


RESEARCH ARTICLE

A survey inquiry into behavioral foundations of hate speech regulations: evidence from Japan

Kentaro Hirose^{1*} , Hae Kim² and Masaru Kohno³

¹Department of International Studies and Regional Development, University of Niigata Prefecture, Niigata, Japan,

²Department of Education, Chiba University, Chiba, Japan and ³Department of Political Science and Economics, Waseda University, Tokyo, Japan

*Corresponding author. E-mail: hirose1981@gmail.com

(Received 3 March 2022; revised 25 May 2022; accepted 12 August 2022)

Abstract

This paper highlights the concept of dignity as the cornerstone that justifies hate speech regulations in democratic societies. In political theory and constitutional law, the primacy of dignity as the moral and legislative justification for regulating hate speech has already been addressed by dignitarianism, especially in the course of debate with free speech advocates. We aim to augment this important claim in the normative literature with empirical data. Specifically, based on our survey conducted in Japan, where its first national anti-hate speech law had only recently been enacted and ordinary citizens were thus less predisposed of the debate, we show that citizens' concerns about the dignity of a targeted victim lead them to support regulations. Our analysis further clarifies the possible mechanisms of the dignitarian rationale, revealing not only the people's public-centered expectation regarding the societal consequences of hate speech, which dignitarians emphasized, but also the importance of more individual-based judgments regarding morality and justice, in shaping their regulatory attitudes.

Keywords: Dignity; hate speech regulations; mediation analysis; offense; survey

1. Introduction

What justifies regulating hate speech in democratic societies?¹ Whether and how to regulate hate speech has been a subject of controversy for decades among jurists, activists, and political theorists around the world (Baker, 1989; Post, 1991; Coliver, 1992; Matsuda *et al.*, 1993; Weinstein, 1999; Lewis, 2007; Hare and Weinstein, 2009; Herz and Molnar, 2012). In many countries, including the UK, Canada, and Germany, laws have been established to regulate and penalize hate speech, but their legislative contents vary considerably in what sort of act against whom should be regulated as well as the severity of applicable criminal sanctions.² In the USA, there is no equivalent law, as free speech is most ardently protected by the First Amendment to the US Constitution with courts only

¹Hate speech is now generally understood as messages intended to incite hatred and/or encourage violence toward a person on the basis of membership in a particular social group. As indicated below, however, there is no universally accepted definition for the term.

²Public Order Act in the UK (section 18) stipulates that 'A person who uses threatening, abusive or insulting words or behaviour, or displays any written material which is threatening, abusive or insulting, is guilty of an offence if (a) he intends thereby to stir up racial hatred, or (b) having regard to all the circumstances racial hatred is likely to be stirred up thereby.' Criminal Code in Canada (section 319) stipulates that 'Everyone who, by communicating statements in any public place, incites hatred against any identifiable group where such incitement is likely to lead to a breach of peace is guilty of an indictable offence and is liable to imprisonment for a term not exceeding two years.' Penal Code in Germany (section 130) stipulates that 'Whoever, in a manner that is capable of disturbing the public peace: 1. incites hatred against segments of the population or calls for violent or arbitrary measures against them; or 2. assaults the human dignity of others by insulting,

allowing for so-called ‘content-neutral’ restrictions (Stone, 1987); this regulatory absence is often described as ‘exceptionalism’ (Krotoszynski, 2006) or ‘American way’ (Rosenfeld, 2003) in constitutional approaches to hate speech.³ Given the enduring disagreement and observed variety, it must be tempting for scholars to drift away from a quest for a uniform theory on hate speech regulations. Indeed, the author of a recent review article suggests that ‘the largely muddled debate over hate speech needs to be broken down into discrete analytical stages’ (Howard, 2019: 95).⁴

How can one challenge this seeming theoretical inertia? In this paper, we revisit an argument advanced by Jeremy Waldron in his acclaimed book, *The Harm in Hate Speech* (Waldron, 2012b). At the core of his thesis lies the distinction between the two kinds of harm that hate speech may incur against the targeted individual, namely ‘undermining dignity’ and ‘causing offense.’ According to Waldron, offense, ‘however deeply felt, is not a proper object of legislative concern,’ since it is ‘inherently a subjective reaction,’ and the law in modern era is never meant to protect anybody’s feelings. Dignity, on the other hand, merits due protections; ‘not dignity in the sense of any particular level of honor or esteem (or self-esteem), but dignity in the sense of a person’s basic entitlement to be regarded as a member of society in good standing.’⁵ In this respect, the harm done against the individual by undermining his/her/their dignity becomes the harm done to the ‘public good’ of that society, the provision of which must be assured by law even as balanced against the importance of free speech principle.⁶

We seek to augment this normative argument, now referred to as ‘dignitarian rationale’ or simply ‘dignitarianism’ in the literature, by demonstrating, through comprehensive empirical evaluations, that the concept of dignity does serve as the cornerstone for justifying regulations of hate speech. For this task, we take advantage of the data from a survey we conducted among ordinary Japanese citizens in 2018. As explained below, Japan then, having only recently (in 2016) adopted the nation’s first anti-hate speech law, was like an incubating ground where norms and interpretations on the subject were in the making. This setting provided a uniquely suited opportunity for our research, unlike Canada and European countries where the regulations had already been in place for some time, and unlike the USA where the opinions for and against governmental regulations are so deeply entrenched in the respective ideological camps. Moreover, Japan offers an additional advantage for our purpose in that, because of the relative homogeneity of its population compared with other countries, few survey respondents would be expected to belong to minority group, that is, potential target of hate speech. This provides us with an ideal situation for examining the attitudes of *ordinary* citizens toward hate speech regulations. To be specific, in Japan, the most widely reported victims of hate speech are Koreans residing in Japan (*Zainichi* Koreans); it is estimated that they constitute less than 0.5% of Japan’s entire population.

We design our survey, most importantly, to juxtapose and test the relative significance of the perceived level of ‘offense’ and that of ‘indignity’ caused by hate speech as potential sources for respondents’ support for hate speech regulations. Our findings confirm that respondents’ concerns about the dignity of a targeted victim lead them to support regulations far more strongly and consistently than their concerns over whether the victim is offended. After verifying this, our research design further allows us to explore possible causal pathways through which these perceptions affect their regulatory attitudes. Such an analysis consolidates the bridge between normative theorization like Waldron’s and our own empirical findings, clarifying the underlying psychological mechanism at work for ordinary citizens who support hate speech regulations. In the dignitarian discussion, it is simply assumed that the degrading of the society, and hence the erosion of the ‘public good,’ constitutes regulatory

maliciously maligning, or defaming segments of the population, shall be punished with imprisonment from three months to five years.’

³Staunch First Amendment defenders reject these characterizations. See, for example, Baker (2012: 58–60).

⁴For conceptual and regulatory varieties, see also Herz and Molnar (2012) and Brown (2015).

⁵See especially Chapter 3 of Waldron (2012b). Direct quotes here are from pages 105–107.

⁶See also Waldron (2012a). Other earlier works that recognized dignity as potential regulatory justification include Heyman (2009) and Tsesis (2009), though they did not explicitly juxtapose indignity against offense. See also Jones (2011) for a relevant assertion that offense should not be recognized as grounds for justifying hate speech regulations.

justification, but this claim has never been subject to empirical scrutiny. It turns out, according to our mediation analysis, that ordinary citizens' regulatory attitudes are shaped not only by such concerns for societal consequences but also by their intrinsic senses of justice and morality. Although the external validity of our findings is ultimately confined to the Japanese case at hand, the implications drawn from the analysis in this paper do speak to the generalizability of the dignity-based argument, which, in our conclusion, provides an important insight, if not the basis, for any discussion on hate speech regulations.

The rest of this paper is organized as follows. In Section 2, we highlight the main characteristics of the dignitarian rationale by reviewing Waldron's original discussion with some critical annotations. Section 3 reviews extant empirical approaches to hate speech. Section 4 presents testable hypotheses. Sections 5 and 6 describe our research design and findings. Section 7 concludes by drawing broad implications and pointing to directions of future research.

2. Dignitarianism and its critics

In any democracy, particularly under the American context where the vast First Amendment jurisprudence has accumulated, it would be controversial to claim that certain speech and speech acts, be it hate speech aimed at minorities, publications of pornography, or the spreading of so-called 'fake news' through the internet, should be banned because they are unworthy of protection by the principle of free speech. In advancing his defense for hate speech regulations, Jeremy Waldron does not make such a claim. He is a consequentialist and rather sees the matter from a balancing perspective: 'We recognize, in general, that the considerations which argue in favor of the broad importance of free-speech *do* extend to speech attempting to stir up racial or religious hatred; but we say nevertheless such speech must be regulated, and in extreme cases prohibited because of the harm it does' (Waldron, 2012b: 147, emphasis original). But this position, of course, begs the question: what harm?

In actuality, the harm caused by hate speech can and do take various forms, as identified and listed elsewhere in the previous literature (e.g., Delgado, 1993; Brown, 2015). Hence, any plausible justification for regulating hate speech must be premised on a well-reasoned argument that not only discerns different aspects of the harm caused but addresses their respective behavioral and regulatory implications. For example, an approach that exclusively focuses on physical aspects of the harm caused by hate speech would not constitute a distinct justification, since such an approach would be equating the act of hate speech with assault. Alternatively, an approach that excessively emphasizes psychological aspects of the harm caused by hate speech would be difficult to serve as a viable regulatory justification, because of subjective elements inherent in human feelings and attitudes which are bound to vary considerably across individuals.

In this context, Waldron's approach stands out in the literature for its innovative and nuanced prodding, as he draws the crucial distinction between the two likely consequences that hate speech may incur against the targeted individual: 'undermining dignity' and 'causing offense.' This distinction, he explained, 'is in large part between objective or social aspects of a person's standing in society and subjective aspects of feeling, including hurt, shock, and anger' (Waldron, 2012b: 106). Waldron argues that the latter does not justify regulations, since the modern laws are never meant to protect anybody's feelings. The former, however, provides the cornerstone for hate speech regulations, because a 'democratic society cannot work, socially or politically, unless its members are respected in their character as equals, and accorded the authority associated with their vote and their basic rights' (Waldron, 2012b: 109). Thus, Waldron's dignity-centered argument marks a departure from the orthodox liberal tradition, advancing the agenda of hate speech regulations from simply a matter of balancing individual rights to something that involves 'public good' at stake.⁷

Dignitarianism, of course, has since been criticized. The criticisms range from the most predictable First-Amendment advocacy (e.g., McConnell, 2012) to a more sophisticated counter-argument that

⁷For a different appraisal, see, for example, Jones (2015: esp. 682).

restriction of free speech would weaken democratic legitimacy (cf. Dworkin, 2009, 2012; Weinstein, 2017), and to nuanced assessments which, while basically sympathetic, point to some specific weaknesses in its logic, such as the inability to make a stronger causal argument (Barendt, 2019) or its failure to account for social hierarchy (Simpson, 2013).⁸ None, however, to our knowledge, has ever questioned the validity of the distinction between offense and indignity. The absence is surprising because the dignitarian rationale for regulating hate speech is premised upon this distinction. Even when some critics point out that making such distinction deems difficult (e.g., Leiter, 2012: 6–7), the issue is not pursued further, as if the difficulty is taken for granted by both sides of the debate.

Indeed, Waldron himself concedes the difficulty when he describes the seamless chain of reactions that take place in the mind of a targeted victim. By admitting this ‘psychological complexity,’ he is forced to acknowledge his inability to identify ‘the lawfulness and unlawfulness of certain speech acts on the basis of a case-by-case analysis’ (Waldron, 2012*b*: 113). How then can one confront two alleged victims and determine whether a punishable act was inflicted on one, both, or neither? Again, Waldron admits the difficulty, and by this time his defense becomes circular: ‘I am not proposing a complicated legal test for distinguishing hate speech from speech that merely offends. I am only suggesting that in defending (or arguing about) such a distinction, we should be willing to come to terms with psychological complexity’ (Waldron, 2012*b*: 115).

This retreat is problematic because the dignitarian argument that hate speech must be regulated when and only when it harms the target’s dignity is not simply a philosophical proposition; it is also a policy proposal. It is one thing to uphold the importance of the conceptual distinction between indignity and offense in the abstract. To show such distinction can be utilized justifiably in drafting, implementing and adjudicating laws is quite another. Waldron does not theorize his argument on the basis of public support for it. We believe that it is this lack of behavioral foundation that limits the persuasiveness of his argument. If the dignitarian proposition is to be presented as a policy proposal in democratic societies, what enhances its plausibility must be direct evidence that a majority of citizens actually support the proposition/proposal. It is precisely such evidence that our survey aims to reveal.

3. Empirical approaches to hate speech

The empirical literature on hate speech extends over many academic disciplines, and our research is certainly not the first to take advantage of survey or survey-experimental methods. Social-psychologists, criminologists as well as scholars in racial and cultural studies have conducted numerous voluntary-based interviews and small-scale experiments in their respective fields. These previous studies, however, have focused on the actual or potential victims of hate speech, such as gays, lesbians, and bisexuals in California (Herek *et al.*, 2002), Jewish and gay students on college campuses (Leets, 2002), Asian-American university students (Boeckmann and Liew, 2002), and indigenous and minority ethnic communities in Australia (Gelber and McNamara, 2016). We maintain that an inquiry into the public support for governmental regulations of hate speech requires a sample of respondents that are randomly drawn from ordinary citizens in a given democratic society.

Our approach also differs from the previous studies in terms of the main dependent variable measure. Typically, as in the works of Gloria Cowan and her co-authors, psychological surveys and survey experiments include a battery of value-related questions to gauge their perception about the harm of hate speech along with their attitudes toward the principle of free speech (Cowan and Hodge, 1996; Cowan *et al.*, 2002; Cowan and Khatchadourian, 2003; Downs and Cowan, 2012). Since we are interested in the respondents’ legislative preference, we aim to measure directly their attitudes toward governmental regulations, while treating the perceived harms of hate speech and attitudes toward the principle of free speech as key explanatory variables and control variables, respectively. As far as we know, this is the first empirical research to explain ordinary citizens’ attitudes toward hate speech regulations.

⁸See, also, Heinze (2013), Zivi (2014), and Seglow (2016).

4. Hypotheses

Our survey investigates whether the dignitarian argument has an empirical foundation. To achieve this goal, we test the validity of several hypotheses. The first set of hypotheses speaks to the core argument of dignitarianism that hate speech should be regulated when it harms the target's dignity and not when it harms the target's feelings. For this argument to stand as a plausible policy or policy guideline in democratic societies, we need to observe the following two empirical patterns:

Hypothesis 1. Citizens' attitudes toward hate speech regulations are shaped independent of how much harm they perceive is done to the victims' feelings.

Hypothesis 2. Citizens' support for hate speech regulations is higher when they perceive that the greater harm is done to the victims' dignity.

To be sure, these two hypotheses are consistent with, but they are not logically derived from, dignitarianism, since the proponents of this normative theory do not elaborate on how the members of a democratic society think about hate speech regulations. The above hypotheses are presented here because we believe that their empirical tests serve as a touchstone for the feasibility of the dignitarian proposition.

If the empirical patterns consistent with Hypotheses 1 and 2 are confirmed, the next step is to ask why. To answer this question, we consider two possible explanations. The first explanation is originally offered by dignitarian theorists themselves, namely, the idea of 'public good': the government should regulate dignity-harming hate speech because such speech will move the society in a bad direction, whereas feelings-harming hate speech should not be regulated because such speech will not erode the 'public good' of the society. To examine whether ordinary citizens believe in the same way as this consequentialist logic, we test the following three hypotheses:

Hypothesis 3. Citizens' belief that hate speech makes their society worse is shaped independent of how much harm they perceive is done to the victims' feelings.

Hypothesis 4. Citizens' belief that hate speech makes their society worse is stronger when they perceive that the greater harm is done to the victims' dignity.

Hypothesis 5. Citizens' support for hate speech regulations is higher when they more strongly believe that hate speech makes their society worse.

In addition to this original logic about societal consequences, we suspect that ordinary citizens may also be motivated to support hate speech regulations based on their intrinsic senses of justice and morality. This second reasoning can be considered as an ethics-driven, deontological (as opposed to consequentialist) mechanism: citizens believe that the government should regulate dignity-harming hate speech because such speech is intrinsically wrong and unjust (whereas feelings-harming hate speech is not). We thus contrast the above set of hypotheses with the following set of alternative hypotheses:

Hypothesis 6. Citizens' belief that hate speech is unjust is shaped independent of how much harm they perceive is done to the victims' feelings.

Hypothesis 7. Citizens' belief that hate speech is unjust is stronger when they perceive that the greater harm is done to the victims' dignity.

Hypothesis 8. Citizens' support for hate speech regulations is higher when they more strongly believe that hate speech is unjust.

5. Research design

In this section, we describe the regulatory background in Japan where our survey was conducted, and then explain our sample and basic design.

5.1 Background

Our survey was conducted through internet with adult residents in Japan in March 2018, less than 2 years after Japan's parliament passed the bill called 'The Act on the Promotion of Efforts to Eliminate Unfair Discriminatory Speech and Behavior against Persons Originating from Outside Japan.'⁹ Prior to the enactment of this law, there had been no formal regulation on hate speech or hate speech act in Japan. The conservative political establishments, including senior members of the ruling Liberal Democratic Party as well as the Ministry of Justice, had long been reluctant to endorse any governmental regulation, which would infringe the constitutionally guaranteed right of free speech and expression.

A rather abrupt momentum toward establishing a new law was born, at least in part, in response to the increased pressures from the international community, especially from the United Nations, which came in the form of various reports and recommendations. It was in the spring of 2013 when The United Nations Committee on Economic, Social, and Cultural Rights for the first time acknowledged the existence of problems in this country by explicitly using the term 'hate speech' in its third periodic report on the implementation of the International Covenant on Economic, Social, and Cultural Rights. In the summer 2014, the two prominent organizations at the United Nations, The Office of the UN High Commissioner for Human Rights and The UN Committee on the Elimination of Racial Discrimination, expressed concerns over the rising racist demonstrations against ethnic Koreans residing in Japan (*Zainichi* Koreans).¹⁰ It was only after the international pressure grew that the domestic media finally started to cover these demonstrations in its report; prior to 2013, even nationally circulated newspapers, such as *Yomiuri* and *Asahi*, had never used the word 'hate speech' in their printed articles. It can thus be assumed that the very concept of hate speech had not previously been known to ordinary citizens in Japan; even today, this concept is expressed in Katakana syllabary, a component of Japanese writing system specifically used for transcribing words of foreign language origins.

The enacted law is often criticized because of its exclusive focus on foreigners residing in Japan, particularly *Zainichi* Koreans. Hate speech against other potentially targeted minority groups, such as gays and lesbians, elderlies, people with disabilities, and holders of certain religious beliefs or political ideologies, are not covered by this legislation. Further, human-rights advocates have expressed their dissatisfactions with the law, particularly because it neither criminalizes any speech act nor includes specific enforcement procedures (cf. Kotani, 2018). These limitations notwithstanding, some prefectural and city governments in Japan have since used and relied upon the spirit of this national law to issue harder restrictions in their respective localities, sometimes establishing their own ordinances with criminal sanctions and extending the coverage of protections to other minority groups.

⁹This law was enacted on 24 May, and came into effect on 3 June 2016. For English translation, see http://www.moj.go.jp/ENGLISH/m_jinken04_00001.html.

¹⁰See, The Office of the United Nations High Commissioner for Human Rights, 'Concluding observations on the sixth periodic report of Japan,' CCPR – International Covenant on Civil and Political Rights, 111 Session (7 July 2014–25 July 2014), and The United Nations Committee on the Elimination of Racial Discrimination, 'Concluding observations of the Committee on the Elimination of Racial Discrimination, Japan,' U.N. Doc. CERD/C/JPN/CO/7-9 (2014). *Zainichi* Koreans include those ethnic Koreans who possess permanent residency status in Japan, those whose immigration to Japan originated before 1945, and those who are descendants of those immigrants. See Matsui (2016).

In sum, it is fair to characterize Japan in 2018, when our survey was conducted, as not having developed stable norms or interpretations on the subject of hate speech regulations: ordinary Japanese citizens were in the midst of developing new norms and interpretations on the subject. Japan then was like an incubating ground, which we believe provided a uniquely suited opportunity to engage our survey inquiry.

5.2 Sample

To administer our survey, we contracted with Nikkei Research, one of the major online survey companies in Japan. The company sent invitation emails to some of the people pre-registered at it and recruited 390 respondents for this research. A monetary incentive was provided in return for their participation, and the amount of money provided to each respondent was determined by the individual contract between the company and the respondent. A stratified random sampling procedure was used when sending out invitation emails so that the appropriate number was assigned to each category of gender and geographical regions in proportion to the actual demographic data reported in Japan's latest edition of *Juminkihondaicho* (Basic Residence Register).

Although invitation emails were sent randomly, the pool of pre-registered individuals from which our respondents were recruited is not a representative sample of the population. Due to the skewed distribution of internet users among the elderly, our sample is limited to those between 20 and 69 years of age. The median respondent in the sample is thus slightly younger than the median Japanese resident. It was also explained to us that, in comparisons with the actual population in Japan, the Nikkei Research samples are generally skewed to higher levels of income and education. We will discuss these data problems later in the paper.

To ensure that our analysis concentrates on valid responses from attentive survey takers, we included two attention-check questions in the survey to filter out inattentive respondents, or 'satisficers.' These questions were simple instructed-response items that anyone paying attention should be able to answer. Respondents were excluded from subsequent analysis if they answered either of these questions incorrectly. After filtering them out, we retain 309 respondents for our analysis.

5.3 Measurement

Our key explanatory variables are the perceived levels of offense and indignity that may influence the respondents' attitudes toward hate speech regulations. To measure these variables, we asked them the following questions:

Recently in Japan, in some areas or on the internet, you can find hate speech, that is, insults and incitements to violence against *Zainichi* Koreans.

- Do you think such hate speech, that is, insults and incitements to violence, against *Zainichi* Koreans would make them feel offended? Or do you think that it would not make them feel offended?
- Do you think such hate speech, that is, insults and incitements to violence, against *Zainichi* Koreans would harm their dignity? Or do you think that it would not harm their dignity?

For each of these questions, the answer options were provided on a 5-point scale with larger values indicating greater perceived harms to the victims of hate speech. In the case of the question about the perceived harm to feelings, the options were: (1) Absolutely would not make them feel offended; (2) Probably would not make them feel offended; (3) Neither; (4) Probably would make them feel offended; and (5) Absolutely would make them feel offended. For the question about the perceived harm to dignity, the options were: (1) Absolutely would not harm their dignity; (2) Probably

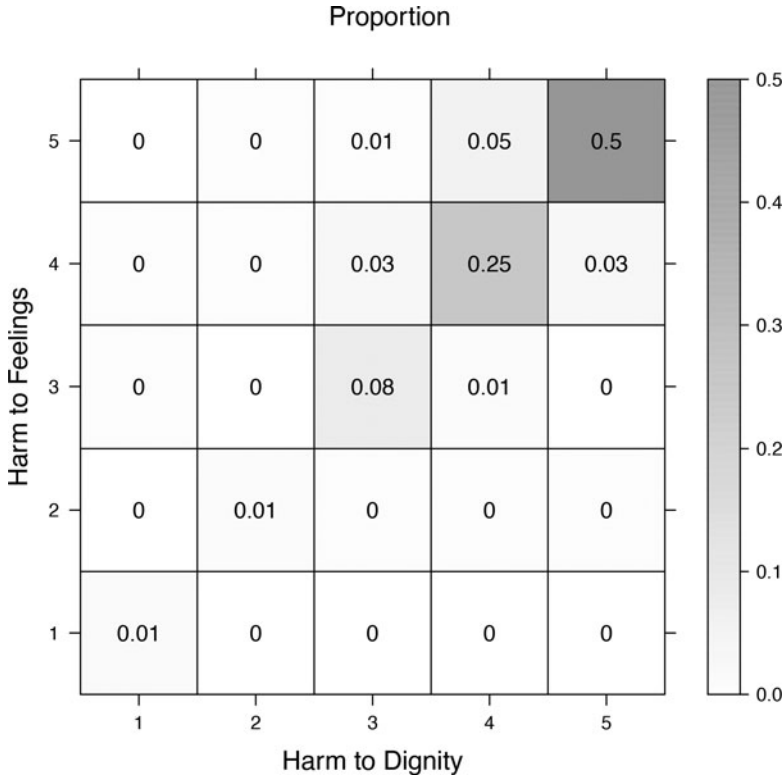


Figure 1. Correlation between the perceived harm to feelings and the perceived harm to dignity.

would not harm their dignity; (3) Neither; (4) Probably would harm their dignity; and (5) Absolutely would harm their dignity. The order of these two questions was randomized across the respondents. To ensure a consistent definition of *Zainichi* Koreans – the most frequent target of hate speech in Japan and the main group to which Japan’s anti-hate speech law is intended to provide protection – a brief sentence was added in the vignette to explain who *Zainichi* Koreans are.¹¹

Figure 1 shows the distributions of as well as the correlation between the perceived harm to feelings and the perceived harm to dignity. One prominent feature of these two variables is that most respondents perceive that hate speech against *Zainichi* Koreans would harm their feelings (87%) and dignity (84%), respectively. Another key feature is a strong positive correlation between the two variables. For example, 85% of the respondents are on the 45-degree line. We emphasize here, however, that despite the strong correlation, the two variables in terms of their effects on the outcome and intervening variables turn out to be clearly distinguishable, as our empirical analyses below demonstrate.

For descriptive purposes, it may also be interesting to know what types of respondents perceive that hate speech causes offense and indignity, respectively. Table 1 shows the results of linear models that regress the perceived harms of hate speech on the respondent’s gender, age, education, income, 11-point scale liberal-conservative ideology measure with greater values indicating conservative political positions, 4-point scale measure of free speech with greater values indicating that the respondent respects the freedom of speech more strongly, and relationship with victims of hate speech, that is, a binary variable of whether the respondent or persons close to him/her were victimized by hate speech.

¹¹The sentence added to explain *Zainichi* Koreans reads: ‘*Zainichi* Koreans mean Korean nationals who reside in Japan (including special long-term residents who have lived in Japan before the Second World War or those who are their decedents).’

Table 1. Sources of the perceived harms of hate speech

	Dependent variable	
	Harm to feelings (1)	Harm to dignity (2)
Male	-0.081 (0.106)	-0.064 (0.110)
Age	0.005 (0.004)	0.007 (0.004)
Ideology	-0.022 (0.024)	-0.056 (0.029)
Free speech	-0.038 (0.064)	-0.048 (0.070)
Victim	0.047 (0.311)	0.196 (0.309)
Four-year college or above	0.079 (0.110)	-0.017 (0.114)
Median income or above	0.010 (0.117)	0.129 (0.131)
Constant	4.431*** (0.311)	4.457*** (0.334)
Observations	247	246

Note: Linear regression models are used. Robust standard errors are in parentheses.

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

Interestingly, none of these covariates helps us explain variation in the perceived harm to feelings and the perceived harm to dignity, respectively.

The outcome variable in our research is the level of support for governmental regulations on hate speech. To measure this variable, we asked the respondents the following question:

Next, we ask you how the government should respond to this issue. Do you think that the government should impose restrictions on such hate speech, that is, insults and incitements to violence, against *Zainichi* Koreans? Or do you think that the government should not impose any restrictions?

For this question, the answer options were provided on a 7-point scale: 0 meaning 'Should not impose any restrictions,' 3 meaning 'Cannot say one way or the other,' and 6 meaning 'Should impose thorough restrictions.' Figure 2 shows the distribution of this variable. There are far more respondents who support hate speech regulations (57%) than those who do not (17%), but there are also a significant number of respondents who stand in the middle (25%).

As the intervening variables that may link the explanatory and outcome variables specified above, we contrast two possible mechanisms noted earlier, namely the societal-concern mechanism and the sense-of-injustice mechanism. We believe that these two possible mechanisms are not mutually exclusive, although the dignitarian rationale does not fully address the second pathway. In order to probe the saliency of the two mechanisms, we used the relevant questions included in our survey:

- What effect do you think such hate speech, that is, insults and incitements to violence against *Zainichi* Koreans would have on the Japanese society? Do you think it would move the Japanese society in a good direction or in a bad direction?
- Do you think such hate speech, that is, insults and incitements to violence against *Zainichi* Koreans would be just? Or do you think that it would be unjust?

For each of these questions, the answer options were provided on a 5-point scale with larger values indicating greater perceived harms. In the case of the question about societal concern, the options

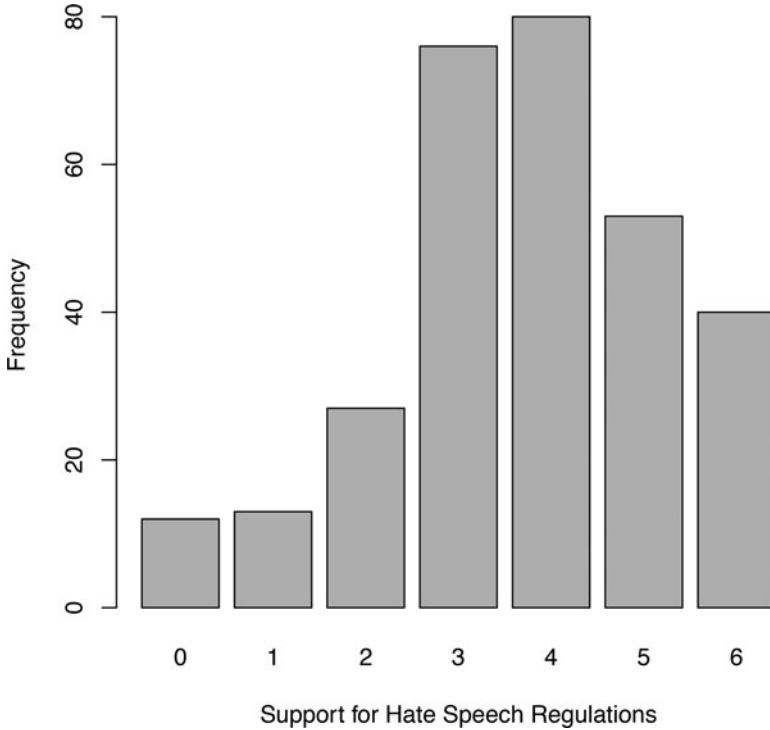


Figure 2. Distribution of support for hate speech regulations.

were: (1) Absolutely would make it worse; (2) Would rather make it worse; (3) Neither; (4) Would rather make it better; and (5) Absolutely would make it better. For the question about the sense of injustice, the options were: (1) Absolutely would not be unjust; (2) Would not rather be unjust; (3) Neither; (4) Would rather be unjust; and (5) Absolutely would be unjust. As shown in Figure 3, most respondents believe that hate speech against *Zainichi* Koreans would make the society worse (73%) and be unjust (70%), and societal concern and the sense of injustice are positively correlated with one another (61% of the respondents are on the 45-degree line).¹²

5.4 Models

To investigate how the respondents' attitudes toward hate speech regulations is shaped by the perceived harms of hate speech, we use the following linear regression model:

$$\begin{aligned} \text{Support for Hate Speech Regulations} &= \lambda_0 + \lambda_1(\text{Harm to Feelings}) \\ &+ \lambda_2(\text{Harm to Dignity}) + \sum_k \lambda_k Z_k + \nu, \end{aligned} \tag{1}$$

where Z represents a set of control variables including the respondent's gender, age, education, income, ideology, attitude toward free speech, and relationship with victims of hate speech.

¹²We did ask the respondents if hate speech would make the society worse or be unjust, but did not ask whether hate speech regulations would make the society worse or be unjust. Although we believe that the latter question is not relevant for our purpose of testing the feasibility of the dignitarian policy, it might have been helpful for more fully describing people's attitudes toward hate speech and hate speech regulations. We thank one of the reviewers for the journal whose comment made us realize this point.

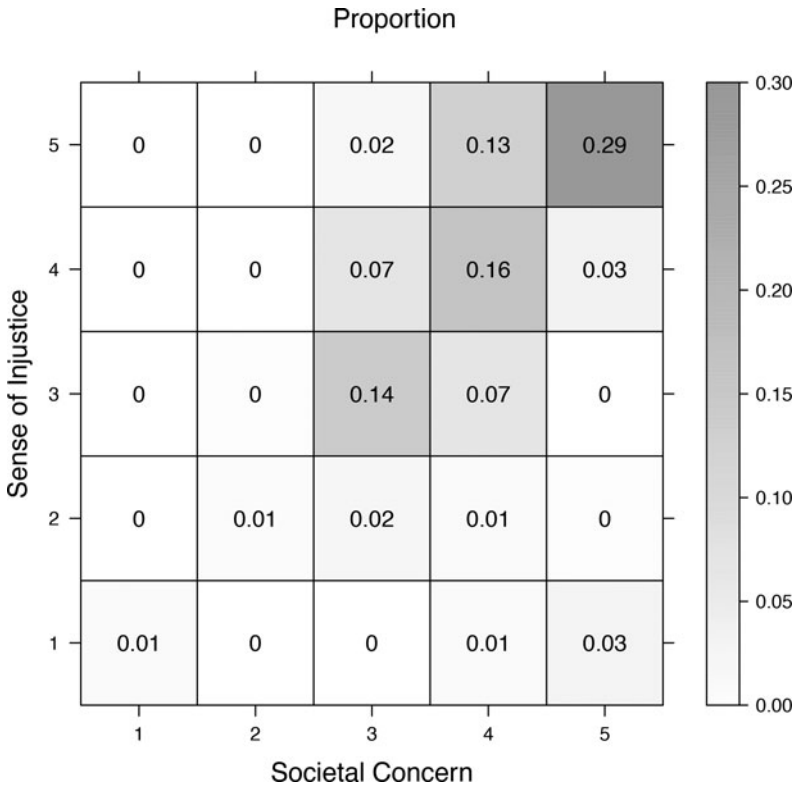


Figure 3. Correlation between societal concern and the sense of injustice.

To examine the mechanisms underlying the associations between the explanatory and outcome variables, we follow the conventional procedure of mediation analysis, by fitting the following three linear regression models:

$$\begin{aligned}
 &\text{Societal Concern} \\
 &= \alpha_0 + \alpha_1(\text{Harm to Feelings}) \\
 &\quad + \alpha_2(\text{Harm to Dignity}) + \sum_k \alpha_k Z_k + e
 \end{aligned} \tag{2}$$

$$\begin{aligned}
 &\text{Sense of Injustice} \\
 &= \gamma_0 + \gamma_1(\text{Harm to Feelings}) \\
 &\quad + \gamma_2(\text{Harm to Dignity}) + \sum_k \gamma_k Z_k + w
 \end{aligned} \tag{3}$$

$$\begin{aligned}
 &\text{Support for Hate Speech Regulations} \\
 &= \beta_0 + \beta_1(\text{Harm to Feelings}) \\
 &\quad + \beta_2(\text{Harm to Dignity}) \\
 &\quad + \beta_3(\text{Societal Concern}) \\
 &\quad + \beta_4(\text{Sense of Injustice}) + \sum_k \beta_k Z_k + u.
 \end{aligned} \tag{4}$$

Model 2 investigates how the respondents' societal concern is shaped by their perceived harm to feelings (α_1) and their perceived harm to dignity (α_2). Model 3 analyzes how the respondents' sense of injustice is shaped by their perceived harm to feelings (γ_1) and their perceived harm to dignity (γ_2). Model 4 explores how the respondents' regulatory attitude is shaped by their societal concern (β_3) and their sense of injustice (β_4) while adjusting for the unmediated association between the regulatory attitude and the perceived harm to feelings (β_1) as well as the unmediated association between the regulatory attitude and the perceived harm to dignity (β_2).

Since we use linear regression models, the association between the perceived harm to feelings and the attitude toward hate speech regulations mediated by the societal-concern mechanism is a product of α_1 and β_3 ; the association between the perceived harm to dignity and the attitude toward hate speech regulations mediated by the societal-concern mechanism is a product of α_2 and β_3 ; the association between the perceived harm to feelings and the attitude toward hate speech regulations mediated by the sense-of-injustice mechanism is a product of γ_1 and β_4 ; and the association between the perceived harm to dignity and the attitude toward hate speech regulations mediated by the sense-of-injustice mechanism is a product of γ_2 and β_4 .

6. Findings

Column (1) of Table 2 shows how the attitude toward hate speech regulations is shaped by the perceived harm to feelings and the perceived harm to dignity, respectively. The respondents' perceived harm to feelings has no statistically significant association with their regulatory attitudes, whereas the association between the perceived harm to dignity and support for hate speech regulations is positive in a statistically significant manner. For example, a shift from the neutral view to the view that hate speech against *Zainichi* Koreans would absolutely harm their dignity is, on average, associated with a 1.39-point increase in the 7-point-scale measure of support for hate speech regulations. These results are consistent with Hypotheses 1 and 2, implying that the dignitarian argument that hate speech should be regulated when and only when it harms the target's dignity may be a feasible policy in democratic societies.

Having established a strong positive association between the perceived harm to dignity and support for hate speech regulations, we now proceed to the next 'why' question. While there may be numerous pathways that link these two variables, we contrast two mechanisms as described earlier: the societal-concern mechanism and the sense-of-injustice mechanism.

Column (2) of Table 2 shows results consistent with Hypotheses 3 and 4: while the respondents' concerns over which direction the society is heading are shaped independently of their levels of the perceived harm to feelings, societal concern is positively and statistically significantly associated with the perceived harm to dignity. For example, a shift from the neutral view to the view that hate speech against *Zainichi* Koreans would absolutely harm their dignity is, on average, associated with a 1.22-point increase in the 5-point-scale measure of societal concern.

Column (3) of Table 2 shows results consistent with Hypotheses 6 and 7: the respondents' beliefs that hate speech is unjust is shaped independently of their levels of the perceived harm to feelings, but the sense of injustice is positively and statistically significantly associated with the perceived harm to dignity. For example, a shift from the neutral view to the view that hate speech against *Zainichi* Koreans would absolutely harm their dignity is, on average, associated with a 1.23-point increase in the 5-point-scale measure of the sense of injustice.

Column (4) of Table 2 shows results consistent with Hypotheses 5 and 8: support for hate speech regulations is positively and statistically significantly associated with societal concern and the sense of injustice. For example, a shift from the neutral view to the view that hate speech against *Zainichi* Koreans would absolutely make the society worse is, on average, associated with a 1.42-point increase in the 7-point-scale measure of support for hate speech regulations. Similarly, a shift from the neutral view to the view that hate speech against *Zainichi* Koreans would absolutely be unjust is, on average, associated with a 0.51-point increase in support for hate speech regulations.

Table 2. Perceived harms, support for regulations, and mechanisms

	Dependent variable			
	Support for regulations (1)	Societal concern (2)	Sense of injustice (3)	Support for regulations (4)
Harm to feelings	-0.050 (0.193)	0.037 (0.081)	-0.039 (0.097)	-0.063 (0.170)
Harm to dignity	0.693*** (0.202)	0.611*** (0.079)	0.613*** (0.108)	0.109 (0.218)
Societal concern				0.709*** (0.162)
Sense of injustice				0.253* (0.127)
Male	-0.073 (0.182)	0.050 (0.088)	0.107 (0.129)	-0.110 (0.173)
Age	-0.004 (0.006)	0.003 (0.003)	-0.0002 (0.005)	-0.006 (0.006)
Ideology	-0.197*** (0.056)	-0.047 (0.024)	-0.102* (0.044)	-0.139** (0.050)
Free speech	0.095 (0.128)	0.037 (0.061)	0.069 (0.095)	0.043 (0.109)
Victim	-0.067 (0.483)	-0.227 (0.261)	-0.623 (0.740)	0.227 (0.354)
Four-year college or above	-0.179 (0.198)	-0.047 (0.090)	-0.173 (0.137)	-0.103 (0.185)
Median income or above	0.111 (0.206)	-0.131 (0.100)	0.250 (0.192)	0.157 (0.202)
Constant	1.843* (0.799)	1.297** (0.397)	1.633*** (0.492)	0.471 (0.727)
Observations	245	243	244	243

Note: Linear regression models are used. Robust standard errors are in parentheses.

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

Figure 4 illustrates these results graphically. A unit increase in the perceived harm to dignity leads to a 0.611 unit increase in societal concern and a 0.613 unit increase in the sense of injustice. And, a unit increase in societal concern brings about a 0.709 unit increase in support for hate speech regulations. Also, a unit increase in the sense of injustice gives rise to a 0.253 unit increase in the level of support. Therefore, a unit increase in the perceived harm to dignity is associated with a $0.611 \times 0.709 \approx 0.433$ unit increase in support for hate speech regulations through the societal-concern mechanism (with a 95% confidence interval of [0.219, 0.66]). It is also associated with a $0.613 \times 0.253 \approx 0.155$ unit increase in the level of support through the sense-of-injustice mechanism (with a 95% confidence interval of [0.006, 0.34]).¹³ Although the logic of societal concern has more explanatory power than that of the sense of injustice, both mechanisms are estimated to be statistically significant. This suggests that the original rationale presented by dignitarian theorists may be insufficient: the link between the perceived harm to dignity and support for hate speech regulations can be explained not only by the citizens' concerns regarding the societal consequences but also by the alternative, ethics-driven logic based on their intrinsic senses of justice and morality.

Before concluding this section, we discuss some potential problems of our empirical investigation. We also report briefly some results from additional tests we performed for robustness checks.

First, the analyses presented above exclude from the sample satisficers, that is, the respondents who did not answer the attention-check questions correctly. We reanalyzed the regression models of Table 2 with the data including those 'careless' respondents, but the results did not change qualitatively except that the association between the sense of injustice and support for hate speech regulations

¹³The 95% confidence intervals for the mediation effects are derived from 1,000 quasi-Bayesian simulations using the *mediation* package of R.

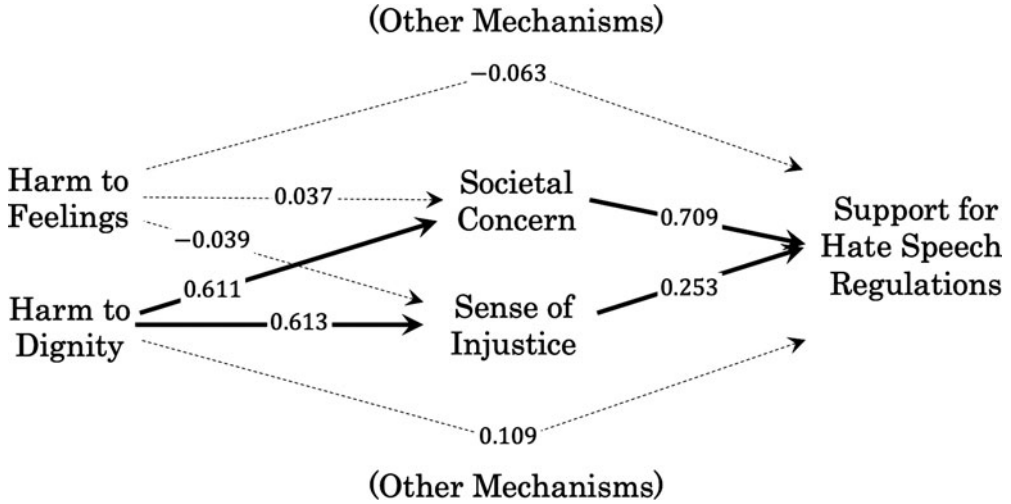


Figure 4. Mechanisms underlying the associations between the perceived harms of hate speech and support for hate speech regulations. The numbers represent the coefficients shown in columns (2)–(4) of [Table 2](#). The thick lines indicate statistically significant associations.

loses its statistical significance (see [Table A1](#) in the online Appendix). We leave to the readers the decision of which results are more convincing.

Second, our regression models include as control variables the respondents’ liberal-conservative ideology measures and attitudes toward free speech, but these two variables may be shaped by their perceived harms to feelings and dignity, respectively. To deal with this possibility of post-treatment bias, we reanalyzed the regression models of [Table 2](#) without controlling for ideology and free speech. The results did not change qualitatively from the original results (see [Table A2](#) in the online Appendix).

Third, online-survey samples generally tend to have richer and more educated respondents than random samples of the population. Although the distribution of the respondents in our sample is only slightly skewed to higher incomes, there are far more educated respondents in our sample compared with the population data (see [Figure A1](#) in the online Appendix). For example, 58% of our respondents graduated from or are students of 4-year college or above, whereas the population proportion is only about 20%. Existing studies do not help us infer the bias derived from this divergence, as this is the first research explaining the association between the perceived harms of hate speech and support for hate speech regulations. Hence, as a second-best solution to this problem, we attempt to measure the bias using our sample data. As already shown in [Table 1](#), education and income do not affect the perceived harms of hate speech, implying that the divergence between the sample and population distributions of income and education may not lead to a significant divergence between the sample and population distributions of the perceived harms of hate speech. Moreover, as shown in [Figure A2](#) in the online Appendix, varying levels of education and income do not significantly change the relationship between the perceived harms of hate speech and support for hate speech regulations. Of course, if other samples derived from other survey studies were used, we might obtain different results. However, at least the results derived from our sample data suggest that the sample-and-population divergence in terms of education and income does not seriously impair the validity of our core arguments.

Lastly, the ages of our respondents are limited to the range between 20 and 69. Unfortunately, we do not have existing studies and sample data that help us infer the bias derived from this age restriction. Instead of presenting our unfounded conjectures about this bias, we accept this limitation, which is inherent in almost all internet-based surveys, and leave the solution to this problem to future research.

7. Conclusions

The concept of dignity is critically important for providing justifications for hate speech regulations, as Jeremy Waldron advocated most famously among other political theorists and constitutional scholars. In this paper, we have presented a set of findings based on the original survey conducted in Japan, which shows unequivocally that ordinary citizens consider the perceived harm of ‘undermining dignity,’ rather than that of ‘causing offense,’ as more valid grounds for governmental regulations. To our knowledge, this sort of empirical evidence has never been presented. Given that Japan’s first anti-hate speech law had only recently enacted, it is fair to claim that our survey was conducted in the environment where the respondents were not as prejudiced or predisposed about the issue as those elsewhere. For this reason, we believe that our findings do speak to the generality, beyond Japan, of the dignitarian rationale and its importance in hate speech debate. Further research of course is warranted to confirm whether the public elsewhere also regards the undermining of dignity as reasonable and legitimate justification for governmental regulations.

More broadly, this paper has been an effort to substantiate one of the well-established normative arguments from a behavioral standpoint. While our focus in this paper has been on the citizens’ attitudes toward hate speech *per se*, we believe that similar studies can be conducted to explore public support for regulations and policies related to the problems of discrimination and redistribution more generally. The importance of bridging the normative and empirical subfields is increasingly recognized in the discipline of political science. However, there still remain skepticisms, and some critics may, for example, regard our endeavor guilty of David Hume’s ‘naturalistic fallacy,’ committing the deduction of an *ought*, a normative proposition, from an *is*, a descriptive statement about the state of the world. We believe this criticism does not apply. It is true that we are deducing the prescriptive, or even policy-related, proposition that the concept of dignity must be placed at the center of hate speech regulations. This proposition is derived, not from simply observing the state of the world, but from the judgments made by the respondents themselves, who represent ordinary citizens in an established democracy. The point of conducting our survey was not to verify a normative claim made by some famous theorist, but to probe whether the citizens themselves would make the evaluations parallel to that claim. We certainly do not maintain that the regulations of hate speech should reflect the status quo, or what we observe as the state of the world.

In this paper, we have sought not only to determine whether dignity matters, but also to clarify how it matters, by investigating the possible mediation mechanisms. Waldron, in his original formulation, did not fully address this issue, seemingly presupposing a kind of public-centered logic about people’s expectation regarding the societal consequences of hate speech. While not denying this causal path, our findings also reveal the importance of the citizens’ more individual-based moral judgments in shaping their regulatory attitudes. Understandably, for normative theorists, whether dignitarianism, or any argument that justifies speech regulation, is rendered as departure from the liberal orthodoxy may be a critical topic worthy of elaborate discussion. From our behavioral standpoint, we simply note that the concerns over societal good and the senses of justice/injustice do seem to go hand-in-hand in the minds of ordinary citizens as far as their attitudes over hate speech regulations are concerned.

Supplementary material

The supplementary material for this article can be found at <https://doi.org/10.1017/S146810992300004X>.

Acknowledgements. We would like to thank Kiichiro Arai, Erik Bleich, Kazunori Inamasu, Go Murakami, Miwa Nakajo, Yoshitaka Nishizawa, Yayo Okano, Gill Steel, Michael Strausz, and other panel participants at the 2019 Japanese Political Science Association Annual Meeting as well as those at the 2020 American Political Science Association Annual Meeting for their helpful comments. We also acknowledge that we benefitted greatly from critical comments of the three anonymous referees on our earlier draft. This study was financially supported by JSPS KAKEN no. 17KT0005.

Conflict of interest. The authors declare none.

References

- Baker CE** (1989) *Human Liberty and Freedom of Speech*. New York: Oxford University Press.
- Baker CE** (2012) Hate speech. In Hertz M and Molnar P (eds), *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. New York: Cambridge University Press, pp. 57–80.
- Barendt E** (2019) What is the harm of hate speech? *Ethical Theory and Moral Practice* **25**, 1–15.
- Boeckmann RJ and Liew J** (2002) Hate speech: Asian American students' justice judgements and psychological responses. *Journal of Social Issue* **58**, 363–381.
- Brown A** (2015) *Hate Speech Law: A Philosophical Examination*. New York: Routledge.
- Coliver S** (ed) (1992) *Striking a Balance: Hate Speech, Freedom of Expression and Non-Discrimination*. London: University of Essex Press.
- Cowan G and Hodge C** (1996) Judgements of hate speech: the effects of target group, publicness, and behavioral responses of the target. *Journal of Applied Social Psychology* **26**, 355–374.
- Cowan G and Khatchadourian D** (2003) Empathy, ways of knowing, and interdependence as mediators of gender differences in attitudes toward hate speech and freedom of speech. *Psychology of Women Quarterly* **27**, 300–308.
- Cowan G, Resendez M, Marshall E and Quist R** (2002) Hate speech and constitutional protection: priming values of equality and freedom. *Journal of Social Issue* **58**, 247–263.
- Delgado R** (1993) Words that wound: a tort action for racial insults, epithets, and name calling. In Matsuda MJ, Lawrence III CR, Delgado R and Crenshaw KW (eds), *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*. Boulder: Westview, pp. 89–110.
- Downs DM and Cowan G** (2012) Predicting the importance of freedom of speech and the perceived harm of hate speech. *Journal of Applied Social Psychology* **42**, 1353–1375.
- Dworkin R** (2009) Forward. In Hare I and Weinstein J (eds), *Extreme Speech and Democracy*. New York: Oxford University Press.
- Dworkin R** (2012) Reply to Jeremy Waldron. In Hertz M and Molnar P (eds), *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. New York: Cambridge University Press, pp. 341–344.
- Gelber K and McNamara I** (2016) Evidencing the harms of hate speech. *Social Identities* **22**, 324–341.
- Hare I and Weinstein J** (eds) (2009) *Extreme Speech and Democracy*. New York: Oxford University Press.
- Heinze E** (2013) Review essay: Hate speech and the normative foundations of regulation. *International Journal of Law in Context* **9**, 590–617.
- Herek GM, Cogan JC and Gillis JR** (2002) Victim experiences in hate crimes based on sexual orientation. *Journal of Social Issue* **58**, 319–339.
- Herz M and Molnar P** (eds) (2012) *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. New York: Cambridge University Press.
- Heyman SJ** (2009) Hate speech, public discourse, and the first amendment. In Hare I and Weinstein J (eds), *Extreme Speech and Democracy*. New York: Oxford University Press, pp. 158–181.
- Howard JW** (2019) Free speech and hate speech. *Annual Review of Political Science* **22**, 93–109.
- Jones P** (2011) Religious belief and freedom of expression: is offensiveness really the issue? *Res Publica* **17**, 75–90.
- Jones P** (2015) Dignity, hate and harm. *Political Theory* **43**, 678–686.
- Kotani J** (2018) Proceed with caution: hate speech regulation in Japan. *Hasting Constitutional Law Quarterly* **45**, 603–622.
- Krotoszynski Jr RJ** (2006) *The First Amendment in Cross-Cultural Perspective: A Comparative Legal Analysis of the Freedom of Speech*. New York: New York University Press.
- Leets L** (2002) Experiencing hate speech: perceptions and responses to anti-semitism and antigay speech. *Journal of Social Issue* **58**, 341–361.
- Leiter B** (2012) Waldron on the regulation of hate speech. *Chicago Public Law and Legal Theory Working Paper* 398.
- Lewis A** (2007) *Freedom for the Thought That We Hate: A Biography of the First Amendment*. New York: Basic Books.
- Matsuda MJ, Lawrence III CR, Delgado R and Crenshaw KW** (1993) *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*. Boulder: Westview.
- Matsui S** (2016) The challenge to multiculturalism: hate speech ban in Japan. *University of British Columbia Law Review* **49**, 427–484.
- McConnell MW** (2012) You can't say that. *New York Times (Sunday Book Review on 22 June)*.
- Post R** (1991) Racist speech, democracy, and the first amendment. *William and Mary Law Review* **32**, 267–327.
- Rosenfeld M** (2003) Hate speech in constitutional jurisprudence: a comparative analysis. *Cardozo Law Review* **24**, 1523–1567.
- Seglow J** (2016) Hate speech, dignity and self-respect. *Ethical Theory and Moral Practice* **19**, 1103–1116.
- Simpson RM** (2013) Dignity, harm, and hate speech. *Law and Philosophy* **32**, 701–728.
- Stone GR** (1987) Content-neutral restrictions. *University of Chicago Law Review* **54**, 46–118.
- Tsesis A** (2009) Dignity and speech: the regulation of hate speech in a democracy. *Wake Forest Law Review* **44**, 497–532.
- Waldron J** (2012a) *Dignity, Rank, and Rights*. New York: Oxford University Press.

Waldron J (2012b) *The Harm in Hate Speech*. Cambridge: Harvard University Press.

Weinstein J (1999) *Hate Speech, Pornography, and the Radical Attack on Free Speech Doctrine*. Boulder, CO: Westview Press.

Weinstein J (2017) Hate speech bans, democracy, and political legitimacy. *Constitutional Commentary* **32**, 527–583.

Zivi K (2014) Doing things with hate speech: a response to Jeremy Waldron's the harm in hate speech. *Contemporary Political Theory* **13**, 94–100.