Dangerous Speech: A Cross-Cultural Study of Dehumanisation and Revenge

Jordan Kiper[1], Christine Lillie[3], Richard A. Wilson[2], Brock Knapp[3], Yeongjin Gwon[4], and Lasana T. Harris[5]

[1]Department of Anthropology, University of Alabama at Birmingham

[2]University of Connecticut School of Law

[3]Department of Psychology and Neuroscience, Duke University

[4]Department of Biostatistics, University of Nebraska Medical Center

[5]Division of Psychology and Language Sciences, University College London

Author Note

We have no known conflict of interest to disclose.

Correspondence concerning this article should be address to Jordan Kiper, Department of Anthropology, University of Alabama at Birmingham, University Hall, 1402 10th Avenue South, UH 3165, Birmingham, Alabama 35294-1241

Abstract

Dehumanization is routinely invoked in social science and law as the primary factor in explaining how propaganda encourages support for, or participation in, violence against targeted outgroups. Yet the primacy of dehumanization is increasingly challenged by the apparent influence of revenge on collective violence. This study examines critically how various propaganda influence audiences. Although previous research stresses the dangers of dehumanizing propaganda, a recent study found that only revenge propaganda significantly lowered outgroup empathy. Given the importance of these findings for law and the behavioral sciences, this research replicates that recent study with two additional samples that were culturally distinct from the original, finding again that only revenge propaganda was significant. To explore this effect further, we also conducted a facial electromyography (fEMG) among a small set of participants, finding that revenge triggered significantly stronger negative emotions against outgroups than dehumanization.

Keywords

Bayesian regression – dehumanization – facial electromyography (fEMG) – international criminal tribunals – propaganda – speech crime trials

1. Introduction

Several studies have recently considered how various forms of propaganda[1] contribute to the spread of misinformation (West & Bergstrom, 2021; Zerback et al., 2021) and motivations for intergroup violence (Cremin & Popescu, 2021; Yanagizawa-Drott, 2014). One such study (Kiper, Gwon, & Wilson, 2020) investigated how nine types of propaganda – which expert witnesses in international criminal trials identified as contributing to mass crimes– relate to support for violence against targeted outgroups. Notably, the findings from that recent study failed to support prior predictions that exposure to dehumanizing propaganda increases support for outgroup violence. Contrary to extant theories, the authors found that only revenge propaganda — and not dehumanization —predicted lowered empathy for a targeted outgroup. Although lowering outgroup empathy is not the same as prompting violence, researchers have found a correlation between reduced empathy and a propensity to violence (Gao et al., 2009; Siever, 2008). These results, along with similar findings on support for demagogic or authoritarian leaders (Petersen, 2020) and participation in violent political movements (Badar, 2016; Fujii, 2009; Straus, 2015), suggest that current theoretical frameworks, which consider dehumanization as the most dangerous form of propaganda, may require revision.

To determine whether dehumanization deserves its primacy in legal and social science theories of propaganda, it is necessary to evaluate critically findings that may falsify theoretical predictions

[1] We note here that while "propaganda" has lacked a unifying definition, speech crime trials have given it greater precision. In those trials, propaganda is consistently described as a persuasion technique that is based on emotional appeals. These negative emotional appeals, according to expert witness Anthony Oberschall (2006), include negative outgroup stereotyping, appeals to victimhood, dehumanization, nationalism, religion, justice, past atrocities, if not conspiracies or paranoia by which the propagandist creates a sense of threat and a demand for violent action (see also Prosecutor v. Šešelj, T. 2054, as cited in Wilson, 2016, p. 737). Other terms, such as "inciting language" (Wilson, 2017) and especially "dangerous speech" (Leader Maynard & Benesch, 2016), have been used interchangeably with "propaganda" in recent legal literature. In the present study, we use the term to refer to one of nine types of speech acts that are recognized in law as constitutive of hate propaganda, which we delineate below.

that inform current speech crime laws (see Dojčinović, 2012, 2019) and the burgeoning study of dangerous speech (Leader Maynard & Benesch, 2013). Two means of doing so, which are especially important for the current replication crisis, are replication and exploratory research using diverse methodologies. Accordingly, this study entailed two conceptual replications of the research conducted by Kiper, Gwon, and Wilson (2020) to determine if their original findings would hold across different audiences and cultural contexts. We also conducted an exploratory facial electromyography (fEMG) study to compare the influence of dehumanizing propaganda and revenge propaganda on negative emotional reactions toward a targeted outgroup. Our results indicate that propaganda exposure did not predict support for violence, all else being equal, but exposure to revenge propaganda significantly lowered outgroup empathy. We thus conclude by discussing the importance of these now replicated findings for evaluating speech crimes, atrocity prevention, and studies of propaganda.

2. Speech Crimes Trials

Since the Nuremberg Trials (1945 – 1946), over thirty propagandists have been prosecuted in international criminal courts for speech crimes. These crimes include persecutory hate speech, instigating violence, disseminating propaganda as part of a joint criminal enterprise, aiding and abetting mass atrocity crimes, and directly and publicly inciting genocide (Dojčinović, 2012, 2019). In each case, prosecutors argued that the speech acts of a defendant played a critical role in causing violence against an outgroup (Wilson, 2016, 2017). Furthermore, social scientists appearing as expert witnesses regularly reinforce the claim that dehumanization is the most dangerous type of propaganda (Oberschall, 2006, 2012).

In the United States, convictions of individuals for hate crimes, which parallel mass atrocity crimes on a smaller scale, have also emphasized the role of dehumanizing language as increasing the chances of inter-group violence (Smith, 2018). Moreover, when considering first amendment rights, legal scholars routinely draw the line at restricting speech that dehumanizes others, since such language is considered as likely to contribute to group defamation (Strossen, 2018) and justify an ingroup's attacks on outgroup members (Waldron, 2012).

Due to a surprising lack of data on the "causal link" between propaganda and violence in law (Benesch, 2012; Wilson, 2016), there is growing attention in determining whether these legal certainties are supported by scientific evidence (Badar & Florijančič, 2020). The evidence to date raises far more questions than answers. Post-conflict ethnographies of mass atrocity crimes have found that perpetrators report being influenced not by propaganda but rather immediate social factors such as using violence instrumentally to reverse local hierarchies, to attain material incentives, and to enact revenge against neighbors (citation needed here, to your work?). Scholarship on authoritarian movements suggests that propaganda functions less by manipulation and more as a signal around which individuals can build political coalitions, such that even seemingly dangerous propaganda may have little influence on would-be coalitional members (Petersen, 2020). Accordingly, several questions about how propaganda works have recently become prominent not only for legal theorists but also scientists across a range of disciplines.

3. Dehumanization

Our central theoretical question is whether the claim that *dehumanizing propaganda is the most dangerous type of propaganda* is in fact true. Expert witnesses in criminal trials have recurrently advanced this claim, and social scientists have argued for over half a century that dehumanization contributes more to outgroup prejudice, discrimination and hatreds than any other content (Herman & Chomsky, 1988; Jowett & O'Donnell, 2018; Pratkanis & Aronson, 2001). Scholars in the humanities, such as philosophers, have likewise claimed that the denial of another's humanness is a necessary condition for outgroup cruelty, persecution, and indifference (Smith, 2018). However, it was not until the last decade that direct empirical research offered clearer evidence for the outcomes predicted by these scholars. Most notably, psychologists have found that blatant dehumanization (in which a person or group is portrayed as less human than oneself) predicts negative attitudes toward a targeted outgroup (cite lasana on dehumanization here? Bruneau et al., 2018), while infrahumanization (in which one's ingroup is portrayed as more human than others) predicts overall reduced empathy for outgroups in general (Kteily et al., 2015).

Nevertheless, empirical studies have left the question regarding the effects of dehumanization on outgroup violence relatively unanswered until recently – and these recent developments stem largely from contributions made by the authors. Specifically, Author (2006, 2009) originally proposed that violence resulting from dehumanization most likely works by reducing social cognitive processing for others, where an outgroup is perceived as having low-warmth and low-competence. Since then, this relationship has been tentatively documented among participants exposed to scantily-clad women (Cikara, Fiske, & Eberhardt, 2010), free riders (Beyer et al., 2013), people in labor markets (Author et al., 2014), and stigmatized groups (Author, 2017). In arguably the most impactful studies to date on dehumanization and violence, Rai and colleagues

(2017; Fiske & Rai, 2014) found that exposure to dehumanization increased the likelihood of accepting instrumental violence against an outgroup – that is, supporting violence not out of hateful beliefs but instead to advance the ingroup's goals. In line with these results, recent studies suggest that dehumanization may not motivate violence directly but rather indirectly by decreasing social cognition or moral regard for a targeted outgroup and legitimating violence to mitigate ingroup threats (see Slovic et al., 2020). Although this is slightly different from the account of propaganda in law which stresses the direct effects of propaganda, it still supports the theoretical claim that dehumanization is the most dangerous type of propaganda.

4. Summary of Kiper, Gwon, and Wilson (2020)

One such challenge comes from Kiper, Gwon, and Wilson (2020), who did not seek to contest dehumanization itself but rather to investigate how exposure to one of nine types of propaganda, as identified in law as increasing support for violence, differentially effected participants' social cognition, and thus moral judgments about others. To do so the authors adapted the latest propaganda typology used by Anthony Oberschall, an expert witness in the most recent speech crime trial, namely, that of Vojislav Šešelj, a Serb ultranationalist, at the ICTY. The nine types were:

> *Direct threat or paranoia:* conveying a threatening message about the outgroup that arouses fears or public demand for action to reduce the threat.
> *Past atrocities:* referencing historical or recent atrocities against the ingroup (whether genuine or fabricated) to justify violent acts against the outgroup.
> *Victimization:* referring to past or ongoing victimization and stressing that unless the ingroup acts, the population will be victimized again.
> *Justice:* attempting to create a consensus that actions against the outgroup are just and consistent with laws or customs.
> *Revenge:* claiming that the ingroup bears no responsibility for violence against the outgroup since the ingroup is merely retaliating for unpunished crimes committed against them.

> *Religion:* using religious language to construct a moral or spiritual principle for the ingroup's actions.
> *Nationalistic speech:* arguing that because the ingroup and state are congruent, members of the ingroup are justified in defending the "nation's" traditions, lands, ancestry, language, and culture.
> *Negative outgroup stereotyping:* generalizing or labelling everyone from the outgroup according to the ingroup's most negative and oversimplified images or ideas about the outgroup.
> *Dehumanization:* depicting the outgroup as animals, pests, diseases, or otherwise harmful to the ingroup and not fully human (as cited in Kiper, Gwon, and Wilson, 2020, p. 409).

For the full description and examples of these types, materials are available in the Supplementary File. Insofar as this typology is also found in the propaganda of terrorist organizations, hate groups, and violent political movements (see Kiper, Gwon, & Wilson, 2020, p. 409; see also Badar & Florijančič, 2020), it suggests that legal experts are indeed justified in identifying these types as central to the repertoire of violent coalitions.

4.1 *Increased justifications for using violence against outgroups*

The first goal of Kiper, Gwon, and Wilson (2020) was to explore whether any of the nine types induced a shift in moral judgments about the legitimacy of inflicting violence on an outgroup. Such a shift was expected, insofar as expert witnesses regularly base their claims on the *information processing model of mass manipulation* (or "mass-manipulation theory" for sake of brevity; see Jowett & O'Donnell, 2019). Mass-manipulation theory predicts that individuals exposed to propaganda that targets an identifiable population as a threat will experience increased indifference or animosity towards that population (see Haslam, 2006). For Oberschall (2006:12-38), hate propaganda shifts an individual's mindset from a "peacetime frame" of mutualism to a “crisis frame” in which individuals disregard the outgroup or considers them a threat, resulting in tacit or open support for violence against them.

To test this prediction, Kiper, Gwon, and Wilson (2020) used a series of vignettes to prompt participants, drawn from a Serbian sample, into identifying with a fictitious ingroup that was facing threats from a fictitious outgroup. This was followed by exposure to one of nine types of propaganda adapted for the study in the form of a speech by an ingroup’s influencer (see methods section below). Kiper, Gwon, and Wilson found that none of the types predicted changes to participants’ moral judgements about violence. While this may be due to the absence of important contextual factors, such as peer pressure or cultural milieu, the authors inferred from this finding that “onetime exposure to propaganda is unlikely to induce support for violence” (p. 423).

Kiper, Gwon, and Wilson employed Bayesian regression to predict the likely effects of propaganda exposure given the language of “likelihood” and “probability” in speech crime trials (Benesch, 2012; Carver, 2000; Wilson, 2015). However, it is not at all apparent that participant responses might shift toward violence if the experiment were repeated. Developments in cognitive science indicate that propaganda does not change people’s beliefs, but rather aligns like-minded individuals behind a sociopolitical movement (see Petersen, 2020). Furthermore, the pattern of these results suggest that in terms of Bayesian priors, a repetition of exposure to propaganda should result in a lowered posterior probability for outgroup violence and, if the coalitional account of propaganda is true, a greater affinity for the ingroup. Put simply, replication should demonstrate that propaganda alone does not predict violence but does enhance coalitional identities – a claim recently stressed by scholars of epistemic vigilance (Mercier, 2020).

If so, demonstrating this with propaganda would discredit a common assumption that the dangers of propaganda can be inferred from the type of language used by the propagandist. One consequence of this is that researchers must situate propaganda in a cultural context – and in a particular set of socio-political conditions – that predicts violence. We are not concerned here with those finer distinctions but instead the specific effects, ceteris paribus, of different types of propaganda on moral judgments and emotions.

Our working theory is that if any one type of propaganda is more dangerous than another, it is an emotional appeal to revenge and not necessarily dehumanization (see Everett & Worthington, 2020). This is because revenge would have deterred social transgressions in ancestral environments (McCullough, Kurzban, & Tabak, 2013), and the cross-cultural experience of entertaining thoughts or feelings of revenge after a perceived transgression (e.g., Chester & Martelli, 2020) suggest that revenge may have been adaptive in our evolutionary past. Therefore, the current study sought to critically examine both the effects of repetition on support for violence and the effects of revenge as compared to dehumanizing propaganda on negative attitudes toward outgroups.

4.2 *Increased empathy for the ingroup*

With aim of replication in mind, we sought to identify any changes to empathy after propaganda exposure, as documented by Kiper, Gwon, and Wilson. The second goal was to examine if any of the propaganda types would be differentially associated with ingroup empathy. As an exploratory study, the authors did not anticipate that any type would have greater effects than the others. However, based again on the coalitional theory of propaganda, we expected that if propaganda had a general effect on audiences, it would be to increase ingroup empathy. What the original study overlooked is this: it is extremely difficult to mobilize individuals behind a violent movement without prior beliefs that such mobilization would be beneficial, but it is much easier to elicit sympathy for members of a seemingly maligned ingroup (Boyer, 2018; Tooby, Cosmides, & Price, 2006). Critically, increased empathy does not deter violence, but it does strengthen ingroup connections (Vollberg, Gaesser, & Cikara, 2021); and when one feels a heightened connection to their group, they are more likely than otherwise to make sacrifices on behalf of their group, including outgroup violence (Newson, 2017; Purzycki & Lang, 2019).

Kiper, Gwon, and Wilson's (2020) study demonstrated that not all propaganda increases ingroup empathy. For instance, appeals to religion and negative stereotypes about a directly threatening outgroup failed to elicit empathy. However, propaganda that centered on victimization, past atrocities, nationalism, and revenge significantly increased ingroup empathy. These responses are not unexpected when considering that the experiment was conducted with Serbian participants who perhaps identified more with a threatened and historically victimized ingroup than other audiences would have (needs a citation since readers may not know Serb history and longstanding self-image of victimization). The "integrated approach" to dangerous speech (Leader Maynard & Benesch, 2016) argues that emotional appeals and historical experiences together predict whether propaganda will be effectual. As such, appeals to themes such as historical victimization and past atrocities may have resonated with a Serbian audience whose history include foreign occupation, ethnoreligious persecution, and genocide (citation). These themes also appeal to other audiences and thereby increase ingroup empathy, in general. Additionally, the priors in the original study suggest that eliciting ingroup empathy is far easier than lowering outgroup empathy or inciting violence. By replicating the study, the current research may support this prediction – and specifically the coalitional account that all propaganda types should reliably increase the likelihood of empathizing with the portrayed ingroup.

4.3 *Decreased empathy for the outgroup*

The third goal of the original study was to explore how exposure to propaganda types lowered outgroup empathy. Previous scholarship predicts that dehumanization would be the strongest. However, dehumanization was not found to be significantly associated with lowered outgroup empathy. Moreover, only one type of propaganda was significant: an emotional appeal to revenge (Kiper, Gwon, & Wilson, 2020, p. 422).

This result carries two important implications. The first is lack of empirical support for the effects of dehumanization on both support for outgroup violence and lowered outgroup empathy. The latter is especially surprising since previous and widely cited research predicts that blatant dehumanization should not only increase negative attitudes for a targeted outgroup but, in the least, reduce outgroup empathy. However, when we consider that other studies have found that it is disgust-relevant social categorization – and not dehumanization alone – that facilitates disregard for an outgroup (Buckels & Trapnell, 2013; Valtora et al., 2021), the lack of significance for dehumanization by means of a vignette is not so surprising. To address this limitation, we decided to investigate the effects of revenge propaganda and dehumanizing propaganda using an f(EMG) study, allowing us to detect subtle signs of both changes to outgroup empathy and levels of disgust for a targeted outgroup.

The second implication is that revenge propaganda – and not dehumanization – significantly predicted lowered outgroup empathy. Beyond noting the impact of revenge on ingroup and outgroup empathies, Kiper, Gwon, and Wilson (2020) concluded that over time, collective feelings of revenge could increase desires or justifications for punishing an outgroup. This also motivated closer examination of potential alternative explanations. Specifically, we thought the idea that such propaganda could serve as "the groundwork for persecution through the enhancement of apparent divisions between groups" (p. 423) merited further attention, given that courts and scholars have sought to predict when propaganda is likely to be dangerous (Wilson, 2015). If revenge consistently reduces outgroup empathy and that reduction is accompanied by strong emotions such as disgust, then it is likely that revenge propaganda may indeed be the groundwork for persecution while dehumanizing propaganda serves another purpose, such as the justification for perpetration in real time or after the fact (Kelman, 2017).

The current study re-examines the relationship between revenge speech and empathy found in Kiper, Gwon, and Wilson (2020) by replicating the original. Adding to this, the current study focuses closely on the effects of both revenge propaganda and dehumanizing propaganda to determine which, if any, elicited significant negative emotions, including disgust, for the targeted outgroup.

5. Replicating the original design and an exploratory f(EMG)

In keeping with the original study design, the research questions (RQs) were as follows. RQ1: How do harmful messages – those recognized in international law as likely to induce violence – increase hostility toward outgroups? RQ2: Does exposure to these messages, notwithstanding cultural variability, effect populations in similar ways? Most importantly and relatedly, the three-part hypothesis from the original study was:

H1: Harmful messages will differentially contribute to (a) increased justifications for using violence to resolve political conflict, (b) increased empathy for the ingroup, and (c) decreased empathy for the outgroup.

However, unlike the original study, we were not interested in the interaction effects of prior levels of nationalism with propaganda, the results of which were inconclusive. Instead, we took note of a more remarkable outcome: again, that exposure to revenge propaganda significantly predicted lowered empathy for outgroups, while dehumanizing propaganda did not. This outcome alone has potential repercussions for psychological theories of intergroup violence and speech crimes. Hence, we investigate this finding further by focusing closely on the differential emotional effects of exposure to revenge propaganda or dehumanizing propaganda using an f(EMG) study, hypothesizing that:

H2: Participants exposed to harmful messages (a) will express a significantly negative emotional reaction to the outgroup, and (b) dehumanizing propaganda will induce a significantly lowered negative reaction in the form of disgust toward an outgroup as compared to revenge propaganda.

Keeping with the original study, we focused primarily on testing the effects of propaganda using a web-based experimental survey. Our team translated the survey from Serbian to English and recruited U.S. participants from M-Turk, since the latter closely approximated the general U.S. population and offered an accessible sample for comparisons in outcomes, especially as effect size increased (something that is critical to predictions using Bayesian models; see Hahn, Murray, & Carvalho, 2020). We first ran the survey independently in the U.S. (U.S. Study 1), repeated the survey with a separate U.S. population (U.S. Study 2), and administered the exploratory f(EMG) study. In what follows, we discuss the design and results of the experimental survey and f(EMG) experiment.

6. The Web-based Experimental Survey

6.1 Participants

To understand the priors for this research, 399 Serbian participants were included in the final analysis of the original survey, the majority of whom were female (56%), ranging in age from 18 to 29 (59%), having completed a high school education (51%), living in a middle-income household (43%), and identifying as somewhat liberal (37%). Participants in our first replication consisted of 408 MTurk workers from the U.S. Of these, only 392 were included in the final analysis because 16 failed to complete or answer the attention question correctly. Most were female (58%), between the ages of 18-29 (34%), were low to middle income (42%), and identified as somewhat liberal (31%). For the second replication, we used the same exclusion criteria, resulting in 339 participants from M-Turk, the majority of whom were male (60%), between 30 to 39 years of age (42%), were middle income (52%), and somewhat liberal (30%). Altogether, 1,130 participants were surveyed, ranging in age from 18 to 29 (41%), were relatively even in terms of gender (52% female) and identified as having a middle income (32%) and somewhat liberal (33%).

6.2 Measures and Procedure

Participants first completed a questionnaire about individual characteristics and were then instructed to imagine themselves as a member of a fictitious community known as "East Margolia." Participants then completed the experimental component which consisted of reading one of nine propaganda excerpts about the fictitious community and its relationship to a similar but contentious outgroup known as "West Margolians." This was followed by twenty questions about justifications for violence and altered intergroup perceptions. Detailed information about the selection of measures, including the propaganda treatments and their use in prior research, can be found in the Supplementary File. Here, we briefly outline the specific factors, treatments, and variables used.

Individual factors. Each study began with a series of randomized questions from survey instruments. Following the original study, violent media exposure was assessed by two questions on a six-point scale (e.g., from "I never watch violent TV shows or movies" to "7 or more hours a day"), which overall had fair internal consistency (α = .75). Authoritarianism was based on 12 questions with a 9-point scale (see Robinson, Shaver, & Wrightman, 1991), which had good internal consistency (α = .81). Just world beliefs were measured on the abbreviated 6-item scale from Collins (1974) and demonstrated very good consistency (α = .88), while questions on religious strength from Koenig and Büssing (2010) had an excellent consistency (α = .96). Given the importance of disgust for dehumanization's probable effects, the replicated studies added measures from the Disgust Sensitivity Scale (Haidt, McCauley, & Rozin, 1994), which even though it had low internal consistency (α = .52), we retained it since the scale itself is so widely used (and to avoid the "file drawer effect").

Propaganda treatments. After answering the above questions, participants read one of nine treatments, which were randomly assigned. Each treatment was based on propaganda by Vojislav Šešelj, who was convicted for international speech crimes (Badar & Florijančič, 2020). Our selections were identical to Kiper, Gwon, and Wilson (2020), who relied on the coding system of Oberschall (2006). Before each treatment (see Supplementary File), participants were first given a short description about East and West Margolia, including an explanation about the context of rising tensions between the two countries, and then instructed – using Caprariello, Cuddy, and Fiske's (2009) scenario-depictions method – to imagine themselves as an East Margolian.[2] Keeping with this method which has been used for measuring the effects of media on prejudice and hate crimes (Cramer et al., 2014), participants were then randomly assigned to a speech by an East Margolian leader, which served as the manipulated structural predictor and, thus, the treatment for the survey.

Outcome variables. After reading the randomly assigned propaganda treatment, participants were then asked to answer two sets of randomized questions that served as outcome variables. The first

[2] As a replicated method, scenario-depictions are a form of vignettes that allow for safely examining cognition or changes therein by using hypothetical scenarios without inflating effect sizes (Capariello, Cuddy, & Fiske, 2009).

set included seven general questions about violence (e.g., “Do you think violence can be justified?”) while the second set included ten specific questions about West and East Margolia (e.g., “To what extent do you think you could understand West Margolian’s point of view?”). After completing the study, participants were debriefed about the nature of the research and told that we used fictitious countries so as not to alter participants’ opinions towards actual people.

6.3 Data Analysis Strategy

To determine sample size, we first conducted a power analysis, choosing a small Cohen’s effect size ($f^2 = 0.05$) and to achieve at least 80% statistical power, while also controlling for a type 1 error of 5% and an anticipated 20% attrition rate. The necessary threshold for detecting an effect in each sample was determined to be 335 participants, and thus we had a sufficient number for each sample. Next, we repeated the analysis strategy of the original study by conducting an exploratory factor analysis (using principal components analysis with varimax rotation) on the outcome variables, Spearman’s correlation to explore relations between the total set of composites derived from individual factors and the factor analysis, and Bayesian multiple linear regression to estimate the predicted outcomes. Bayesian methods were appropriate given that trial judgments prioritize the likelihood of propaganda having direct effects on audiences. With that in mind, and to most accurately account for sampling, stratification, and treatment effects using Bayesian regression, we analyzed data in the following order. Prior probabilities were based on the original study (Sample 1), while the final posterior probabilities – and thus the predicted effects of propaganda types – were based on the additional outcomes identified by U.S. Study 1 (Sample 2) and then U.S. Study 2 (Total Sample).

6.4 Results

Turning first to the factor analysis, we verified the factors identified in the original study and also replicated those factors with each U.S. sample. Given these results, the Kaiser-Meyer-Olkin measure of each study verified a “meritorious” sampling adequacy, namely, in the final sample, with KMO = .82 (Dodge, 2008), and KMO values greater than .69, which is above the acceptable limit of .5 (Field, 2013). The significance of the Bartlett’s test (χ2 (136) = 8,028.02, p < .001) indicated clear patterns in participants’ responses, and four factors had eigenvalues over Kaiser’s criterion of 1 and together explained 64.01% of the variance (table 1 shows the factor loadings with rotation). The clustered items were Factor 1 (F1): justifications for violence, Factor 2 (F2): ingroup empathy, Factor 3(F3): outgroup empathy, and Factor 4 (F4): intergroup blame. In terms of reliability, justification for violence (α = .87), ingroup empathy (α = .84) and outgroup (α = .84) were highly reliable, while intergroup blame was minimally reliable (α = .73). For our primary analyses, we retained the original study’s inference in naming the reliable factors “justifications for violence,” “ingroup empathy,” and “outgroup empathy” (for further analyses, see Supplementary file 1: Appendix).

[INSERT TABLE 1 HERE]

We then examined the descriptive (table 2) and correlational (table 3) relationships of all data for comparisons. Descriptives thus include the original and total sample composition. For the latter, the most notable correlations were that justifications for violence moderately correlated with violent media exposure (r = .381, p < .001), while ingroup empathy was moderately correlated with disgust sensitivity (r = .269, p < .001). Similarly, outgroup empathy moderately correlated with justifications for violence (r = .288, p < .001), violent media exposure (r = .272, p < .001), and just world beliefs (r = .227, p < .001). We provide additional correlations for each particular sample in the Appendix.

[INSERT TABLE 2 HERE]

[INSERT TABLE 3 HERE]

We then examined whether exposure to propaganda increased the likelihood of justifying violence, increasing ingroup empathy, or decreasing outgroup empathy. To that end we conducted posterior inference with 5,000 Markov Chain Monte Carlo (MCMC) samples taken from every tenth iteration and after a burn-in of 5,000 iterations, which allowed us to compute highly accurate posterior estimates. The MCMC convergence was checked using diagnostic procedures including trace and autocorrelation plots. The statistical significance of each treatment was determined based on the 95% highest posterior density (HPD) interval. If the interval did not include the value zero, the predictor was statistically significant for the outcome. Finally, because replication and the posteriors thereof offer the most predictive outcomes, our narrative focuses on the results from the total sample, but we include prior samples in each table for comparisons and to indicate changes in effect sizes. Additional analyses and an expanded description of the methods used, including the data files and key SAS coding, are provided in the supplementary file.

Table 3 indicates that of the nine types of propaganda, none predicted justifications for violence, while Table 4 displays the predictors for ingroup empathy. Our analysis showed that past victimization (*b* = 0.832, SD = .143), revenge (*b* = .787, SD = .142), nationalism *(b* = .742, SD = .146), dehumanization (*b* = .580, SD = .144), religion (*b* = .507, SD = .145), justice (*b* = .432, SD = .144), stereotypes (*b* = 0.403, SD = .147), and past atrocities (*b* = .529, SD = .145) were statistically significant and positively associated with ingroup empathy. For outgroup empathy (Table 5), only revenge propaganda (*b* = -.356, SD = .136) was statistically significant for predicting decreased empathy for the targeted outgroup.

[INSERT TABLE 4 HERE]

[INSERT TABLE 5 HERE]

[INSERT TABLE 6 HERE]

6.5 Discussion

In sum, we found that no propaganda type increased justifications for violence (H1.a), every propaganda type predicted increased ingroup empathy (H1.b), while only revenge predicted decreased outgroup empathy (H1.c). These results indicate that exposure to propaganda has predictable effects on audiences, even in cultural settings as distinct as Serbia and the USA. As we anticipated, nearly any propaganda type will likely increase one's affinity for their ingroup, while revenge propaganda is most likely to decrease regard for a targeted outgroup.

7. Exploratory fEMG Study

To investigate further whether there were differences between dehumanizing propaganda and revenge propaganda, and to address the lack of data on emotional variance generated with propaganda exposure, we designed an additional exploratory facial electromyography (fEMG) study. As a common method used in psychology, fEMG measures facial muscle activity commonly displayed when people experience different emotion states. Accordingly, it provided a non-invasive — yet non-explicit measure — of emotional reactions by participants when exposed to propaganda.

7.1 Participants

To determine sample size, we first conducted a sensitivity analysis using G*Power (Faul et al., 2017), which revealed that a sample of 30 participants would be sufficient to detect small effects of $f = 0.27$ assuming $\alpha = 0.05$ and $\Omega = 0.95$ (mean correlation among repeated measures = 0.5). We thus collected data from 37 participants who enrolled in the f(EMG) study at Anonymous University-1. However, because data from seven participants resulted in recording errors, our final sample consisted of 30 participants, which although was low powered, was just enough to detect effects.

7.2 Materials, Apparatus, and Procedure

Following previous f(EMG) studies (Lundqvist, Flykt, & Öhman, 1998), we selected still images of male faces displaying neutral expressions from the Karolinska Directed Emotional Faces database as visual stimuli to accompany the propaganda. To complete the task, participants used a 13-inch computer monitor connected to a PC, on which an E-Prime 2.0 software presented the stimuli and recorded participants' responses. An adjacent monitor and Mac Mini recorded the fEMG data using AcqKnowledge 4 software. The fEMG was sample at 1000Hz for the corrugator supercilia (CS) and levator labii superioris (LLS) muscles. Critically, the CS is engaged during anger facial expressions and negative affect (e.g., frustration), while the LLS is engaged during disgust facial expressions[3] (Ekman & Friesen, 1978).

As with the survey, participants were given the context of rising tensions between East Margolia and West Margolia and told that they were reading a speech from an East Margolian leader. Each participant was presented with either revenge propaganda, dehumanizing propaganda, or a control speech without either theme. After reading one of these, participants were told to maintain focus on the screen while a series of pictures of faces were presented to them. The participants were first shown a crosshair cue, followed by the nationality of the subsequent person (East Margolia or West Margolia), and lastly the picture of a face displayed for 4000 milliseconds (ms), which was identified as either a West Margolian (outgroup) or East Margolian (ingroup). The faces were randomized across trials and the sequence was repeated 30 times, such that participants saw 15 in-group and 15 out-group faces.

7.3 Data Analysis

Before conducting analyses, the raw fEMG data were integrated, rectified, and log transformed. We then conducted a pair of mixed 3 (*condition*: dehumanization, revenge, control speeches) X 2 (*group*: East, West Margolian faces) ANOVAs independently for each muscle, with condition as the between-participant variable, and group as a within-participants variable. We followed up

[3] While there is no one-to-one mapping of emotion to facial muscle, disgust involves the LLS rather than CS, allowing us to better determine whether disgust occurred rather than simply negative affect.

significant main effects and interactions with LSD post-hoc analyses. We only consider such post-hoc analyses to be robust if the confidence intervals (CI) do not include zero.

7.4 Results

There was no significant difference in our conditions when participants were viewing faces of West Margolians (outgroups; *F* (2, 27) = 3.25, *p* = .055), but after being exposed to revenge compared to control speeches, participants activated the LLS significantly more (M_{diff} = 9.69E-3, SE_{diff} = 4.56E-3, *p* = .043, 95% CI [3.35E-4, 2.00E-2]). Similarly, and most tellingly, we also found more LLS activity (see Figure 1) after participants were exposed to revenge compared to dehumanizing speeches, (M_{diff} = 1.04E-2, SE_{diff} = 4.56E-3, *p* = .031, 95% CI [1.04E-3, 1.98E-2]).

There was no significant differences in the CS, there was significantly increased muscle activity when participants were exposed to revenge speeches compared to dehumanization speeches (M_{diff} = 8.55E-3, SE_{diff} = 4.14E-3, *p* = .049, 95% CI [5.00E-5, 1.71E-2]), and a trending but not robust increased muscle activity when participant were exposed to revenge speeches compared to control speeches (M_{diff} = 7.46E-3, SE_{diff} = 4.14E-3, *p* = .083, 95% CI [-1.60E-2, 1.05E-3]). There was no difference following exposure to dehumanizing and control speeches (M_{diff} = -1.10E-3, SE_{diff} = 4.14E-3, *p* = .793, 95% CI [-9.60E-3, 7.41E-3]).

[INSERT FIGURE 1 HERE]

7.5 Discussion

These data indicate that neither revenge propaganda nor dehumanizing propaganda was significantly related to changes in negative emotional reactions toward the outgroup (H2.a). However, the f(EMG) did show that participants displayed significantly more LLS muscle responses for an outgroup after exposure to revenge propaganda than dehumanizing propaganda, consistent with a disgust response (H2.b). Furthermore, the lack of similar differences for ingroup faces suggests that propaganda itself does not activate this disgust response, but rather that disgust is reserved for persons who warrant it, and propaganda channeled that emotion in this experiment.

8. General Discussion

With the goal of identifying predictive associations between types of propaganda in law, the results of the present research replicated and advanced the original study's findings. Although no single instance of propaganda contributed to justifications for violence, every type of propaganda increased ingroup empathy while only revenge decreased outgroup empathy and when compared to dehumanization, was stronger in eliciting disgust for an outgroup. Accordingly, these findings entail important implications for law and the behavioral sciences.

When it comes to violence, speech crime trials have sought to link propaganda causally to collective violence, but our findings raise doubts about such a direct connection. Propaganda alone may not be necessary nor sufficient for compelling a coalition to support or engage in violence against an outgroup. Instead, propaganda likely provides the groundwork for groups who are already committed to violence to coordinate attacks against others (see also Atran, 2021). Our findings not only offer preliminary support for this coalitional view of propaganda (Petersen, 2020) but also trends in law that argue for indicting propagandists not because they directly cause violence, but because, along with other factors, they elevate the risk of hate crimes or mass atrocities (Wilson, 2017). This implication is significant for speech crimes. If dangerous speech contributes to harmful behaviors by laying the groundwork for violence rather than directly causing it, then it should be considered dangerous for preparing violence among groups. In other words and using legal jargon, data reported here support treating speech crimes as inchoate offenses (that prepare for other crimes) rather than complete offences (crimes in themselves).

Nevertheless, our findings are not without limitations. Specifically, the measures we used for violence were not very different from participants' general worldviews about violence, to which participants overall showed low prior commitments. It may be the case that for persons with higher priors for violence, or whose identity is fused with a group committed to violence (Atran, 2021), will experience an increase in support for violence after propaganda exposure. Thus, the present findings – in combination with the lack of data on propaganda and violence – point toward the need for more research specifically on propaganda exposure among violent coalitions.

Turning to effects on ingroup empathy, it is not surprising that every type elicited stronger feelings for the ingroup. Our findings suggest that it is easier to get people to sympathize with a threatened ingroup than to motivate feelings of dislike or violence for an outgroup. The types of propaganda explored here are therefore unlikely to be the only meaningful categories for influencing ingroup empathy. Research by Choi and Bowles (2007) further supports this interpretation, as they found that humans are widely altruistic but parochial in their altruism, predicting that humans will show diligence for ingroup cooperation but also for threats to the ingroup. Future research should investigate how propaganda relates to such parochial altruism.

Our findings suggest that social researchers and courts might revise their views on the primacy of dehumanization among factors motivating violence and recognize that calls for revenge have more significant effects on an audience. Revenge propaganda was the only type significantly associated with lowered outgroup empathy, and when compared to dehumanizing propaganda, revenge was significantly stronger at increasing feelings of disgust for the outgroup. This aligns with psychological research that suggests it is not per say dehumanization that increases the likelihood of disregarding others but rather disgust-relevant social categorization (Rai et al., 2017). Further, our results align with evolutionary theories of revenge which posit that vengeful feelings are proximate motivations for punishing free-riding or uncooperative behaviors, and thus function to deter current and future transgressions, which was adaptive response in ancestral environments (McCullough, Kurzban, & Tabak, 2013). It seems likely, then, that revenge propaganda, when compared to other types of propaganda identified by law as potentially dangerous, would evoke the strongest response, and that this response would include disgust for the potential transgressors.

Still, there are limits to how far we can speculate about revenge propaganda. By replicating Kiper, Gwon, and Wilson's (2020) original study, this research was purposefully limited to researching the direct effects of distinct types of propaganda identified by law, all else being equal. Yet, all else is not equal when it comes to viewing actual propaganda in the real world. As previously noted, current research on dangerous speech asserts that propaganda and culture are together necessary for evaluating the effects of would-be speech crimes. Insofar as the legal typology that we used here is based on mid-twentieth century theories of mass manipulation, the main shortcoming of this study may be its strength. That is, we show that revenge propaganda is in fact the most likely type of propaganda to lower outgroup empathy. Along similar lines, it should be noted that neither the original nor the current study considered the social identities of groups. It may be that if participants were also given the religious, political, or ethnic identities of the outgroup along with propaganda, their judgments would reflect different or even stronger outcomes.

Our findings support a chorus of researchers and social reformers who are asking courts and corporations to (re)consider the dangers of propaganda. Both international courts and social media companies, such as Facebook/Meta, have prioritized dehumanization in their policies and procedures that regulate content. Our results point towards significant effects of other types of propaganda as well, namely, revenge narratives. On this basis, we do not wish for our findings to be the basis of policy but instead to encourage the ongoing creation of evidence-based measures by criminal tribunals, governments, and corporations, lest their regulation of speech be empirically uninformed and misguided.

References

Atran, S. (2021). Psychology of transnational terrorism and extreme political conflict. Annual Review of Psychology, 72, 471-501.

Badar, M.E. (2016). The road to genocide: The propaganda machine of the self-declared Islamic State (IS). International Criminal Law Review, 16(3), 361-411.

Badar, M.E., & Florijančič, P. (2020). The Prosecutor v. Vojislav Šešelj: A symptom of the fragmented institutional criminalization of hate and fear propaganda. International Criminal Law Review, 1-87.

Benesch, S. (2012). The ghost of causation in international speech crime cases. In P. Dojčinović, (Ed.), Propaganda, war crimes trials and international law: From speakers' corner to war crimes (pp. 254-268). Routledge.

Beyer, F., Münte, T. F., Erdmann, C., & Krämer, U. M. (2013). Emotional reactivity to threat modulates activity in mentalizing network during aggression. Social Cognitive and Affective Neuroscience, 9(10), 1552-1560.

Boyer, P. (2018). Minds Make Societies: How Cognition Explains the World Humans Create. Yale University Press.

Bruneau, E., Jacoby, N., Kteily, N., & Axe, R. (2018). Denying humanity: The distinct neural correlates of blatant dehumanization. Journal of Experimental Psychology: General, 147(7), 1078-1093.

Carver, R. (2000). Broadcasting and political transition: Rwanda and beyond. In J. Currey (Ed.), African Broadcast Cultures: Radio Transition (pp. 189-197). Oxford University Press.

Chester, D.S., & Martelli, A.M. (2020). Why revenge sometimes feels so good. In E. Worthington Jr. & N.G. Wade (Eds.), Handbook on Forgiveness, Second Edition (pp. 43-51). New York: Routledge.

Choi, J., & Bowles, S. (2007). The coevolution of parochial altruism and war. Science, 318(5850), 636-640.

Cikara, M., Eberhardt, J. L., & Fiske, S. T. (2011). From agents to objects: Sexist attitudes and neural responses to sexualized targets. Journal of Cognitive Neuroscience, 23(3), 540-551.

Cramer, R., Clark, J., Kehn, A., Burks, A., & Wechsler, H. (2014). A mock juror investigation of blame attribution in the punishment of hate crime perpetrators. International Journal of Law and Psychiatry, 37(6), 551-557.

Cremin, M.R., & Popescu, B.G. (2021). Sticks and stones? Connecting insurgent propaganda with violent outcomes. Journal of Conflict Resolution, 1-25.

Dojčinović, P. (2012). Propaganda, War Crimes Trials and International Law: From Speakers' Corner to War Crimes. Routledge.

Dojčinović, P. (2019). Propaganda and International Criminal Law: From Cognition to Criminality. Routledge.

Everett, L., Worthington, J.R., & Wade, N.G. (Eds.). (2020). Handbook of Forgiveness, Second Edition. Routledge.

Fiske, A.P., & Rai, T.S. (2014). Virtuous Violence: Hurting and Killing to Create, Sustain, End, and Honor Social Relationships. Cambridge, MA: Cambridge University Press.

Fujii, L.A. (2009). Killing Neighbors: Webs of Violence in Rwanda. Ithaca, NY: Cornel University Press.

Gao, Y., Glenn, A.L., Schug, R.A., Yang, Y., & Raine, A. (2009). The neurobiology of psychopathy: A neurodevelopmental perspective. The Canadian Journal of Psychiatry, 54, 813-823.

Hahn, R., Murray, J., & Carvalho, C.M. (2020). Bayesian regression models for causal inference: Regulation, confounding, and heterogenous effects. Bayesian Analysis, 15(3), 965-1056.

Haslam, N. (2006). Dehumanization: An integrative review. Personality and Social Psychology Review, 10, 252-264.

Herman, E., & Chomsky, N. (1988). Manufacturing Consent: The Political Economy of the Mass Media. Pantheon.

Jowett, G.S., & O'Donnell, V. (2019). Propaganda and Persuasion, Seventh Edition. Los Angeles: Sage.

Kelman, H.C. (2017). Violence without moral restraint: Reflections on the dehumanization of victims and victimizers. Journal of Social Issues, 29, 25-61.

Kiper, J., Gwon, Y., & Wilson, R.A. (2020). How propaganda works: Nationalism, revenge, and empathy in Serbia. Journal of Cognition and Culture, 20, 403-431.

Kteily, N., Bruneau, E., Waytz, A., & Cotterill, S. (2015). The ascent of man: Theoretical and empirical evidence for blatant dehumanization. Journal of Personality and Social Psychology, 109(5), 901.

Leader Maynard, J., & Benesch, S. (2016). Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. Genocide Studies and Prevention, 9(3), 70-95.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska directed emotional faces (KDEF). CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institute, 91(630), 2-2.

McCullough, M., Kurzban, R., & Tabak, B. (2013). Cognitive systems for revenge and forgiveness. Behavioral and Brain Sciences, 36, 1-15.

Mercier, H. (2020). Not Born Yesterday: The Science of Who We Trust and What We Believe. Princeton University Press.

Newsom, M. (2107). Football, fan violence, and identity fusion. International Review for the Sociology of Sport.

Oberschall, A. (2006). Vojislav Seselj's nationalist propaganda: Contents, techniques, aims and impacts, 1990–1994. United Nations.

Oberschall, A. (2012). Propaganda, Hate Speech and Mass Killing. In In Predrag Dojcinovic (Ed.), Propaganda, War Crimes Trials and International Law From Speakers' Corner to War Crimes (pp. 171–200). Routledge.

Petersen, M.B. (2020). The evolutionary psychology of mass mobilization: How disinformation and demagogues coordinate rather than manipulate. Current Opinion in Psychology, 35, 71-75.

Pratkanis, A., & Aronson, E. (2001). The Age of Propaganda: The Everyday Use and Abuse of Persuasion. Holt.

Purzycki, B., & Lang, M. (2019). Identity fusion, outgroup relations, and sacrifice: A cross-cultural test. Cognition, 186, 1-6.

Rai, T.S., Valdesolo, P., & Graham, J. (2017). Dehumanization increases instrumental violence, but not moral violence. PNAS, 114, 8511-8516.

Siever, L.J. (2008). Neurobiology of aggression and violence. American Journal of Psychiatry, 165, 429-442.

Slovic, P., Mertz, C.K., Markowitz, D., Quist, A., & Västfjall, D. (2020). Virtuous violence from the war room to death row. PNAS, 117(34), 20474-20482.

Smith, D. L. (2018). Less than human: Why we demean, enslave, and exterminate others. St. Martin's Press.

Straus, S. (2015). Making and Unmaking Nations: The Origins and Dynamics of Genocide in Contemporary Africa. Ithaca, NY: Cornell University Press.

Tooby, J., Cosmides, L., & Price, M. (2006). Cognitive adaptations for n-person exchange: The evolutionary roots of organizational behavior. Managerial and Decision Economics, 27, 103-129.

Valtora, R., Baldissarri, C., Andrighetto, L., & Volpato, C. (2021). Seeing others as a disease: The impact of physical (but not moral) disgust on biologization. International Review of Social Psychology, 34(1), 1-17.

Vollbgerg, M., Gaesser, B., & Cikara, M. (2021). Activating episodic simulation increases effective empathy. Cognition, 209, 104558.

West, J.D., & Bergstrom, C.T. (2021). Misinformation in and about science. PNAS. 118(15), 1-8.

Wilson, R. A. (2015). Inciting Genocide with Words. Michigan Journal of International Law, 36, 278–320. https://doi.org/10.2139/ssrn.2439325.

Wilson, R. A. (2016). Propaganda and History in International Criminal Trials. Journal of International Criminal Justice, 14(3), 519.

Wilson, R. A. (2017). Incitement on Trial: Prosecuting International Speech Crimes. Cambridge University Press.

Yanagizawa-Drott, D. (2014). Propaganda and conflict: Evidence from the Rwandan genocide. The Quarterly Journal of Economics, doi:10.1093/qje/qju020.

Zerback, T., Töpfl, F., & Knöpfle, M. (2021). The disconcerting potential of online disinformation. New Media & Society, 23(5), 1080-1098.