Journal of Social and Political Psychology
jspp.psychopen.eu | 2195-3325

Original Research Reports

Framing Hate: Moral Foundations, Party Cues, and (In)Tolerance of Offensive Speech

Grant M. Armstrong[a], Julie Wronski*[a]

[a] Department of Political Science, University of Mississippi, University, MS, USA.

Abstract

One of the most controversial elements of political tolerance concerns support for hate speech. We argue that there are two factors that can reduce tolerance for hate speech: 1) moral foundations and 2) party cues. U.S. citizens' tolerance of hate speech will be reduced when it is framed as a violation of a specific moral foundation, opposed by a political party, or when the morality violation is utilized by party elites. Using two survey experiments, we manipulated the target of hate speech (i.e. Muslims or the American flag), whether the speech violated a moral foundation (i.e. harm or loyalty), and which political party supported or opposed the hate speech in question. For flag burning, moral frames and party cues on their own reduced U.S. citizens' tolerance relative to a non-political control, while moral frames and party cues were successful in reducing tolerance of anti-Muslim speech compared to a free speech appeal. Partisans were generally responsive to cues from the in-party. We also found instances of moral repackaging, where morally incongruent appeals from the in-party reduced tolerance of flag burning among Democrats. Among Republicans, harm morality decreased tolerance of anti-Muslim speech when invoked by the in-party, but increased tolerance when used by the out-party – an indication of the power of party cues to repackage moral arguments and to trigger backlash. These results provide a better understanding of what factors can affect tolerance for hate speech, providing political leaders and social justice advocates with a roadmap to alleviate this problem.

Keywords: hate speech, political tolerance, moral foundations, party cues

Non-Technical Summary

Background
One of the foundational components of a democratic society is political tolerance. Conceptualized as the willingness to extend the rights of citizenship to a social group, political tolerance typically involves allowing members of unpopular or marginalized groups to teach in public schools, publish provocative materials, compete for positions of political power, and freely discuss their values and beliefs in public. Yet, there exists a dark side to political tolerance: support for hate speech as a form of free speech. When a society writ large tolerates hateful or offensive speech, it victimizes its most vulnerable members, catalyzes discrimination and oppression of social groups, and ultimately threatens that society's social justice principles.

Why was this study done?
We seek to identify methods of reducing tolerance towards harmful speech within the American political context, in an effort to address this threat to social justice. We identify two frames through which public tolerance for hate speech can be diminished 1) Moral Foundations Theory, and 2) Partisan Cue-Taking. Yet, the specific methods by which moral foundations and partisan cues can shape tolerance of hate speech are unclear. Can party cues and moral frames reduce tolerance for hate speech in isolation, or do they work in tandem? Is moral reframing capable of reducing tolerance for hate speech? Are cues from the in-party ubiquitously successful in reducing tolerance of hate speech, even when they

employ atypical moral arguments? Conversely, can cues from the out-party ever reduce tolerance, or do they produce backlash and greater tolerance of hate speech?

What did the researchers do and find?
We examine how moral frames and party cues affect tolerance for U.S. flag burning and anti-Muslim Muhammad cartoon drawings. Using two survey experiments conducted via Amazon's Mechanical Turk platform, we manipulated the target of hate speech (i.e. Muslims or the American flag), whether the speech violated a specific moral foundation (i.e. harm or loyalty), which party supported or opposed the hateful speech, and whether a party invoked a particular moral frame when stating their opposition. Depending on the type of harmful speech, moral frames and party cues on their own reduced U.S. citizens' tolerance of hate speech. Unsurprisingly, partisans were responsive to cues from the in-party, and occasionally exhibited decreased tolerance towards hate speech when it was presented as a moral frame violation that resonated with their party's values (i.e. Democrats and harm; Republicans and loyalty). We also found some instances of successful moral reframing. While morally congruent arguments from the out-party had no effect on reducing tolerance, morally incongruent appeals from the in-party (i.e., Democratic Party embracing loyalty) reduced tolerance of flag burning speech among Democrats. Interestingly, among Republicans, harm/care morality decreased tolerance of anti-Muslim speech when invoked by the in-party in one study, but *increased* tolerance among Republicans when used by Democrats in the other study – an indication of the power of party cues to repackage moral arguments and to trigger backlash.

What do these findings mean?
Overall, our results provide a more nuanced understanding of what factors can reduce (or intensify) tolerance for hate speech. Our findings carry important normative implications for promoting social justice by illuminating methods which political leaders and advocates can employ in order to undercut tolerance for exceptionally uncivil speech. In the current American political environment where a new type of vitriol has been normalized, party leaders can generate opposition toward certain forms of hate speech, even when that speech is targeted at relatively unpopular ethnic or religious groups (such as Muslim-Americans). Such a shift in attitudes about hate speech might lessen negative attitudes toward these groups. In this respect, party leaders can facilitate a more tolerant society by discussing hate speech as a type of morality violation, or simply and more effectively, by condemning such speech without moral appeals.

Journal of Social and Political Psychology, 2019, Vol. 7(2), 695–725, https://doi.org/10.5964/jspp.v7i2.1006
Received: 2018-07-12. Accepted: 2019-06-27. Published (VoR): 2019-09-26.
Handling Editor: Mark J. Brandt, Tilburg University, Tilburg, The Netherlands
*Corresponding author at: Department of Political Science, University of Mississippi, P.O. Box 1848, University, MS 38677, USA. E-mail: jwronski@olemiss.edu

One of the foundational components of a democratic society is political tolerance. Conceptualized as the willingness to extend the rights of citizenship to a social group (Gibson 1992, 2005), political tolerance typically involves allowing members of unpopular or marginalized groups to teach in public schools, publish provocative materials, compete for positions of political power, and freely discuss their values and beliefs in public. Yet, there exists a dark side to political tolerance: support for hate speech as a form of free speech (Brison, 2013; Lambe, 2004; Peffley, Knigge, & Hurwitz, 2001). When a society writ large tolerates hateful or offensive speech, it victimizes its most vulnerable members (Boeckmann & Liew, 2002; Leets, 2002), catalyzes discrimination and oppression of social groups (Tsesis, 2002), and ultimately threatens that society's social justice principles.

Hate speech emerged as a controversial topic of political debate following the end of World War II. According to the American Bar Association, it is defined as "speech that offends, threatens, or insults groups, based on race, color, religion, national origin, sexual orientation, disability, or other traits."[i] While some European nations have balanced freedom of expression with respect and equality for individuals by forbidding the utterance of hate speech (Cohen, 2014), the United States' Supreme Court has consistently upheld the constitutionality of speech that possesses an offensive nature or low value, even when that speech is intentional and emotionally harmful to vulnerable groups (Brison, 2013; Cohen, 2014). Though the debate on tolerating hate speech is decades old, it has recently reached a fever pitch on college campuses and in the context of "political correctness" (Altman, 1993; Delgado & Yun, 1994; Lukianoff & Haidt, 2015; Massaro, 1991). More recently, this conflict between tolerating hateful speech and fighting for social justice was embodied in the August 2017 Charlottesville, VA protests, and President Trump's subsequent comment that there were "some very fine people on both sides."[ii] Against this backdrop, we seek to identify methods of reducing tolerance towards harmful speech within the American political context, in an effort to address this threat to social justice.

Prior research demonstrates that political tolerance of hate speech is largely contingent upon the *framing* of such speech (Nelson, Clawson, & Oxley, 1997). Building upon this work, we integrate social psychology and political science theories to identify two additional frames through which public tolerance for hate speech can be diminished: 1) Moral Foundations Theory (Haidt, 2012) and 2) Partisan Cue-Taking (Zaller, 1992). Yet, the specific methods by which moral foundations and partisan cues can shape tolerance of hate speech are unclear. Can party cues and moral frames reduce tolerance for hate speech in isolation, or do they work in tandem? Is moral reframing (Feinberg & Willer, 2013) capable of reducing tolerance for hate speech? Are cues from the in-party ubiquitously successful in reducing tolerance of hate speech, even when they employ atypical moral arguments? Conversely, can cues from the out-party ever reduce tolerance, or do they produce backlash and greater tolerance of hate speech?

To answer these questions, we examine how moral frames and party cues affect tolerance for two distinct instances of hate speech: 1) U.S. flag burning, which likely should be deemed more offensive by Republicans who hold national loyalty moral values (Graham & Haidt, 2012; Haidt & Graham, 2007), and 2) anti-Muslim Muhammad cartoon drawings, which likely should be considered more offensive by Democrats who wish to protect minority groups from harm (Ditto & Koleva, 2011; Graham, Haidt, & Nosek, 2009; Koleva et al., 2012). Using two survey experiments conducted via Amazon's Mechanical Turk platform, we manipulated the target of hate speech (i.e. Muslims or the American flag), whether the speech violated a specific moral foundation (i.e. harm or loyalty), which party supported or opposed the hateful speech, and whether a party invoked a particular moral frame when stating their opposition. Depending on the type of harmful speech, moral frames and party cues on their own reduced U.S. citizens' tolerance of hate speech. In the case of flag burning, we observed this pattern relative to a non-political control, while moral frames and party cues were successful in reducing tolerance of anti-Muslim speech when pitted against a free speech appeal.

Unsurprisingly, partisans were responsive to cues from the in-party, and occasionally exhibited decreased tolerance towards hate speech when it was presented as a moral frame violation that resonated with their party's values (i.e. Democrats and harm; Republicans and loyalty). We also found some instances of successful moral reframing. While morally congruent arguments from the out-party had no effect on reducing tolerance, morally incongruent appeals from the in-party (i.e., The Democratic Party embracing loyalty) reduced tolerance of flag burning speech among Democrats. Interestingly, among Republicans, harm/care morality decreased tolerance of anti-Muslim

speech when invoked by the in-party in one study, but *increased* tolerance when used by the out-party in the other study – an indication of the power of party cues to repackage moral arguments and to trigger backlash. We did not observe such backlash among Democrats, as loyalty frames used by the Republican Party had no effect. Overall, our results provide a more nuanced understanding of what factors can reduce (or intensify) tolerance for hate speech, providing a roadmap that political leaders and social justice advocates can use to alleviate this problem.

Political Tolerance, Moral Foundations, and Party Cues

Political tolerance refers to one's degree of willingness to extend the rights of citizenship to other groups, particularly marginalized or unpopular groups (Gibson, 1992, 2005). These levels of tolerance typically are determined by public support for allowing marginalized groups to become teachers in schools, to hold positions of government authority, and, of relevance to the present study, to speak freely in public (Schwadel & Garneau, 2014; Twenge, Carter, & Campbell, 2015). In the U.S., while political tolerance has increased overall during the past half century (Schafer & Shaw, 2009), tolerance towards different groups has varied, where certain groups, like Muslims, have remained marginalized, while other groups, including the LGBT community, have experienced increased tolerance (Gibson, 2008; Schwadel & Garneau, 2014).

Much of the literature examines tolerance through the lens of inter-group conflict, and describes an inverse effect, such that, when fear of a particular group is high, political tolerance exhibited towards that group is low (Davis, 2007; Davis & Silver, 2004; Gibson, 1988; Stone, 2004; Sullivan et al., 1981). For instance, political intolerance towards Communists rose during the McCarthy era of the Cold War, and civil liberty guarantees were minimized for those of Muslim and Arabic backgrounds following the 9/11 terrorist attacks (Davis & Silver, 2004; Gibson, 2008). Thus, a perceived threat from a particular group can reduce public support for the political speech of that group, and even increase support for speech targeting that group (i.e. hate speech). Even more, hate crimes tend to increase with an induction of threat. For instance, crimes against Muslim-Americans significantly increased following the 9/11 attacks (Byers & Jones, 2007). In the case of political speech, these group-based fears are activated by public safety concerns, which, in turn, decrease tolerance. Notably, public support for a Ku Klux Klan rally decreased when this speech was framed as a public disturbance (Nelson, Clawson, & Oxley, 1997).

Building upon this research, we argue that tolerance for certain forms of political speech, and in particular hateful, offensive speech, can be similarly diminished when that speech is framed as violating key moral foundations. According to Haidt (2012), individuals employ five core moral foundations – care, fairness, loyalty, authority, and sanctity – when reacting to political stimuli and forming policy preferences (see also Ditto & Koleva, 2011; Koleva et al., 2012). When a policy is portrayed as either upholding or violating a particular moral foundation, public support for that policy can be swayed. For instance, on the issue of same-sex marriage, opposition is largely driven by purity and sanctity morals that non-traditional marriages violate, while support is garnered by viewing same-sex marriage through the lens of care and fairness towards the LGBT community (Haidt, 2012; Haidt & Graham, 2007). Moral foundation appeals can, furthermore, convince citizens to take atypical policy positions and re-evaluate preferred candidates by repackaging the topic using a different moral lens – a process defined as "moral reframing" (Day et al., 2014; Feinberg & Willer 2013, 2015; Voelkel & Feinberg, 2018). For example, conservatives can hold more liberal, pro-environment attitudes when policies are couched in arguments about loyalty, authority and purity (Feinberg & Willer, 2013; Kidwell, Farmer, & Hardesty, 2013; Wolsko, Ariceaga, &

Seiden, 2016), while framing military spending and English as the United States' official language in terms of fairness makes liberals more likely to take conservative positions on these issues (Feinberg & Willer, 2015).

In similar fashion, tolerance of hate speech should depend on which moral value is being emphasized, with the expectation that depicting offensive speech as a morality violation will reduce support for that type of speech. While the effectiveness of any particular moral foundation in moving political attitudes varies across ideologues (per Day et al., 2014; Graham, Haidt, & Nosek, 2009), with some foundations more potent for liberals and others for conservatives, moral arguments can nonetheless serve as a method of reducing *aggregate* tolerance for hate speech. This should especially be the case when moral frames are employed as a foil against free speech appeals, building upon Nelson et al. (1997).

In addition to moral foundation appeals, party heuristics (Campbell et al., 1960) should be able to shape opinions regarding tolerance of hate speech through the process of elite cue-taking (Lupia, 1994). These cues signal to partisans where they should stand on a variety of issues, including social welfare programs, immigration reform, and energy policies (Druckman, Peterson, & Slothuus, 2013; Evans & Pickup, 2010; Gerber & Huber, 2010). These cues can influence citizens' evaluations of political events and candidates (Bartels, 2002; Gerber, Huber, & Washington, 2010; Goren, 2005), and move attitudes on divisive issues such as abortion (Achen & Bartels, 2016) or non-political judgments (Iyengar & Westwood, 2015). Thus, citizens form their political attitudes through the judgments of those individuals with whom they share a partisan identity, even to the extent that the positions of their party can override evaluations of substantive policy content (Cohen, 2003). Moreover, party cues provide the contextual information necessary for citizens to construct a relationship between a message and their dispositional beliefs (Zaller, 1992). In the case of hate speech, party cues should alert citizens as to which moral foundation is being violated by the offensive speech, and help them connect their moral values with their tolerance of that speech. Put simply, when a political party opposes a form of hate speech under the premise that the speech undermines a core moral value of the citizenry, this endorsement legitimizes the moral violation frame, and enhances its ability to reduce tolerance of hate speech. Since moral frames and party cues should have the ability to reduce tolerance towards hate speech in isolation, as described above, we expect:

1. Moral Frame Alone Hypothesis
2. Party Cues Alone Hypothesis

We also should observe nuanced variation when moral frames and party cues are combined. Specifically, individuals should react to moral foundation violation frames and party cues on the basis of 1) the type of harmful speech in question, 2) which political party opposes that speech, and 3) individuals own partisan affiliations. We can, furthermore, tease apart the efficacy of moral violation frames when they are endorsed by the in-party or out-party, and when they utilize moral reframing. This leads us to propose four pathways by which the combination of moral and party cues can influence political tolerance of hate speech:

1. Congruent Moral Frame from In-Party Hypothesis
2. Congruent Moral Frame from Out-Party Hypothesis
3. Incongruent Moral Frame from In-Party Hypothesis
4. Backlash Hypothesis

We examine two contemporary forms of hate speech that allow us to explore these nuances: United States flag burning, and anti-Muslim Muhammad cartoon drawings.

Tolerance for Flag Burning and Anti-Muslim Speech

The United States' constitution provides generous protections on free speech. Moreover, court rulings consistent therewith seem to be consonant with public opinion. A 2016 Pew Research poll shows that Americans are more tolerant of free speech than any other nation (Wike, 2016). Generally, Americans believe that people should be able to say what they want, even if those utterances may be offensive or harmful to members of certain social groups (i.e. hate speech). Bolstering this acceptance of hate speech, 67% of Americans believe "that people should be allowed to make public statements that are offensive to minority groups," while 77% of Americans "support the right of others to make statements that are offensive to their own religious beliefs" (Pew Research Center, 2017). Yet, there exists some heterogeneity in political tolerance for hate speech. Republicans and Democrats tolerate different groups, wherein Republicans demonstrate greater intolerance toward pro-gay and pro-choice activists, while Democrats exhibit greater intolerance toward pro-life and anti-gay activists (Crawford & Pilanski, 2014).

Because Republicans and Democrats display varying levels of tolerance for certain social groups (Crawford & Pilanski, 2014) and rely upon different moral frameworks (Haidt, 2012), we can locate examples of offensive speech that Republicans should be more likely to tolerate than Democrats, and vice versa. Two examples are flag burning and anti-Muslim speech. For Republicans, burning the American flag should be deemed offensive since they tend to be more patriotic and hold stronger national loyalty moral values (Graham & Haidt, 2012; Haidt & Graham, 2007). Anti-Muslim speech, in contrast, should be considered hateful by Democrats who wish to protect minority groups from harm (Ditto & Koleva, 2011; Graham, Haidt, & Nosek 2009; Koleva et al., 2012). While both forms of speech are constitutionally protected, tolerance for each should differ across Democrats and Republicans.

Flag burning was ruled constitutionally protected speech in the 1989 Supreme Court case of *Texas v Johnson*. Nevertheless, public support for such speech has remained low. A 2006 Gallup poll showed that a majority of Americans would favor a constitutional amendment banning the desecration and burning of the flag (Carroll, 2006). There is also evidence that Republicans are more accepting of banning flag burning than Democrats. Depending on the framing of the issue, Republicans were 12%-17% more likely than Democrats to favor banning this type of speech (Carroll, 2006). Academic research complements these findings (Graham & Haidt, 2012; Haidt, 2012), concluding that Republicans, driven by loyalty and authoritarianism, are less likely than Democrats to be tolerant of flag burning.

Moral Frame Alone Hypothesis and Party Cues Alone Hypothesis

Inextricably tied to tolerance of flag burning is the moral foundation of loyalty/betrayal, defined by strong in-group attachments with devotion to a group, institution, or state (Haidt, 2012). While conservatives, and by extension via partisan sorting (Levendusky, 2009; Mason, 2015) Republicans,[iii] value all five foundations almost equally, they tend to rely upon the socially "binding" foundations of loyalty, authority, and purity when assessing political and social issues (Graham, Haidt, & Nosek, 2009). In contrast, liberals and Democrats favor the "individualizing" foundations of care and fairness, and place little weight on loyalty values when forming political attitudes. As such, messages framing flag burning as a violation of the loyalty moral foundation (*Moral Frame Alone Hypothesis*) and detailing Republican Party opposition to it (*Party Cues Alone Hypothesis*) should reduce tolerance among Repub-

licans, thus entrenching their existing disapproval of this form of offensive speech in the presence of a morally congruent appeal (per Day et al., 2014).

While U.S. public opinion toward Muslims has been cold since 9/11, Muslims recently have been viewed more favorably by the American public (Pew Research Center, 2017). Their favorability rating has increased since 2014 to about 48 on a 100-point feeling thermometer scale (Pew Research Center, 2017). Despite the recent increase, it is still the lowest rating for any religious group, falling behind even atheists. Further, Republicans are more likely than Democrats, according to these Pew polls, to believe that the Islamic religion is more likely than others to encourage violence among its believers, and to view Muslims as more extreme and less American. In turn, stemming from their emphasis on care/harm morality and concern to shield vulnerable groups in society, Democrats should consider anti-Muslim speech offensive and exhibit less tolerance towards it than Republicans (Ditto & Koleva, 2011; Graham, Haidt, & Nosek, 2009; Koleva et al., 2012). As a result, message frames that employ "individualizing" harm morality (*Moral Frame Alone Hypothesis*) and Democratic Party disapproval (*Party Cues Alone Hypothesis*) in opposition to anti-Muslim speech should reduce already lower tolerance for this speech among Democrats. Given past research on the inefficacy of incongruent moral appeals (Day et al., 2014; Feinberg & Willer 2013, 2015; Kidwell et al., 2013; Wolsko, Ariceaga, & Seiden, 2016), loyalty and harm appeals on their own should have a weaker impact on reducing tolerance among Democrats and Republicans, respectively.[iv]

Congruent Moral Frame and In-Party Cues Hypothesis

Individuals tend to show particular deference to authority figures from their in-group (Frimer, Gaucher, & Schaefer, 2014), suggesting that they should be more responsive to frames and cues coming from their own party. With the Republican Party's reverence of national symbols and patriotism, and the Democratic Party's commitment to diversity and inclusion of Muslim-Americans, party cues should be the most effective in reducing tolerance for harmful speech when addressing an in-group owned issue. This means that partisans should reduce their tolerance for hate speech in lock-step with their party's opposition, particularly when the Republican Party opposes flag burning and the Democratic Party opposes anti-Muslim speech, as these acts represent moral violations to in-group members (per Voelkel & Brandt, 2019). Further, Republicans should respond to Republican Party cues opposing flag burning based on a loyalty frame, while Democrats should respond to Democrat Party cues opposing anti-Muslim speech based on a harm morality frame, because these cues are, in essence, speaking the group's moral language (per Lakoff, 2010).

Congruent Moral Frame From Out-Party Hypothesis and Incongruent Moral Frame From In-Party Hypothesis

Party cues may also act as a repackaging agent for moral frames in order to reduce tolerance for offensive speech among partisans who would otherwise be supportive of protecting flag burning or anti-Muslim speech. Such moral reframing can decrease tolerance for hate speech by either 1) targeting members of the out-party and presenting them with a moral argument that resonates with their ideological values (in line with Day et al., 2014; Feinberg & Willer, 2013, 2015; Kidwell, Farmer, & Hardesty, 2013; Wolsko, Ariceaga, & Seiden, 2016), or 2) targeting members of the in-party and advancing a moral argument that is ideologically incongruent with their members' beliefs (a novel advancement of the moral reframing theoretical framework). In the former (*Congruent Moral Frame from Out-Party Hypothesis*), Democrats should decrease their tolerance of anti-Muslim speech when the Republican Party frames it as a harm morality violation, and Republicans should be less tolerant of flag burning when the Democratic Party decries it as a betrayal of the nation. In the latter (*Incongruent Moral Frame from In-Party Hy-*

pothesis), Democrats should decrease their tolerance of flag burning when the Democratic Party frames it as a loyalty morality violation, and Republicans should express intolerance towards anti-Muslim speech when the Republican Party claims that it harms Muslim-Americans.

Backlash Effect Hypothesis

Yet, recent research on "negative partisanship" (Abramowitz & Webster, 2016) has shown that individuals can react to the positions of out-party elites, with disdain towards the out-group driving backlash. It is therefore also possible that individuals may respond to moral foundation violation frames and party cues by *increasing* their tolerance of hateful speech that the out-party opposes. This likely will be pronounced when the instance of hate speech opposed by out-party generates antipathy, such as when Republicans see the Democratic Party opposing anti-Muslim speech and Democrats see the Republican Party opposing flag burning. Thus, when the Democratic Party uses harm morality in an effort to oppose anti-Muslim speech, Republicans might actually support such speech so as to distance themselves from Democrats. Democrats might respond similarly to the Republican Party's use of loyalty moral frames to increase their tolerance and support of flag burning.

As party cues vary in their persuasiveness across issue salience (Ciuk & Yost, 2016), the current research affords an ideal context for ascertaining the role of parties in moral reframing, and a general understanding of how partisan and moral appeals can shape tolerance for hate speech.

Method

To examine how moral foundations and party cues can reduce political tolerance towards offensive speech, we employed two online survey experiments that manipulated information about a hypothetical instance of hate speech in a mock news story format.[v] The first (Study 1) was administered in June 2017, while the second (Study 2) was conducted in November 2018, with both utilizing participants from Amazon's Mechanical Turk (Mturk) platform and administered through Qualtrics' online survey software. By conducting our studies in different time periods with two different samples, we ameliorated the potential pre-treatment problems inherent in framing experiments (as noted by Druckman & Leeper, 2012). The details of the design and specific measures vary slightly across the two studies, but both include the experimental treatment, a measure of support for protecting various types of hate speech, and measures assessing participants' moral foundations, political preferences, and demographic background.

Experimental Treatment

Each of our experimental studies manipulated information about two hypothetical instances of hate speech that were based on actual events: 1) the burning of the United States flag, and 2) a Muhammad cartoon-drawing contest. The story about U.S. flag burning was based on a demonstration in Brooklyn, New York in 2015, where protesters who burned the flag claimed police brutality and systematic racism within the justice system.[vi] The anti-Muslim example was based on an event in Garland, Texas where attendees, including certain prominent right-wing political figures, drew different depictions of Islam's prophet Muhammad in an overt affront to members of the Muslim community.[vii]

Using these fictionalized news stories detailing an instance of hate speech, we implemented a between subjects factorial design, such that each respondent was randomly assigned to read a single news story.[viii] Study 1 employed a 2x4 factorial design with 8 possible conditions, whereas Study 2 used a 2x5 factorial plus control design such that participants were randomly assigned to one of 11 conditions (see Figure A1 in the Supplementary Materials).[ix] The first manipulated factor in both studies was the target of hateful speech, where respondents read a story about either 1) a demonstration where the American flag was burned, or 2) an event featuring a Muhammad cartoon contest (hereafter referred to as the "flag burning" and "Anti-Muslim" conditions, respectively). Study 2 also included a pure control condition that described a non-controversial, non-political event, in order to establish a baseline measure of tolerance for flag burning and anti-Muslim speech.[x] Each story was paired with a relevant picture.

The second manipulated factor was the moral frame and party's response to the hate speech. In Study 1 respondents were informed that the event in question was either an example of 1) free speech, 2) speech that violates a moral foundation (loyalty in the case of flag burning, and harm for anti-Muslim speech), 3) speech that the Republican Party deems a moral foundation violation and opposes, but the Democratic Party does not oppose, or 4) speech that the Democratic Party deems a moral foundation violation and opposes, but the Republican Party does not oppose.

For the free speech and moral foundation conditions in Study 1, respondents were presented the following text, after being introduced to the flag burning or anti-Muslim event:

> Flag Burning condition: "Burning the U.S. flag, as shown in the picture, is [an expression of free speech that cannot be taken away by the government/ destruction of a national symbol and it betrays patriotic Americans]."
>
> Anti-Muslim condition: "Drawing cartoons of the prophet Muhammad, as shown in the picture, is [an expression of free speech that cannot be taken away by the government/offensive and can be emotionally harmful and intimidating to Muslim-Americans]."

In the party cues conditions, respondents were presented with information regarding whether each party supported or opposed that form of speech:

> Flag Burning condition: "The [Democratic/Republican] Party criticized the event, saying that burning the U.S. flag is destruction of a national symbol and it betrays patriotic Americans. The [Republican/Democratic] Party, however, disagreed and stressed that burning the U.S. flag, as shown in the picture, is an expression of free speech that cannot be taken away by the government."
>
> Anti-Muslim condition: "The [Democratic/Republican] Party criticized the event, saying that drawing cartoons of the prophet Muhammad, as shown in the picture, is offensive and can be emotionally harmful and intimidating to Muslim-Americans. The [Republican/Democrat] Party, however, disagreed and stressed that drawing cartoons of the prophet Muhammad, as shown in the picture, is an expression of free speech that cannot be taken away by the government."

These treatments, however, provided information from both parties simultaneously. As such, Republican opposition to hate speech was stated in tandem with Democratic tolerance of hate speech, and vice versa. The use of both parties in the treatments, while increasing the overall amount of party cue information to the respondent, inhibits our ability to discern whether the effects stem from the endorsement by the respondent's in-party or by the opposition of the respondent's out-party. Further, party opposition was always coupled with a moral foundation violation.

While the coupling of party cues with a moral foundation violation frame should produce the greatest decrease in tolerance for hate speech (the additive effect of the *Moral Frame Alone Hypothesis* and the *Party Cues Alone Hypothesis*), the experimental design in Study 1 prevents us from estimating the effect of party cues independent of the moral frame.

In order to disentangle the effects between in-party and out-party cues, and estimate effects of party cues independent of any moral foundation framing, Study 2 manipulated the discussion of the offensive event in question as an example of 1) speech that violates a moral foundation (loyalty in the case of flag burning, and harm for anti-Muslim speech), 2) speech that the Democratic Party opposes, 3) speech that the Republican Party opposes, 4) speech that the Republican Party opposes on the basis of it being a moral foundation violation, and 5) speech that the Democratic Party opposes on the basis of it being a moral foundation violation.[xi] The moral foundation violation frame used the same text as Study 1. The Democratic and Republican Party opposition conditions read as follows:

> Flag Burning condition: "Recently, in a large U.S. city, there was a U.S. flag-burning demonstration led by a group of protesters. The [Democratic/Republican] Party believes that burning the U.S. flag, as shown in the picture, should *not* be protected by the First Amendment."
>
> Anti-Muslim condition: "Recently, in a U.S. city, there was an event where cartoons of the prophet Muhammad were drawn by a group of protesters. The [Democratic/Republican] Party believes that drawing cartoons of the prophet Muhammad, as shown in the picture, should *not* be protected by the First Amendment."

Finally, the Democratic and Republican Party opposition using a moral foundation framework conditions were displayed in the following manner:

> Flag Burning condition: "The [Democratic/Republican] Party criticized the event, saying that burning the U.S. flag is destruction of a national symbol and it betrays patriotic Americans."
>
> Anti-Muslim condition: "The [Democratic/Republican] Party criticized the event, saying that drawing cartoons of the prophet Muhammad, as shown in the picture, is offensive and can be emotionally harmful and intimidating to Muslim-Americans."

Mturk Samples

Study 1 was conducted in June 2017, while Study 2 was fielded in November 2018, both using Mturk. Of the 1461 respondents who started Study 1, 1410 received at least the experimental treatment and answered the associated protected free speech opinion item. Our realized sample excludes six respondents who noted in the comments that the party cues were incorrect, or were in the free speech condition and hostile towards the study (e.g. made disparaging remarks about the institution conducting the research). Similarly, in Study 2, 2226 respondents began the study, but only 1870 completed the free speech opinion item and were not identified as fraudulent respondents (per Kennedy, Clifford, Burleigh, Waggoner, & Jewell, 2018). As such, we dropped 188 respondents who were suspect foreign IP locations, VPS users, or came from the same remote geolocation. By omitting the reversed party cues conditions, and any mention of a political party supporting hate speech, Study 2 did not include any respondents who noticed the "wrong" party cues or were otherwise hostile towards the research.[xii]

Across both surveys, respondents took approximately twelve minutes, and no respondent took less than 90 seconds, to complete the entire survey. Like most Mturk samples, ours were more white (78% in Study 1 and 80% in Study 2), educated (63% in Study 1, and 62% in Study 2 had at least a college degree), and secular (47% in Study 1 and 44% in Study 2 never attend religious services, and 36% in Study 1 and 35% in Study 2 described themselves as non-religious), while gender (54% female in Study 1, 58% female in Study 2) was more representative of the general population. Further, as is typical with Mturk samples, respondents were more Democratic (57% Democratic, 12% Independent, and 31% Republican in Study 1, 56% Democratic, 13% Independent, 31% Republican in Study 2), and Liberal (47% Liberal, 26% Moderate, and 27% Conservative in Study 1, 47% Liberal, 25% Moderate, and 28% Conservative in Study 2) than the general population. Each respondent in both surveys was paid fifty cents for their participation, commensurate with Mturk rates for 10-15-minute studies. While less representative than more traditional internet surveys, Mturk is more representative than "in person convenience samples" (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011). More importantly, experimental treatment effects of Mturk samples are generalizable (Coppock, 2019; Levay, Freese, & Druckman, 2016; Mullinix, Leeper, Druckman, & Freese, 2015).

Procedures

Before entering the survey, respondents in both studies answered three screener questions about their gender, race, and religious denomination,[xiii] and consented to participation in the study. Immediately following the screening, the experimental mock news story treatment was presented to respondents in order to prevent any priming effects. Each treatment included a question asking respondents to state their support for protecting the type of hate speech featured in the story.[xiv] Respondents then answered questions about their party identification, ideological and political preferences, affect towards various social groups, moral foundations, and demographics. In Study 2, respondents also answered questions regarding how much they found flag burning and anti-Muslim speech to be generally considered hateful speech in American society, and personally offensive to them, immediately following the tolerance item.

Political Tolerance of Hate Speech

Our main dependent variable is political tolerance for hate speech, namely, the type of hate speech featured in the experimental treatment. Tolerance was determined by asking respondents, immediately following the mock news story: "In your opinion, do you agree or disagree that the First Amendment should protect this kind of speech?"[xv] Respondents reported their level of support on a 6-point Likert scale that ranged from "strongly agree" to "strongly disagree," which was recoded 0-1 such that higher values indicate greater support for protecting, or tolerating, that form of harmful speech (Study 1: $M = 0.68$, $SD = 0.32$; Study 2: $M = 0.63$, $SD = 0.32$).

Offensiveness of Hate Speech

It is possible that respondents did not deem the two instances of speech employed in our treatments (burning the U.S. flag and drawing anti-Muslim cartoons) offensive, or even considered them as forms of hate speech, in which case our treatments are unable to provide pathways for diminishing tolerance of hate speech. In order to ensure that respondents actually considered these actions as constituting hate speech and to understand whether this determination was made based on personal feelings or perceptions of public sentiment, Study 2 included two items immediately following the experimental treatment and associated measure of tolerance. The first item asked respondents to rate the extent to which they considered the type of speech in question offensive to Americans as a whole, on a 4-point scale coded from 0="definitely not offensive" to 1="definitely offensive" (Flag burning: $M =$

0.77, *SD* = 0.30; Anti-Muslim: *M* = 0.73, *SD* = 0.29). The second asked them to rate how personally offended they were by that type of speech: not offended at all, somewhat offended, or extremely offended, recoded 0-1 such that higher values represented greater offense (Flag burning: *M* = 0.55, *SD* = 0.42; Anti-Muslim: *M* = 0.42, *SD* = 0.39).[xvi] In the non-political control condition, 86% reported that burning the U.S. flag was either definitely or probably offensive to Americans as a whole, with 74% stating it was either extremely or somewhat personally offensive to them. Likewise, 80% of respondents in the control condition thought drawing anti-Muslim cartoons was definitely or probably offensive to American society, and 65% found it somewhat or extremely personally offensive. Taken together, respondents overwhelmingly considered these two instances forms of hateful speech.

Party Identity and Control Variables

In both studies, we measured respondents' partisan identities using the traditional 7-point self-placement item, scaled from 0-1, where 0=Strong Democrat and 1=Strong Republican (Study 1: *M* = 0.41, *SD* = 0.34; Study 2: *M* = 0.40, *SD* = 0.35). Based on this scale, we created two variables for partisan identifiers only: 1) a party identity dummy, coded 0=Democrats and 1=Republicans, which combines strong partisans with leaners, and 2) partisan strength operationalized as the folded 7-point scale (Study 1: *M* = 0.56, *SD* = 0.40; Study 2: *M* = 0.63, *SD* = 0.34). We also assessed individual-level political preferences, moral foundations, and demographics including ideology, political interest, race, gender, education, and church attendance (see Supplementary Materials, Table A1, for details).

Analytical Strategy

Since our aim is to decipher which combinations of moral frames and partisan cues successfully decrease tolerance of flag burning or anti-Muslim speech, we analyze our experimental treatment effects for each type of speech separately using STATA 15. In Study 1, which has four treatment categories per type of speech, we estimate a series of Ordinary Least Squares (OLS) models that each regress tolerance for speech on the treatment factorial variable, alternating which treatment serves as the excluded category (Tables 1 and 3). In Study 2, which contains six treatment categories per type of speech, we estimate a 2x3 (Moral Frame X Party Cues) factorial design using OLS, where Moral Frame has factor levels 1) No Moral Frame provided, and 2) Moral Frame provided; and Party Cues has factor levels 1) No Party Cues provided, 2) Democratic Party Cues provided, and 3) Republican Party Cues provided (Table 2). To estimate our treatment effects conditional upon respondents' partisan identification, we use a 4x2 (Treatment X Party Identity dummy) factorial design in Study 1, and a 2x3x2 (Moral Frame X Party Cues X Party Identity dummy) factorial design in Study 2, where tolerance of speech is regressed (again using OLS) onto each factor and their interactions. We then use the "margins" package for post-estimation, generating predicted values of tolerance across treatments (Figures 2 and 4) and marginal treatment effects for Republican and Democratic respondents, with associated significance tests (Figures 3 and 5, and discussed in text as appropriate).

Results

The mean level of support for First Amendment freedom of speech protections – our measure of political tolerance – for U.S. flag burning (top panels) and anti-Muslim speech (bottom panels) across all experimental conditions and both studies is displayed in Figure 1 (Study 1 on left panels, Study 2 on right panels).

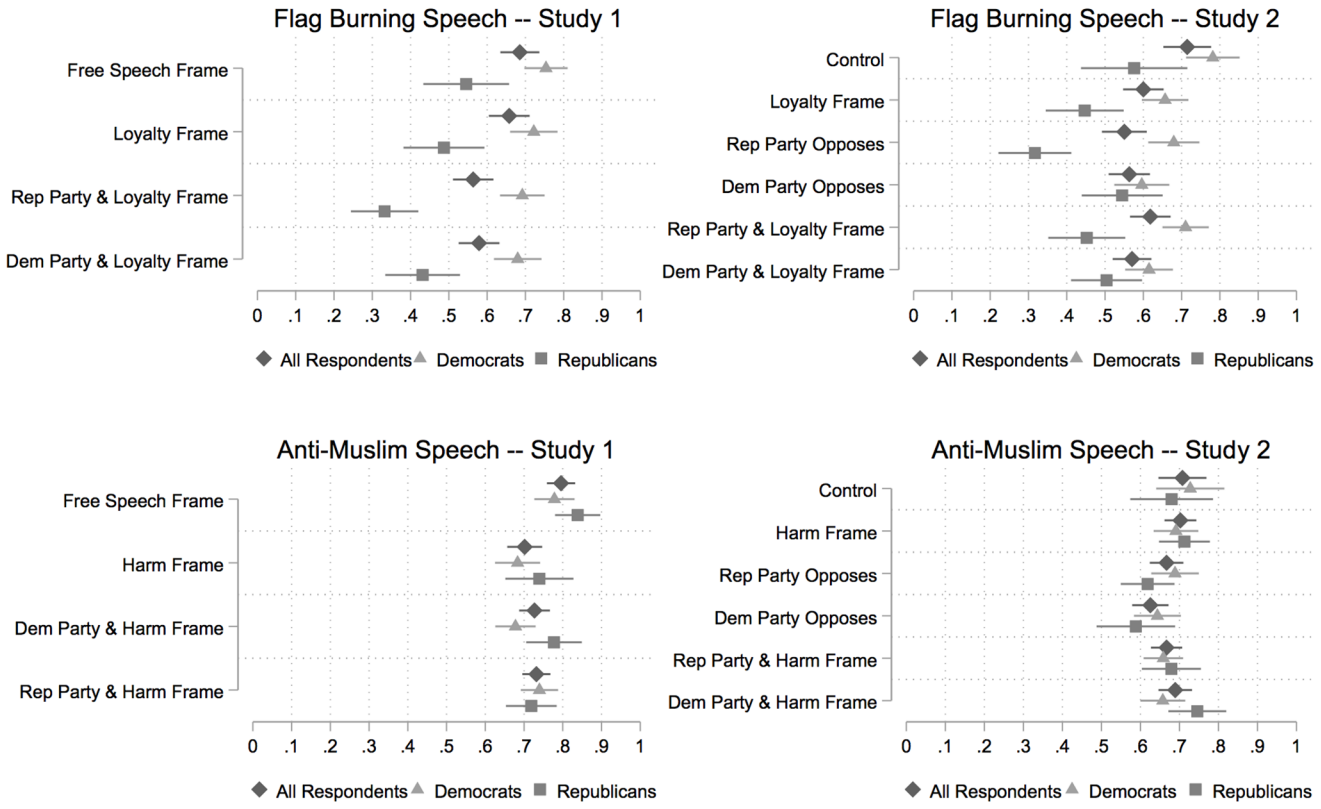

Figure 1. Mean tolerance of hateful speech by condition.

Note. In Study 2, the non-political control condition is split into two, where respondents who answered the flag burning item first comprise the flag burning control condition, and respondents who answered the anti-Muslim item first become the anti-Muslim control condition.

First, consistent with contemporary public opinion on these issues, overall tolerance of anti-Muslim speech is higher than that of flag burning when examining all respondents and collapsing across all conditions (Study 1: *M* anti-Muslim speech = 0.74, *M* flag burning = 0.62, *M* difference = 0.12 on the 0-1 scale, $t(1408) = 7.06$, $p < .01$; Study 2: *M* anti-Muslim speech = 0.67, *M* flag burning = 0.60, *M* difference = 0.07 on the 0-1 scale, $t(2035) = 4.90$, $p < .01$). Yet, within each type of speech, we can visually inspect the conditions under which moral foundation frames and party cues lower tolerance, viewing generally that tolerance of flag burning is more varied than anti-Muslim speech. Further, a broad pattern of decreased tolerance emerges for all respondents (diamond markers), Democrats (triangle markers), and Republicans (square markers). As such, we treat the free speech condition as our baseline level of tolerance in Study 1, and the non-political control condition as the baseline in Study 2.

There is, however, a noteworthy deviation from this pattern. Average tolerance of anti-Muslim speech in Study 2's non-political control condition reflects a spillover ordering effect, such that tolerance of anti-Muslim speech was marginally *lower* when asked immediately after the flag burning item ($M = 0.622$ v. $M = 0.708$, $t(165) = 1.805$, $p = .073$ on a two-tailed test). As a result, we split Study 2's control condition into two, where respondents who answered the flag burning item first comprise the flag burning control condition ($M = 0.715$, see Figure 1 top right panel) and respondents who answered the anti-Muslim item first become the anti-Muslim control condition ($M = .708$, see Figure 1 bottom right panel). Even with this revised baseline control, we still do not clearly discern decreased anti-Muslim speech tolerance stemming from our treatments, despite finding equivalent levels of tolerance

among respondents in the harm frame condition in both studies (Study 1 *M* = 0.702, 95% CI [0.66, 0.75]; Study 2 *M* = 0.702, 95% CI [0.66, 0.74]). Yet, those in the anti-Muslim speech control condition revealed less tolerance than those in Study 1's free speech condition by approximately 9% of the 0-1 scale.[xvii] On the surface, decreasing tolerance for anti-Muslim speech seems best accomplished through priming non-political considerations or other instances of offensive political speech. Nonetheless, we explore the efficacy of our various moral foundation and party cue appeals, particularly as they affect partisans, in the context of decreasing tolerance for flag burning and anti-Muslim speech.

Decreasing Flag Burning Tolerance

We begin examining support for protecting flag burning speech aggregating across all respondents. Though such aggregation may wash out party cue treatment effects, especially in Study 1 where opposing stances from both parties are presented simultaneously, these analyses still provide important information regarding the effectiveness of moral foundation frames and party positioning in moving mass public opinion. Foremost, we can discover whether moral frames can decrease tolerance for harmful speech, *en masse*, regardless of citizens' party affiliations (or lack thereof), ideological values, or personal morality positions (per the *Moral Frame Alone Hypothesis*). Next, we can assess whether opposition from *any* political party is sufficient to decrease tolerance, as evidence of more generalized political elite cueing (per Zaller, 1992, and the *Party Cues Alone Hypothesis*). And finally, we can reveal if a particular political party's opposition exerts greater influence over tolerance, thus providing evidence of party ownership over certain forms of harmful speech.

Turning to Study 1 (Table 1), relative to the baseline free speech frame, framing flag burning as a violation of loyalty morals has no effect on tolerance for this speech (*B* = -0.028, n.s., Column 1).[xviii] However, when *either* the Republican or the Democratic party endorses the loyalty moral violation in opposition of flag burning, this additional party cue significantly decreases political tolerance of the speech (*B* = -0.079, $p < .05$ when Democratic Party endorses, *B* = -0.094, $p < .05$ when Republican Party endorses, see Column 2) compared to the level of tolerance when the loyalty frame alone is presented. This combination leads to a significant reduction in tolerance of flag burning speech relative to the baseline free speech frame of about 10% of the 0-1 scale (*B* = -0.107, $p < .05$ when Democratic Party endorses, *B* = -0.122, $p < .05$ when Republican Party endorses, see Column 1).

In Study 2 (Figure 2, and Table 2, Columns 1 through 3), use of the loyalty moral value frame alone *did* reduce tolerance of flag burning compared to the control condition (*B* = -0.115, $p < .01$), in contrast to Study 1. Respondents also exhibited lower levels of flag burning tolerance when either the Democratic *or* Republican Party opposed flag burning when compared to the control condition, by about 15% and 16% of the 0-1 scale, respectively (*B* = -0.151, $p < .01$ when Democratic Party opposes, *B* = -0.164, $p < .01$ when Republican Party opposes). While we also see significantly less tolerance when the party's opposition to flag burning is coupled with the loyalty moral appeal relative to the control in Figure 2, these effects are smaller than when the loyalty frame or party cues were utilized independently of one another (given the significant interaction coefficients of *B* = 0.122, $p < .05$ for the Democratic Party, and *B* = 0.182, $p < .01$ for Republican Party in Table 2, Column 1). Further, compared to the loyalty frame alone or the party opposition alone treatments, the pairing of party cues with moral frames had no effect on flag burning tolerance, suggesting that the negative effects found for the combined party cue/moral frame treatments in Study 1 are driven almost exclusively by the additional partisan information. Across studies, these aggregate findings provide mixed support for *Moral Frame Alone Hypothesis* and consistent support for the *Party Cues Alone*

Hypothesis, and appear consistent with Zaller (1992), in that a frame is generally more persuasive than no frame, and elite appeals are more powerful than no appeals.

Table 1

Experimental Treatment Effects for Flag Burning (Study 1)

	All Respondents[a]						Partisans Only[b]					
	1		2		3		4		5		6	
Effect	***B***	***SE***	***B***	***SE***	***B***	***SE***	***B***	***SE***	***B***	***SE***	***B***	***SE***
Free Speech Frame	–	–	0.028	0.038	0.122**	0.037	–	–	0.032	0.043	0.062	0.041
Loyalty Frame	-0.028	0.038	–	–	0.094*	0.038	-0.032	0.043	–	–	0.030	0.043
Dem Party using Loyalty Frame	-0.107**	0.038	-0.079*	0.038	0.015	0.038	-0.074‡	0.043	-0.042	0.045	-0.012	0.043
Rep Party using Loyalty Frame	-0.122**	0.037	-0.094*	0.038	–	–	-0.062	0.041	-0.030	0.043	–	–
Party ID							-0.208**	0.063	-0.235**	0.062	-0.360**	0.054
Free Speech Frame X Party ID							–	–	0.026	0.089	0.151‡	0.083
Loyalty Frame X Party ID							-0.026	0.089	–	–	0.125	0.082
Dem Party Loyalty Frame X Party ID							-0.040	0.086	-0.014	0.085	0.111	0.079
Rep Party Loyalty Frame X Party ID							-0.151‡	0.083	-0.125	0.082	–	–
Constant	0.686**	0.026	0.658**	0.027	0.564**	0.027	0.754**	0.029	0.722**	0.032	0.692**	0.030

Note. OLS regression estimates with robust standard errors. In Columns 1 and 4, the "Free Speech Frame" is the excluded treatment category, in Columns 2 and 5 the "Loyalty Frame" is the excluded treatment category, and in Columns 3 and 6 the "Republican Party using Loyalty Frame" is the excluded treatment category. Party ID is a dichotomous variable coded 0=Democrats, and 1=Republicans. All analyses are confined to the subset of respondents who received one of the flag burning treatments (i.e. respondents viewing an anti-Muslim speech treatment are excluded from the above models). These same models with demographic controls are provided in the Supplementary Materials, Table A3. Table A4 (Supplementary Materials) estimates the models shown in Columns 1 and 4, including an interaction between each treatment and respondents' loyalty morality to check for any heterogeneous treatment effects.

[a]N = 704, R^2 = 0.020. [b]N = 623, R^2 = 0.157.

‡p < .10. *p < .05. **p < .01.

We next turn to how Republicans and Democrats responded to our flag burning treatments. Democrats are, on average, more tolerant of flag burning than Republicans in Study 1's baseline free speech condition (Democrats M = 0.754 v. Republicans M = 0.545, $t(146) = 3.62$, p < .01, see Figure 1, top left panel) and in Study 2's control condition (Democrats M = 0.781 v. Republicans M = 0.576, $t(77) = 2.9$, p < .05, see Figure 1, top right panel), comporting with partisan divides in public opinion on the matter. Yet, both Democratic and Republican respondents decreased their tolerance towards flag burning, with their patterns of reduced tolerance conditional upon how moral foundation values, party cues, and their combination were used to frame the issue.

Similar to all respondents, in Study 1 (see Figure 3, and Table 1, Columns 4 through 6) the loyalty moral foundation frame alone had no effect on reducing tolerance of flag burning relative to the free speech baseline condition among partisans (B = -0.032 for Democrats, B = -0.058 for Republicans, both n.s., see Figure 3, top left panel, and Table 1, Column 4). However, in Study 2 (see Figure 2, and Table 2, Column 2), Democrats (but not Republicans[xix]) reduced their average tolerance in the loyalty frame condition relative to the pure control (B = -0.125, p < .01). This inconsistent support for the *Moral Frame Alone Hypothesis* is probably due to differing baselines, where Study 1 used the free speech frame instead of a true control as the point of comparison. By including the pure control in Study 2, it is clear that moral frames alone can reduce tolerance for flag burning.

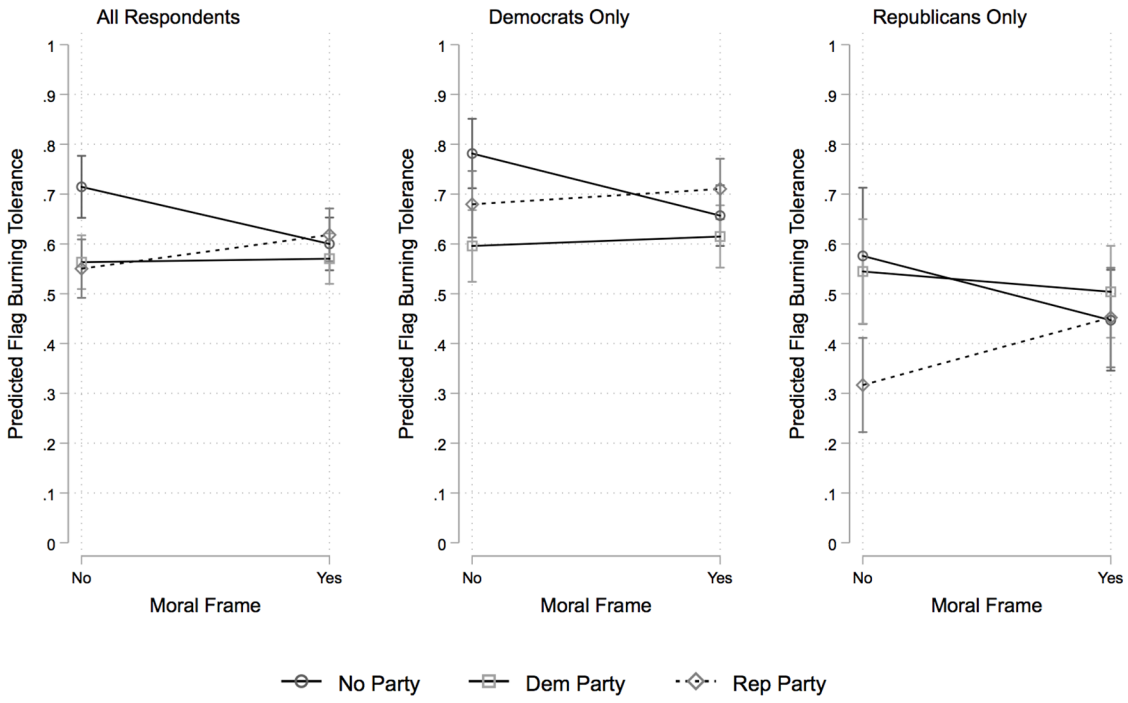

Figure 2. Experimental treatment effects on flag burning tolerance (Study 2).

Note. Open dots represent the estimated predicted values for tolerance of speech in each of the flag burning speech conditions plus control condition, with 95% confidence interval bars. All estimates are derived from OLS regression models that employ a factorial design (see Table 2). Estimates for "All Respondents" are derived from the 2x3 factorial design without demographic controls estimated in Table 2, Column 1. Estimates for "Democrats Only" and "Republicans Only" are derived from the 2x3x2 factorial design without demographic controls estimated in Table 2, Column 2.

When examining tolerance towards flag burning in Study 2 we find support for the *Party Cues Alone Hypothesis* among Democrats, while among Republicans party cues alone decrease tolerance only when coming from the in-party (see Figure 2, middle and right panels). Democrats receiving information that either the Democratic or Republican Party opposed flag burning are on average 10% to 19% less tolerant of this speech than Democrats in the control group (B = -0.186, p < .01 Democratic Party cues; B = -0.102, p < .05 Republican Party cues). In contrast, when the Republican Party opposes flag burning, tolerance among Republican respondents typically decreases by about 25% of the 0-1 scale (B = -0.259, p < .01), compared to Republicans in the control, with no significant effect of Democratic Party cues on Republican respondents.

The *Congruent Moral Frame from In-Party Hypothesis* predicts that partisans will be particularly receptive to cues from the in-party when they invoke a moral foundation that is found in the party's lexicon (i.e. the Republican Party and loyalty moral values). We observe this pattern of effects in Study 1. When the Republican Party uses the loyalty moral violation in opposition to flag burning, this typically reduces tolerance relative to the loyalty frame alone by 15% of the scale (B = -0.15, p < .05, see Figure 3 bottom middle panel, and Table 1, Column 5), and to the free speech frame by over 20% of the scale (B = -0.21, p < .05, see Figure 3 top right panel, and Table 1, Column 4), the largest effects on flag burning tolerance gleaned in Study 1. However, the experimental design used in Study 1 enmeshed cues from both parties with moral frameworks, making it unclear whether these effects were due to the Republican Party's opposition, their use of the loyalty moral violation frame, or the Democratic Party's support of flag burning as a form of free speech.

Table 2

Experimental Treatment Effects for Flag Burning and Anti-Muslim Speech (Study 2)

	Flag Burning Speech				Anti-Muslim Speech			
	1[a]		2[b]		3[c]		4[d]	
Effect	***B***	***SE***	***B***	***SE***	***B***	***SE***	***B***	***SE***
Moral Frame	-0.115**	0.042	-0.125**	0.047	-0.005	0.038	-0.037	0.053
Democratic Party Cues	-0.151**	0.042	-0.186**	0.051	-0.082*	0.039	-0.084	0.054
Republican Party Cues	-0.164**	0.044	-0.102*	0.049	-0.041	0.038	-0.039	0.054
Moral Frame X Democratic Party Cues	0.122*	0.056	0.144*	0.068	0.069	0.050	0.051	0.068
Moral Frame X Republican Party Cues	0.182**	0.058	0.155*	0.066	0.005	0.048	0.007	0.067
Respondent's Party Identity			-0.205**	0.078			-0.048	0.069
Moral Frame X Party Identity			-0.005	0.099			0.069	0.082
Democratic Party Cues X Party Identity			0.154	0.102			-0.008	0.091
Republican Party Cues X Party Identity			-0.157	0.098			-0.022	0.084
Moral Frame X Democratic Party Cues X Party Identity			-0.055	0.131			0.074	0.112
Moral Frame X Republican Party Cues X Party Identity			0.109	0.129			0.021	0.105
Constant	0.715**	0.032	0.781**	0.036	0.708**	0.031	0.728**	0.044

Note. OLS regression estimates with robust standard errors. Columns 1 and 3 provide estimates from a 2x3 factorial design, and Columns 2 and 4 provide estimates from a 2x3x2 factorial design. In each model, Moral Frame has factor levels 1) No Moral Frame provided, and 2) Moral Frame provided; Party Cues has factor levels: 1) No Party Cues provided, 2) Democratic Party Cues provided, and 3) Republican Party Cues provided; and Party Identity has factor levels 1) Democratic identifying respondents, and 2) Republican identifying respondents. As such, coefficients reported in the "Constant" are mean levels of tolerance for speech in the pure control condition (for each type of speech), and are also Democrats in the 3-way factorial. Analyses are confined to the subset of respondents who received only that certain type of speech treatment, notated by column titles "Flag Burning Speech" and "Anti-Muslim Speech," respectively. These same models with demographic controls are provided in the Supplementary Materials, Table A7. Table A8 (Supplementary Materials) estimates these models, including an interaction between each treatment and respondents' loyalty morality to check for any heterogeneous treatment effects.

[a]N = 898, R^2 = 0.018. [b]N = 788, R^2 = 0.111. [c]N = 972, R^2 = 0.009. [d]N = 845, R^2 = 0.017.

‡p < .10. *p < .05. **p < .01.

Disentangling these effects in Study 2, we find that, while Republicans reveal less tolerance when the Republican Party uses the loyalty frame to oppose flag burning compared to the control (B = -0.124, p < .05), this effect is smaller than when Republican opposition was provided without any loyalty moral justification (see Figure 2 right panel). In fact, compared to the Republican opposition frame alone, the Republican Party's usage of loyalty marginally *increases* flag burning tolerance (B = 0.136, p = .053). Further, there is no significant difference in Republicans' average tolerance to flag burning when the Democratic Party opposes using the loyalty frame, relative to the control, the loyalty frame alone, or Democratic Party opposition alone. This suggests that out-party cues, even when employing the correct moral packaging, are still ineffective. Combined, loyalty moral values do little to buttress the effectiveness of party cues for Republicans, regardless of which party endorses the loyalty frame, failing to support the *Congruent Moral Frame from In-Party Hypothesis* or the *Congruent Moral Frame from Out-Party Hypothesis*. Instead, in-party cues are more powerful than loyalty in reducing tolerance for flag burning among Republicans.

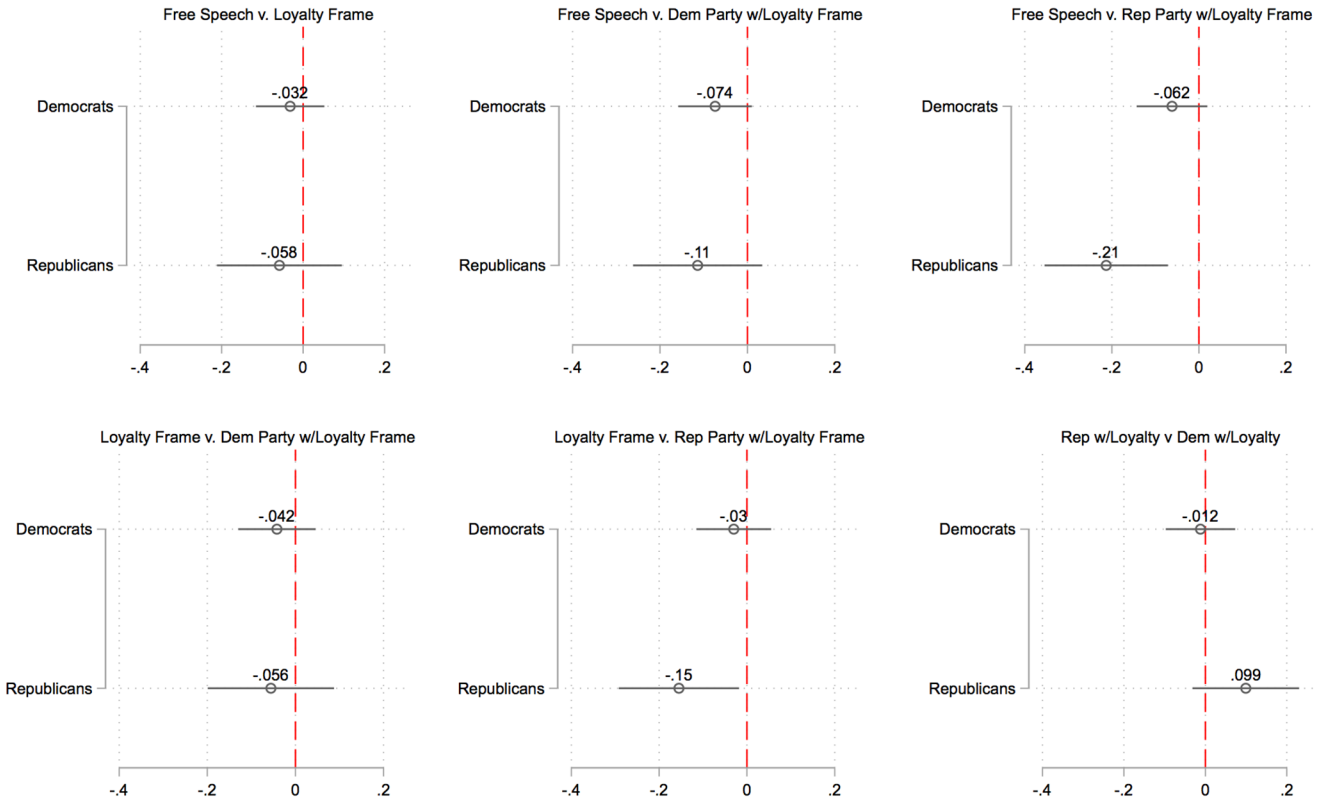

Figure 3. Experimental treatment effects on flag burning tolerance by party ID (Study 1).

Note. Open dots represent the estimated marginal effects of the experimental treatment (the comparison described in each panel's title) for Democrats and Republicans, with lines representing 95% confidence intervals. The three panels in the top row present marginal effects generated from the OLS regression model without demographic controls provided in Table 1, Column 4. The left and middle panels in the bottom row present marginal effects generated from the OLS regression model without demographic controls provided in Table 1, Column 5. The right panel in the bottom row presents marginal effects generated from the OLS regression model without demographic controls provided in Table 1, Column 6.

Since the Democratic Party typically tolerates flag burning and does not rely on loyalty values when speaking to its base, our treatments allow us to assess the effects of partisan endorsement of unorthodox moral values in line with the *Incongruent Moral Frame from In-Party Hypothesis*. Compared to tolerance in Study 2's pure control, Democrats decreased their tolerance for flag burning, on average, by 17% when told that the Democratic Party opposes flag burning on the grounds it is a betrayal to patriotic Americans (B = -0.166, p < .05, see Figure 2 middle panel), which is evidence of successful moral repackaging wherein party cues shape attitudes using incongruent moral appeals. However, given that average flag burning tolerance is statistically indistinguishable between the loyalty frame alone, Democratic Party opposition alone, and Democratic Party with loyalty frame conditions (see Figure 2 middle panel), the repackaged moral argument is no more effective at reducing tolerance than party cues or moral frames in isolation. We therefore consider this weak support for the *Incongruent Moral Frame from In-Party Hypothesis*. We also find no evidence for the *Backlash Hypothesis*, since relative to the control or Republican opposition alone, Democrats' average tolerance of flag burning either decreased or remained the same when presented with Republican opposition using the loyalty frame (see Figure 2 middle panel).

In sum, moral value frames and party cues decreased tolerance for flag burning among all respondents combined, as well as among Democrats. Republicans were responsive to loyalty frames alone, and to flag burning opposition from the Republican party, but not from the Democratic Party. Thus, even attitudes on an issue that resonates with Republicans' preferences for national loyalty can be moved given the right approach. Interestingly, the pairing of party cues and loyalty moral values did not reduce tolerance to any greater degree than party cues alone, regardless of the party identity of the respondent, indicating the primacy and power of party leaders to mold attitudes that possess moral roots.

Decreasing Anti-Muslim Tolerance

When assessing tolerance for anti-Muslim speech, specifically an instance of drawing the prophet Muhammed, we first examine our treatment effects aggregated across all respondents (for the same reasons provided in the case of flag burning tolerance).[xx] In Study 1 (Table 3, Column 1), we find that reduced tolerance for anti-Muslim speech emerges in the presence of a harm moral violation frame alone (B = -0.094, p < .01), and when this frame is endorsed by the Democratic (B = -0.068, p < .05) or Republican Party (B = -0.064, p < .05), compared to the free speech frame. In Study 2 (Table 2, Column 3), Democratic Party opposition to anti-Muslim speech typically reduces tolerance by about 8% of the 0-1 scale (B = -0.082, p < .05), otherwise our treatments fail to decrease tolerance for anti-Muslim speech relative to the control condition. As such, we find mixed support for the *Moral Frame Alone Hypothesis* (Study 1 only), and the *Party Cues Alone Hypothesis* (when the Democratic Party opposes, Study 2). Together, support for protecting anti-Muslim speech can be reduced within the mass public when it is framed as a violation of the harm moral foundation but *only* when used to counter a free speech appeal.

Looking exclusively at partisan respondents, mean tolerance for anti-Muslim speech is marginally higher among Republicans than Democrats in Study 1's baseline free speech condition (Democrats M = 0.778 v. Republicans M = 0.838, t(152) = 1.41, p = .081 on a one-tailed test), but is equivalent in Study 2's baseline control condition (Democrats M = 0.728 v. Republicans M = 0.68, t(64) = 0.69, p = 0.49 on a two-tailed test). Though, compared to that of flag burning, the variation in tolerance of anti-Muslim speech across conditions for both sets of partisans appears relatively small (see bottom portion of Figure 1). As such, even when we examine partisans in isolation (Figure 5, and Table 3, Columns 4 through 6 for Study 1; Figure 4 middle and right panels, and Table 2, Column 4 for Study 2), our treatments have minimal (and occasionally adverse) effects on reducing tolerance towards anti-Muslim speech. While in Study 1 the harm morality violation frame decreases tolerance for anti-Muslim speech among Democrats (B = -0.095, p < .05) and marginally among Republicans (B = -0.099, p = .065) relative to the free speech frame condition (see Figure 5, top left panel, and Table 3, Column 4), this moral frame fails to move Democrats (B = -0.037, n.s., see Figure 4 middle panel) or Republicans (B = 0.033, n.s., see Figure 4 right panel) compared to the control condition in Study 2. We again find mixed support for the *Moral Frames Alone Hypothesis* depending upon which baseline comparison we employ.

Table 3

Experimental Treatment Effects for Anti-Muslim Speech (Study 1)

	All Respondents[a]						Partisans Only[b]					
	1		2		3		4		5		6	
Effect	***B***	***SE***	***B***	***SE***	***B***	***SE***	***B***	***SE***	***B***	***SE***	***B***	***SE***
Free Speech Frame	–	–	0.094**	0.030	0.068*	0.027	–	–	0.095*	0.040	0.101**	0.037
Harm Frame	-0.094**	0.030	–	–	-0.025	0.031	-0.095*	0.040	–	–	0.005	0.040
Dem Party using Harm Frame	-0.068*	0.027	0.025	0.031	–	–	-0.101**	0.037	-0.005	0.040	–	–
Rep Party using Harm Frame	-0.064*	0.026	0.030	0.029	0.005	0.027	-0.039	0.036	0.056	0.038	0.062‡	0.036
Party ID							0.060	0.040	0.056	0.053	0.099*	0.045
Free Speech Frame X Party ID							–	–	0.004	0.067	-0.039	0.060
Harm Frame X Party ID							-0.004	0.067	–	–	-0.043	0.070
Dem Party Harm Frame X Party ID							0.039	0.060	0.043	0.070	–	–
Rep Party Harm Frame X Party ID							-0.081	0.057	-0.077	0.068	-0.121*	0.061
Constant	0.796**	0.019	0.702**	0.023	0.727**	0.020	0.778**	0.026	0.683**	0.030	0.678**	0.027

Note. OLS regression estimates with robust standard errors. In Columns 1 and 4, the "Free Speech Frame" is the excluded treatment category, in Columns 2 and 5 the "Harm Frame" is the excluded treatment category, and in Columns 3 and 6 the "Democratic Party using Harm Frame" is the excluded treatment category. Party ID is a dichotomous variable coded 0=Democrats, and 1=Republicans. All analyses are confined to the subset of respondents who received one of the anti-Muslim speech treatments (i.e. respondents viewing a flag burning treatment are excluded from the above models). These same models with demographic controls are provided in the Supplementary Materials, Table A5. Table A6 (Supplementary Materials) estimates the models shown in Columns 1 and 4, including an interaction between each treatment and respondents' loyalty morality to check for any heterogeneous treatment effects.

[a]$N = 706$, $R^2 = 0.017$. [b]$N = 618$, $R^2 = 0.032$.

‡$p < .10$. *$p < .05$. **$p < .01$.

Given Democrats' disdain for harm towards marginalized groups like Muslims (Haidt, 2012; Koleva et al., 2012), we anticipated that when the Democratic Party opposed anti-Muslim speech (per the *Party Cues Alone Hypothesis*), or the Democratic Party utilized the harm frame in opposition to anti-Muslim speech (per the *Congruent Moral Frame from In-Party Hypothesis)*, Democratic respondents would respond in kind by reducing tolerance for this harmful speech. On one hand, the combination of Democratic Party opposition and harm frame does produce a 10% reduction of tolerance compared to the free speech frame in Study 1 ($B = -0.101$, $p < .01$, see Table 3, Column 4), as expected. But, the additional Democratic Party cue does *not* further decrease Democrats' tolerance compared to the harm frame alone in Study 1 ($B = -0.005$, n.s., see Table 3, Column 5), and in Study 2 Democrats' tolerance remains virtually unchanged when provided information about the Democratic Party's position on offensive anti-Muslim speech.[xxi]

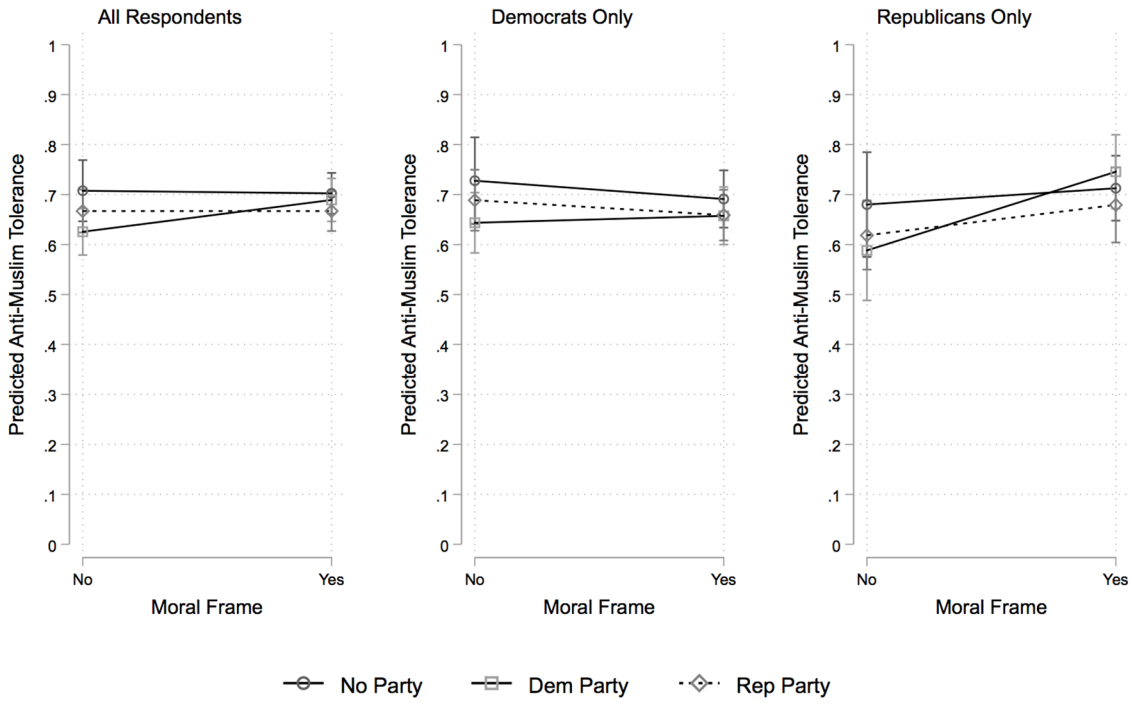

Figure 4. Experimental treatment effects of anti-Muslim tolerance (Study 2).

Note. Open dots represent the estimated predicted values for tolerance of speech in each of the anti-Muslim speech conditions plus control condition, with 95% confidence interval bars. All estimates are derived from OLS regression models that employ a factorial design (see Table 2). Estimates for "All Respondents" are derived from the 2x3 factorial design without demographic controls estimated in Table 2, Column 3. Estimates for "Democrats Only" and "Republicans Only" are derived from the 2x3x2 factorial design without demographic controls estimated in Table 2, Column 4.

Even though Democrats emphasize the value of care/harm to a greater degree than Republicans, we find support for the *Congruent Moral Frame from In-Party Hypothesis* only when these moral appeals are pitted against a free speech argument (Study 1). Since Study 1 presents the position of both parties simultaneously, we cannot determine whether Democrats reacted to the morally congruent appeal from their party, or in resistance to the free speech position of the out-party. Thus, our results are minimally supportive of the *Congruent Moral Frame from In-Party Hypothesis* hypothesis. Democrats, furthermore, were unmoved by any Republican appeal condemning anti-Muslim speech in either study,[xxii] even when Republican Party leaders invoked the moral value of care/harm. As such, we lack support for the *Congruent Moral Frame from Out-Party Hypothesis.*

Republican respondents, on the other hand, reacted to their party's endorsement of harm morality violations in the context of anti-Muslim speech by reducing their tolerance of this form of hate speech in Study 1 – supportive evidence of the *Incongruent Moral Frame from In-Party Hypothesis*. Specifically, when the Republican Party discussed anti-Muslim speech as being harmful to Muslims, this message resonated with Republicans, resulting in a 12% decrease, on average, in tolerance compared to the free speech frame (B = -0.12, p < .01, see Figure 5 top right panel and Table 3, Column 4). Our results demonstrate that Republicans can exhibit intolerance for anti-Muslim speech when that speech is framed as a moral violation paired with in-party opposition.

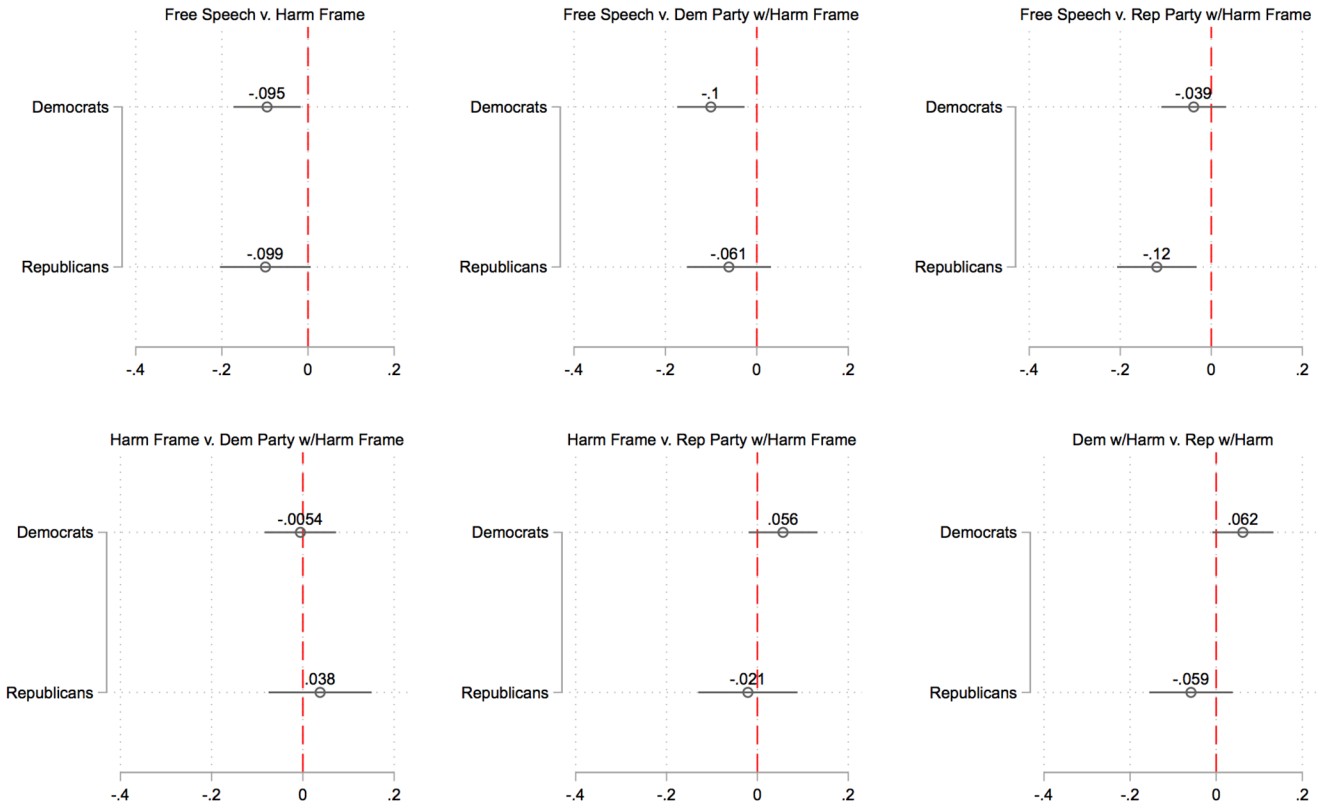

Figure 5. Experimental treatment effects on anti-Muslim tolerance by party ID (Study 1).

Note. Open dots represent the estimated marginal effects of the experimental treatment (the comparison described in each panel's title) for Democrats and Republicans, with lines representing 95% confidence intervals. The three panels in the top row present marginal effects generated from the OLS regression model without demographic controls provided in Table 3, Column 4. The left and middle panels in the bottom row present marginal effects generated from the OLS regression model without demographic controls provided in Table 3, Column 5. The right panel in the bottom row presents marginal effects generated from the OLS regression model without demographic controls provided in Table 3, Column 6.

While this finding may provide evidence of the normatively optimistic conclusion that members of the political right can be softened in their views towards Muslims, it is possible that this result instead reflects a "negative partisanship"-style backlash (per Abramowitz & Webster, 2016) to the Democratic Party's support of anti-Muslim speech on free speech grounds in Study 1. Study 2 provides support for this *Backlash Hypothesis* among Republican respondents. When reading that the Democratic Party opposes anti-Muslim speech because it harms Muslim-Americans, Republicans became more tolerant of it than they were when hearing about Democratic Party opposition without any moral framing ($B = 0.16$, $p < .05$, see Figure 4 right panel). Thus, it is not simply the endorsement by the out-party, but the out-party's explicit use of the care/harm frame as justification for opposition that drives Republicans' increasing tolerance of anti-Muslim speech.

In sum, the harm moral foundation alone as a tool to reduce tolerance of anti-Muslim speech has limited utility. While it can decrease tolerance among all respondents, both Democrats and Republicans, when contrasted against free speech appeals, it is ineffective when used against a non-political control. Encouragingly, among Republicans the combination of the harm moral violation with in-party opposition significantly reduced anti-Muslim

tolerance in Study 1. But the converse combination, out-party opposition using a harm moral violation, actually increased tolerance of anti-Muslim speech in Study 2.

Discussion and Implications

Across two survey experiments each employing two examples of hate speech, we demonstrate that the effects of framing on tolerance for hate speech vary depending on the political target group and the baseline condition. While not uniform, we found the most consistent evidence in support of the *Moral Frames Alone* and *Party Cues Alone Hypotheses*. In the case of flag burning, we observed this pattern relative to a non-political control among all respondents. Partisans were also responsive to cues from the in-party, and occasionally exhibited decreased tolerance towards hate speech when it was presented as a moral frame violation that resonated with their party's values (i.e., Republicans and loyalty) – supporting the *Congruent Moral Frame from In-Party Hypothesis*. However, there were no real additive effects of moral frames and party cues. The combination of Republican Party and loyalty morality did not decrease tolerance more than either loyalty frames or Republican Party cues on their own. There also was mixed, albeit limited, evidence of a party's ability to repackage moral values among fellow partisans. While morally congruent arguments from the out-party had no effect on reducing tolerance, morally incongruent appeals from the in-party (i.e., Democratic Party embracing loyalty) reduced tolerance of flag burning speech among Democrats – evidence of the *Incongruent Moral Frame from In-Party Hypothesis*. There is no consistent evidence that harm morality appeals can decrease tolerance for Anti-Muslim speech in the absence of a free speech counter-argument. Instead, Republicans became more tolerant of anti-Muslim speech when it was condemned by the Democratic Party as a violation of the care/harm value, in support of the *Backlash Hypothesis*. Thus, party cues have the ability, with or without attachment to moral values, to shape opinions of hate speech either positively or negatively, therefore demonstrating the formidable nature of party identification.

What could explain these inconsistent results? First, the differences between the two studies might be attributable to the distinct baseline used in each study. In Study 1, we employed a free speech frame wherein the speech in question was defended as being protected by the First Amendment as our baseline condition. This baseline was not necessarily a true control because it likely primed political considerations related to civil liberties. In Study 2, contrastingly, we adopted a neutral, non-political topic for the control, thus avoiding any potential political priming. Second, Study 2 afforded us the opportunity to tease apart the effects of party and morality using more nuanced treatments. In Study 1, when respondents were provided with an elite cue, it came from both the Republican Party and the Democratic Party simultaneously, and always referenced a morality violation. While our significant party and morality effects in Study 1 seemingly suggest that these appeals work best in conjunction, Study 2 shows this is not the case. When we peel apart moral values and party cues in Study 2 our findings suggest that party cues work on their own, and coupling moral values and party cues is actually less effective in reducing tolerance for flag burning than party cues alone. This evidence of party cue persuasion is quite remarkable given the importance and salience of the flag in American political culture (and in contrast to what would be expected per Ciuk & Yost, 2016).

Regarding the receptiveness of partisans to their respective party's endorsement of an atypical moral value, the evidence for moral repackaging is limited. On one hand, Democrats did respond to a moral value not typically associated with their political party. When compared to the control in Study 2, Democrats were more likely to oppose flag burning when the value of loyalty was invoked by their party leaders. Republicans, however, ignored appeals

highlighting harm whether or not it was employed by their party leaders, except when it was pitted against a free speech appeal made by the Democratic Party in Study 1. In no instance of speech in either study did morally congruent appeals from the out-party decrease tolerance for harmful speech – in contrast to the main theoretical expectation of the moral reframing literature. Instead of adjusting policy arguments to match the moral values of the target group (per Feinberg & Willer, 2013, 2015), party cues alone had the ability to change moral value receptiveness among their partisans. Such a finding provides an important contribution to moral reframing theory, and a possible avenue for future research on this topic.

Several aspects of the paper deserve particular attention. First, it is important to note that the question of restricting hate speech might not be fully captured with a single Likert-scale measure. Second, while our studies include large samples, the number of Republicans in each treatment condition is limited. Based upon a G*Power analysis we should have at least 90 Republicans per treatment in order to discern our small effects (ranging from .1 to .15) with an alpha of 0.05. Instead, we obtain anywhere from 44 to 68 Republicans per treatment across both studies. As a result, we have insufficient power to find significant treatment effects among Republican respondents (see also FN19). Third, opinions might have been different if we had included the names of actual party leaders instead of Republican and Democratic Party labels. For instance, if we had chosen President Trump as the Republican figure, answers might have been driven more by his persona than the party cue itself. Future studies could insert real Republican and Democratic elites and compare the results with those found in this study. Last, what constitutes hate speech is debatable, illustrating the inherent subjectivity of speech. Indeed, definitions of what is considered "hateful" often are broad and vague and, as a result, open to interpretation (Brink, 2001; Brison, 2013). Nonetheless, the ability to frame the message is a powerful tool, and future studies should examine the fluidity of opinions regarding other targets of hate speech, including the ongoing controversy regarding the kneeling of athletes during the national anthem.

Despite these potential drawbacks, our findings carry important normative implications for promoting social justice by illuminating methods which political leaders and advocates can employ in order to undercut tolerance for exceptionally uncivil speech. In the current American political environment where a new type of vitriol has been normalized, party leaders can generate opposition toward certain forms of hate speech, even when that speech is targeted at relatively unpopular ethnic or religious groups (such as Muslim-Americans). Such a shift in attitudes about hate speech might lessen negative attitudes toward these groups. In this respect, party leaders can facilitate a more tolerant society by discussing hate speech as a type of morality violation, or simply and more effectively, by condemning such speech without moral appeals.

On a more melancholic note, party leaders could use their influence to exacerbate existing tensions and promote certain stereotypes and prejudices. Equally troubling, a party's opposition to hateful speech could actually generate greater intolerance from the other side. We found that among Republicans, tolerance of anti-Muslim speech actually increases when the Democratic Party employed the moral value of harm to oppose anti-Muslim speech. In that instance, a backlash was generated among Republicans, who became more tolerant of cartoons that are deeply offensive to Muslims. This finding extends the research on negative partisanship (Abramowitz & Webster, 2016), which finds that animus toward the out-party is a stronger motivator of political attitudes than affect for the in-party.

But controlling the debate, and setting its terms is paramount. In fact, the best methods to lower public support for protecting anti-Muslim hate speech may be to prime non-political topics, or other forms of offensive speech

(i.e. flag burning). As such, it is helpful to recognize that each issue might need to be framed and debated in different terms, since employing a free speech counterargument can decrease tolerance for hate speech in some instances, while other times the most effective arguments against hate speech are apolitical.

Social justice issues that conflict with civil liberty principles present a formidable challenge to message framing. This challenge can be overcome, however, in some instances, by party cues alone. If, however, different issues or even different forms of hate speech were studied, the results might differ, and opinions might be less malleable or more contingent on a coupling of moral values and party cues. Further, Democrats might demonstrate the negative partisanship pattern seen here by Republicans on the issue of anti-Muslim speech. Ongoing shifts, such as the deepening schism between the two major political parties, might further change this party-morality dynamic, particularly in relation to combating hate speech. Regardless, it is likely that partisan cues will continue to color judgments on the issue of acceptable speech.

Finally, some might view reducing hateful speech as an evolutionary stage of an advanced democracy, a process that will produce a more tolerant, more considerate society. Others, however, might read these results with great trepidation and caution. Strong proponents of free speech could look at these findings as evidence of the public's lack of appreciation of the First Amendment. Irrespective of the position taken in the debate over offensive speech, opinions on such issues will continue to be influenced by framing, partisan-elite rhetoric, and moral appeals.

Notes

i) Individuals within a society may vary in the extent to which they are offended, threatened, or insulted by any particular instance of speech that meets this technical definition. Thus, for our purposes we are interested in speech that a majority of United States citizens would characterize as harmful and offensive to society as a whole.

ii) https://www.theatlantic.com/politics/archive/2017/08/trump-defends-white-nationalist-protesters-some-very-fine-people-on-both-sides/537012/

iii) See also Supplementary Materials, Table A2.

iv) While this statement indicates an interaction between a message's moral frame and an individual's party identity, it is not to say that such moral appeals have absolutely no effect in the aggregate. It is possible that loyalty and harm moral arguments can reduce tolerance among an ideologically incongruent target audience (and we find in some instances that they do), but these effects should be less prevalent than those of morally congruent appeals. Further, the primary purpose of this study is to identify the types of messaging strategies that can make Democrats, Republicans, and citizens as a whole less tolerant of hate speech, *not* to pinpoint interaction effects relative to moral reframing. Therefore, while we statistically test the moderating effects of party identity on message frames, we do not offer any explicit hypotheses related to these tests.

v) The survey was approved by the Institutional Research Board at The University of Mississippi.

vi) For further details on the actual events behind our fictionalized news stories, see:
https://www.nbcnewyork.com/news/local/American-Flag-Burning-Disarm-NYPD-Fort-Greene-Park-Brooklyn-Racism-Confederate-Death-New-York-Crime-311252831.html
https://www.nbcnews.com/news/us-news/draw-muhammad-shooting-who-was-behind-cartoon-contest-n353081

vii) The anti-Muslim speech referenced in the story involved a violent encounter. This violent encounter was omitted from our treatments so as not to bias respondents who may have reacted to the violence instead of to the speech.

viii) In both Study 1 and Study 2, random assignment successfully balanced subjects across conditions on the basis of their party identity, ideology, race, gender, education, and church attendance.

ix) In both studies, we used Qualtrics' randomizer function to present each respondent with one of the treatment conditions, allowing for unequal assignment (i.e. we did not specify balancing across conditions).

x) The topic discussed in the control condition is the obscure sport of Kaninhop, also known as bunny jumping. Similar to the tactic employed by Feinberg and Willer (2015), we selected this topic because of its uncontroversial nature and lack of familiarity. The sport of bunny jumping should not have primed considerations related to hate speech, as less than 10% of respondents rated it as socially or personally offensive. The text of this condition can be found under the description of bunny jumping found in this list: https://list25.com/the-25-most-obscure-sports-in-the-world/2/

xi) To avoid further increasing the number of conditions, we chose not to include conditions where either party defended the speech in question. Thus, the conditions in Study 2 have each party taking a stance against hate speech, but not defending it. This meshes with the paper's focus to determine how tolerance for hate speech can be reduced.

xii) When accounting for the responses to our demographic variables, we obtained a realized N = 1402 in Study 1, and N = 1870 in Study 2.

xiii) These items were used to screen Muslim respondents who may have been offended by the Muhammed cartoon contest story and associated picture, resulting in 19 respondents dropped from Study 1, and 21 from Study 2.

xiv) Respondents in Study 2's control condition were asked their support for protecting flag-burning and anti-Muslim speech on the screen immediately following the story about bunny jumping, with the order in which these two items were presented randomized to account for ordering effects. Item order had no effect on flag burning tolerance (mean when shown first = 0.715, mean when shown second = 0.656, $t(166) = 1.19$, $p = 0.24$ on a two-tailed test), though it did impact tolerance for anti-Muslim speech (see results section for discussion of how these order effects are incorporated into the analyses).

xv) The text of these items in Study 2's control condition read as follows: "In your opinion, do you agree or disagree that the First Amendment should protect [flag/burning/drawing anti-Muslim cartoons]?" followed by the same six response options provided in the other conditions.

xvi) The text of these questions read as follows: 1) “For each type of speech listed, please rate how much you think this type of speech *reflects the definition of hate speech* provided above. To be clear, we are not asking for your personal opinion, but rather whether you think *Americans as a whole* would consider each of the following hate speech based on the above definition." 2) “For each type of speech listed, please rate how much you are *personally offended* by it.”

xvii) Study 1 free speech condition M = 0.796, 95% CI [0.76,0.83]; Study 2 control condition, anti-Muslim item first M = 0.708, 95% CI [0.65, 0.77].

xviii) Unless otherwise specified, all reported p-values for beta coefficients are based upon a two-tailed t-test. There are no significant heterogeneous treatment effects across levels of loyalty/betrayal moral values in Study 1 (see Supplementary Materials, Table A4) or Study 2 (see Supplementary Materials, Table A8, Columns 1 and 2).

xix) The effect of the loyalty frame alone for Republican respondents is in the hypothesized negative direction, but does not meet the threshold for statistical significance (B = -0.13, p = .136). We suspect that the lack of significance in this instance may result from the low number of Republican respondents in Study 2 (see discussion and implications section for power limitations in our studies).

xx) There are no significant heterogeneous treatment effects across levels of care/harm moral values in Study 1 (see Supplementary Materials, Table A6) or Study 2 (see Supplementary Materials, Table A8, Columns 3 and 4).

xxi) Interestingly in Study 2, information about Democratic Party opposition marginally reduced anti-Muslim speech tolerance among Democrats, compared to the control, when controlling for respondents' political interest, race, gender, church attendance, and education (B = -0.104, p = .052, see Supplementary Materials, Table A7, Column 4).

xxii) See overlapping estimates in Figure 4 middle panel, and Democrats' non-significant marginal effects in Figure 5 top right and bottom right panels.

Funding

Funding was provided by the University of Mississippi Department of Political Science.

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgments

We would like to thank Conor Dowling, Jon Winburn, and attendees of the American Politics Working Group at the University of Mississippi for helpful comments on this project. An early version was presented at the 2017 Southern Political Science Association Annual Meeting, and we are thankful to David Ciuk and panel attendees for their feedback. We would also like to thank Mark Brandt and our anonymous reviewers for their extremely valuable guidance and suggestions.

Ethics Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Data Availability

For this study, supplementary materials are freely available (see the Supplementary Materials section).
Please contact the corresponding author for replication data.

Supplementary Materials

Further information regarding our experimental designs, treatments, survey measures, and additional statistical tests are provided in the Supplementary Materials. (For access see Index of Supplementary Materials below.)

Index of Supplementary Materials

Armstrong, G. M., & Wronski, J. (2019). *Supplementary materials to "Framing hate: Moral foundations, party cues, and (in)tolerance of offensive speech".* PsychOpen. https://doi.org/10.23668/psycharchives.2596

References

Abramowitz, A. I., & Webster, S. (2016). The rise of negative partisanship and the nationalization of US elections in the 21st century. *Electoral Studies, 41*, 12-22. https://doi.org/10.1016/j.electstud.2015.11.001

Achen, C. H., & Bartels, L. M. (2016). *Democracy for realists: Why elections do not produce responsive government.* Princeton, NJ, USA: Princeton University Press.

Altman, A. (1993). Liberalism and campus hate speech: A philosophical examination. *Ethics, 103*, 302-317. https://doi.org/10.1086/293497

Bartels, L. M. (2002). Beyond the running tally: Partisan bias in political perceptions. *Political Behavior, 24*(2), 117-150. https://doi.org/10.1023/A:1021226224601

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*(3), 351-368. https://doi.org/10.1093/pan/mpr057

Boeckmann, R. J., & Liew, J. (2002). Hate speech: Asian American students' justice judgments and psychological responses. *Journal of Social Issues, 58*(2), 363-381. https://doi.org/10.1111/1540-4560.00265

Brink, D. O. (2001). Millian principles, freedom of expression, and hate speech. *Legal Theory, 7*(2), 119-157. https://doi.org/10.1017/S1352325201072019

Brison, S. (2013). Hate speech. In H. LaFollette (Ed.), *The International Encyclopedia of Ethics* (pp. 2332-2342). Malden, MA, USA: Wiley-Blackwell.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3-5. https://doi.org/10.1177/1745691610393980

Byers, B. D., & Jones, J. A. (2007). The impact of the terrorist attacks of 9/11 on anti-Islamic hate crime. *Journal of Ethnicity in Criminal Justice, 5*(1), 43-56. https://doi.org/10.1300/J222v05n01_03

Campbell, A., Converse, P., Miller, W., & Stokes, D. (1960). *The American voter*. Chicago, IL, USA: University of Chicago Press.

Carroll, J. (2006, June 29). *Public support for Constitutional Amendment on Flag Burning.* Retrieved from http://news.gallup.com/poll/23524/public-support-constitutional- amendment-flag-burning.aspx

Ciuk, D. J., & Yost, B. A. (2016). The effects of issue salience, elite influence, and policy content on public opinion. *Political Communication, 33*(2), 328-345. https://doi.org/10.1080/10584609.2015.1017629

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology, 85*(5), 808-822. https://doi.org/10.1037/0022-3514.85.5.808

Cohen, R. (2014). Regulating hate speech: Nothing customary about it. *Chicago Journal of International Law, 15*(1), 229-255.

Coppock, A. (2019). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods, 7*(3), 613-628. https://doi.org/10.1017/psrm.2018.10

Crawford, J. T., & Pilanski, J. M. (2014). Political intolerance, right and left. *Political Psychology, 35*(6), 841-851. https://doi.org/10.1111/j.1467-9221.2012.00926.x

Davis, D. W. (2007). *Negative liberty: Public opinion and the terrorist attacks on America*. New York, NY, USA: Russell Sage Foundation.

Davis, D. W., & Silver, B. D. (2004). Civil liberties vs. security: Public opinion in the context of the terrorist attacks on America. *American Journal of Political Science, 48*(1), 28-46. https://doi.org/10.1111/j.0092-5853.2004.00054.x

Day, M. V., Fiske, S. T., Downing, E. L., & Trail, T. E. (2014). Shifting liberal and conservative attitudes using moral foundations theory. *Personality and Social Psychology Bulletin, 40*(12), 1559-1573. https://doi.org/10.1177/0146167214551152

Delgado, R., & Yun, D. H. (1994). Pressure valves and bloodied chickens: An analysis of paternalistic objections to hate speech regulation. *California Law Review, 82*(4), 871-892. https://doi.org/10.2307/3480935

Ditto, P. H., & Koleva, S. P. (2011). Moral empathy gaps and the American culture war. *Emotion Review, 3*(3), 331-332. https://doi.org/10.1177/1754073911402393

Druckman, J. N., & Leeper, T. J. (2012). Learning more from political communication experiments: Pretreatment and its effects. *American Journal of Political Science, 56*(4), 875-896. https://doi.org/10.1111/j.1540-5907.2012.00582.x

Druckman, J. N., Peterson, E., & Slothuus, R. (2013). How elite partisan polarization affects public opinion formation. *American Political Science Review, 107*(1), 57-79. https://doi.org/10.1017/S0003055412000500

Evans, G., & Pickup, M. (2010). Reversing the causal arrow: The political conditioning of economic perceptions in the 2000–2004 US presidential election cycle. *The Journal of Politics, 72*(4), 1236-1251. https://doi.org/10.1017/S0022381610000654

Feinberg, M., & Willer, R. (2013). The moral roots of environmental attitudes. *Psychological Science, 24*(1), 56-62. https://doi.org/10.1177/0956797612449177

Feinberg, M., & Willer, R. (2015). From gulf to bridge: When do moral arguments facilitate political influence? *Personality and Social Psychology Bulletin, 41*(12), 1665-1681. https://doi.org/10.1177/0146167215607842

Frimer, J. A., Gaucher, D., & Schaefer, N. K. (2014). Political conservatives' affinity for obedience to authority is loyal, not blind. *Personality and Social Psychology Bulletin, 40*(9), 1205-1214. https://doi.org/10.1177/0146167214538672

Gerber, A. S., & Huber, G. A. (2010). Partisanship, political control, and economic assessments. *American Journal of Political Science, 54*(1), 153-173. https://doi.org/10.1111/j.1540-5907.2009.00424.x

Gerber, A. S., Huber, G. A., & Washington, E. (2010). Party affiliation, partisanship, and political beliefs: A field experiment. *American Political Science Review, 104*(4), 720-744. https://doi.org/10.1017/S0003055410000407

Gibson, J. L. (1988). Political intolerance and political repression during the McCarthy Red Scare. *American Political Science Review, 82*(2), 511-529. https://doi.org/10.2307/1957398

Gibson, J. L. (1992). The political consequences of intolerance: Cultural conformity and political freedom. *American Political Science Review, 86*(2), 338-356. https://doi.org/10.2307/1964224

Gibson, J. L. (2005). Parsimony in the study of tolerance and intolerance. *Political Behavior, 27*(4), 339-345. https://doi.org/10.1007/s11109-005-7408-4

Gibson, J. L. (2008). Intolerance and political repression in the United States: A half century after McCarthyism. *American Journal of Political Science, 52*(1), 96-108. https://doi.org/10.1111/j.1540-5907.2007.00301.x

Goren, P. (2005). Party identification and core political values. *American Journal of Political Science, 49*(4), 881-896. https://doi.org/10.1111/j.1540-5907.2005.00161.x

Graham, J., & Haidt, J. (2012). Sacred values and evil adversaries: A moral foundations approach. In M. Mikulincer & P. R. Shaver (Eds.),*The social psychology of morality: Exploring the causes of good and evil* (pp. 11-31). Washington, DC, USA: American Psychological Association.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029-1046. https://doi.org/10.1037/a0015141

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York, NY, USA: Vintage Books.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*(1), 98-116. https://doi.org/10.1007/s11211-007-0034-z

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science, 59*(3), 690-707. https://doi.org/10.1111/ajps.12152

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P., & Jewell, R. (2018). *The shape of and solutions to the MTurk quality crisis.* Retrieved from SSRN repository: https://ssrn.com/abstract=3272468

Kidwell, B., Farmer, A., & Hardesty, D. M. (2013). Getting liberals and conservatives to go green: Political ideology and congruent appeals. *Journal of Consumer Research, 40*(2), 350-367. https://doi.org/10.1086/670610

Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of Research in Personality, 46*(2), 184-194. https://doi.org/10.1016/j.jrp.2012.01.006

Lakoff, G. (2010). *Moral politics: How liberals and conservatives think*. Chicago, IL, USA: University of Chicago Press.

Lambe, J. L. (2004). Who wants to censor pornography and hate speech? *Mass Communication and Society, 7*(3), 279-299. https://doi.org/10.1207/s15327825mcs0703_2

Leets, L. (2002). Experiencing hate speech: Perceptions and responses to anti-Semitism and antigay speech. *Journal of Social Issues, 58*(2), 341-361. https://doi.org/10.1111/1540-4560.00264

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of Mechanical Turk samples. *SAGE Open, 6*(1). https://doi.org/10.1177/2158244016636433

Levendusky, M. (2009). *The partisan sort: How liberals became Democrats and conservatives became Republicans*. Chicago, IL, USA: University of Chicago Press.

Lukianoff, G., & Haidt, J. (2015, September). The coddling of the American Mind. *The Atlantic*. Retrieved from https://www.theatlantic.com/magazine/archive/2015/09/the-coddling-of-the-american-mind/399356

Lupia, A. (1994). Shortcuts versus encyclopedias: Information and voting behavior in California insurance reform elections. *American Political Science Review, 88*(1), 63-76. https://doi.org/10.2307/2944882

Mason, L. (2015). “I disrespectfully agree”: The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science, 59*(1), 128-145. https://doi.org/10.1111/ajps.12089

Massaro, T. M. (1991). Equality and freedom of expression: The hate speech dilemma. *William and Mary Law Review, 32*(2), 211-265.

Mullinix, K., Leeper, T., Druckman, J., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science, 2*(2), 109-138. https://doi.org/10.1017/XPS.2015.19

Nelson, T. E., Clawson, R. A., & Oxley, Z. M. (1997). Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review, 91*(3), 567-583. https://doi.org/10.2307/2952075

Peffley, M., Knigge, P., & Hurwitz, J. (2001). A multiple values model of political tolerance. *Political Research Quarterly, 54*(2), 379-406. https://doi.org/10.1177/106591290105400207

Pew Research Center. (2017, July 26). 7. How the U.S. general public views Muslims and Islam. Retrieved from https://www.pewforum.org/2017/07/26/how-the-u-s-general-public-views-muslims-and-islam/

Schafer, C. E., & Shaw, G. M. (2009). Trends—Tolerance in the United States. *Public Opinion Quarterly, 73*(2), 404-431. https://doi.org/10.1093/poq/nfp022

Schwadel, P., & Garneau, C. R. (2014). An age–period–cohort analysis of political tolerance in the United States. *The Sociological Quarterly, 55*(2), 421-452. https://doi.org/10.1111/tsq.12058

Stone, G. R. (2004). *Perilous times: Free speech in wartime from the Sedition Act of 1798 to the war on terrorism*. New York, NY, USA: W. W. Norton & Company.

Sullivan, J. L., Marcus, G. E., Feldman, S., & Piereson, J. E. (1981). The sources of political tolerance: A multivariate analysis. *American Political Science Review, 75*(1), 92-106. https://doi.org/10.2307/1962161

Tsesis, A. (2002). *Destructive messages: How hate speech paves the way for harmful social movements*. New York, NY, USA: NYU Press.

Twenge, J. M., Carter, N. T., & Campbell, W. K. (2015). Time period, generational, and age differences in tolerance for controversial beliefs and lifestyles in the United States, 1972–2012. *Social Forces, 94*(1), 379-399. https://doi.org/10.1093/sf/sov050

Voelkel, J. G., & Brandt, M. J. (2019). The effect of ideological identification on the endorsement of moral values depends on the target group. *Personality and Social Psychology Bulletin, 45*(6), 851-863. https://doi.org/10.1177/0146167218798822

Voelkel, J. G., & Feinberg, M. (2018). Morally reframed arguments can affect support for political candidates. *Social Psychological & Personality Science, 9*(8), 917-924. https://doi.org/10.1177/1948550617729408

Wike, R. (2016, October 12). Americans more tolerant of offensive speech than others in the world. Retrieved from Pew Research Center website: https://www.pewresearch.org/fact-tank/2016/10/12/americans-more-tolerant-of-offensive-speech-than-others-in-the-world/

Wolsko, C., Ariceaga, H., & Seiden, J. (2016). Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology, 65*, 7-19. https://doi.org/10.1016/j.jesp.2016.02.005

Zaller, J. (1992). *The nature and origins of mass opinion*. Cambridge, United Kingdom: Cambridge University Press.