Indignity or Offense?
A Survey-Experimental Inquiry into Behavioral Foundations of Hate Speech Regulations

Kentaro Hirose* Hae Kim† Masaru Kohno‡

Keywords:
Hate Speech Regulation; Dignity; Survey Experiment; Mediation Analysis

Last Update: August, 2020

ABSTRACT

What justifies regulating hate speech in democratic societies? In this paper, we conduct a survey experiment and highlight the concept of dignity as the cornerstone for such regulations. 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 Jeremy Waldron and other "dignitarians," especially in the course of debate with free speech advocates. We aim to consolidate this important claim in the normative literature on behavioral grounds. Specifically, based on our survey experiment 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 strongly and consistently across a variety of treatment conditions. Our experiment further clarifies the possible mediation mechanisms of the dignitarian rationale, revealing not only the people's public-centered expectation regarding the societal consequences of hate speech, which Waldron himself emphasized, but also the importance of more individual-based judgements regarding morality and justice, in shaping their regulatory attitudes.

*Department of Political Science and Economics, Waseda University. Email: khirose@aoni.waseda.jp.
†Department of Education, Chiba University. Email: kimhae@chiba-u.jp
‡Department of Political Science and Economics, Waseda University. Email: kohno@waseda.jp

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 United Kingdom, 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 United States, there is no equivalent law, as free speech is most ardently protected by the First Amendment to the U.S. Constitution with courts only allowing for so-called "content-neutral" restrictions (Stone 1987); this regulatory absence is often described as "exceptionalism" (Krotoszynski Jr. 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, *p*.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 2012*b*). At the core of his thesis lies the distinction between the two kinds of harm

[1] Hate 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.

[2] Public Order Act in the United Kingdom (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, maliciously maligning, or defaming segments of the population, shall be punished with imprisonment from three months to five years."

[3] Staunch First Amendment defenders reject these characterizations. See, for example, Baker (2012, *pp*.58-60).

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

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 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 consolidate on behavioral grounds 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 experiment we conducted among ordinary Japanese citizens in the year 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 United States where in their absence the opinions for and against governmental regulations are so deeply entrenched in the respective ideological camps.

Our experiment yields a set of novel findings. First and foremost, the experiment reveals solid behavioral foundations for dignity-based regulations of hate speech. That is, our findings confirm that, across a variety of treatment conditions, citizens' 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 or not. We further show,

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

[6]See also Waldron (2012*a*). 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 assertation that offense should not be recognized as grounds for justifying hate speech regulations.

through a standard cross-validation technique (e.g., Hastie, Tibshirani and Friedman 2009), that these main results are robust across numerously repeated sub-sample simulations, estimating that our measure of harming dignity improves out-of-sample predictive performance at a rate as high as the liberal-conservative ideology, the most basic predictor of people's attitude toward hate speech regulations. Second, despite this overwhelming tendency, we also observe such cases where the citizens' regulatory attitudes are not affected by the perceived harm against the victims' dignity. These exceptions occur when citizens are primed that the targeted victims are "those with disabilities," as opposed to the other three treatment conditions where the victims are identified as Koreans residing in Japan ("*Zainichi* Koreans") or "minorities" in general, or remain unidentified. This anomaly, while open to different interpretations, at least points to the possibility that the dignitarian rationale may not uniformly justify regulating hate speech against all types of potential victims. Third, beyond simply verifying the importance of the perceived harm to dignity, our survey experiment further reveals two causal pathways through which such perceptions affect the citizens' regulatory attitudes, namely one through their concerns over the society's degrading and the other through their individual senses of justice and morality. While Waldron's original conception was limited to the former, we suggest that the two mechanisms are not mutually exclusive and that they both shape the citizens' regulatory attitudes. 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 the next section, we highlight the main characteristics of the dignitarian rationale by reviewing Waldron's original discussion with some critical annotations. In the third section, we introduce our approach and describe our survey experiment. In the fourth section, we present our findings and elaborate on their significance. In the fifth section, we provide robustness checks. The final section concludes by drawing broad implications and pointing to directions of further 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 intrinsically 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 2012*b*, *p*.147, emphasis original). But this position, of course, begs the question: what harm?

Here draws Waldron 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 2012*b*, *p*.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 2012*b*, *p*.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 restriction of free speech would weaken democratic legitimacy

[7]For a different appraisal, see, for example, Jones (2015, esp. *p*.682).

(Dworkin 2012; Weinstein 2017; cf. Dworkin 2009), 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, *pp.*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*, *p.*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*, *p.*115). This retreat is problematic because the dignitarian rationale is not simply a philosophical proposition but is offered also as legislative justifications. 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.

Here we submit that it would be wrong to treat the loss of dignity as just one of many facets of the psychological process that takes place on the part of any targeted victim. Such a treatment would belittle the idea most innovative about dignitarianism, namely the contrast between the dignity's "objective or social aspects" and the offense's (feeling's) "subjective aspects." Rather we argue that the validity and legislative usefulness of the concept of dignity must be evaluated in accordance with the opinions of ordinary citizens at large, not what goes on in the mind of any particular victim. Neither analytical philosophy nor

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

diagnostic psychology can be relied upon for such determination. For this task, we turn to a survey experiment that investigates how ordinary citizens regard the perceived harms of hate speech and its consequences.

3 Survey Experiment

In this section, we introduce how our approach differs from previous studies, describe the regulatory background in Japan where our survey experiment was conducted, and explain our sample and basic design.

3.1 Approach

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, Cogan and Gillis 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 large-scale sample of respondents better representing national population.

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 as causal variables.

Finally, the most original feature of our approach lies in the time and location, or the real-world context, in which the survey experiment was conducted, namely Japan in 2018. Japan then, as detailed below, had only recently (in 2016) adopted the nation's first anti-hate-speech law, and ordinary Japanese citizens were in the midst of developing new norms and interpretations on the subject. Thus, it can be assumed that our sample of respondents had less prejudices and predispositions about hate speech regulations than would those respondents assembled elsewhere.

3.2 Background

Our survey experiment was conduced through internet with adult residents in Japan in March 2018, less than two 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

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

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, disabled persons, 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.

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

3.3 Sample

To administer our survey experiment, we contracted with Nikkei Research, one of the major online survey companies in Japan. Based on a stratified random sampling procedure, our

[10] See, 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 (07 Jul 2014 - 25 Jul 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).

respondents were sampled from the pool of monitors registered at this firm so that the appropriate number was assigned to each category of gender as well as geographical regions of their residencies in proportion to the actual demographic data reported in Japan's latest edition of *Jūminkihondaichō* (Basic Residence Register). 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 generally skew slightly higher income and more educated. Given these features of the sample, we intend to conduct thorough robustness checks, as provided below.

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 5,068 respondents for our analysis.

3.4 Experimental Design

Our survey experiment is a two-factor, between-respondents design. Each factor has four different conditions, and respondents were thus randomly assigned to one of the sixteen conditions.

The first of the two factors concerns with the content of the hate speech, varying in terms of the level of implied violence and whether the term "hate speech" itself appears in the description of the relevant act. We included these variations, considering that ordinary Japanese citizens were unlikely to have uniform understandings about what constitutes hate speech. Thus, specifically, the four content conditions include: "insults" (Content 1), "insults and incitements to violence" (Content 2), "hate speech, that is, insults" (Content 3), and "hate speech, that is, insults and incitements to violence" (Content 4).

The second factor concerns with the types of groups against whom hate speech is targeted. We divided the entire sample into four: "*Zainichi* Koreans" (Target 1), "those with

disabilities" (Target 2), "minorities" (Target 3), and no specification of targeted group (Target 4). To ensure a consistent definition of what each of these categories stands for, a brief sentence was added in the vignette to explain, for example, who "*Zainichi* Koreans" are.[11] Although in the Japanese context, as noted earlier, the most widely reported victims are *Zainichi* Koreans, we thought it was necessary to include these other types, again because of the likely absence in Japan of common understanding regarding what constitutes hate speech.

Our key explanatory variables are the perceived levels of offense and indignity that may influence the respondents' attitudes toward regulating hate speech. To measure them, the respondents assigned to Content 4 and Target 1, for example, were presented with a short leading passage and questions, as follows:

> 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 offense, for example, 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. The order of these two questions was randomized across respondents.

[11] The 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)." The other definitional sentences read, respectively: "minorities mean those who belong to some minority group that can be distinguished from others in terms of religion, political beliefs, physical and mental disabilities, sexual orientations, etc"; and "those with disabilities mean persons with some physical, intellectual, or mental disabilities."

Lastly, we asked each respondent the dependent variable question regarding whether he/she supports governmental regulations on hate speech. In this question, the content and target-type specifications were explicitly repeated. Thus, for the respondents assigned to Content 4 - Target 1 condition, the question was asked as follows:

> 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 A1 in Online Appendix shows the distribution of the answers to this question under Content 4 - Target 1. In that condition, 57% of the respondents supported hate speech regulations partially or thoroughly, while only 17% of them opposed to the idea, and 25% of them had a neutral stance on this issue.

4 Main Findings: Harm to Dignity as Basis for Regulation

In this and the following sections, we present a set of findings from our experiment. Let us begin by focusing on the subset of respondents who were assigned and asked about their attitudes toward "hate speech, that is insults and incitements to violence" against "*Zainichi* Koreans." Of all the sixteen conditions in our experimental design, we regard this Content 4 - Target 1 combination to be the proto-type of hate speech situations in Japan. The explicit inclusion of "hate speech" and "violence" in the wording makes it more likely for the respondents' expectations to converge into believing that some harmful consequences are involved. Further, in Japan, as noted earlier, *Zainichi* Koreans are undoubtedly the most frequent target of hate speech in actuality, and it is this group to whom Japan's anti-hate speech law is intended to provide protection. As we realize this subset represents only a small part of our sample, however, thorough robustness checks will follow.

4.1 Regression Analysis

To estimate the effect of the perceived harm to the victims' feelings and the perceived harm to their dignity, respectively, on the respondents' attitude regarding hate speech regulations, we use the following linear regression model:

$$\begin{aligned} &\text{Hate-Speech Regulation} \\ &= \lambda_0 + \lambda_1(\text{Harm to Feelings}) + \lambda_2(\text{Harm to Dignity}) + \sum_k \lambda_k Z_k + v \end{aligned} \tag{1}$$

where Z represents a set of control variables including: (1) the respondent's gender; (2) age; (3) 11-point scale liberal-conservative ideology measure with greater values indicating conservative political positions; and (4) subjective value of free speech with greater values indicating that he/she respects the freedom of speech more strongly.

Figure 1 demonstrates that a strong positive association exists between the respondents' perceptions of the harm inflicted against victims' dignity and their support for hate speech regulations, even after controlling for the perceived harm to their feelings as well as other covariates. For example, a respondent with the view that hate speech against *Zainichi* Koreans will least likely harm their dignity is predicted to think that the government should not impose any restrictions on hate speech against *Zainichi* Koreans. In contrast, a respondent with the view that hate speech will most likely harm their dignity is predicted to think that the government should impose some restrictions.

[Figure 1 about here]

With regard to the respondents' perceptions about the harm inflicted against victims' feelings, on the other hand, the estimated results are not statistically significant. There is almost no difference in the level of support for hate speech regulations between a respondent who thinks that hate speech against *Zainichi* Koreans will absolutely make them feel offended and a respondent who thinks the completely opposite. These results corroborate the argument that hate speech regulations should be justified on the basis of the harm to dignity, but not the harm to any feelings.

4.2 Out-of-Sample Prediction

The association reported above between the harm to dignity and the attitude toward hate speech regulations remains inconclusive in light of the possibility that the results of our regression analysis may be due to overfitting to a particular sample of respondents. To assess the substantive importance of the variable of our interest and also to ascertain the generalizability of the results to other un-surveyed citizens, we have conducted out-of-sample prediction, or so-called cross-validation (see, for example, Hastie, Tibshirani and Friedman (2009)). A detailed description of this procedure is included in Online Appendix.

Figure 2 reports the out-of-sample prediction improvement rates of the harm to dignity. The improvements rates of other predictors are also reported for comparison. Adding into our regression model the perceptions about the harm against victims' feelings does not improve our ability to forecast the respondents' regulatory attitudes; indeed, doing so rather worsens it, implying that the concern for whether the victim is offended is simply irrelevant for predicting the citizens' attitudes toward hate speech regulations. In contrast, our forecasting ability increases by 2-4% once we take into account the respondents' perceptions about the harm to victims' dignity. This improvement rate is equivalent to, or even slightly higher than, that achieved by adding the liberal-conservative ideology, presumably the most basic predictor for the citizens' regulatory attitudes.

[Figure 2 about here]

Hence, the concern for the harm to the victims' dignity is not only associated with the attitude toward hate speech regulations; it also serves as an important determinant of such attitude. We realize that the two independent variables, the perceptions about victims being offended and the perceptions about their dignity being harmed, are strongly correlated with each other (see Figure A2 in Online Appendix). However, given the remarkable consistency with which we have found the positive effects of indignity on the dependent variable, our finding of a non-association between the perceived offense and the attitude toward hate speech regulations is not likely to be a statistical artifact generated by such multicollinearity.

4.3 Testing the Mechanism

Having established a strong positive association between the harm to dignity and the attitude toward hate speech regulations, we now proceed to the next question: why? While there may be numerous causal pathways through which the perceived indignity influences the public's regulatory attitude, we contrast two mechanisms in particular. The first mechanism is something originally offered by Waldron, that is, the idea of "public good": the harm inflicted against victims' dignity induces ordinary citizens to support hate speech regulations because they expect that it will move their society in a bad direction. The second mechanism we consider is more individual based: ordinary citizens oppose hate speech detrimental to the victims' dignity, because they think that harming someone's dignity is morally wrong or not consistent with their sense of justice. We believe that these two causal possibilities are not mutually exclusive, although Waldron does not fully address the second pathway in his formulation of dignitarian rationale.

In order to probe the saliency of the public good mechanism and the individual justice mechanism respectively, we use the following 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. As shown in Figure A3 in Online Appendix, the two concepts, societal concern and sense of injustice, are positively correlated with each other, although the degree of the correlation is not as strong as the one found between offense and indignity.

In order to estimate the two mediating effects, one through the public good mechanism and the other through the individual justice mechanism, we followed the conventional pro-

cedure of mediation analysis, by fitting the three linear regression models:

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

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

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

where Z represents a set of control variables such as the respondent's gender, age, ideology, and subjective value of free speech. Figure 3 schematizes the paths for illustration.

[Figure 3 about here]

The first model estimates the marginal effect, on the respondents' societal concern, of the perceived harm to feelings (α_1) and of the perceived harm to dignity (α_2), respectively. Similarly, the second model estimates the marginal effect, on their sense of injustice, of the perceived harm to feelings (γ_1) and of the perceived harm to dignity (γ_2), respectively.

As shown in Figure A4 in Online Appendix, the respondents' perceptions about the harm against victims' dignity is positively associated with their concern over which direction the society is heading. For example, a respondent who thinks that hate speech against *Zainichi* Koreans will least likely harm their dignity is predicted to think that such hate speech will make the society better, whereas a respondent who thinks that hate speech against *Zainichi* Koreans will most likely harm their dignity is predicted to think that such hate speech will make the society worse. By contrast, the perceived harm to victims' feelings does not have any impact on societal concern.

Similarly, as shown in Figure A5 in Online Appendix, the perceived harm to victims' dignity is positively associated with the sense of injustice. For example, a respondent who thinks that hate speech against *Zainichi* Koreans will least likely harm their dignity is predicted to think that such hate speech will be just, whereas a respondent who thinks that

hate speech against *Zainichi* Koreans will most likely harm their dignity is predicted to think that such hate speech will be unjust. Meanwhile, the respondents' perceptions about the harm against victims' feelings does not yield any significant change in their sense of justice/injustice.

The third model estimates the marginal effect, on the regulatory attitude, of societal concern (β_3) and the sense of injustice (β_4), while also estimating the "residual" effect of the harm to feelings (β_1) and the harm to dignity (β_2) on the attitude mediated through channels other than the above two causal paths. As confirmed in Figure A6 in Online Appendix, respondents' support for hate speech regulations is positively associated with both their societal concern and their sense of justice. In contrast, after accounting for these two mediating channels, the attitude toward hate speech regulations is no longer associated with either the perceived indignity or the perceived offense, suggesting that there is no "residual" effect mediated through channels other than the public good mechanism and the justice mechanism.

Since we have used linear regression models, the effect of the perception of indignity on the attitude toward hate speech regulations mediated by the public good mechanism can be computed as a simple product of two coefficients α_2 and β_3. Similarly, the effect mediated by the individual justice mechanism is simply a product of γ_2 and β_4. Figure 4 reports these mediation effects of indignity and offense. All confidence intervals of the mediation effects of the perceived offense include zero, implying that it does not influence the regulatory attitude through either mediating channel. In contrast, the positive effect of the indignity mediated through the public good mechanism indicates that the perceived harm to victims' dignity increases the respondents' concern about negative consequences on the society, which in turn makes the citizen more likely to support hate speech regulations.

Yet Figure 4 also shows that the concern for negative consequences on the society is not the only channel through which the harm to dignity affects the attitude toward hate speech regulations. The positive effect of indignity mediated through the justice mechanism suggests that the harm to dignity also leads to individual-based moral judgement that hate speech is unjust, which in turn influences the attitude toward hate speech regulations. There is almost no difference in the magnitude of each mediation effect, implying that both mechanisms are

equally important when explaining why the perceived harm to dignity affects the attitude toward hate speech regulations.

[Figure 4 about here]

5 Robustness Checks: How Universal Is Dignitarian Rationale?

The analysis and findings presented in the previous section surely lends support for the dignitarian logic of hate speech regulations, but they were based on a specific subset of the respondents primed with both a certain content and a certain target of hate speech. In attempt to uphold the logic's generalizability, we now investigate whether the above results robustly hold when we analyze other sets of respondents assigned to different experimental conditions.

5.1 Varying Content of Hate Speech

We begin by analyzing the effect of each covariate on the attitude toward hate speech regulations by varying the content of hate speech (Content 1 to 4), while holding constant the targeted victim as *Zainichi* Koreans. Figure 5 shows the t-values of the covariates derived from the analysis based on these different sub-samples of respondents. The perceived harm against victims' feelings has no statistically significant association with the regulatory attitude regardless of types of hate speech content, suggesting a robust non-association between the offense and hate speech regulations. In contrast, the perceived harm inflicted against victims' dignity is positively associated across all types of hate speech contents in a statistically significant manner, indicating a robust association between indignity and hate speech regulations.

[Figure 5 about here]

Figure 6 reports the results of out-of-sample prediction for these alternate sub-sample analyses. Adding into our regression model the respondents' perceptions about the harm

against victims' feelings does not improve its ability to predict their regulatory attitude under any type of hate speech content. On the other hand, the predictive performance of our model always improves whenever we take into account the respondents' perceptions about harm to victims' dignity. In particular, as for the condition for "hate speech, that is, insults," the rate of prediction improvement goes up to nearly 7% on average. Hence, overall, the cross-validation lends further support for the dignitarian argument, confirming the positive association in the case of the perceived indignity and the absence of such association in the case of the perceived offence.

[Figure 6 about here]

Finally, Figure 7 summarizes the results of mediation analysis which we replicated for each of these sub-samples. The respondents' perceptions about the harm against victims' feelings does not influence the regulatory attitude toward any type of hate speech content through either of the hypothesized mediating channels. On the other hand, the mediation effects of the harm to dignity through the public good mechanism are statistically significant for all subsamples, except the respondents who were asked about "hate speech, that is, insults" (but, even in this specific subsample, too, the t-value is only slightly below the conventional significance level). The mediation effects of indignity through the justice mechanism are statistically significant only in two subsamples, suggesting that the justice mechanism is not as robust as the public good mechanism as causal pathway. Furthermore, in two subsamples, we detect statistically significant "residual" effects of the perceived harm to dignity, implying that there might exist systematic causal pathways other than the public good mechanism and the justice mechanism through which the harm to dignity influences the attitude toward hate speech regulations.

[Figure 7 about here]

In sum, the dignitarian rationale holds up in most cases investigated here. Our respondents, regardless of various contents of hate speech, do distinguish the two kinds of harm, offense and indignity, in a clear and consistent manner, only attributing the latter as justification for stricter governmental regulations.

5.2 Varying Target of Hate Speech

We have further performed the robustness checks by replicating the analysis for all experimental conditions, varying not only the content but also the target of hate speech. The left panel of Figure 8 shows the t-values of the marginal effects of the perceived harm to the victims' feelings on the regulatory attitude. Out of the sixteen experimental conditions, only three sub-samples yielded statistically significant results. In contrast, the perceived harm to the victims' dignity is positively associated with the attitude in all types of contents in cases where the targeted victims were primed as *Zainichi* Koreans (Target 1) and where they remained unspecified (Target 4). In the case where the victims were identified simply as "minorities" (Target 3), the perceived harm to dignity is positively associated for three of the four hate speech contents (Content 1, Content 3, and Content 4) with only Content 2 being the exception (namely, when the respondents were primed with the wording "insults and incitements to violence").

[Figure 8 about here]

While these results lend support for the generalizability of the digintarian argument, they also point to a notable exception. As shown in Figure 8, the robust association between the harm to dignity and the attitude toward hate speech regulations does not hold up in the subsample of respondents primed with "those with disabilities" as the target (Target 2): out of the four treatment conditions with varying content of hate speech within this sub-sample, a statistically significant association shows up in only one (Content 1). As discussed below in our concluding section, we believe that this anomaly is open to contrasting interpretations. From a strictly empirical standpoint, however, this finding does provide an important reservation, pointing to the possibility that the dignitarian justification for hate speech regulations may or should not apply universally to all applicable targeted groups.

6 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 experiment 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 experiment 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 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. 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 judgements made by the respondents themselves, who represent ordinary citizens in an established democracy. The point of conducting our survey experiment 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 judgements 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.

In closing our paper, we must discuss the implications of the important anomaly, namely the case of those disabled. As noted above, when the persons with disablity were explicitly referred to as the targeted victims, the perceptions of the harm against their dignity do not influence the citizens' attitudes toward hate speech regulation. This finding is open to two contrasting interpretations. On the one hand, the disappearance of the effect of the perceived indignity may be taken as suggesting that our search must continue until we find an alternative justifiable cause for regulating hate speech against this particular minority group. This interpretation implies that the normative situations for those disabled are worse than others, since even the perceived indignity does not viably serve as a basis around which public expectation can converge into endorsing governmental regulation. On the other hand, it is possible to interpret that the absence of the dignity effect rather indicates that the normative situations surrounding those disabled are as not as bad as, or perhaps even better than, those surrounding other minority groups. According to this interpretation, the dignity effect is absent precisely because a kind of taken-for-granted norm already exists around which public expectations can converge. The difference between these two interpretations raises a difficult question as to how we can measure whether such a pre-existing norm is consequential in shaping human behavior, a question beyond the scope of the present paper.

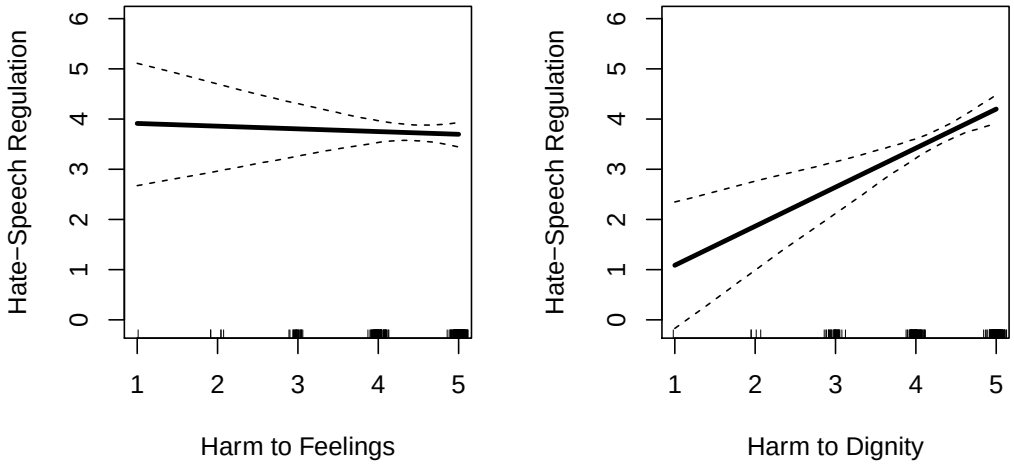

Figure 1: Associations between the harms of hate speech and the attitude toward hate speech regulations. The results are based on the linear regression model that includes as covariates the harm to feelings, the harm to dignity, the respondent's gender, age, ideology, and subjective value of free speech. The dashed lines represent 95% confidence intervals based on robust standard errors. The sample of the respondents assigned to Content 4 ("hate speech, that is, insults and incitements to violence") and Target 1 (*Zainichi* Koreans) is used.

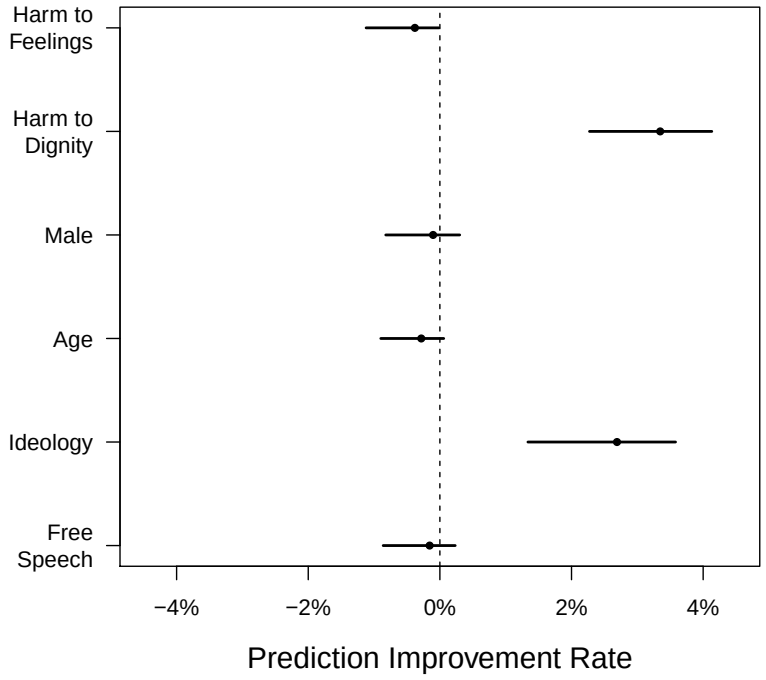

Figure 2: Out-of-sample prediction performance. Each horizontal line represents the 95% simulated distribution of the prediction improvement rate, measured by the root mean squared error, from adding a certain predictor to the linear regression model that excludes that predictor. The sample of the respondents assigned to Content 4 ("hate speech, that is, insults and incitements to violence") and Target 1 (*Zainichi* Koreans) is used.

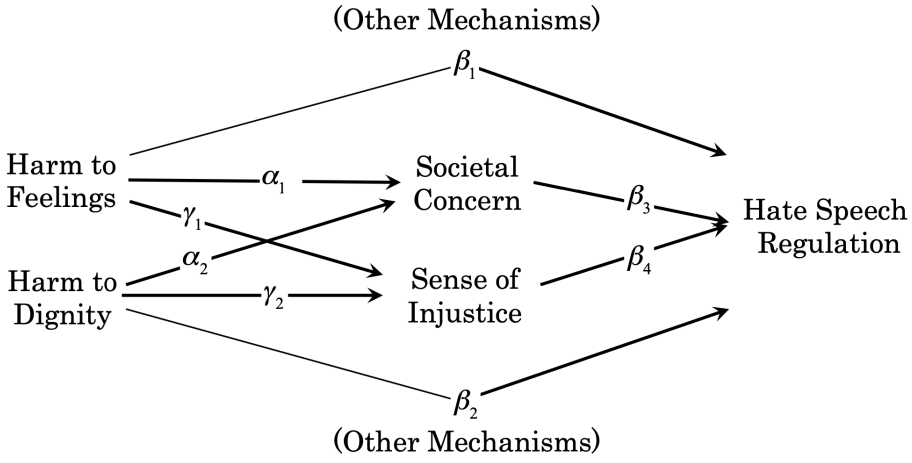

Figure 3: Mediating channels through which the harm to dignity as well as the harm to feelings may influence the attitude toward hate speech regulations.

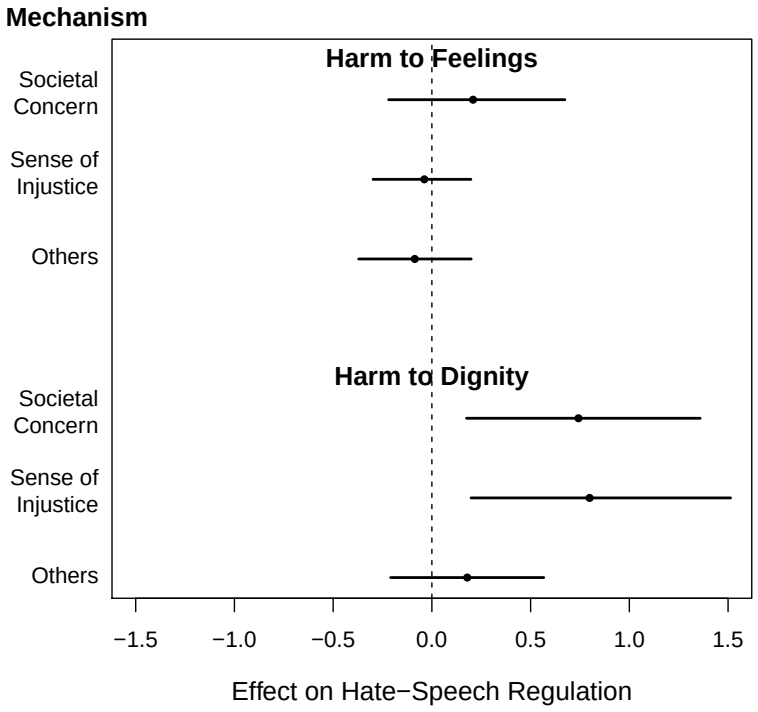

Figure 4: Effects of the harm to dignity as well as the harm to feelings on the attitude toward hate speech regulations mediated through various mechanisms. Each horizontal line represents the 95% confidence interval of a mediation effect based on robust standard errors. The sample of the respondents assigned to Content 4 ("hate speech, that is, insults and incitements to violence") and Target 1 (*Zainichi* Koreans) is used.

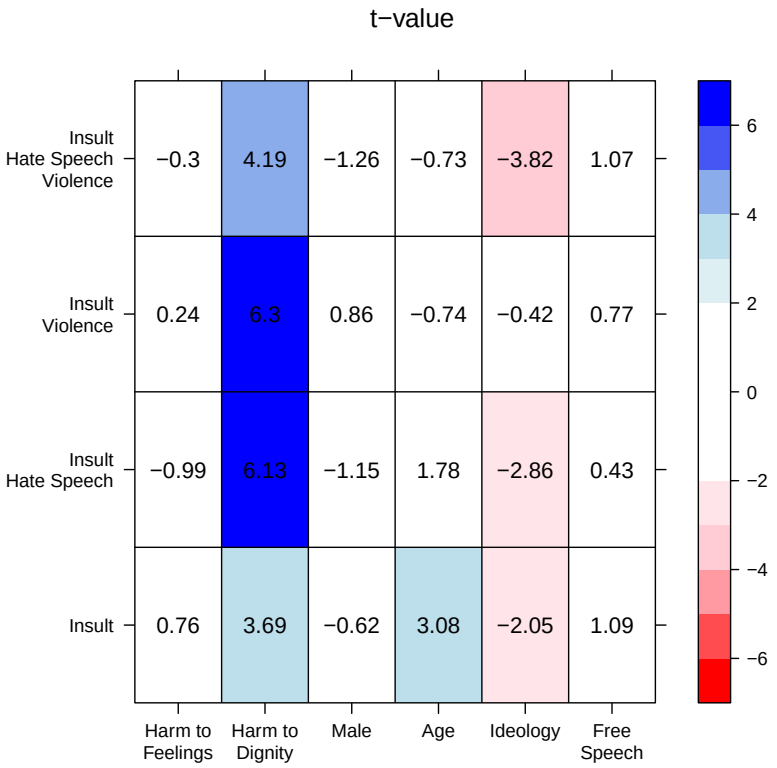

Figure 5: t-values of the estimated coefficients for the predictors of the attitude toward governmental regulations across four different types of hate speech contents. The results are based on the linear regression model that includes as covariates the harm to feelings, the harm to dignity, the respondent's gender, age, ideology, and subjective value of free speech. Robust standard errors are used to compute the t-values. The sample of the respondents assigned to Target 1 (*Zainiichi* Koreans) is used.

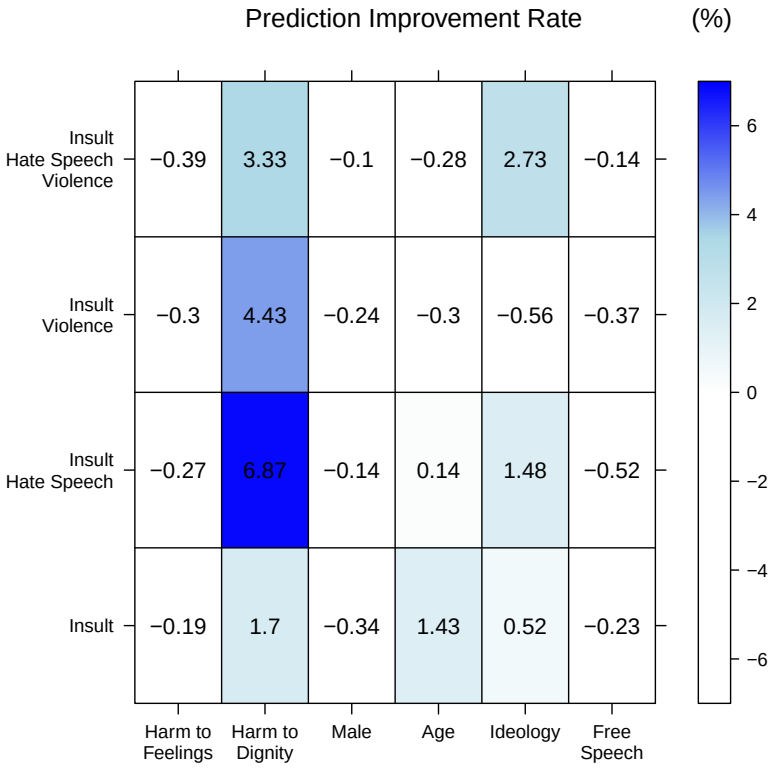

Figure 6: Out-of-sample prediction performance across four different types of hate speech contents. The prediction improvement rate of a certain predictor is measured by the root mean squared error derived from the model with and without that predictor. The sample of the respondents assigned to Target 1 (*Zainiichi* Koreans) is used.

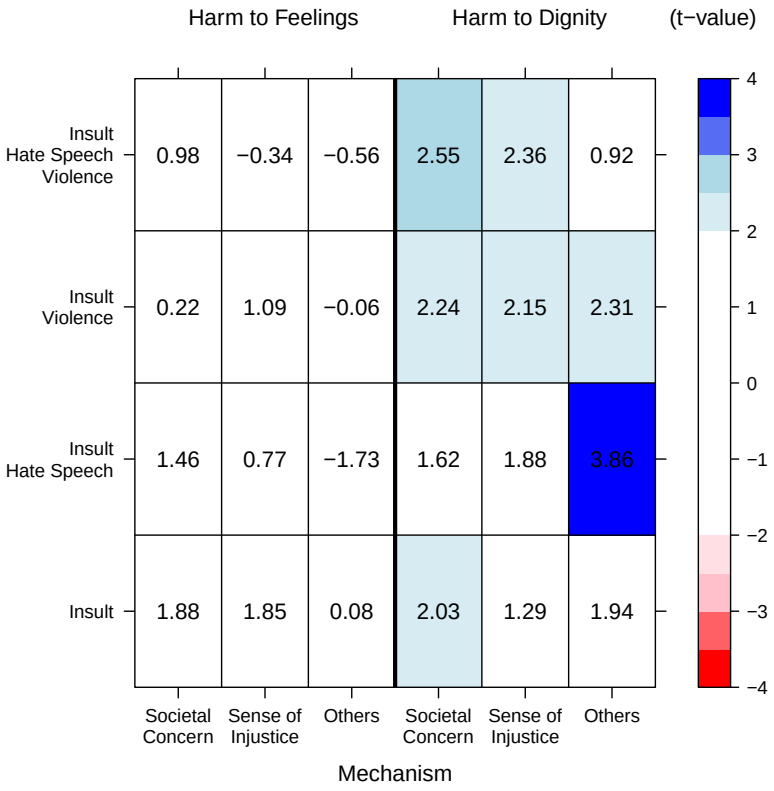

Figure 7: t-values of the estimated effects of the harm to dignity as well as the harm to feelings on the attitude toward hate speech regulations mediated through various mechanisms across four different types of hate speech contents. The t-values are computed from robust standard errors. The sample of the respondents assigned to Target 1 (*Zainiichi* Koreans) is used.

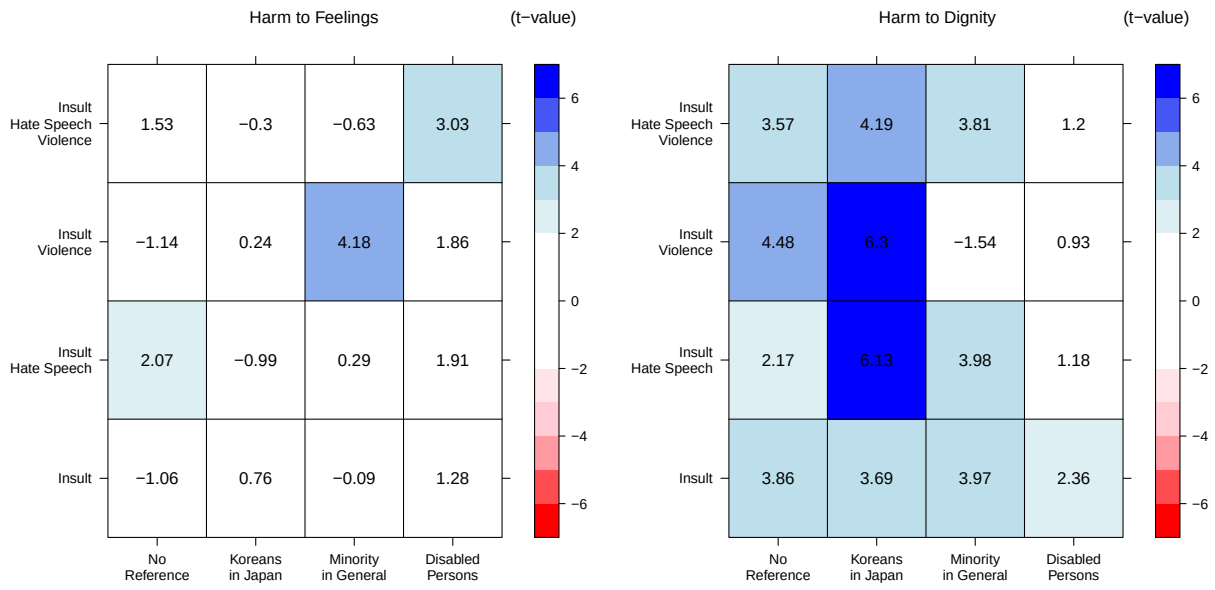

Figure 8: t-values of the estimated coefficients for the predictors of the attitude toward governmental regulations across different types and targets. The results are based on the linear regression model that includes as covariates the harm to feelings, the harm to dignity, the respondent's gender, age, ideology, and subjective value of free speech. Robust standard errors are used to compute the t-values.

References

Baker, C. Edwin. 1989. *Human Liberty and Freedom of Speech*. New York: Oxford University Press.

Baker, C. Edwin. 2012. Hate Speech. In *The Content and Context of Hate Speech: Rethinking Regulation and Responses*, ed. Michael Hertz and Peter Molnar. New York: Cambridge University Press.

Barendt, Eric. 2019. "What is the Harm of Hate Speech?" *Ethical Theory and Moral Practice* 25:1–15.

Boeckmann, Robert J. and Jeffrey Liew. 2002. "Hate Speech: Asian American Students' Justice Judgements and Psychological Responses." *Journal of Social Issue* 58(2):363–381.

Brown, Alexander. 2015. *Hate Speech Law: A Philosophical Examination*. New York: Routledge.

Coliver, Sandra, ed. 1992. *Striking a Balance: Hate Speech, Freedom of Expression and Non-Discrimination*. London: University of Essex Press.

Cowan, Gloria and Cyndi Hodge. 1996. "Judgements of Hate Speech: The Effects of Target Group, Publicness, and Behavioral Responses of the Target." *Journal of Applied Social Psychology* 26(4):355–374.

Cowan, Gloria and Desiree Khatchadourian. 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(4):300–308.

Cowan, Gloria, Miriam Resendez, Elizabeth Marshall and Ryan Quist. 2002. "Hate Speech and Constitutional Protection: Priming Values of Equality and Freedom." *Journal of Social Issue* 58(2):247–263.

Downs, Daniel M. and Gloria Cowan. 2012. "Predicting the Importance of Freedom of Speech and the Perceived Harm of Hate Speech." *Journal of Applied Social Psychology* 42(6):1353–1375.

Dworkin, Ronald. 2009. Forward. In *Extreme Speech and Democracy*, ed. Ivan Hare and James Weinstein. New York: Oxford University Press.

Dworkin, Ronald. 2012. Reply to Jeremy Waldron. In *The Content and Context of Hate Speech: Rethinking Regulation and Responses*, ed. Michael Hertz and Peter Molnar. New York: Cambridge University Press.

Gelber, Katharine and Luke McNamara. 2016. "Evidencing the Harms of Hate Speech." *Social Identities* 22(3):324–341.

Hare, Ivan and James Weinstein, eds. 2009. *Extreme Speech and Democracy*. New York: Oxford University Press.

Hastie, Trevor, Robert Tibshirani and Jarome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Heinze, Eric. 2013. "Review Essay: Hate Speech and the Normative Foundations of Regulation." *International Journal of Law in Context* 9(4):590–617.

Herek, Gregory M., Jeanine C. Cogan and J. Roy Gillis. 2002. "Victim Experiences in Hate Crimes Based on Sexual Orientation." *Journal of Social Issue* 58(2):319–339.

Herz, Michael and Peter Molnar, eds. 2012. *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. New York: Cambridge University Press.

Heyman, Steven J. 2009. Hate Speech, Public Discourse, and the First Amendment. In *Extreme Speech and Democracy*, ed. Ivan Hare and James Weinstein. New York: Oxford University Press.

Howard, Jeffrey W. 2019. "Free Speech and Hate Speech." *Annual Review of Political Science* 22(93-109).

Jones, Peter. 2011. "Religious Belief and Freedom of Expression: Is Offensiveness Really the Issue?" *Res Publica* 17(1):75–90.

Jones, Peter. 2015. "Dignity, Hate and Harm." *Political Theory* 43(5):678–686.

Kotani, Junko. 2018. "Proceed with Caution: Hate Speech Regulation in Japan." *Hasting Constitutional Law Quarterly* 45(3):603–622.

Krotoszynski Jr., Ronald J. 2006. *The First Amendment in Cross-Cultural Perspective: A Comparative Legal Analysis of the Freedom of Speech*. New York: New York University Press.

Leets, Laura. 2002. "Experiencing Hate Speech: Perceptions and Responses to Anti-Semitism and Antigacy Speech." *Journal of Social Issue* 58(2):341–361.

Leiter, Brian. 2012. "Waldron on the Regulation of Hate Speech." *Chicago Public Law and Legal Theory Working Paper* 398.

Lewis, Anthony. 2007. *Freedom for the Thought That We Hate: A Biography of the First Amendment*. New York: Basic Books.

Matsuda, Mari J., Charles R. Lawrence III, Richard Delgado and Kimberle Williams Crenshaw. 1993. *Words That Would: Critical Race Theory, Assualtive Speech, and the First Amendment*. Boulder: Westview.

Matsui, Shigenori. 2016. "The Challenge to Multiculturalism: Hate Speech Ban in Japan." *University of British Columbia Law Review* 49:427–484.

McConnell, Michael W. 2012. "You Can't Say That." *New York Times (Sunday Book Review on June 22)* .

Post, Robert. 1991. "Racist Speech, Democracy, and the First Amendment." *William and Mary Law Review* 32:267–327.

Rosenfeld, Michael. 2003. "Hate Speech in Constitutional Jurisprudence: A Comparative Analysis." *Cardozo Law Review* 24:1523–1567.

Seglow, Jonathan. 2016. "Hate Speech, Dignity and Self-Respect." *Ethical Theory and Moral Practice* 19(5):1103–1116.

Simpson, Robert Mark. 2013. "Dignity, Harm, and Hate Speech." *Law and Philosophy* 32(6):701–728.

Stone, Geoffrey R. 1987. "Content-Neutral Restrictions." *University of Chicago Law Review* 54:46–118.

Tsesis, Alexander. 2009. "Dignity and Speech: The Regulation of Hate Speech in a Democracy." *Wake Forest Law Review* 44:497–532.

Waldron, Jeremy. 2012*a*. *Dignity, Rank, and Rights*. New York: Oxford University Press.

Waldron, Jeremy. 2012*b*. *The Harm in Hate Speech*. Cambridge: Harvard University Press.

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

Weinstein, James. 2017. "Hate Speech Bans, Democracy, and Political Legitimacy." *Constitutional Commentary* 32:527–583.

Zivi, Karen. 2014. "Doing Things with Hate Speech: A Response to Jeremy Waldron's The Harm in Hate Speech." *Contemporary Political Theory* 13(1):94–100.

Online Appendix

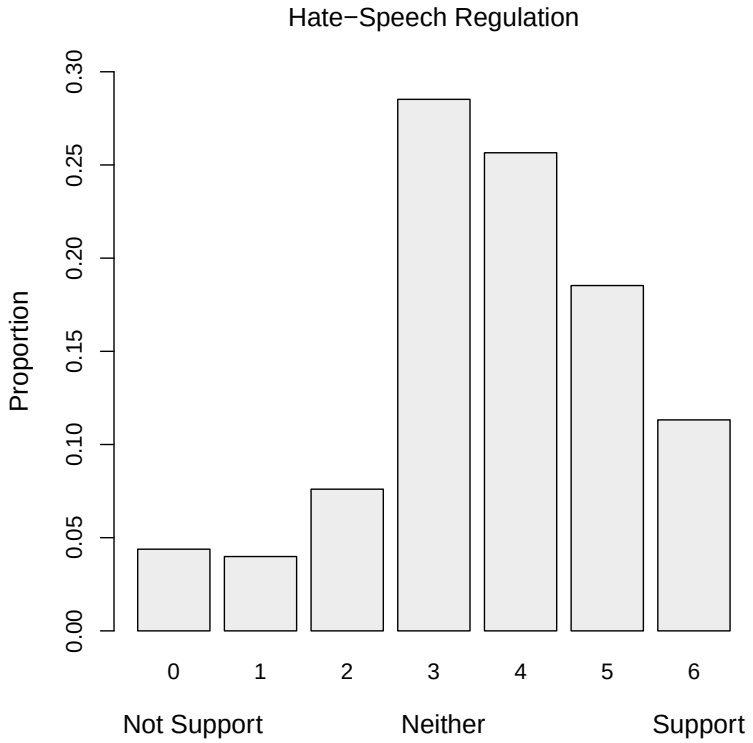

Figure A1: Distribution of the respondents' attitudes toward hate speech regulations. The sample of the respondents assigned to Content 4 ("hate speech, that is, insults and incitements to violence") and Target 1 (*Zainichi* Koreans) is used.

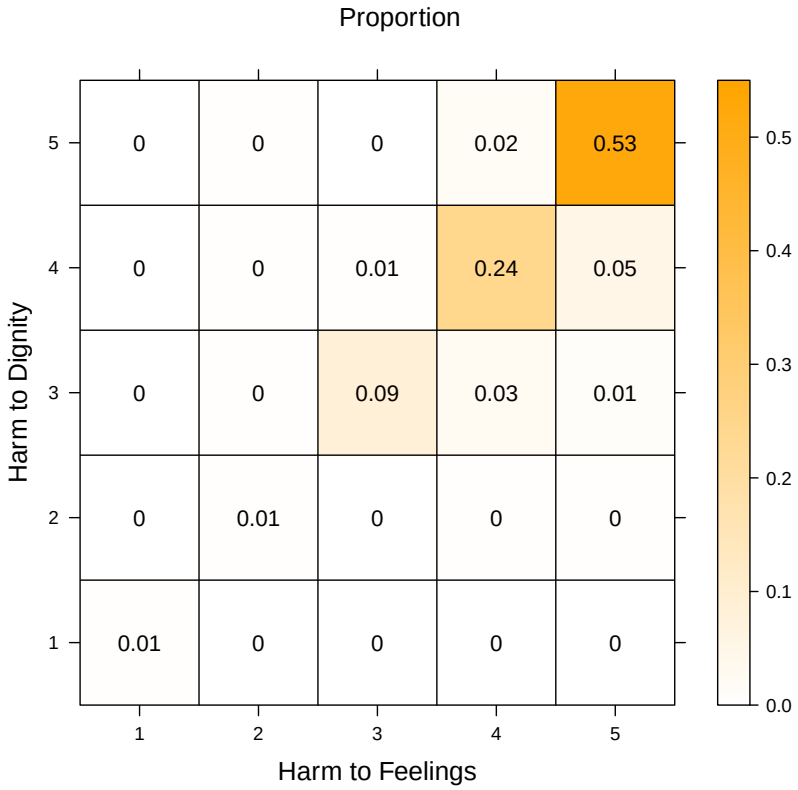

Figure A2: Positive association between the perceived level of the harm to feelings and the perceived level of the harm to dignity. The sample of the respondents assigned to Content 4 ("hate speech, that is, insults and incitements to violence") and Target 1 (*Zainichi* Koreans) is used.

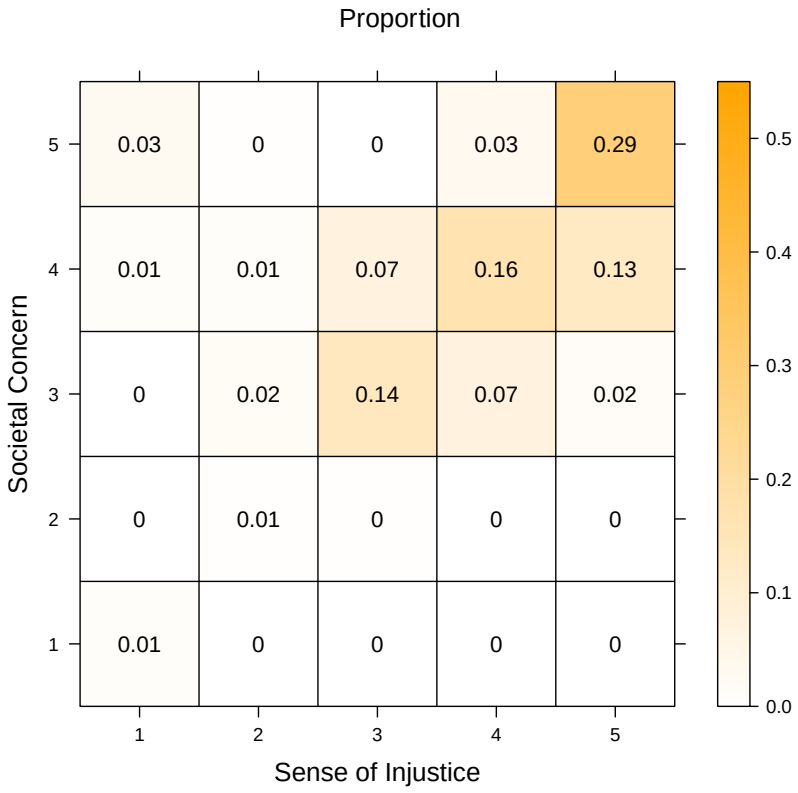

Figure A3: Positive association between societal concern and the sense of injustice. The sample of the respondents assigned to Content 4 ("hate speech, that is, insults and incitements to violence") and Target 1 (*Zainichi* Koreans) is used.

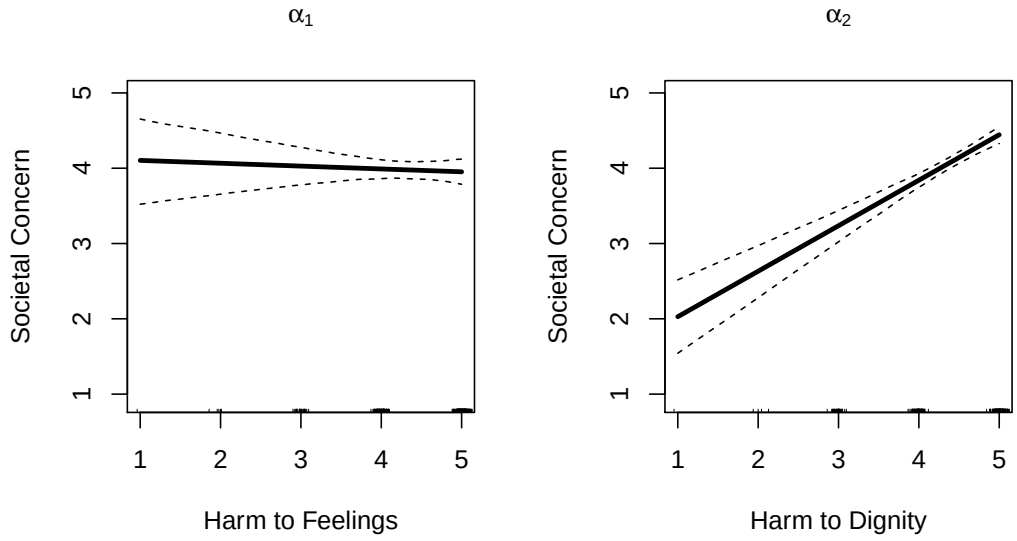

Figure A4: Associations between societal concern and the harm to dignity as well as the harm to feelings. The results are based on the linear regression model that includes as covariates the harm to feelings, the harm to dignity, the respondent's gender, age, ideology, and subjective value of free speech. The dashed lines represent 95% confidence intervals based on robust standard errors. The sample of the respondents assigned to Content 4 ("hate speech, that is, insults and incitements to violence") and Target 1 (*Zainichi* Koreans) is used.

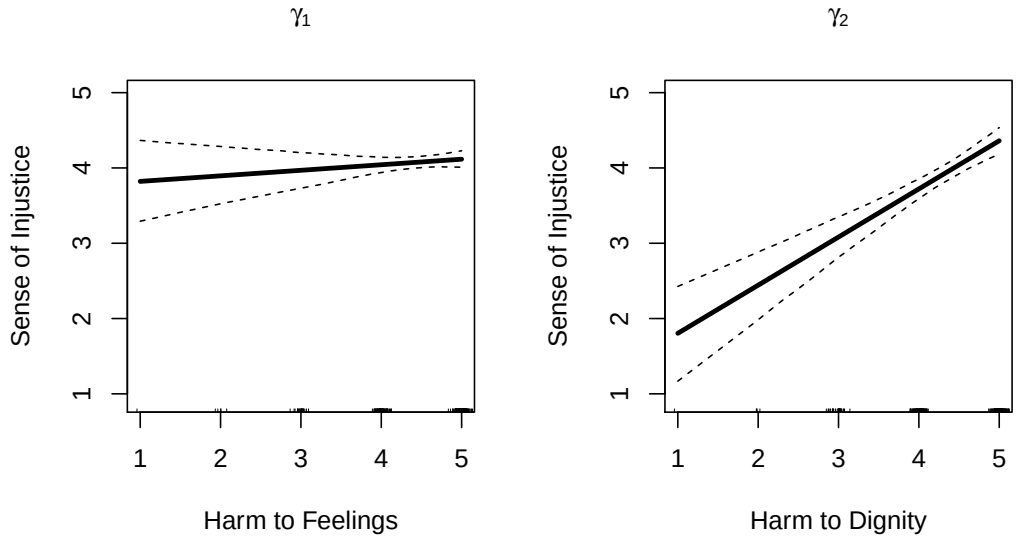

Figure A5: Associations between the sense of injustice and the harm to dignity as well as the harm to feelings. The results are based on the linear regression model that includes as covariates the harm to feelings, the harm to dignity, the respondent's gender, age, ideology, and subjective value of free speech. The dashed lines represent 95% confidence intervals based on robust standard errors. The sample of the respondents assigned to Content 4 ("hate speech, that is, insults and incitements to violence") and Target 1 (*Zainichi* Koreans) is used.

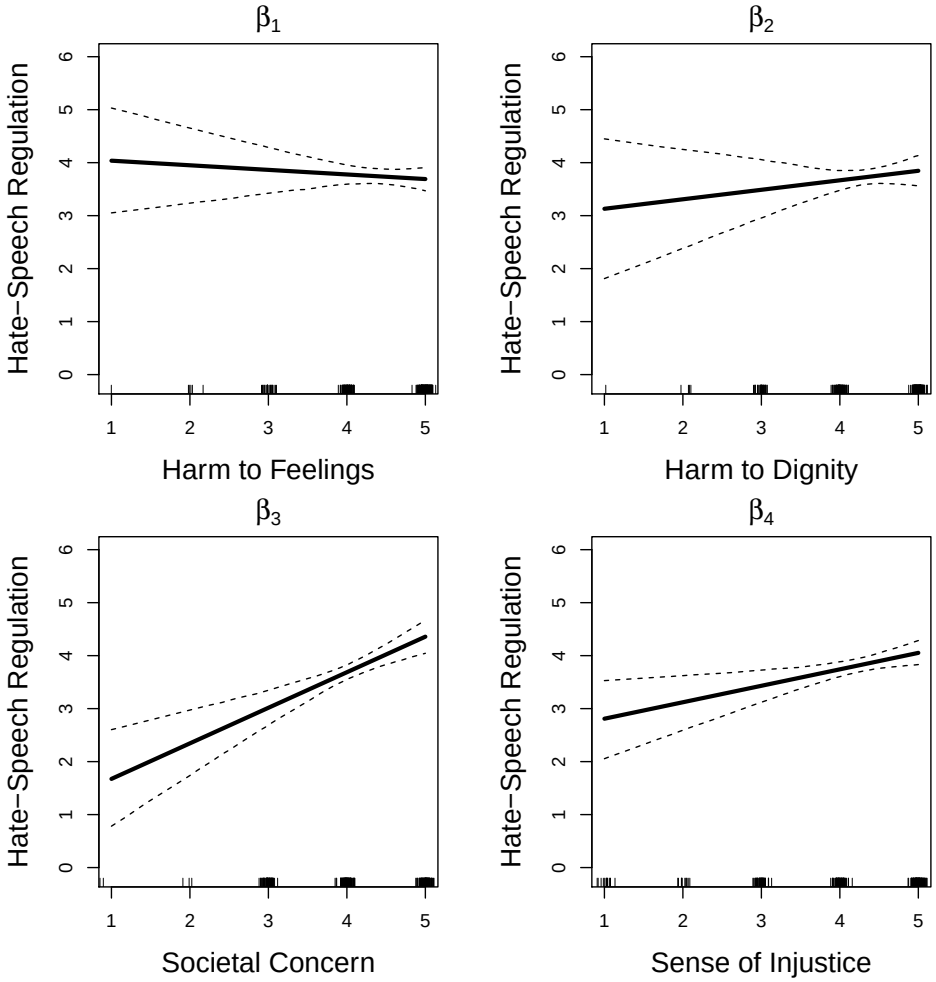

Figure A6: Associations between the harms of hate speech and the attitude toward hate speech regulations. The results are based on the linear regression model that includes as covariates the harm to feelings, the harm to dignity, societal concern, the sense of injustice, the respondent's gender, age, ideology, and subjective value of free speech. The dashed lines represent 95% confidence intervals based on robust standard errors. The sample of the respondents assigned to Content 4 ("hate speech, that is, insults and incitements to violence") and Target 1 (*Zainichi* Koreans) is used.

Procedure for Out-of-Sample Prediction

1 Randomly split the respondents into five groups.

2 Retain one of the five sub-samples as the out-of-sample data set and fit a regression model to the pooled respondents of the remaining four sub-samples.

3 Using the estimated coefficients derived from the regression model fitted to the pooled data set, predict the regulatory attitudes of the out-of-sample respondents.

4 Compute the root mean squared error (RMSE) of the out-of-sample prediction, that is, the magnitude of the prediction errors represented by the differences between the actual and predicted attitudes toward hate speech regulations in the out-of-sample data set.

5 To produce the average of the RMSEs, repeat Steps 2-4 five times by selecting each time a different sub-sample as the out-of-sample data set.

To measure the degree to which the harm to dignity improves forecasting performance, we conducted cross-validation for two regression models: (1) one with the perceived level of the harm to feelings and the four control variables described in the text as the predictors of the attitude toward hate speech regulations and (2) one with these five covariates and the perceived level of the harm to dignity as an additional predictor. We then computed the percentage improvement obtained by adding the perceived level of the harm to dignity to the model with the other five covariates. Specifically, we computed

$$\left(\frac{\widehat{\mathrm{RMSE}}_2 - \widehat{\mathrm{RMSE}}_1}{\widehat{\mathrm{RMSE}}_1} \right) \times 100\%$$

where $\widehat{\mathrm{RMSE}}_1$ and $\widehat{\mathrm{RMSE}}_2$ represent the average RMSEs obtained from the model with and without the harm to dignity, respectively. To take into account randomness-induced uncertainty about the estimate of the prediction improvement rate, we repeated Steps 1 to 5 1,000 times.