University of Connecticut

Department of Economics Working Paper Series

Words that Kill? Economic Perspectives on Hate Speech and Hate Crimes

Dhammika Dharmapala
University of Connecticut

Richard H. McAdams
University of Illinois at Urbana-Champaign

Working Paper 2003-05

January 2003

341 Mansfield Road, Unit 1063
Storrs, CT 06269–1063
Phone: (860) 486–3022
Fax: (860) 486–4463
http://www.econ.uconn.edu/

Abstract

This paper analyzes the conditions under which the level of hate speech (expressing hostility towards racial and other minorities) in society can influence whether individuals commit hate crimes against minorities. More generally, we explore the conditions under which speech can influence behavior by revealing social attitudes. We propose a model in which potential offenders care not only about the intrinsic benefits from the crime and the expected costs of punishment, but also about the esteem conferred by like-minded individuals. The number of such individuals is uncertain, but can (in certain circumstances) be inferred from the level of hate speech. We assume that individuals trade off their expressive utility from voicing their true opinions against the costs imposed by formal and/or informal sanctions on hate speech. We characterize the separating and pooling equilibria of this asymmetric information game, and show that the costs of engaging in speech affect what views are expressed in equilibrium. Then, we specify a set of conditions where individuals have common prior beliefs, engage in Bayesian inference, and are risk-neutral in esteem under which speech is neutral, i.e. has no effect on behavior. Then, we relax these assumptions, taking into account the relevant psychological evidence, and derive the impact of hate speech on hate crime using a variety of different formulations. We conclude that those assumptions that appear to have the strongest empirical support (the correspondence bias in inference and the concavity of utility in esteem) imply that raising the costs of engaging in hate speech will deter hate crime.

Journal of Economic Literature Classification: K42

Acknowledgments: We would like to thank Geoff Brennan, Bob Cooter, Rachel Croson, Bruno Frey, Nuno Garoupa, Steve Heyman, Heidi Hurd, Keith Hylton, Louis Kaplow, Neal Katyal, Dorothea Kübler, Timur Kuran, Gerry Mackie, Michael Moore, Philip Pettit, Eric Rasmusen, Bruno Verbeek, participants at the 2002 American Law and Economics Association meetings, and seminar participants at Harvard University for helpful discussions and comments. Any remaining errors are, of course, our own. Dharmapala acknowledges the financial support of the John M. Olin Foundation.

1) Introduction

In the summer of 1999, a former Indiana University undergraduate named Benjamin Smith embarked on a series of drive-by shootings, targeting minorities in Illinois and Indiana. He began by firing on a group of Orthodox Jews in Chicago, injuring several. Then, he drove to Evanston, IL, where he shot and killed Ricky Byrdsong, an African-American college coach out walking with two of his children. Later that day, Smith proceeded to Bloomington, IN, where he shot and killed Won-Joon Yoon, a Korean doctoral student at Indiana University. Smith eventually committed suicide following a police chase.[1] That same summer, Buford Furrow, Jr., opened fire on a Jewish community center in Los Angeles, injuring several children (an act that he described as a "wake-up call for Americans to kill Jews"). Later, he shot and killed a Pilipino-American mailman, before eventually giving himself up to police.[2] Less than a year later, Richard Baumhammers shot racial minorities in the Pittsburgh area, killing five. His victims were of African-American, Jewish, Chinese, Indian, and Vietnamese origin.[3] In the fall of 2001, a few weeks after the September 11 attacks, Mark Stroman shot and killed Vasudev Patel, an Indian immigrant gas station attendant, in Mesquite, Texas, apparently believing him to be a Muslim. He is also charged with the earlier crimes of murdering a Pakistani immigrant, Waqar Hasan, four days after the September 11 attack, and wounding a Bangladeshi immigrant, Rais Bhuiyan, the following week.[4]

Such deadly rampages are rare, but less serious hate crimes (i.e. crimes motivated by racial or other hatred) are quite common. The FBI reported 8063 hate crimes in 2000, mostly property offenses and assaults. However, this data suffers from gaps and inconsistencies in reporting by local law enforcement agencies, as well as from general underreporting, and other estimates are much higher.[5]

[1] See e.g. E. Ferkenhoff and M. Ko, "Killer's Trail of Blood," *Chicago Tribune*, July 5, 1999.

[2] See e.g. M. Lait and N. Zamichow, "Valley Shooting Suspect Surrenders, Confesses," *Los Angeles Times*, August 12, 1999, and B. Rector, "The Legacy of Hate Crime Is Passion with which to Fight It," *Los Angeles Times*, February 4, 2001.

[3] See e.g. K. E. Finkelstein, "5 People Are Shot to Death, and a Lawyer Is Arrested," *New York Times*, April 29, 2000.

[4] See, e.g., Associated Press, "Immigrant Store Owner's Killer Gets Death Sentence," *The Commercial Appeal*, April 5, 2002; Tim Wyatt, "Killer of Gas Clerk Gets Death Penalty," *Dallas Morning News*, April 5, 2002. Stroman allegedly confessed to all three crimes, *id*., and allegedly told others he committed nine crimes in the time period and was planning more. See Tim Wyatt, "Gas Clerk Killer is Convicted," *Dallas Morning News*, April 3, 2002.

[5] See Federal Bureau of Investigation, *Hate Crime Statistics*, 2000, and Southern Poverty Law Center, "Discounting Hate," *Intelligence Report*, Winter, 2001, available online at:

The killing rampages, however, provide particular insight into the difficulty these crimes pose for economic analysis. The central focus of the economics of crime is on deterrence (e.g. Becker, 1968), yet the perpetrators of these crimes seem to have been undeterred both by a high probability of detection and by approximately maximal sanctions (death or life imprisonment).[6,7] Nor does economic analysis seem to "explain" such crimes, except in the almost tautological and generally unhelpful manner of saying that these individuals have a strong "taste" for committing these violent acts.[8] While it is undoubtedly true that tastes play a major role in this type of behavior, we argue below that incorporating additional variables can provide insights into the policy tools available to combat these types of crimes.

What the conventional view ignores, we believe, is the degree to which the perpetrators of these crimes are driven by the desire for fame among those who share their ideology. Smith, Furrow, and Stroman belonged to organized hate groups (the World Church of the Creator, Aryan Nations, and Aryan Brotherhood, respectively), and Baumhammers was the self-styled leader of an anti-immigration political party.[9] Although these offenders are likely to have expected approval for their crimes from their close associates, the opportunity for fame exists among a much wider network than one's immediate circle. Given the scope of modern communications, individuals committing hate crimes can expect to gain widespread acclaim from sympathetic strangers who live far from the locations of the crimes.[10]

To capture this notion of fame as a motivation, we draw on the "esteem theory" (Pettit, 1990; McAdams, 1997; Brennan and Pettit, 2000; Cowen, 2002) from the economics of social

http://www.tolerance.org/news/article_hate.jsp?id=341 (estimating 50,000 hate crimes annually). The Center has been tracking hate crimes since 1981.

[6] Furrow, Baumhammers, and Stroman have all been convicted and sentenced to life in prison or death. See the articles cited supra note 4, H. Weinstein, "Furrow Gets 5 Life Terms for Racist Rampage," *Los Angeles Times*, March 27, 2001, and Anonymous, "Racist Killer of 5 Gets Death Sentence," *New York Times*, May 12, 2001.

[7] To a lesser extent, economic analysis discusses incapacitation as a means of preventing crime, but incapacitation is at best a partial solution because there is no way to predict who is going to commit such crimes.

[8] The attribution of such crimes to the insanity of the perpetrators is a variant of the "tastes" explanation, and thus is equally unhelpful.

[9] See the articles cited *supra* notes 1-4, and also e.g. A. Beeler and E. Osnos, "Puzzling Path Down Road to Racism," *Chicago Tribune*, July 6, 1999 and Anonymous, "Suspect Saw Immigration as Disastrous to Whites," *Pittsburgh Post-Gazette*, May 3, 2000.

[10] In sentencing Baumhammers, the judge explicitly drew attention to the role of racist web sites in inspiring the rampage – see "Statement by Judge Manning in Baumhammers' Sentencing," *Pittsburgh Post-Gazette*, September 7, 2001. Stroman claimed to have expected broad support for his post 9/11 killings. He said in a television interview that he "did what every American wanted to do but didn't" because they "didn't have the nerve." Robert E. Pierre, "Victims of Hate, Now Feeling Forgotten," *Washington Post*, Sept. 14, 2002. His defense lawyer told the jury hearing Stroman's case that he "thought he was being a hero. . . He thought that America would praise him and pin a medal on his chest." Tim Wyatt, "Killer of Gas Clerk Gets Death Penalty," *Dallas Morning News*, April 5, 2002.

norms.[11] This theory posits that the esteem of others, like conventional consumption goods, enters into individuals' utility functions as an intrinsic motivation. However, the arguments we develop are essentially unaffected if esteem is assumed to be valued instrumentally because it secures more consumption goods (see McAdams (2000, pp. 346-47)).[12]

While the introduction of esteem as a motivation may involve some loss of theoretical parsimony, we argue that it enables us to explain phenomena that are otherwise puzzling, and to generate significant new insights. In particular, it enables us to extend the standard economic theory of crime in a direction that can encompass the most extreme examples of hate-motivated criminal behavior, of the kind highlighted above. Our other aim is to contribute to the economic literature on speech and its regulation (Posner, 1986; Loury, 1994; Kuran, 1995; Rasmusen, 1998; Morris, 2001). We argue that, when individuals care about esteem, speech can influence conduct in certain circumstances by providing information about what behavior others esteem or disesteem. We elaborate the model in the context of hate speech,[13] identifying conditions under which the prevalence of hate speech may affect the expected utility from committing hate crimes.

In addition to the esteem assumption, the model we propose assumes that individuals have incomplete information about the basis on which others confer esteem (or disesteem), and that individuals gain "expressive utility" from articulating their true views. Thus, in the absence of speech costs, individuals' decisions produce a separating equilibrium, where individuals express their views and everyone can perfectly infer the amount of esteem (or disesteem) that a given behavior will produce. In the presence of speech costs, however, some speech may be deterred, the result being a (partially or perfectly) pooling equilibrium that creates uncertainty

[11] The esteem theory is typically used as an explanation for the voluntary private provision of public goods, an issue that we do not focus on in this paper (see McAdams (1995)).

[12] In this latter interpretation, the esteem assumption can simply be regarded as a reduced-form representation of effects that would hold even in the absence of esteem considerations. For instance, it is possible that a higher estimate of the number of esteemers may lead potential offenders to believe that they are less likely to be caught or punished (as witnesses, the police, and the jury may be more sympathetic), and so lead to more hate crimes. While it would be possible to develop such a model with no explicit reference to esteem, we believe that our formulation leads to more general insights about the effects of speech on conduct.

[13] Hate speech has been variously defined. Examples include: "speech designed to promote hatred on the basis of race, religion, ethnicity or national origin" (Rosenfeld, 2001, p. 2); or "expression that abuses or degrades others on account of their racial, ethnic, or religious identity." (Heyman, 1996, p. ix). We focus our discussion on racist speech, though the analysis applies to the broader category (including, for instance, hate speech based on sexual orientation). In addition, we argue below that our analysis supports defining hate speech with an emphasis on that which conveys, directly or indirectly, the message that the speaker will esteem one who commits a criminal offense against members of the targeted group.

about true social attitudes. We begin by characterizing these speech equilibria, showing that the costs of speech can affect what views are expressed in equilibrium.

To clarify the conditions under which speech can influence behavior by signaling attitudes, we first derive a neutrality result. If individuals have common prior beliefs, engage in Bayesian inference, and are risk-neutral in esteem, then speech will not affect conduct. This result can be extended to cases where individuals have randomly dispersed priors or private signals. However, any further departure from these assumptions leads, in general, to the non-neutrality of speech; thus, the costs of engaging in hate speech can have an impact on the level of hate crime.

The paper then presents a number of alternative formulations. Particularly because the nature of the hate crimes discussed above may call standard rationality assumptions into question, we place considerable emphasis on the psychological evidence regarding such issues as cognitive biases in inference. Here, we briefly summarize only the results of the most plausible alternative assumptions: If we assume that individuals' priors or private signals are positively correlated with their tastes for committing hate crimes, then raising the costs of hate speech will *increase* the level of hate crime. If we assume that individuals deviate from Bayesian inference by exhibiting what psychologists term the "correspondence bias," then raising the costs of hate speech will *reduce* hate crimes (whether or not priors are common, random, or correlated). Finally, if individuals are risk-averse in esteem, then raising the costs of hate speech will, by creating greater uncertainty, also reduce hate crimes.

These results are intended to be illustrative of the likely range of possibilities, rather than to be conclusive. We hope that one contribution of this paper is to clarify and sharpen the empirical questions that are relevant for understanding the effects of speech on conduct, and on which future research on this topic within economics, psychology, and law should focus. It should also be emphasized that none of these results, by themselves, entail that government or private regulation of hate speech is beneficial, as we do not address the costs associated with public or private regulation of speech.

The paper proceeds as follows. Related literature is discussed in Section 2. Section 3 presents the paper's basic model and derives the neutrality result. Section 4 then relaxes the model's assumptions one at a time, and discusses the various results, emphasizing the most plausible alternative assumptions. Section 5 discusses some practical issues related to public or private efforts to increase the costs of hate speech, and Section 6 concludes.

2) Related Literature

The growing literature by legal scholars on hate speech focuses primarily on doctrinal issues, particularly the extent to which hate speech regulation can be reconciled with existing First Amendment jurisprudence (e.g. Heyman, 1996). Rosenfeld (2001) adopts a comparative perspective, noting that many democracies regulate hate speech more extensively than does the United States, and concludes that the US approach is less justifiable than the competing models. Another strand of the legal literature (e.g. Matsuda *et al*., 1993) focuses on the psychological and dignitary harms suffered by the targets of hate speech. In contrast, our focus is on the effects of hate speech on the behavior of perpetrators of hate crimes.[14]

Within economics, a literature on speech analyzes how reputational effects help determine what messages individuals are willing to send in equilibrium; prominent examples include Loury (1994), Kuran (1995) and Morris (2001). However, there is virtually no economic literature examining hate speech and its possible consequences. Exceptions include Hylton (1996), who draws on Mill's utilitarian philosophy to discuss the circumstances in which hate speech should be regulated. He argues that, in contrast to a perspective that enshrines freedom of speech as a "natural right," utilitarianism can countenance restrictions on speech when the benefits from such regulation (or the harms averted) are sufficiently large. Cooter (2000, p. 323) argues that hate speech should receive a lower level of constitutional protection than other kinds of speech in certain circumstances. Moreover, the general issue of hatred and related concerns about violent activism have been attracting growing interest from economists (see Glaeser (2002) and Stamland and Shogren (2002)). None of this research, however, addresses the particular issue analyzed in this paper.

Perhaps the most closely related literature is that on the economics of speech regulation. Posner (1986) analyzes a landmark First Amendment case,[15] using an approach that weighs the benefits from regulation (i.e. the harm averted, appropriately discounted) against the costs (such

[14] McKinnon (1993) and others have argued that there exists a link between pornography and violence, a claim that may seem analogous to ours. However, these authors do not specify a precise mechanism through which such an effect may occur.

[15] *United States v. Dennis*, 183 F. 2d 201 (2d. Cir. 1950).

as the losses from legal error). While offering some intriguing suggestions (such as the role of appeals to listeners' self-interest), Posner does not provide a systematic theory of how speech leads to action. Rasmusen (1998) applies an economic approach to the regulation of the desecration of symbols (such as flags). He argues that such desecration should be banned whenever those who experience psychic costs from it would be willing to pay more to prevent it than the perpetrators would pay to carry it out. In contrast, our focus is on the effects of hate speech on *conduct*; we therefore set aside the psychic utility and disutility that may be experienced by the speakers and targets.

3) The Basic Model

3.1) Assumptions

3.1.1) The Model Of Speech

Assume that there is a continuum of individuals belonging to the majority ethnic group in society. Let [0, 1] represent the space of racial views held by these individuals. Individual *i*'s racial views are denoted by $x_i^* \in [0, 1]$; individuals differ only in their racial views, and an individual's x_i^* will be referred to as her "type" in the analysis that follows. Higher values of x_i^* are assumed to represent views that are relatively more hostile to minorities. There is a subset $[\gamma, 1]$ of views that represent neo-Nazi or other racist positions; $\gamma \in (0, 1)$ is a threshold view, above which individuals approve of hate crimes against minorities. Let $\rho \in (0, 1)$ be the proportion of the population that has views $x_i^* \geq \gamma$ (i.e. who approve of hate crimes). The basic informational assumption is that the x_i^*'s, and hence ρ, are not publicly observed. All that is observable is the view that each individual chooses to express publicly. Individual *i*'s expressed view is denoted by $v_i \in [0, 1]$; this is the message space, and is assumed to be identical to the type space.[16]

In choosing a view to express, each individual is assumed to maximize a payoff function that consists of three components. The first of these is an "expressive" utility one gains by expressing the view one holds (as discussed informally in Kuran (1995, pp. 30-35)). This is

[16] The assumption that the type space and message space are identical may appear to be restrictive. However, allowing individuals a wider range of expression than [0, 1] would not fundamentally affect the nature of the outcomes, given that the assumption of "expressive utility" tends to anchor opinions expressed in equilibrium close to the true type space. In such a model, though, the levels of sanctions required to induce pooling equilibria may be higher than in the results below.

captured by the quadratic preferences – $(v_i - x_i^*)^2$. Thus, the farther is one's expressed view from one's true position, the greater the expressive loss. This is somewhat analogous to the use of spatial preferences in political economy models to represent the "ideological" loss when the implemented policy diverges from one's ideal point. There, though, the policy is a public good, whereas here it is an individual-specific choice of expression that is compared to the ideal point. Note also that the assumption of expressive utility entails that this is not a cheap-talk model (unlike that in Morris (2001)).

The other components of individuals' utility are the formal and informal sanctions that are imposed for certain kinds of speech.[17] A more complete analysis would endogenize the determination and enforcement of these sanctions; here, they are simply taken as exogenous. The level of formal sanctions is denoted by C^F. It is assumed that the government can condition formal sanctions only on the views *expressed* by a particular individual. That is, it can outlaw particular forms of speech, but not the corresponding thoughts (either because of constitutional restrictions on the kinds of laws that can be passed, or due to the difficulty of proof). Moreover, we assume that constitutional constraints also prevent the government from outlawing speech to the left of γ (i.e. it cannot ban speech that does not belong to the most extremist subset of feasible expression). Thus, $C^F = C^F(v_i)$ for each individual i. On the other hand, informal sanctions (imposed by private actors) can be conditioned on all the observables (v_i and the expressed views of all other individuals, denoted v_{-i}) and on any inferences about true views that can be drawn from these observations. In particular, the informal sanctions imposed on i depend on the posterior belief held by other individuals about i's true views. This belief is denoted by μ_i, where $\mu_i \equiv \Pr[x_i^* \geq \gamma \mid v_i, v_{-i}]$. Informal sanctions on individual i are denoted by $C^I(\mu_i(v_i, v_{-i}))$. To simplify the analysis, it is assumed that the informal sanctions are only applied when $\mu_i = 1$ (although this is not fundamental to the results below).

The total payoff of individual i, denoted U_i, can be expressed as:

$$U_i = -(v_i - x_i^*)^2 - C^F(v_i) - C^I(\mu_i(v_i, v_{-i})) \qquad (1)$$

[17] In Kuran's terminology (1995, pp. 26-30), these two components together constitute one's "reputational utility." We omit discussion of what Kuran terms "intrinsic utility," where an individual gains utility from having his expression influence public policy, because there is a continuum of individuals and each can be assumed to be negligible with respect to the determination of public policy.

A strategy for each individual is simply a choice of view to express: $v_i \in [0, 1]$. Whenever possible, each individual also forms a belief about the true views of every other individual. The game thus has two stages:

1) Each individual (noncooperatively) chooses a view $v_i \in [0,1]$ to express;
2) The government and private actors impose formal and informal sanctions (if any), respectively, on each individual.[18]

An equilibrium can be defined as follows:

Definition of Equilibrium: *An equilibrium is a profile of expressed views (v_i's) and beliefs (μ_i's) such that each individual's payoff is maximized – i.e.*

$$v_i = arg\ max\ U_i(x_i^*;\ v_i,\ v_{-i})$$

and beliefs are formed using Bayes' Rule whenever possible. In particular (whenever it is possible to form nontrivial beliefs),

$$\mu_i = 1 \text{ only if } x_i^* \in [\gamma,\ 1] \text{ and } \mu_i = 0 \text{ only if } x_i^* \in [0,\ \gamma]$$

Given an equilibrium of this nature (i.e. a choice of v_i by each individual, and the consequent inferences concerning the true racial attitudes of the population), there will exist a (possibly degenerate) probability distribution for ρ. Let F: $(0, 1) \rightarrow [0, 1]$ be the cdf of this distribution, and $f(\cdot)$ be its pdf. For the moment, we assume that these distributions are common knowledge, and that no individual receives any signals or information regarding the true value of ρ, other than that summarized by these distributions. Thus, each individual has the same beliefs concerning ρ. These assumptions will be relaxed in later sections of the paper.

3.1.2) The Potential Offender's Decision

Now consider the decision problem faced by a potential offender who is contemplating whether to engage in a hate crime.[19] Let B be the intrinsic utility she derives from committing the act, even in the absence of esteem from others. Let C be the disutility from the costs associated with being detected (such as imprisonment), discounted by the probability of apprehension. If this were the entire story, then (assuming that the default payoff from not committing the crime is 0) the individual would commit the crime whenever $B > C$. Such a simple account would leave

[18] The sanction is obviously costly for the individual on whom it is imposed, and it is to these costs that C^F and C^I refer. Clearly, there are also costs of enforcing sanctions, but these, and the associated free rider problems with the enforcement of informal sanctions, are not addressed here.

[19] Typically, such an individual would presumably be drawn from among those with views in $[\gamma, 1]$, but this does not matter for the analysis.

no scope for policy interventions of any kind, assuming that the probability of detection is already high, that the punishment is already approximately maximal, and that it is not possible to manipulate preferences to influence *B*.

However, the novel element of the theory developed in this paper is to introduce the assumption that potential offenders care not only about *B* and *C*, but also about the *esteem* they anticipate receiving from like-minded individuals if they commit the crime. That is, a potential offender is assumed to derive utility from the esteem of those who approve of hate crimes. While an individual is likely to value more highly the esteem received from close associates than from strangers, we assume that individuals also place some positive value on the esteem of strangers ("fame"). Moreover, it is this "stranger esteem" that is theoretically significant. The theory developed here focuses on uncertainty regarding esteem and it is the number of esteeming strangers that is not known with certainty, so a potential offender must use the public information derived from the expressed views of individuals to estimate ρ (i.e. the number of esteemers). Because there is little or no uncertainty regarding esteem from close associates, we generally ignore "associate esteem." To the extent that there is any uncertainty over the esteem one receives from associates, then the analysis of this paper applies to that esteem as well.

For simplicity, it is assumed that each esteemer confers a fixed, known "amount" of esteem, normalized to 1.[20] The utility derived by the potential offender is denoted by $u(\rho)$, where

$$u(\rho) = \int_0^1 u(z) f(z) dz \tag{2}$$

No particular assumptions are made at this stage about the shape of $u(\rho)$, other than that it is increasing (i.e. $u'(\cdot) > 0$); this is virtually definitional, being simply equivalent to positing that esteem is a good, in the sense of being positively valued. More specific assumptions will be made in later sections of the paper.

When the potential offender cares about the esteem she anticipates receiving, her net payoff from the crime (assuming a 0 default payoff if the individual does not commit the crime), denoted by V, is:[21]

[20] The results would not change fundamentally if the "amount" of esteem conferred were also subject to uncertainty.

[21] This formulation requires that the utility function is separable in the intrinsic (net) utility ($B - C$) and the utility from esteem. Note that, as ($B - C$) is defined in utility terms, it can implicitly accommodate risk aversion or risk preference over these net gains. The approach adopted here is to take the probability of detection and the sanction as given, and to focus solely on the effects of changes in the expected utility of esteem. However, in a more general setting where all policy instruments are chosen simultaneously, there may be some interactions between the risk to

$$V = B - C + \int_0^1 u(z)f(z)dz \qquad (3)$$

Thus, she will commit the crime whenever $V > 0$. This concludes the description of the basic model. The next subsection examines the consequences of speech regulation.

3.2) Speech Equilibria

Having described the assumptions of the basic model, we now turn to the equilibria of the speech game. It should be noted that the aim below is not to characterize the equilibria in all possible circumstances, but to highlight those of greatest relevance for the argument of the paper.

3.2.1) Equilibrium with No Sanctions

Consider first the case where there are neither government nor private sanctions on hate speech. It follows straightforwardly that:

Remark 1: Suppose that $C^F = C^I = 0$. Then, the equilibrium involves perfect separation:

$$v_i = x_i^* \; \forall x_i^* \in [0, 1]$$

$$\mu_i = 1 \text{ if } v_i \in [\gamma, 1] \text{ and } \mu_i = 0 \text{ otherwise}$$

Thus, the equilibrium involves sincere expression: each individual expresses her true viewpoint, and there is no uncertainty in equilibrium about the number of haters.

3.2.2) Equilibrium with Formal Sanctions Only

Suppose that, as before, $C^I = 0$, but that $C^F > 0$. Specifically, the government imposes a strictly positive sanction on any individual who expresses a view $v_i \in (\gamma, 1]$.[22] It is assumed that the penalty is sufficiently large to deter all violations in equilibrium. Then, the "sincere expression" equilibrium characterized above is modified as follows:

Remark 2: Suppose that $C^F > 0$ and $C^I = 0$. Then, the equilibrium is:

$$v_i = x_i^* \; \forall x_i^* \in [0, \gamma)$$

$$v_i = \gamma \; \forall x_i^* \in [\gamma, 1]$$

$$\mu_i = 1 \text{ if } v_i = \gamma, \text{ and } \mu_i = 0 \text{ otherwise}$$

the potential offender generated by the choice of sanction and detection probability on the one hand, and uncertainty about racial attitudes on the other. For instance, it may be optimal to set these so as to increase the overall risk borne by a (risk-averse) potential offender and thus maximize deterrence. Alternatively, the risk borne by the potential offender may be viewed as a social cost (see Polinsky and Shavell (1979)). These issues are not pursued here.

[22] This is formulated as an open set in order to avoid technical complications; this does not fundamentally affect the intuition.

Thus, those individuals whose x_i* is within the permitted range of speech engage in sincere expression. The haters, on the other hand, make the most extreme statement (γ) that is consistent with the law. Any rational observer can infer straightforwardly that those making statement γ are all haters; thus, the fraction of haters in the population, ρ, is revealed in this equilibrium. However, the *distribution* of true views among these haters remains hidden because the haters pool among themselves.

3.2.3) Equilibria with Informal Sanctions Only

Now suppose that $C^F = 0$, while $C^I > 0$. Recall the assumption that these sanctions are strictly positive only when $\mu_i = 1$; thus, the analysis here focuses on characterizing certain equilibria of this game for particular ranges of values of $C^I(1)$. The first point to note is that sincere expression is not an equilibrium when $C^I(1) > 0$:

Remark 3: Suppose that $C^F = 0$ and $C^I(1) > 0$. Then, no perfectly separating equilibrium exists.

Proof: Recall the equilibrium belief in the perfectly separating equilibrium characterized in Remark 1: $\mu_i = 1$ if $v_i \in [\gamma, 1]$ and $\mu_i = 0$ otherwise. Given this, for any $C^I(1) > 0$, individuals with $x_i* = \gamma$ will always be able to find an ε sufficiently small that they prefer to deviate to $v_i = \gamma - \varepsilon$ (which gives a payoff of $-\varepsilon^2$ that is arbitrarily close to zero, while the separating equilibrium gives a payoff of $- C^I(1) < 0$).

As there does not exist any perfectly separating equilibrium, what kinds of equilibria do exist? First, consider relatively low values of $C^I(1)$. The following result can be derived:

Proposition 1: Suppose that $C^F = 0$ and that $C^I(1) > 0$ is sufficiently small, with $C^I(1) = (x_R - \gamma)^2$ for some $x_R \in [0, \gamma)$. Then, there exists a partially separating equilibrium, where:

$$v_i = x_i* \text{ if } x_i* \in [0, x_R] \text{ or if } x_i* \in [\gamma, 1]$$

$$v_i = x_R \text{ if } x_i* \in (x_R, \gamma)$$

$$\mu_i = 1 \text{ if } v_i > x_R, \text{ and } \mu_i = 0 \text{ otherwise}$$

Proof: Consider individuals with $x_i* = \gamma$. The equilibrium payoff is $- C^I(1)$; the payoff from deviating to some $v_i > x_R$ is clearly lower. The payoff from deviating to $v_i = x_R$ is $-(x_R - \gamma)^2$. This equilibrium thus requires that $C^I(1) \leq (x_R - \gamma)^2$. If this holds, then, *a fortiori*, any individual with $x_i* > \gamma$ will also satisfy the equilibrium.

Consider a nonhater whose view is arbitrarily close to $x_i* = \gamma$ on the left (i.e. $x_i* = \gamma - \varepsilon$). Such an individual has equilibrium payoff $-(x_R - \gamma + \varepsilon)^2$; the payoff from deviating to v_i

$= \gamma - \varepsilon$ is $- C^I(1)$. As $\varepsilon \to 0$, the equilibrium requires that $C^I(1) \geq (x_R - \gamma)^2$. If this holds, then, *a fortiori*, any individual with $x_i^* < \gamma - \varepsilon$ will also satisfy the equilibrium.

The condition above ($C^I(1) = (x_R - \gamma)^2$) ensures that both these requirements are satisfied simultaneously.

The intuition here is that both those with views farthest from those of the haters, and the haters themselves, speak sincerely. However, those nonhaters whose views are closer to γ express a position that differs from their true views. By moving their expression just far enough that haters find it unpalatable to dissimulate and pool with them, they avoid being confused with haters. Given that they will be identified in equilibrium anyway, haters simply engage in sincere expression. Thus, low levels of informal sanctions do not lead to greater uncertainty about hatred.

The equilibrium characterized above ceases to exist when $C^I(1)$ is sufficiently large. Suppose that $x_R = 0$; then, individuals with $x_i^* = \gamma$ will deviate when $C^I(1) > \gamma^2$, as the deviation payoff ($- \gamma^2$) exceeds the equilibrium payoff ($- C^I(1)$). For larger values of $C^I(1)$, the equilibrium is:

Proposition 2: Suppose that $C^F = 0$ and that $C^I(1)$ is sufficiently large (in particular, $C^I(1) > 1$). Then, there exists a perfect pooling equilibrium:

$$v_i = 0 \ \forall x_i^* \in [0, 1]$$

$$\mu_i = \mu_p \text{ if } v_i = 0, \text{ and } \mu_i = 1 \text{ if } v_i > 0$$

Proof: Consider the most extreme haters (individuals with $x_i^* = 1$). Their equilibrium payoff is $- (0 - 1)^2 = - 1$; their highest possible payoff from deviating is by setting $v_i = 1$, which yields a payoff of $- C^I(1) < - 1$. Thus, they will satisfy the equilibrium. This holds *a fortiori* for any $x_i^* < 1$.

The intuition here is that the informal sanctions are so costly to those believed to be haters that even the most extreme haters prefer to send the message $v_i = 0$; thus, the entire population pools. It was assumed for convenience that no informal sanction is imposed unless $\mu_i = 1$. Note, though, that everyone in this equilibrium is perceived as having a strictly positive probability of being a hater ($\mu_i > 0$). Even if some sanctions were imposed on everyone (or all nonhaters suffered some cost from even a small probability of being thought to be a hater), the equilibrium above would continue to hold, as long as this cost were small *relative* to the sanction $C^I(1)$ imposed on those known (with certainty) to be haters.

3.2.4) Equilibria with both Formal and Informal Sanctions

Now consider the case where there exist both formal and informal sanctions for hate speech. First, note that Remark 3 still applies in this context – there does not exist any perfectly separating equilibrium. The partially separating equilibrium of Proposition 1 is modified as follows:

Proposition 3: Suppose that $C^F > 0$ is imposed on speech in $(\gamma, 1]$ and is sufficiently large to deter all violations; suppose also that $C^I(1) > 0$ is sufficiently small (as in Proposition 1). Then, the following partially separating equilibrium exists:

$$v_i = x_i^* \text{ if } x_i^* \in [0, x_R]$$

$$v_i = x_R \text{ if } x_i^* \in (x_R, \gamma)$$

$$v_i = \gamma \text{ if } x_i^* \in [\gamma, 1]$$

$$\mu_i = 1 \text{ if } v_i > x_R, \text{ and } \mu_i = 0 \text{ otherwise}$$

Proof: Analogous to proof of Proposition 1.

This is identical to the equilibrium in Proposition 1, except that now the haters pool among themselves, all expressing the most extreme position consistent with the law (as in the equilibrium in Remark 2). Thus, while the number of haters is revealed perfectly, the intensity of hatred is now hidden.

When informal sanctions are large, the same equilibrium as in Proposition 2 holds. The existence of formal sanctions does not make any difference here, as the entire population wishes to pool on $v_i = 0$ in response to the informal sanctions alone. In general, formal sanctions only have an independent effect on behavior when informal sanctions are relatively small. The main impact of formal sanctions is to prevent separation *among* haters, thus depriving potential offenders of information about the distribution of true views within the extremist community. When sufficiently large, informal sanctions can also create uncertainty about the number of haters.

3.3) A Neutrality Result

We have seen in the previous subsection that the degree to which speech is regulated (whether by formal or informal sanctions) helps to determine what views individuals are willing to express in public, and hence affects the degree of uncertainty about social attitudes. The central insight of this paper is that, when individuals care about the esteem they receive from others, the amount and variety of speech (through its effect on this uncertainty) will in general affect conduct. In the particular application developed in this paper, the costliness of hate speech (and the consequent level of uncertainty concerning ρ) will generally affect the level of hate crime. However, we begin by specifying a (quite restrictive) set of circumstances in which speech is neutral, in order to provide a benchmark model against which those of Section 4 can be compared. Intuitively, a set of sufficient conditions for neutrality are that potential offenders have common (correct) priors concerning the distribution of ρ, have an unbiased estimator of ρ (i.e. their estimate of the mean of the distribution is the same, regardless of the costliness of hate speech), and utility from esteem is linear (i.e. potential offenders are risk-neutral in esteem).

In terms of our model, a speech regime can be characterized simply by the levels of formal and informal sanctions (C^F and C^I). Each such regime will give rise to a probability distribution over ρ. Earlier, the cdf of this probability distribution was denoted by $F(\cdot)$. Let the expected utility to a potential offender from committing a hate crime be denoted by V^F, where:

$$V^F = B - C + \int_0^1 u(z) f(z) dz \tag{4}$$

Now, suppose that a different speech regime gives rise to a distinct probability distribution, with cdf $G(\cdot)$ and pdf $g(\cdot)$. Then, the expected utility to a potential offender from committing a hate crime is V^G, where:

$$V^G = B - C + \int_0^1 u(z) f(z) dz \tag{5}$$

We make the following assumptions (all of which will be relaxed in subsequent sections of the paper):

A1: ("Common Priors and Public Signals") In each speech regime, the *only* information about ρ available to any potential offender is given by $F(\cdot)$ (in the first regime) or by $G(\cdot)$ (in the other regime).

This entails that no individual receives any private signals pertaining to the value of ρ. In addition, it requires that the prior beliefs of all individuals (before observing the views expressed in equilibrium) concerning ρ are the same. As discussed further below, A1 is not strictly necessary for the neutrality result; however, it is a convenient starting point for the analysis.

A2: ("Unbiased Estimation") $F(\cdot)$ and $G(\cdot)$ have the same expected value – i.e.

$$\int_0^1 zf(z)dz = \int_0^1 zg(z)dz \tag{6}$$

This assumption entails that potential offenders have available an unbiased estimator of ρ: a different speech regime may give rise to a probability distribution over ρ with a different variance, but the mean will be unaffected. For instance, in the case where greater uncertainty is caused by higher costs of hate speech, potential offenders will realize that the lower volume of hate speech is in fact due (at least in part) to the higher cost. Consequently, adjusting for this effect, the estimate of the expected value of ρ should be unchanged.

A3: ("Risk-Neutrality") Potential offenders' utility from esteem $u(\rho)$ is linear; i.e.

$$u(\rho) = \alpha\rho + \beta$$

where $\alpha > 0$ and β are arbitrary parameters.

This entails that potential offenders care only about the expected value of ρ, and the other moments of the distribution are irrelevant.

These assumptions lead straightforwardly to the following neutrality result:

Proposition 4: Given assumptions A1, A2, and A3, it follows that $V^G = V^F$; i.e. the potential offender's incentives to commit the hate crime are unaffected by the speech regime.

Proof: $$V^F = B - C + \int_0^1 u(z)f(z)dz \qquad \text{(from 4)}$$

$$= B - C + \int_0^1 (\alpha z + \beta)f(z)dz \qquad \text{(using A3)}$$

$$= B - C + \alpha\int_0^1 zf(z)dz \ + \ \beta\int_0^1 f(z)dz$$

Using A2 and the property that $F(1) = 1$:

$$= B - C + \alpha\int_0^1 zg(z)dz \ + \beta$$

$$= V^G$$

This result establishes a set of conditions under which speech does not matter for conduct. In this particular context, the costliness of hate speech does not influence the incentives to commit hate crimes.

3.4) Extending the Neutrality Result: Dispersed Priors and Private Signals

Assumption A1 entails that every individual begins the game with the same prior distribution of beliefs over the possible values of ρ (the proportion of haters in society), and that no individual receives any informative signal about ρ that is not publicly available to all individuals. As noted earlier, A1 is not strictly necessary for the neutrality result; here, we extend the result to certain types of cases with dispersed priors and private signals. First, consider relaxing the common priors assumption, while maintaining the assumption that all signals are publicly observable. Given A3, potential offenders care only about the expected value of the random variable representing the number of racists. Thus, suppose that each individual begins the game with a prior belief, denoted φ, about the mean number of haters. As before, individuals are also heterogeneous in their intrinsic tastes for committing the crime (i.e. their values of (B - C)), but suppose for now that the distribution of priors is independent of the distribution of tastes.

The assumption of dispersed priors entails that each individual starts with a (possibly) biased estimate of ρ. However, suppose that the priors are correct on average – that is, the distribution of prior beliefs across the population corresponds to the prior distribution over ρ (given by the cdf $F(\cdot)$) assumed in the earlier analysis. In other words, the fraction of the population that holds a prior belief that $\rho < x_0$ (i.e. for whom $\varphi < x_0$) is the same as the probability mass associated (in the earlier model) with values of $\rho < x_0$; i.e. $F(x_0)$.

Given these assumptions about dispersed priors, consider a situation where hate speech is sufficiently costly that a perfect pooling equilibrium occurs (where all individuals send the same message). Then, there can be no updating of the prior beliefs (the φ's) with which individuals entered the game. A potential offender who is contemplating committing a hate crime will thus use her prior belief φ to decide whether to commit the crime. Using Eq. (4) and the reasoning in the proof of Proposition 4, a potential offender with tastes given by (B - C) will commit the crime if:

$$B - C + \alpha\varphi + \beta > 0$$

Rearranging this expression, it follows that the *ex ante* probability that this individual will commit the crime, given her value of (B – C), is:

$$\Pr\left[\varphi > \frac{C - B - \beta}{\alpha}\right] = 1 - F\left(\frac{C - B - \beta}{\alpha}\right) \qquad (7)$$

Now suppose that instead of a pooling equilibrium, we have a perfectly separating equilibrium. Then, the true value of ρ is revealed precisely to all individuals. A potential offender with tastes (B – C) will then update her prior beliefs and use the true ρ. Thus, she will commit the crime if:

$$B - C + \alpha\rho + \beta > 0$$

A government or private actor choosing whether to regulate hate speech has to decide on a policy before it knows the true value of ρ - all it knows at the decisionmaking stage is the prior distribution of ρ, given by $F(\cdot)$. Thus, the *ex ante* probability that an individual with tastes (B – C) will commit the crime when there is complete (public or private) regulation is given by Eq. (7) above. The probability that the same individual will commit the crime under a regime of no (public or private) regulation is:

$$\Pr\left[\rho > \frac{C - B - \beta}{\alpha}\right] = 1 - F\left(\frac{C - B - \beta}{\alpha}\right) \quad (8)$$

This, of course, is identical to the probability in Eq. (7). This suggests that, if the dispersed priors are correct on average and are independent of tastes, then, *ex ante*, the expected number of hate crimes is unaffected by speech. In other words, under these circumstances, the neutrality result of Section 3.3 extends to the case of dispersed priors.

This neutrality may appear counterintuitive. After all, more speech leads to more precise public signals, causing those individuals who started with particularly high priors to revise their beliefs downward. Thus, some of these individuals who may have committed the crime as a result of their high priors will not do so when the precise public signals are revealed. However, recall that the distribution of prior beliefs is assumed to be independent of tastes for the crime. Thus, some potential offenders who have a strong taste for committing the crime will start with very low priors. In the absence of further information, they will refrain from committing the crime. If, however, precise public signals are revealed, these potential offenders will revise their beliefs upwards, causing some of them to commit the crime. Thus, greater precision has two offsetting effects; in expectation, they cancel out under the assumptions specified above.

An analogous result arises if we assume common priors, but allow each individual to receive a private signal about ρ that cannot be communicated publicly (say, from their interaction with their immediate circle of peers). As long as the private signals are correct on average, and the distribution of the signals is independent of tastes, the neutrality result applies. However, when the priors, or the private signals, are correlated with tastes, then the neutrality result no longer holds. This theme is taken up in Section 4 below, where we relax each of the assumptions specified in this section in turn, and derive the implications for the effects of speech on hate crime.

4) The Effects of Hate Speech on Hate Crime: Alternative Formulations

In Section 3 above, we derived a neutrality result: under certain restrictive assumptions, the views that are expressed in equilibrium do not affect the aggregate number of hate crimes. In more general terms, our conditions describe a set of circumstances in which conduct is unaffected by speech. In this section, we relax each of our assumptions in turn, and argue that in these more general circumstances, speech can indeed be expected to influence conduct.

4.1) Relaxing the Assumption of Common or Uncorrelated Priors and Private Signals

In Section 3.4, we extended our neutrality result to the case of randomly dispersed priors and random private signals. We now consider the possibility that priors are neither common nor randomly distributed, but are correlated with other characteristics of the individual. In particular, we focus on the psychologically plausible possibility that individuals' prior beliefs are correlated with their tastes for committing hate crimes (i.e. their values of $(B - C)$). In theory, an individual's prior estimates of the mean number of hate crime approvers may be positively or negatively associated with the individual's taste for hate crimes. If priors are positively correlated with tastes, then potential hate offenders will initially tend to have a φ that is higher than ρ – i.e. to overestimate the mean number of hate crime approvers. Conversely, if priors are negatively correlated with tastes, then potential hate offenders will initially tend to underestimate ρ.[23]

The question we seek to answer is how speech will affect behavior under these circumstances. (As discussed below, we believe that the behavioral assumption of correlated priors makes the most sense when combined with a behavioral assumption of non-Bayesian updating. For now, however – following the standard practice of relaxing one assumption at a time – we retain the assumption that individuals engage in Bayesian inference.) Suppose that

[23] We ignore the case of "anti-haters" with preferences against hate crimes because they are inframarginal with respect to the crime. Obviously, if priors are negatively correlated with tastes, anti-haters will initially overestimate the mean number of hate crime approvers; if priors are positively correlated with tastes, anti-haters will tend to underestimate the mean.

haters start with a high estimate φ of ρ, and that we move from a perfect pooling equilibrium to a separating equilibrium. Under the perfect pooling equilibrium, there can be no updating, so haters simply use their prior φ in their decisionmaking process. In a perfectly separating equilibrium, ρ will be revealed precisely. More generally, a reduction in the costliness of hate speech will lead to more hate speech, but it will be clear to haters that there is not as much hate speech as would be consistent with their prior beliefs. Thus, they will revise their beliefs downwards in the direction of the true ρ. Lowering the estimated ρ in turn lowers the expected utility of hate crimes, and therefore, reduces hate crime (of course, there will be some nonhaters who start with low φ's, and update their beliefs upwards, but they are not likely to commit hate crimes). Conversely, if potential hate offenders tend to initially underestimate ρ, greater signaling will cause them to raise their estimate of the mean, thereby raising the expected utility of hate crimes, and therefore the number of hate crimes.

In general, the effect of more speech will be to move beliefs towards the true ρ. Thus, if we relax *only* the assumption of common or uncorrelated priors, the crucial question is whether it is more likely that those priors are positively or negatively associated with tastes. We know of no social science evidence precisely addressing this question, but there is a general body of social psychological research from which one can draw a strong conclusion. The "false consensus effect" and "social projection" refer to the finding that individuals "tend to believe that most other people share their own preferences, habits, or sentiments," (Krueger & Clement, 1997, p. 299), or at least to overestimate the frequency of their preferences, habits, or sentiments in the larger population. Though the strength of the effect varies across circumstances[24] and there are a few exceptional cases where people instead assume dissimilarity, the basic finding is supported across a wide variety of experiments (for reviews, see Krueger (1997); Marks & Miller (1987)). Thus, it would seem that the most plausible assumption regarding correlated priors is that individuals' priors are positively associated with their tastes, with the result that potential hate offenders overestimate the mean number of hate crime approvers. Under this assumption, Bayesian updating will tend to correct the prior by lowering the individual's estimate. As a result, a greater number or variety of signals, including hate speech, will facilitate greater

[24] Some studies find, for example, that individuals are less likely to overestimate their typicality when they are members of manifestly small, socially deviant subgroups, which means that white supremacists may be less likely to make this error (see e.g. Frable (1993)).

updating, lowering the mean estimate, and therefore the expected utility of hate crimes, and hence the number of hate crimes.

A similar argument can be made for the case of private signals. Suppose that each individual receives a private signal concerning ρ, and that haters tend to receive signals indicating a high ρ (for instance, the signal could be derived from the individual's immediate circle of associates, and it could be that haters tend to associate primarily with other racists). Then, more speech (i.e. a more informative *public* signal) will lead, *via* Bayesian updating, to revision of these beliefs downwards, in the direction of the true ρ, and dissuade them from committing the crime. However, there is no offsetting effect of increased crime among those initially low signals about ρ, because these individuals are inframarginal with respect to the crime. Consequently, lowering the costs of hate speech can *reduce* the level of hate crime.

However, we are skeptical about conclusions reached on the basis of an assumption that individuals engage in Bayesian inference, while naively failing to correct for systematic biases in their priors or private signals. That is, if it is indeed the case that prior beliefs or private signals are correlated with tastes, then (given that each individual knows her own tastes), this raises the question of why they do not discount those priors or signals to counteract the bias. For instance, if haters tend to receive particularly high private signals about ρ (perhaps because they interact frequently with other haters), they can adjust their beliefs appropriately to take this into account. Hence, we do not believe that it is sensible to relax one standard assumption – common or uncorrelated priors – based on psychological evidence without also considering how such evidence bears on other elements of the standard rational choice approach. Thus, we next consider the possibility of non-Bayesian updating.

4.2) Relaxing the Assumption of Bayesian Inference

The neutrality result in Section 3 was derived using assumption A2, which gives each individual an unbiased estimator of ρ. As a result, when a change in the cost of signaling one's racial view causes a change in the type or quantity of signals individuals send about their racial views, individual estimates of the mean of ρ do not change. Although increased signals reduce uncertainty in the estimates around the mean, they do not influence estimates of the mean because each individual logically attributes a change in signaling behavior to the change in external costs of the behavior.

Various behavioral studies, however, cast substantial doubt on the assumption of perfect Bayesian rationality. These studies do not suggest, of course, that individuals fail to update at all in the light of new information. Rather, the evidence suggests that cognitive biases cause individuals to deviate systematically from Bayesian inference, making certain predictable errors. The existence of bias raises the possibility that an individual's mean estimate of ρ is sensitive to the type or quantity of signals (the amount of hate speech) even when those signals vary only in accordance with their costs. If so, then the neutrality result no longer holds. Instead, depending on the alternative assumptions one makes, hate speech will raise or lower the mean estimate of ρ and therefore raise or lower the amount of hate crime.

In this section, we explore this issue in detail, focusing on the non-Bayesian assumption most justified by the relevant empirical literature. We first consider the effects of relaxing *only* the assumption of Bayesian inference and then the effect of relaxing that assumption in combination with an assumption of correlated priors.

4.2.1) Introducing the Correspondence Bias (with Common Priors)

Once we depart from the assumption of Bayesian inference, the crucial question is whether an increase in the number and type of signals that accompanies the change from a pooling to a separating equilibrium will systematically influence a potential hate offender's mean estimate of ρ. Two forms of bias are possible: (i) that a greater amount and variety of hate speech will cause potential hate offenders to *increase* their mean estimate of ρ, or (ii) that a greater amount and variety of hate speech will cause potential hate offenders to *decrease* their mean estimate of ρ. Either assumption produces a novel, non-neutral result. The basic result of non-Bayesian inference, therefore, is that any systematic bias produces a link between the amount of hate speech and the number of hate crimes.

Reaching more specific conclusions requires that we decide in which direction inference is likely to be biased. We are not aware of any social science evidence precisely addressing the inferential processes of potential hate offenders. Once again, though, there is a general body of social psychological research from which one can draw a strong conclusion. The "correspondence bias" and "fundamental attribution error" refer to the finding that individuals tend to attribute the behavior of others to their internal dispositions or attitudes (in economic terms: tastes or beliefs), rather than external situational constraints, to a greater degree than is logically warranted. In the words of Jones (1990, p.138), "we see behavior [of others] as corresponding to a disposition more than we should"; moreover, this bias is described as "the most robust and repeatable finding in social psychology" (*loc. cit.*). Miller & Prentice (1996, p. 803) emphasize the "enormous support" the finding has received over several decades of research, while Gilbert & Malone (1995, p.22) conclude that: "In scores of experiments, subjects have violated attribution theory's logical canon by concluding that an actor was predisposed to certain behaviors when, in fact, those behaviors were demanded by the situations in which they occurred."

We will use an early example from the literature to illustrate. Jones & Harris (1967) showed subjects an essay on Cuba that was either pro-Castro or anti-Castro. In two experiments, the researchers told the subjects that the essays were written for an exam or for a debate. The researchers then told some subjects that the essay writer did not choose the position taken in the essay, but had been assigned that position either by the course instructor giving the exam or the debate coach. Contrary to the researcher's expectations, in this "no-choice" condition, the subjects evaluated those who had written the pro-Castro essay as holding views that were significantly more pro-Castro than those who had written the anti-Castro essay. In other words, even though there was a fully sufficient external explanation of the essay's content – to avoid the penalty imposed on those who deviated from the demands of the course instructor or debate coach – the subjects still tended to attribute that content to the actual attitudes and beliefs of the author. Subsequent studies have replicated this finding using a wide variety of essay topics, and have also sought to rule out various competing interpretations. Researchers have concluded that the correspondence bias survives various modifications to the experimental design.[25]

[25] For example, some critics contended that there was still significant choice in the "no choice" condition – an essayist could have refused the assignment – or that the researcher-created essays demonstrated the kind of argumentative expertise that made it appear more likely to reflect the writer's true position. Snyder & Jones (1974) dealt with these two concerns by initially assigning some subjects to write an essay contrary to their own position, and then having the subjects evaluate other subjects' essays. First, those subjects who agreed to write an essay contrary to their own view (as everyone who was asked did; no one refused their assignment) should perceive that the external constraint of an assignment is fully sufficient to cause someone to produce such an essay. Second, essays written by subjects who held contrary views should not contain argumentative expertise correlated with actually holding such a view. Notwithstanding these manipulations, however, the researchers still found a significant correspondence bias, even among subjects who had written an essay contrary to their own position and who had then evaluated an essay contrary to the author's own position. That is, when the researchers gave a "no-choice" essay to a subject who had written a "no-choice" essay, the latter subject erroneously evaluated the author as holding attitudes corresponding to the position taken in the essay. Researchers did not manage to fully extinguish the correspondence bias until they introduced a condition making it clear to the evaluating subject that the target subject had merely copied by hand an essay written by someone else (Snyder & Jones 1974; Jones, 1979).

Thus, the psychological evidence strongly commends the assumption that individuals are not perfect Bayesians, but are biased in the direction of inferring too strongly that attitudes correspond with behavior. Applied to hate speech, this bias suggests that individuals are likely to over-attribute another individual's hate speech to his racist attitude and likely to over-attribute another individual's failure to engage in hate speech (or his engaging in anti-hate speech) to the absence of a racist attitude. The inference individuals will fail to make, at least to the full extent Bayesian reasoning requires, is that an individual who does not engage in hate speech was deterred from doing so by the external costs of such speech (and that an individual who engages in anti-hate speech was induced to do so by the external benefits of such speech). Thus, even though it may be logical to estimate *the same mean* number of hate crime approvers in the presence of either (1) low hate speech costs and a high level of hate speech and (2) high hate speech costs and a low level of hate speech, the correspondence bias means that individuals will infer a higher level of hate attitudes when there is a higher level of hate speech. Put differently, because the bias is to assume that attitudes "correspond" to behavior to a greater degree than is actually the case, deterring the behavior of hate speech will lower the perceived estimate of individuals with racist attitudes. Lowering the estimate of hate crime approvers in turn lowers the expected utility from committing hate crime, and therefore, lowers hate crime. Thus, if we replace the assumption of Bayesian inference with one of correspondence bias, the result is that hate speech is one cause of hate crime.

4.2.2) The Correspondence Bias in Combination with Correlated Priors

We noted above that we believe that it is more plausible *either* to maintain both standard rationality assumptions – common or uncorrelated priors with Bayesian inference – or to replace both with some empirically defensible alternatives. For the sake of completeness, we have just examined the effect of relaxing the rationality assumptions one at a time: (1) correlated priors with Bayesian inference and (2) common or uncorrelated priors with a correspondence bias. Now we consider the effect of relaxing both original assumptions and assuming (more plausibly) (3) correlated priors with a correspondence bias.

Once one abandons the assumption of Bayesian inference, the implications of correlated priors for the relationship between hate speech and hate crime change dramatically. If priors are positively or negatively correlated, potential hate offenders will, respectively, overestimate or underestimate the number or intensity of hate crime approvers. This bias will influence the level of hate crime; however, our interest is in how this level *changes* in response to changes in the cost of hate speech. In the absence of Bayesian inference, we can no longer assume that more information will cause individuals to update their beliefs in the direction of the true parameter values. If potential hate offenders tend to overestimate the number of hate crime approvers (which, as we noted above, is highly likely), the crucial question remains whether an increase in the quantity and variety of signals – a move to a separating equilibrium – will cause them to overestimate *to a greater or lesser degree*. Conversely, if potential hate offenders tended to underestimate the number of hate crime approvers, the key question is whether an increase in hate speech will cause them to underestimate more or less.

As we just demonstrated, the correspondence bias suggests a clear answer. Where costs deter hate speech, individuals will nonetheless erroneously attribute lower levels of hate speech to a lower frequency of racist attitudes that favor hate crime. Conversely, where decreased costs produce a rise in hate speech, individuals will tend erroneously to attribute higher levels of hate speech to a higher frequency of attitudes favoring hate crimes. To clarify this, consider a situation where haters start with high priors φ, and we move from a perfect pooling equilibrium with no hate speech to a separating equilibrium with some hate speech. If they were to engage in Bayesian inference, haters would revise their beliefs downward (as explained in Section 4.1). When their inferential processes are affected by the correspondence bias, however, they will attribute the increase in hate speech at least partly to ρ being higher than they previously thought (a consequence of failing to take fully into account that the costs of hate speech are now lower). Hence, haters will revise their beliefs *upward* (even further away from the true ρ). Note that if haters were to start with low priors, the updating would also be upwards (though this time in a direction closer to the true ρ).

Thus, even if potential hate offenders overestimate the number of hate crime approvers (as is likely), they will overestimate *more* in the presence of high levels of hate speech than they will in the presence of low levels of hate speech. Conversely, if potential hate offenders underestimate the number of hate crime approvers, they will underestimate *less* with high levels of hate speech than with low levels. Thus, given both behavioral assumptions together – correlated priors and a correspondence bias in updating – hate speech will tend to increase the expected value of, and therefore the amount of, hate crime.

4.3) Relaxing the Assumption of Risk-Neutrality

So far, we have maintained the assumption A3 of risk-neutrality in esteem. In this section, we consider, in turn, the alternative assumptions of risk-aversion (concavity) and risk-preferring behavior (convexity), while reinstating all the other assumptions (A1 and A2) of Section 3.3, and discuss the theoretical and empirical arguments for the different assumptions concerning risk attitudes. The effects of concavity and convexity on the incentives to commit hate crimes are arguably only of second-order compared to factors (discussed in Sections 4.1 and 4.2) that affect the *mean* of the distribution of beliefs. Nonetheless, we consider it important to consider the implications of relaxing A3, particularly because that assumption is non-standard (unlike A1 and A2) and behaviorally implausible.

4.3.1) The Effects of Hate Speech Assuming Concave Utility

Consider the following assumption:

A3′: ("Strict Concavity") $u''(\rho) < 0$

Recall that in Section 3.3 we assumed two different speech regimes, each associated with a probability distribution over ρ, with the cdf's denoted by $F(\cdot)$ and $G(\cdot)$. Now, assume further that (without loss of generality), the speech regime that leads to $G(\cdot)$ is more restrictive of hate speech, and so leads to greater uncertainty about ρ. We maintain the earlier assumption (A2) of unbiased estimation, so that the distributions have the same expected values. However, there is more density in the tails of the distribution given by $G(\cdot)$:[26]

A4: $F(\cdot)$ dominates $G(\cdot)$ by the criterion of second-order stochastic dominance.

This entails that the distribution represented by $G(\cdot)$ is "riskier" than that represented by $F(\cdot)$ (see Rothschild and Stiglitz, 1970; Lippman and McCall, 1981, pp. 215-216; Mas-Colell, Whinston and Green, 1995, pp. 197f). Note that, because we maintain assumption A2 in the analysis that follows, we focus not on the general case where $F(\cdot)$ dominates $G(\cdot)$ by second-order stochastic dominance, but on the special case where the random variables have the same mean – i.e. a mean-preserving spread.

In these circumstances, it proves convenient to represent the new random variable as the sum of the old random variable Z (with cdf $F(\cdot)$) and a random variable R with zero mean

[26] For example, assume that the total population is 10,000. Then, suppose $f(\cdot)$ is the pdf of the uniform distribution over [50, 150] and that $g(\cdot)$ is the pdf of the uniform distribution over [0, 200].

($E(R)$ = 0) and strictly positive variance (as in Mas-Colell *et al.*, 1995, p. 197). Thus, the new random variable $Z + R$ has cdf $G(\cdot)$, while R has cdf $H(\cdot)$ and pdf $h(\cdot)$. The net benefit to a potential offender from committing the crime under the high-cost regime, denoted V^G, can be expressed as:

$$V^G = B - C + \int_0^1 \left[\int_0^1 u(z+r)h(r)dr \right] f(z)dz \tag{9}$$

Denoting the net benefit under the low-cost regime by V^F, it follows straightforwardly that:

Proposition 5: Given A1, A2, A3′ and A4, $V^F > V^G$

Proof: Note that:

$$\int_0^1 (z+r)h(r)dr = z\int_0^1 h(r)dr + \int_0^1 rh(r)dr = z$$

as $E(R) = 0$. Therefore,

$$\int_0^1 u(z)f(z)dz = \int_0^1 u\left(\int_0^1 (z+r)h(r)dr \right) f(z)dz$$

By A3′,

$$u\left(\int_0^1 (z+r)h(r)dr \right) > \int_0^1 u(z+r))h(r)dr$$

Therefore,

$$\int_0^1 u\left(\int_0^1 (z+r)h(r)dr \right) f(z)dz > \int_0^1 \left[\int_0^1 u(z+r))h(r)dr \right] f(z)dz$$

Hence, $V^F > V^G$

Thus, when the utility function of potential offenders is concave in esteem, a higher degree of uncertainty about the number of racists who will confer esteem lowers the net benefits that are expected from committing hate crimes. Thus, a separating equilibrium (associated with low hate speech costs) will cause more hate crime than a partially or perfectly pooling equilibrium (associated with high hate speech costs).[27]

[27] We focus on the creation of uncertainty through (public or private) measures that raise the costs of (i.e. 'tax') hate speech. However, it is also possible that uncertainty could be created by *subsidies* for hate speech: if everyone engaged in hate speech, then there would be just as much uncertainty as when noone does so. While we feel obliged

4.3.2) The Effects of Hate Speech Assuming Convex Utility

Now, consider the assumption of convex utility:

A3″: ("Strict Convexity") $u''(\rho) > 0$

It then follows that:

Proposition 6: Given A1, A2, A3″ and A4, $V^G > V^F$

Proof: Analogous to that of Proposition 5.

Thus, assuming a utility function that is convex in esteem leads to the opposite conclusion to that of Proposition 5: a pooling equilibrium (associated with low hate speech costs) will reduce uncertainty about the number of racists and thereby tend to discourage hate crimes.

4.3.3) The Case for Concavity: Theory and Evidence

Given that the results and policy implications of the assumptions of concavity and convexity differ so dramatically, it is crucial to decide which of these assumptions to adopt. In this section, we discuss the theoretical and empirical grounds for believing that concavity is the more reasonable assumption. As a theoretical matter, concavity is very much the standard assumption in economic analysis, and is supported by a wide range of empirical evidence, as well by the notion of diminishing marginal utility. Despite some argument that those who commit crimes differ from the rest of the population in their risk preferences,[28] it is common in economic models of criminal behavior to assume that potential offenders, like the rest of the population, have utility functions that are concave in ordinary consumption goods (see e.g. Polinsky and Shavell (1979)). Thus, the question is whether esteem is like an ordinary consumption good.

On one view, esteem is merely valued instrumentally, as a means of gaining goods valued for ultimate consumption. Accordingly, esteem is like money. If esteem is valued only as a

to raise this as a theoretical possibility, it should be stressed that if we relax A2 (as in section 4.2), then the correspondence bias would imply that potential offenders would be encouraged to commit hate crimes by the increased level of hate speech induced by a subsidy.

[28] Specifically, criminals are said to be less risk-averse, or even to be risk-preferring, with the latter claim often being made with reference to the idea that criminals respond more to increases in the probability of detection than to increases in the sanction (Becker, 1968). See, e.g., Block and Gerety (1995) (finding that prisoners respond more to increases in probability of detection than increases in sanction, unlike the control group of students). However, this notion of risk-loving behavior (and hence a utility function that is not concave) is thought to apply primarily to nonmonetary sanctions (in particular, incarceration). Thus, risk attitudes towards ordinary consumption goods appear to differ from risk attitudes towards prison time. There is no warrant, however, for extrapolating from attitudes towards prison time to attitudes towards esteem. (In any event, it may be possible to explain a greater responsiveness to increases in the probability of detection than to increases in sanctions even while assuming risk aversion, for instance, using a model of state-dependent utility - see e.g. Neilson and Winter (1997)).

means to gaining ordinary consumption goods, and the individual's utility function for those consumption goods is concave, then it follows that the individual's utility function for esteem is concave (just as it is for money). On another view, however, esteem is valued intrinsically, as an ultimate consumption good (see McAdams, 1997). Because risk attitudes towards specific kinds of goods can differ, there is no warrant for merely assuming that the utility function for intrinsically valued esteem is concave (though there is also no warrant for any other assumption).

There is, however, some psychological evidence suggesting that individuals behave as if their utility functions are concave in esteem. In a series of experiments beginning with Asch (1951), psychologists have found that individuals will conform to group judgments, but that there is a substantial difference in conformity depending on whether the group is otherwise unanimous or otherwise has one dissenter. In the original Asch experiment, a subject was asked to express his perception of which of three lines was closest in length to a fourth line in front of four to six others (who were confederates of the experimenter). When the subject gave his answer after the confederates had given the obviously wrong answer, the subject "conformed" and gave the same obviously wrong answer approximately 32% of the time (Asch, 1951, p. 181). However, in another experimental condition, in which the unanimity was broken by having just one confederate give the correct answer, the degree of the subjects' conformity dropped precipitously to approximately 5.5% (Asch, 1951, p. 186). Many experiments confirm this effect in a variety of settings (for a recent cross-cultural review, see Bond & Smith (1996)). Subsequent studies also demonstrate that the effect is not primarily informational; there is much less conformity when subjects are allowed to report their belief privately (see the studies cited in Aronson (1992, p. 24)). Thus, we would conclude, the desire to gain esteem and avoid disesteem is what causes the conformist behavior.

These experiments were obviously not designed to test whether utility in esteem is concave. Nonetheless, we read the dramatic drop in conformity due to a single dissenter as supporting diminishing marginal utility from esteem. Moving from unanimity to a lone dissenter changes the subject's esteem calculus as the number of potential esteemers for giving the right answer moves from zero to one. Given that over three-quarters of conformity disappears in this single step, we know that subsequent steps - adding more dissenters - cannot have as strong an effect. One interpretation is that the first unit of esteem generates more utility than subsequent units.[29] Thus, the evidence appears to support concavity of the utility function.

4.4) Conclusion

This section demonstrates the possibility of non-neutral results, where speech *does* influence behavior. Under the most plausible assumptions, increasing the costs of hate speech will decrease hate crime. The basic model and the alternative formulations developed above can be extended in a number of directions. For instance, our model focuses on utility from the esteem conferred by like-minded individuals; however, incorporating the disesteem conferred on the offender by the general community does not fundamentally change the conclusions. Due to space constraints, we do not consider some interesting additional possibilities that would strengthen the paper's results.[30]

5) Policy Extensions and Caveats

The basic model reveals a novel mechanism that links speech and conduct. By considering the pursuit of esteem, the model explains how individuals "persuade" one another through speech – by providing information about what behavior will be esteemed or disesteem. As applied to hate speech, the model suggests a possible policy tool for influencing the number of hate crimes: formal or informal sanctions to increase the costs of hate speech.

There are, however, a number of important caveats and qualifications to this policy implication of our analysis. Most importantly, nothing in our analysis necessarily supports regulation of speech because we do not consider here the costs of such regulation. We do note that *non-legal*, private regulation of hate speech appears to have the ability to raise the costs of hate speech. The paper's analysis therefore applies as much to the existing informal social norms and formal organizational rules against hate speech[31] as to any potential governmental rules.

[29] Alternatively (and for our purposes, essentially equivalently), moving from unanimity to a lone dissenter involves moving from n to $(n – 1)$ disesteemers, so the last unit of disesteem generates more disutility than prior units.

[30] For instance, those who commit hate crimes might receive additional utility from the *expression* of approval they receive independent of the number of approvers they believe to exist. It may also be the case that hate speech facilitates a process of preference change by which individuals come to approve hate crimes or come to gain intrinsically from the commission of hate crimes. The expression of hate speech may also increase the perpetrator's *self-esteem*, reinforcing the effect due to the esteem she receives from others.

[31] Organizational rules include prohibitions of hate speech by certain private employers, universities, and internet service providers. For example, America Online requires that users agree not to post or distribute any content that "victimizes, harasses, degrades, or intimidates an individual or group of individuals on the basis of religion, gender, sexual orientation, race, ethnicity, age, or disability." See http://www.aol.com/copyright/rules.html. Yahoo and the broker site eBay have similar provisions: see http://docs.yahoo.com/info/terms/ (paragraph 6a) and http://pages.ebay.com/help/community/png-offensive.html.

Nonetheless, private regulations, while free from First Amendment concerns, still impose costs. We do not offer any analysis of those costs, so we do not claim that the benefits of even private regulation outweigh the costs. We leave these important issues for future research.

Second, even if it were desirable to raise the costs of hate speech, there is a separate question of whether regulation could succeed in doing so. We will not address this issue comprehensively, leaving it too for future research, but we will note some ways in which our model illuminates the practical issues. Finally, we will also address one potential objection to the paper's analysis – a countervailing claim that hate speech makes hate crime less likely.

5.1) Raising the Costs of Hate Speech: Problems of Definition and Anonymity

Conventional criticisms of hate speech frequently focus on the subjective harm it imposes on its targets. This focus, however, creates severe practical obstacles to regulation because it may not be possible to prevent that harm. There are two problems. First, there is an unappealing trade-off in how one defines hate speech: an overly broad definition may burden non-offensive speech, while a narrow definition – one that attempts to raise the costs only for the harmful speech – may allow racists to shift to a different form of expression, arguably causing the same ill effects on targets as the prohibited expression. American history is full of racially coded phrases, whereby one raises racist concerns without explicit references to race. One might claim, therefore, that any narrowly targeted regulation will fail to raise the cost of hate speech, while any broad regulation is excessively restrictive. A second problem is anonymity. Speakers may react to the formal or informal penalties on hate speech by shifting to anonymous speech, writing their hate messages on buildings or sidewalks when no one is watching, or distributing such messages in untraceable flyers. While this makes detection and enforcement very difficult, it may cause the same harm to the targets as would identified (non-anonymous) messages.

These practical problems are less severe, however, when the harm to be avoided is the one we identify: the incentives hate speech gives to potential hate offenders. First, under our approach, the definitional trade-off is less stark because a relatively narrow definition of hate speech may be sufficient to reduce hate crime. For example, to reduce the expected esteem benefits from committing a racially motivated murder, it is only necessary to raise the costs of speech that conveys approval of such murders. Because most people disapprove strongly of murder, it requires strong and explicit language to convince others than one actually approves of it. One can create the benefit we identify merely by raising the costs of this strong and explicit language.

Second, the problem of anonymous hate speech is likely to be irrelevant in our framework. The harm we identify from hate speech is that it conveys credible information about *the number of* individuals who will esteem perpetrators of hate crimes (or do so to a certain intense degree). Overt hate speech, where the speakers are clearly identified, provides more credible information about the number of hate crime approvers than does anonymous hate speech. The reason is that, when one cannot identify the source of many anonymous messages, one usually cannot know *how many* sources there actually are. It is always possible that just one individual produces all the anonymous messages (an anonymous message may of course claim to represent a large number of individuals, but such claims are usually cheap talk). Thus, potential offenders using Bayesian inference will tend to discount anonymous speech in estimating the number of approvers.[32] Although the correspondence bias suggests that people will infer more hate crime approval from more hate speech, it does not suggest any particular bias to this discounting of anonymous speech. Thus, if potential offenders are subject to the correspondence bias, their downward revision of their estimate following a reduction in the level of identified hate speech will not be fully offset even if all or some speakers engage in anonymous speech.

5.2) Are Hate Speech and Hate Crime Substitutes?

Finally, we briefly consider a contrary theory. Our model of the behavior of potential offenders has conceptually separated them from speakers – those whose esteem is sought. However, it is also possible that the same individual's choice of speech may interact with her decision regarding whether to commit the crime. If so, then hate speech and hate crime may be either complements or substitutes. In the latter case, allowing the individual the chance to "blow off steam" by engaging in hate speech may reduce the likelihood that she will also commit a hate crime; this would represent a caveat qualifying some of the claims we have made in this paper. However, it is likely that the distribution of intrinsic utility ($B - C$) across individuals is such that there are very few individuals who would commit a hate crime, even for a high level of esteem, while there are many more individuals who may engage in hate speech if the costs are sufficiently low. Then, most speakers are inframarginal with respect to the choice of whether to commit the crime, while their hate speech does influence the (relatively small number of)

[32] For similar reasons, anonymous hate speech will also fail to substantially reduce the variance of this estimate.

individuals who *are* on the margin with respect to the crime. Moreover, the opportunities for gaining esteem from racists for being one of a relatively large number of individuals engaging in hate speech are severely limited in comparison to the esteem that can be gained by committing hate crimes. Thus, any “substitution effect” is likely to be a very minor factor.

6) Conclusion

This paper has developed an analysis of the circumstances in which hate speech can induce the commission of hate crimes. We have proposed a model in which individuals trade off their “expressive utility” from voicing their true opinions against the costs imposed by formal and/or informal sanctions on hate speech, and where potential offenders care about the esteem they receive from like-minded individuals. We have derived a set of conditions under which speech is neutral, and have systematically explored the consequences of departures from those assumptions.

At the most general level, the key insight of this paper is a novel mechanism for relating speech and conduct, *via* the revelation of information about social attitudes. If one assumes that individuals value esteem, have incomplete information about what is esteemed and by whom, and value expressing their actual views, then, unless our stringent neutrality conditions hold, any form of speech thought to reveal the basis of an individual’s esteem judgments will have systematic effects on behavior. While we conclude that those assumptions that have the strongest empirical support (such as the “correspondence bias” and concavity) imply that raising the costs of engaging in hate speech will deter hate crime, our results are primarily intended to be illustrative. We hope that our central insight can be further developed and tested in the future, leading to more definite conclusions.

References

Aronson, E. (1992) *The Social Animal*, 6th ed., W.H. Freeman: New York, NY.

Asch, S. E. (1951) “Effects of Group Pressure upon the Modification and Distortion of Judgment,” in M. H. Guetzkow (ed.) *Groups, Leadership, and Men*, Carnegie Press: Pittsburgh, PA, pp. 177-190.

Becker, G. S. (1968) “Crime and Punishment: An Economic Approach,” *Journal of Political Economy*, 76, 169-217.

Block, M. K. and V. E. Gerety (1995) “Some Experimental Evidence on Differences between Student and Prisoner Reactions to Monetary Penalties and Risk,” *Journal of Legal Studies*, 24, 123-138.

Brennan, H. G. and P. N. Pettit (2000) “The Hidden Economy of Esteem,” *Economics and Philosophy*, 16, 77-98.

Cooter, R. D. (2000) *The Strategic Constitution*, Princeton University Press: Princeton, NJ.

Cowen, T. (2002) “The Esteem Theory of Norms,” *Public Choice*, 113, 211-224.

Frable, D. E. S. (1993) “Being and Feeling Unique: Statistical Deviance and Psychological Marginality,” *Journal of Personality*, 61, 85-110.

Gilbert, D. T. and P. S. Malone (1995) “The Correspondence Bias,” *Psychological Bulletin*, 117, 21-38.

Glaeser, E. L. (2002) “The Political Economy of Hatred,” Harvard Institute for Economic Research Discussion Paper 1970.

Heyman, S. J. (1996) “Hate Speech and the Theory of Free Expression,” in S. J. Heyman (ed.) *Hate Speech and the Constitution*, Garland Publishing; New York and London, i-xciii.

Hylton, K. N. (1996) “Implications of Mill’s Theory of Liberty for the Regulation of Hate Speech and Hate Crimes,” *University of Chicago Law School Roundtable*, 3, 35-57.

Jones, E. E. (1979) “The Rocky Road from Acts to Dispositions,” *American Psychologist*, 34, 107-117.

Jones, E. E. (1990) *Interpersonal Perception*, W.H. Freeman: New York, NY.

Jones, E. E. and V. A. Harris (1967) “The Attribution of Attitudes,” *Journal of Experimental Social Psychology*, 3, 1-24.

Krueger, J. (1997) “On the Perception of Social Consensus,” *Advances in Experimental Social Psychology*, 30, 163-240.

Krueger, J. and R. W. Clement (1997) “Estimates of Social Consensus by Majorities and Minorities: The Case for Social Projection,” *Personality and Social Psychology Review*, 1, 299-313.

Kuran, T. (1995) *Private Truths, Public Lies: The Social Consequences of Preference Falsification*, Harvard University Press: Cambridge, MA.

Lippman, S and J. McCall (1981) “The Economics of Uncertainty,” in K. J. Arrow and M. D.

Intriligator (eds.) *Handbook of Mathematical Economics*, Vol. 1, North-Holland Publishing Co: Amsterdam.

Loury, G. C. (1994) "Self-Censorship in Public Discourse: A Theory of 'Political Correctness' and Related Phenomena," *Rationality and Society*, 6, 428-461.

Marks, G., and N. Miller (1987) "Ten Years of Research on the False-Consensus Effect: An Empirical and Theoretical Review," *Psychological Bulletin*, 102, 72-90.

McAdams, R. H. (1995) "Cooperation and Conflict: The Economics of Group Status Production and Race Discrimination," *Harvard Law Review*, 108, 1003-1084.

McAdams, R. H. (1997) "The Origin, Development and Regulation of Norms," *Michigan Law Review*, 96, 338-433.

McAdams, R. H. (2000) "An Attitudinal Theory of Expressive Law," *Oregon Law Review*, 79, 339-390.

McKinnon, C. A. (1993) *Only Words*, Harvard University Press; Cambridge, MA.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995) *Microeconomic Theory*, Oxford University Press; New York, NY.

Matsuda, M., C. R. Lawrence III, R. Delgado, and K. W. Crenshaw (1993) *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, Westview Press; Boulder, CO.

Miller, D. T. and D. A. Prentice (1996) "The Construction of Social Norms and Standards," in *Social Psychology: Handbook of Basic Principles* (E. T. Higgins and A. W. Kruglanski, eds.), 799-829, The Guilford Press: New York, NY.

Morris, S. (2001) "Political Correctness," *Journal of Political Economy*, 109, 231-265.

Neilson, W. S. and H. Winter (1997) "On Criminals' Risk Attitudes," *Economics Letters*, 55, 97-102.

Pettit, P. N. (1990) "*Virtus Normativa*: Rational Choice Perspectives," *Ethics*, 100, 725-755.

Polinsky, A. M. and S. Shavell (1979) "The Optimal Tradeoff between the Probability and Magnitude of Fines," *American Economic Review*, 69, 880-891.

Posner, R. A. (1986) "Free Speech in an Economic Perspective," *Suffolk University Law Review*, 20, 1-54.

Rasmusen, E. B. (1998) "The Economics of Desecration: Flag Burning and Related Activities," *Journal of Legal Studies*, 27, 245-270.

Rosenfeld, M. (2001) “Hate Speech in Constitutional Jurisprudence: A Comparative Analysis,” Cardozo Law School, Public Law Research Paper No. 41.

Rothschild, M. and J. Stiglitz (1970) “Increasing Risk: I. A Definition,” *Journal of Economic Theory*, 2, 225-243.

Snyder, M. and E. E. Jones (1974) “Attitude Attribution When Behavior Is Constrained,” *Journal of Experimental Social Psychology*, 10, 585-600.

Stamland, T. and J. Shogren (2002) “Signals, Screens, and Spin: The Use of Information and Misinformation to Reduce Violent Activism,” mimeo.