not necessarily be manifested primarily in the form of increased hate crime, but rather may be manifested through increased levels of discrimination. Similarly, analogous social harms may be caused by

---

11The 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).
other types of disproportionate impacts. For example, the just world effect may cause similar social harms to arise from racial profiling. Suppose that profiling by law enforcement is due (at least partly) to animus towards the disfavored group on the part of law enforcement personnel. 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 (these effects, it should be noted, would not occur if law enforcement were engaging purely in statistical discrimination). Similarly, if some degree of taste-based employment discrimination exists, then the just world effect may lead to a downward revision of the perceived characteristics of the disfavored group, and hence to more discrimination (rather than the elimination of discrimination through competition, as suggested by the classic Becker model).\textsuperscript{12}

Finally, our analysis may also be related to one of the earliest 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 for the disfavored group’s intrinsic negative characteristics, leading 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 make inferences while naively ignoring 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.

6 References


Lawrence, F. 1999. *Punishing Hate: Bias Crime under American Law*


