Author/s:
Abeywickrama, RS;Rhee, JJ;Crone, DL;Laham, SM

Title:
Why moral advocacy leads to polarization and proselytization: The role of self-persuasion

Date:
2020-01-01

Citation:
Abeywickrama, R. S., Rhee, J. J., Crone, D. L. & Laham, S. M. (2020). Why moral advocacy leads to polarization and proselytization: The role of self-persuasion. Journal of Social and Political Psychology, 8 (2), pp.473-503. https://doi.org/10.5964/jspp.v8i2.1346.

Persistent Link:
https://hdl.handle.net/11343/274423

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

Original Research Reports

Why Moral Advocacy Leads to Polarization and Proselytization: The Role of Self-Persuasion

Ravini S. Abeywickrama*[a], Joshua J. Rhee[a], Damien L. Crone[b], Simon M. Laham[a]

[a] Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, Australia. **[b]** Positive Psychology Center, University of Pennsylvania, Philadelphia, PA, USA.

Abstract

This research is the first to examine the effects of moral versus practical pro-attitudinal advocacy in the context of self-persuasion. We validate a novel advocacy paradigm aimed at uncovering why moral advocacy leads to polarization and proselytization. We investigate four distinct possibilities: (1) expression of moral foundational values (harm, fairness, loyalty, authority, purity), (2) reliance on moral systems (deontology and consequentialism), (3) expression of moral outrage, (4) increased confidence in one's advocacy attempt. In Study 1 (N = 255) we find differences between moral and practical advocacy on the five moral foundations, deontology, and moral outrage. In Study 2 (N = 218) we replicate these differences, but find that only the expression of moral foundations is consequential in predicting attitude polarization. In Study 3 (N = 115) we replicate the effect of moral foundations on proselytization. Our findings suggest that practical compared to moral advocacy may attenuate polarization and proselytization. This carries implications for how advocacy can be re-framed in ways which minimize social conflict.

Keywords: polarization, morality, persuasion, advocacy, migration, moral foundations, self-persuasion, attitudes

Non-Technical Summary

Background
Society is becoming increasingly polarized, and conflict between groups holding opposing views is frequent. A potential source of this conflict is people frequently advocating for their opinions in terms of moral values, which tend to be perceived as universal and sacred. However, we do not know the specific consequences of moral advocacy, and how it can lead to negative outcomes (e.g. political polarization). We also do not know if there are other forms of advocacy which may attenuate these negative outcomes.

Why Was This Study Done?
In light of the currently fractured and polarized attitudinal landscape, in which advocacy is frequent, we sought to understand: (1) the consequences of moral advocacy on attitude polarization (attitudes becoming more extreme), and proselytization (people's willingness to persuade others of their own opinion), (2) the specific kinds of moral values or moral content which lead to such negative outcomes, and (3) alternative ways in which advocacy might be encouraged (e.g. advocacy grounded in "practical" or economic arguments as opposed to moral arguments) to reduce the likelihood of such outcomes.

What Did the Researchers Do and Find?
We conducted an online survey in which 588 people were asked to write arguments for their position on migration and climate-change issues. People were randomly divided to advocate in moral or practical terms. We measured their attitudes pre and post-advocacy, and the confidence they placed in their advocacy attempt. We also calculated scores on the kinds of moral content expressed in their arguments, for example, the extent to which they expressed moral

values such as harm, care, purity, and emotions such as anger and disgust. Our findings showed that moral (versus practical) advocacy leads to people's attitudes becoming more extreme post-advocacy, and increased willingness to persuade others of one's opinion. This was due to increased expression of moral values, as opposed to other types of moral content, such as emotions, or confidence in advocating.

What Do These Findings Mean?

Our findings suggest that re-framing advocacy away from moral values, and towards more "objective" practical arguments such as, economic consequences, leads to less polarization and proselytization. This means that by encouraging practical advocacy, we might be able to minimize social conflict, while simultaneously retaining our right to free speech.

Journal of Social and Political Psychology, 2020, Vol. 8(2), 473–503, https://doi.org/10.5964/jspp.v8i2.1346

Received: 2019-10-30. Accepted: 2020-05-28. Published (VoR): 2020-09-02.

Handling Editor: Lucas A. Keefer, University of Southern Mississippi, Hattiesburg, MS, USA

*Corresponding author at: Melbourne School of Psychological Sciences, Level 12, Redmond Barry Building, Grattan Street, Parkville, Victoria, 3010, Australia. E-mail: ravini.abeywickrama@unimelb.edu.au

The US federal government shutdown in response to the election of a border wall around Mexico reflects the failure of society to make compromises on morally-charged issues (Delton, DeScioli, & Ryan, 2020; Ryan, 2017). The current political climate is undoubtedly one of disagreement and polarization, in which conflict between ideologically opposing groups is frequent. Indeed, partisan antipathy in the United States is at its highest since 1994, with both Republicans and Democrats now being more likely to report unfavorable views of the other party, and even perceive the other party's policies as a threat to the nation's well-being (Pew Research Center, 2014). An often cited cause of this social division is the increased prevalence of moral rhetoric as a foundation for political advocacy (Clifford, Jerit, Rainey, & Motyl, 2015; Tappin & McKay, 2019).

Despite the destructive effects of moral rhetoric, little is known about the mechanisms by which moral rhetoric leads to increased social division. Specifically, it is currently not known how relying on moral rhetoric during pro-attitudinal advocacy (advocating for one's own position; see Briñol, McCaslin, & Petty, 2012; Gordijn, Postmes, & de Vries, 2001), changes the advocate's *own* attitudes and communicative intentions. We propose that relying on moral rhetoric (compared to non-moral rhetoric) during pro-attitudinal advocacy leads to greater attitude polarization, and increased proselytization (persuading others of one's opinions), via *self-persuasion* (see Briñol et al., 2012; Tesser, 1978; Wilson, 1990).

Importantly, the current research explores the mechanisms through which pro-attitudinal advocacy leads to polarization and proselytization with the aim of uncovering how advocacy can be re-framed to reduce closed communicative practices. For this purpose, the current research contrasts "moral" versus "practical" expression of one's attitudes in the pursuit of a) validating a novel advocacy task aimed at changing the language people use to advocate for their views, b) exploring how the extent to which people rely on moral advocacy may polarize one's attitude post-advocacy, or increase one's desire to proselytize one's attitudes, and c) exploring four possible

mechanisms – moral value framing (individualizing and binding moral foundations), moral system framing (deontology vs consequentialist reasoning), emotional language use (moral outrage), and meta-cognitive confidence, which may account for polarization and proselytization.

The Problem With Moral Rhetoric

One potential source of political polarization is the introduction of moral rhetoric to justify attitudes. Extant research in moral psychology suggests that when an issue is construed as moral, it evokes moral emotions such as disgust or anger (Haidt, 2003; Tangney, Stuewig, & Mashek, 2007), and comes to be perceived as universal (Skitka, Bauman, & Sargis, 2005; van Bavel, Packer, Haas, & Cunningham, 2012) and beyond compromise (for a recent review see Rhee, Schein, & Bastian, 2019; Ryan, 2017; Skitka, 2010). Such changes in the construal of the moralized attitude affects interpersonal interaction, often causing individuals to try to distance themselves from alternative viewpoints (Frimer, Brandt, Melton, & Motyl, 2019; Wright, Cullum, & Schwab, 2008), and even perceive opinions on the relevant idea in question as more polarized (Anderson et al., 2014; Ryan, 2017), which can in turn lead to erosion of social trust (Rapp, 2016).

This work in moral psychology is consistent with work in standard persuasion (i.e. an external source delivering a persuasive message to a recipient; see Petty & Brinol, 2014 for a recent review), which suggests that moral framing increases polarization (Clifford, 2019; Clifford et al., 2015). For example, Clifford (2019) found that presenting participants with persuasive messages framed in moral terms (e.g. harm) led participants to state that they would be more upset if people close to them disagreed with their position on the issue.

In addition to polarization, research has also demonstrated that moral rhetoric is a strong motivator of proselytization. For example, in a study of articles published in six Chilean outlets, Valenzuela, Piña, and Ramírez (2017) found that news stories were more likely to be shared on social media when they were framed in moral terms, while they were less likely to be shared if framed in terms of economic consequences or conflict. Although morally framed issues are more likely to be shared among one's own ingroup, it has also been found that morally emotive rhetoric is less-likely to be shared with outgroup members, leading to the formation of attitudinal "echo-chambers" (Brady et al., 2017).

Thus, while moral framing of attitudes may be particularly effective in disseminating an attitude among like-minded others, and even mobilizing in-group members to advance a group's position (van Zomeren, 2013; van Zomeren, Postmes, & Spears, 2012; van Zomeren, Postmes, Spears, & Bettache, 2011) it can also be a roadblock to engagement in constructive discussion and compromise with those who hold opposing views.

Pro-Attitudinal Advocacy and Self-Persuasion

Despite research demonstrating the consequences of moral rhetoric on polarization and proselytization, less is understood about the mechanisms by which this occurs. To address this gap in understanding, we focus on the relationships between moral rhetoric, polarization, and proselytization, in the context of *self-generated* persuasion. Given that highly convicted attitudes tend to be difficult to change using standard persuasion (Krosnick & Petty, 1995), an alternative to standard persuasion that has been used over the last few decades (although not examined as frequently), is self-persuasion. While in standard persuasion the source and recipient are two separate entities, in self-persuasion, the message is generated and received by the same person (Briñol et al., 2012; Hovland, Janis, & Kelley, 1953; Wilson, 1990). Thus, while standard persuasion examines the consequences of an external persuasive message on a third-party recipient, self-persuasion examines the consequences of recipients gener-

ating persuasive messages *themselves.* This implies that when one advocates for one's position on a topic (i.e. pro-attitudinal advocacy), one's attitudes and communicative intentions may change due to self-persuasion. For example, one may become more polarized or entrenched in one's opinion towards migration, after advocating for one's position on this topic. We examine the mechanisms and consequences of moral versus practically-framed advocacy appeals in the context of pro-attitudinal advocacy, given its high ecological validity and prevalence in the current political climate.

To our knowledge, this is the first set of studies to explore the effects of moral versus practical advocacy in the context of self-persuasion occurring during pro-attitudinal advocacy. We sought to understand if a) people can self-persuade to take a more or less extreme position (polarize or depolarize) after advocating for their opinion, and b) even if attitudes do not change, how advocacy may inform subsequent communicative intentions, such as intentions to proselytize.

The second gap in the literature we aim to address is creating a specific advocacy paradigm (i.e. moral versus practical advocacy) that can be used in the context of self-persuasion to depolarize attitudes and reduce closed communicative intentions, such as intentions to proselytize. Previous research has shown effects of moral framing on attitude polarization, whereby moral framing leads to greater polarization and unwillingness to compromise (Clifford, 2019; Ryan, 2017). Although a growing body of literature now suggests that the use of moral language in rhetoric can lead to polarization, few studies have investigated methods of persuasion that may depolarize an already polarized attitude (see Feinberg & Willer, 2013 for an exception). Although many have proposed rational argumentation or non-moral framing as a potential solution to reducing polarization (Kovacheff, Schwartz, Inbar, & Feinberg, 2018), this is yet to be tested empirically in the context of advocacy. We suggest that one way to reduce polarization and encourage more open communication between ideological groups, is to change the way in which people communicate their own attitudes and beliefs. That is, we suggest that reframing one's attitudes in "practical" as opposed to moral terms is less likely to lead to depolarization, and lower intentions to proselytize one's attitudes.

Proposed Alternative: Practical Reasoning

Given the potential of moral rhetoric to stimulate social conflict, we explore an alternative method one can use to justify one's attitudes. We term this alternative "practical" reasoning. This type of reasoning is based on economic consequences and cost-benefit analyses. It is commonly described by economic theory, whereby the best course of action is the one that allocates scarce resources to the option that "maximizes expected utility" (Kahneman & Tversky, 1979; Simon, 1979; Tversky & Kahneman, 1981). Compared to moral reasoning, practical reasoning is arguably a more withdrawn, less convicted and more "objective" or value-free way of expressing one's attitudes.

The moral versus practical distinction is important, given that attitudes and arguments grounded in these two kinds of reasoning tend to yield different attitudinal outcomes (Leidner, Kardos, & Castano, 2018; Luttrell, Petty, Briñol, & Wagner, 2016; Luttrell, Phillip-Muller, & Petty, 2019; Wheeler & Laham, 2016). For example, research intersecting standard persuasion and political psychology, shows that presenting arguments grounded in moral values as opposed to those grounded in cost-benefit analyses are more effective in reducing support towards torture in the US (Leidner et al., 2018). Further, research in social psychology examining the consequences of attitude properties, shows that attitudes that are based on morality are more consequential in predicting behavior than those based on non-moral reasoning (Luttrell et al., 2016).

Morally framed messages and attitudes tend to be more persuasive, and likely lead to greater polarization and proselytization compared to non-moral attitudes, in the context of standard persuasion. We test whether this effect generalizes to the context of self-persuasion, and predict that the effect of moral advocacy is driven by the saliency of moral content contained in moral compared to non-moral messages. Unlike previous work examining the effects of attitude bases, or the effects of presenting moral versus non-moral messages, the focus of the current study is on moral versus *practical* message content generated by the advocate *themselves* during pro-attitudinal advocacy. To our knowledge, this is the first study contrasting these two types of message content in the context of self-persuasion. We predict that using moral arguments instead of practical arguments to justify one's attitudes will lead to greater polarization and increased intentions to proselytize one's attitudes, via four possible mechanisms: moral values (moral foundations), moral systems framing (deontology vs consequentialism), moral emotions (moral outrage), or confidence in one's advocacy attempt (meta-cognitive confidence).

Proposed Mechanisms

People use a variety of frames and appeals to both justify and promote their attitudes. A relatively common way in which individuals express their attitudes is through the use of moral rhetoric – that is, framing attitudes in terms of moral values (Luttrell et al., 2019). Another way of expressing one's attitudes is by relying on moral systems, that is, deontological reasoning (based on rules and principles) and consequentialist reasoning (based on outcomes or consequences; see Wheeler & Laham, 2016). Yet another possibility is relying on emotive language, specifically, "other-condemning" emotions such as anger and disgust, which signify moral outrage (see Haidt, 2003). Another possibility that does not capture the language of advocacy, but rather, the confidence placed on one's advocacy attempt, is *meta-cognitive confidence* (see Petty & Brinol, 2015 for a review). We discuss each of these four possibilities in turn.

Moral Foundations Theory

Over the last two decades, the field of moral psychology has made substantial advances in the conceptualization and study of mechanisms underlying moral decision making and moral values (Greene, 2015; Haidt, 2007). A powerful driver of such advances has been the development of new theoretical frameworks which seek to define psychological domains that underlie morality-specific cognition and decision-making (Haidt, 2007). Perhaps the most prominent of these accounts has been Moral Foundations Theory (MFT; Haidt & Graham, 2007) which posits that there are five fundamental domains of moral values which represent innate sources of moral intuitions present in all individuals – namely, harm/care, fairness/reciprocity, ingroup/loyalty, authority/respect, and purity/sanctity. A key tenet of MFT is that, although these five domains of moral intuition are present in all individuals, there may be substantial differences between societies, and subgroups within societies, in the degree to which a given domain may be emphasized or elaborated upon (Graham, Haidt, & Nosek, 2009; Haidt & Graham, 2007; Koleva, Graham, Iyer, Ditto, & Haidt, 2012).

Moreover, MFT proposes that political liberals place greater emphasis on harm/care and fairness/reciprocity (collectively the individualizing foundations), while political conservatives place emphasis on ingroup/loyalty, authority/respect, and purity/sanctity foundations (the binding foundations) (Graham et al., 2009; Haidt & Graham, 2007). For simplicity, we refer to the combination of these two foundation clusters as "moral expressiveness", which captures the extent to which one relies on both the individualizing and binding foundations to justify one's attitude.

Within the context of moral psychology research, MFT approaches have demonstrated particular utility in explaining political, ideological and cultural divides (Graham et al., 2009; Haidt & Graham, 2007). For example, Koleva et al. (2012) found that the degree to which individuals endorse certain moral foundations were strongly predictive of their position on a number of highly polarizing political issues such as same-sex marriage, euthanasia, and cloning. Indeed, many of these positions were more strongly predicted by the level of endorsement of certain foundations (especially purity/sanctity) than by political orientation or religious belief.

The finding that highly polarized positions on political issues tend to coincide with strong endorsement of certain moral foundations raises the possibility that moral foundations endorsement may in fact be an underlying mechanism for polarization. Supporting this idea, Mooijman, Hoover, Lin, Ji, and Dehghani (2018) found that the frequency of moral foundations terms appearing in Tweets sent during the 2015 Baltimore protests (Yan & Ford, 2015) could predict both hourly police arrest numbers in the Baltimore area, and whether the following day was likely to feature a violent or peaceful protest. This suggests that exposure to rhetoric containing moral foundations language led to more extreme endorsement of a given position. In the context of self-persuasion, this also suggests that moral, compared to practical advocacy is likely to lead to greater attitude polarization, and increased willingness to proselytize one's opinions, depending on the extent to which one relies on moral values to justify one's position.

Deontology vs Consequentialism

Research has also found that people may strategically or intuitively appeal to different types of moral systems when seeking to justify their positions related to moral issues. For example, Piazza and Sousa (2014) suggest that people are more likely to rely on *deontological* appeals (duties, principles or rules about right or wrong behavior), than *emotive* (elicitation of emotional reactions) or consequentialist reasoning (outcomes, consequences, or effects) in making moral judgements. Further, using content-based analyses, Wheeler and Laham (2016) found that participants were more likely to appeal to consequentialist reasoning when asked to justify their position on issues relevant to the individualizing foundations than the binding foundations, for example. These findings suggest that moral advocacy is more likely to encourage advocates to use words related to deontology (and potentially consequentialism) compared to practical advocacy.

Moral Emotions

Another possibility that we consider, is that moral advocacy leads to increased polarization and proselytization via greater expression of moral emotions. While moral messages tend to be perceived as more emotional compared to practical messages, in general (Brady et al., 2017; Luttrell et al., 2019), we specifically consider emotions conveying moral outrage such as anger and disgust (Haidt, 2003; Salerno & Peter-Hagene, 2013). Moral outrage refers to emotions that are expressed in response to a perceived moral norm violation (Crockett, 2017; Haidt, 2003; Salerno & Peter-Hagene, 2013). Extant research suggests that people may be particularly motivated to share their expressions of moral outrage via gossip, shaming and punishment as a means of signaling their moral character to other ingroup members (Jordan, Hoffman, Bloom, & Rand, 2016), and reaffirming ingroup norms (Crockett, 2017).

Consistent with this notion, recent research conducted by Brady et al. (2017), found that posts on Twitter regarding a morally laden issue (e.g. same-sex marriage, gun control, climate-change) were more likely to be retweeted when they contained words expressing moral outrage. It was also found that this effect of moral-emotion words on increased sharing was amplified among networks of individuals who shared the same ideological position as the initial Tweet. Taken together, such findings suggest that receiving moral rhetoric containing expressions of

moral outrage may increase proselytization, and also increase polarization by promoting increased message sharing within ideological echo chambers.

Meta-Cognitive Confidence

A final possibility is that moral versus practical advocacy generates different levels of *meta-cognitive confidence.* Meta-cognitive confidence captures the confidence one places in one's thoughts and judgements (see Wagner, Briñol, & Petty, 2012, for a review). For example, one may perceive one's arguments to be strong or weak (Briñol et al., 2012), perceive oneself to have expended high or low effort in generating arguments (Briñol et al., 2012), or perceive oneself to be a knowledgeable or an unknowledgeable source on the topic (Ehret, Van Boven, & Sherman, 2018; Rios, Goldberg, & Totton, 2018). Meta-cognitive confidence is important because it has been shown to be a key driver of polarization not only in the context of self-persuasion more generally (see Clarkson, Tormala, & Leone, 2011; Clarkson, Valente, Leone, & Tormala, 2013), but more specifically during self-persuasion occurring as a result of pro-attitudinal advocacy. Specifically, increasing meta-cognitive confidence as indexed by perceived argument quality and effort, has been shown to lead to greater polarization following pro-attitudinal advocacy (Briñol et al., 2012).

Moreover, research from standard persuasion provides evidence for a "moral-matching" effect, such that, those who possess attitudes grounded in moral concerns are more likely to be persuaded by moral messages (Luttrell et al., 2019). This suggests that attitude basis may moderate the extent to which we observe polarization, due to match-induced processing fluency (Mayer & Tormala, 2010). That is, matching effects may increase the ease of message processing, increase meta-cognitive confidence and thus polarization post-advocacy.

Drawing on this literature, we hypothesize that asking one to advocate for one's position using moral as opposed to practical concerns, likely leads to increased meta-cognitive confidence in one's attempt. Given the emphasis placed on the universality and correctness on one's own moral values, advocates are more likely to experience processing ease and fluency when justifying their position in moral terms. Moral values are highly sacralized and ingrained in our minds, and we use these as a basis to view the world (Kovacheff et al., 2018). People tend to think that their own moral values are both factual and universal (Tetlock, 2003), and thus may have difficulty understanding worldviews which deviate from one's own (Feinberg & Willer, 2013, 2015). Therefore, justifying one's position using moral reasons likely leads to increased meta-cognitive confidence, which may translate into attitude polarization, and/or greater willingness to persuade others of one's own opinion.

The Current Research

The main aim of this research is to explore whether moral versus practical advocacy changes levels of moral language and advocacy-related confidence during pro-attitudinal advocacy related to two compelling socio-political issues: migration and climate-change. One possibility is that moral advocacy increases grounding in foundational moral values. Another is that moral advocacy increases grounding in broader moral frameworks (e.g., deontology vs. consequentialism). A third possibility is that moral advocacy increases emotional grounding, or expression of moral outrage. Yet another possibility is that moral advocacy simply increases confidence in one's advocacy attempt. Thus, we have four possible accounts of the influence of moral advocacy on polarization, which we explicitly test using a new paradigm and text-based analysis.

In Study 1, we consider the extent to which moral versus practical advocacy influences (a) foundational value expression, (b) moral framework language, and (c) moral outrage. In Study 2, we consider the extent to which

these factors account for any effects of framing on polarization, by additionally examining the effects of meta-cognitive confidence. In Study 3, we extend our dependent variables to include communicative practices, such as, intentions to proselytize one's opinions. Across the three studies, we seek to explore which of the four possible mechanisms likely accounts for the effects of moral advocacy on social division.

Study 1

Method

Participants and Design

Two-hundred and fifty-five US residents were recruited via Amazon Mechanical Turk (M = 40.71, SD = 12.45, female = 138). An a priori power analysis using G*Power 3.1 revealed that this sample size was sufficient to detect small-medium effects (f = .20) typically found in social psychology (Richard, Bond, & Stokes-Zoota, 2003), with 80% power in an ANCOVA (Condition: Moral vs Practical) including two covariates (initial stance, political orientation) (Faul, Erdfelder, Lang, & Buchner, 2007).

Materials and Procedure

Participants completed a survey online via Qualtrics (Qualtrics, 2016). Participants completed an attitude questionnaire assessing favorability towards various issues (e.g. carbon emissions, consumer behavior). Our issue of interest was one relating to migration (see the Supplementary Materials, Appendix A for detailed attitude issue description).

Initial stance — Participants were asked to indicate their attitude towards migration using the following item, "Compared to the number of migrants the government usually takes, how many migrants do you think the US government should take?" on a 7-point scale (1 = *much fewer*, 7 = *a lot more*, M = 4.07, SD = 1.79). Those who indicated scale-point 1-3 were classified as initial stance = Anti (coded -1; n = 77), 4 = Neutral (coded 0; n = 78) and 5-7 as initial stance = Pro (coded 1; n = 100).

Following initial attitude measurement, participants were randomly allocated to write advocacy appeals grounded in either moral (n = 123), or practical concerns (n = 132). The instructions for each appeal type can be found in the Supplementary Materials (Appendix B).

Demographics — Finally, participants were presented with a few demographic questions (age, gender, political orientation). Our primary measure of political orientation was a single 7-point Likert scale (1 = *very liberal*, 7 = *very conservative*; M = 3.69, SD =1.86, adapted from Feinberg & Willer, 2015; van Leeuwen & Park, 2009). A secondary political orientation measure was party affiliation with four choice options (Republican party = 34%, Democratic party = 50%, Libertarian party = 4%, other = 12%).

Characterizing Moral Content

Traditionally, moral content analyses have often been facilitated by word frequency-based approaches such as the Moral Foundations Dictionary (MFD; Graham et al., 2009) which identifies a number of keywords that are representative of each of the five foundations. Using moral foundations theory as a criterion for the presence of moral content in text can be a particularly powerful tool, as it allows moral psychology research to be extended

into the study of freely generated text, which is common in online social-media platforms such as Twitter. Indeed, this method has been found to be highly effective in identifying moral content (Sagi & Dehghani, 2014a, 2014b), and also in predicting real-world moral behavior (e.g. Mooijman et al., 2018).

Thus, to measure the moral content of participants' text responses, we used distributed dictionary representations (Garten et al., 2018) covering different kinds of moral content (described below) (see Garten, Boghrati, Hoover, Johnson, & Dehghani, 2016; Sagi & Dehghani, 2014a, 2014b, for more examples). Distributed dictionaries perform a similar function to the more commonly-used raw frequency-based content analysis approach implemented in LIWC (Tausczik & Pennebaker, 2010), in that they measure the presence of pre-defined semantic content domains in a set of texts. However, distributed dictionaries differ from LIWC analyses in that they do not rely on raw word counts (i.e., the presence of keywords in texts). Rather, they rely on the overall semantic similarity of texts to the provided keywords, providing an arguably more sensitive measure of semantic content. Specifically, distributed dictionaries measure semantic content using a vector space model, derived from a large training corpus, that represents words as points in multidimensional space, such that words with similar meanings (e.g., chair, seat) are closer together in that semantic space (i.e., have similar values on the model dimensions) by virtue of the fact that those words tend to occur in similar linguistic contexts in the training corpus. As in Garten et al. (2018), we operationalized semantic similarity as the cosine similarity (ranging from -1 to 1) of two vectors representing (1) a given participants' text response, obtained by computing the average vector for all the words in the response, and (2) a dictionary, obtained by computing the average vector for all the words in a given dictionary category (e.g., the negative pole of the care foundation)[i].

We created dictionary-based measures for three different kinds of moral content as follows. First, we measured semantic similarity of participants' responses to each of the five moral foundations, using dictionaries from previous research, where the positive and negative poles of each moral foundation are represented by four keywords (e.g., the words *suffer*, *cruel*, *hurt* and *harm* representing the negative pole of the care foundation; Garten et al., 2018; Hoover, Johnson, Boghrati, Graham, & Dehghani, 2018). Second, we developed additional moral content dictionaries based on the research of Wheeler and Laham (2016), covering specific moral framings (consequentialist and deontological), and third, capturing the expression of three different moral emotions relating to moral outrage (contempt, anger, disgust). This served to test the kind of moral language which primarily drives the effects of moral versus practical advocacy on polarization and proselytization (see Supplementary Materials, Appendix C for more information on how the dictionaries were constructed).

Indexing moral expressiveness — Moral expressiveness was conceptualized using the two broad categories of the five moral foundations: individualizing and binding foundations. A composite score on the individualizing foundations was created by averaging across *z*-scores of the following dictionary categories: harm, care, fairness, and cheating. A composite score on the binding foundations was created by averaging across *z*-scores of loyalty, betrayal, authority, subversion, purity and degradation. The scores on these measures represent semantic similarity of the text to the concept, such that a score of 0 means average similarity to the concept, negative scores imply relative dissimilarity, and positive scores indicate greater similarity.

Results and Discussion

Differences Between Conditions

A principal components analysis was first conducted to determine if the three moral emotions theorized (anger, contempt, disgust) load onto one component (i.e. a moral outrage component). The analysis revealed that all three moral emotions loaded onto a single component explaining 87% of the variance. This component represents *moral outrage* (component loadings: anger = .90, contempt = .94, disgust = .96), such that higher scores reflect greater moral outrage following one's advocacy attempt.

In order to assess differences between conditions on levels of moral outrage, moral expressiveness, and moral systems framing, a series of One-Way ANCOVAs were conducted with one between-subjects factor (Condition: Moral vs Practical) with initial stance and political orientation as covariates. Significant omnibus tests were followed by Bonferroni-corrected tests of multiple comparisons. Results revealed that our advocacy manipulation was successful - the moral and practical conditions significantly differed on most moral content tested (see Table 1 below). Almost all significant differences represent medium to very large effects according to classification of effect sizes (η_p^2 small = .01, medium = .06, large = .14; Cohen, 2013), indicating substantial differences in the language used across the two advocacy conditions. Correlations between all variables are included in Table 2.

Table 1

Means and SEs for All Dictionary Scores as a Function of Condition (Moral Versus Practical), Controlling for Initial Stance and Political Orientation in Study 1

	Moral		Practical		*F*	
Moral foundation	***M***	***SE***	***M***	***SE***	**(*df* = 251)**	η_p^2
Individualizing	0.29	0.07	-0.27	0.06	37.59**	.13
Binding	0.39	0.06	-0.37	0.07	73.76**	.23
Deontology	0.25	0.08	-0.24	0.08	16.73**	.06
Consequentialism	-0.08	0.09	0.08	0.09	1.67	.01
Moral outrage	0.45	0.08	-0.42	0.08	58.72**	.19

*p < .05. **p < .001. Two-tailed tests.

Table 2

Pearson's Correlations Between Continuous Variables in Study 1

Variable	IS	PO	MO	CONS	DNT	BIND	IND
IND	-.03	-.14*	.52*	.35**	.80**	.87**	–
BIND	.00	-.13*	.52**	.15*	.77**	–	
DNT	-.01	-.14*	.12*	.55**	–		
CONS	.04	.02	-.25**	–			
MO	.01	.03	–				
PO	-.41**	–					
IS	–						

Note. Shorthand notation: Individualizing foundations = IND; Binding foundations = BIND; Deontology = DNT; Consequentialism = CONS; Moral outrage = MO; Political orientation = PO, Initial stance = IS.

*p < .05. **p < .01. Two-tailed tests.

The results of Study 1 demonstrate that explicitly requesting advocates to frame pro-attitudinal advocacy in practical versus moral language changes the moral content of their appeal. Using a novel advocacy paradigm, we show that encouraging advocates to frame their appeal in moral versus practical terms increases the level of moral expressiveness, moral outrage, and deontological framing used to construct one's appeal relating to migration attitudes[ii]. Despite differences in average levels of moral content, we do not know if these differences are consequential in predicting attitudes or communicative intentions post-advocacy.

We attempt to replicate our findings in Study 2 with a few key modifications: a) using a different controversial attitude issue (i.e. issue relating to climate-change), b) a different sample population (college students versus MTurk population), c) including measures of pre and post-advocacy attitudes to assess attitude polarization as a result of advocacy, and d) to examine which of the following mechanisms: moral expressiveness, moral systems framing, moral emotions, or meta-cognitive confidence drives attitude polarization.

Study 2

Method

Participants and Design

Two-hundred and fifty five psychology undergraduates from an Australian university were recruited in exchange for course credit (M = 19.34, SD = 2.10, female = 164, other = 3). While two-hundred participants would have been sufficient to detect small-medium effects (f = .20) with 80% power in an ANCOVA (Condition: Moral vs Practical) with two covariates (initial stance, political orientation), we oversampled to allow for potential exclusions. The sample size was also sufficient to detect small-medium effects (r = .20) in two-tailed correlational analyses with 80% power (Faul et al., 2007).

Materials and Procedure

Similar to Study 1 participants completed the study on Qualtrics (Qualtrics, 2016). Participants were first asked to indicate their Time 1 attitudes and attitude basis.

Initial stance — Participants were presented with information on a hypothetical carbon emissions policy which was modelled based on energy-saving recommendations in Australia (Department of Environment and Energy, 2018). Participants were asked how favorable they felt towards a policy specifying a maximum thermostat temperature of 20 degrees Celcius during winter on a 7-point Likert scale (1 = *extremely unfavorable*, 7 = *extremely favorable*, M = 4.18, SD = 1.79; see Supplementary Materials, Appendix A for full description of attitude issue). Neutral participants (scale point = 4, n = 37) were excluded from the study, leaving a final sample of N = 218 (Anti = 97, Pro = 121). This is because attitude (de)polarization cannot be conceptualized and calculated clearly for those who do not indicate an initial preference (we elaborate on the calculation of attitude depolarization scores below).

Attitude basis — Two items were then used to assess the extent to which participants' attitudes were based on moral and practical concerns ("To what extent is your attitude towards migration based on moral [practical] concerns?", 1 = *not at all based*, 7 = *very much based*; adapted from Teeny & Petty, 2018; moral concerns M = 4.33, SD = 1.74; practical concerns M = 5.21, SD = 1.49). This was to see if the "moral-matching effect" found in standard

persuasion generalizes to the context of self-persuasion (although not a key hypothesis), whereby moral messages should be more persuasive for those whose attitudes are based on moral concerns (see Luttrell et al., 2019).

Advocacy task — Participants were randomly allocated to either moral (n = 108) or practical advocacy conditions (n = 110), and presented with the same instructions as Study 1 with minor re-wording to account for the new attitude issue.

Thought-listing — Participants listed up to eight thoughts they had in response to generating their arguments (taken from Clark, Wegener, Sawicki, Petty, & Briñol, 2013) before listing their post-advocacy attitude.

Time 2 attitude — Participants indicated their Time 2 attitude towards the carbon emissions policy (M = 4.46, SD = 1.82). Next, participants completed the following meta-cognitive measures related to their arguments in a randomized order: perceived effort, perceived argument quality and source credibility.

Perceived effort — Participants indicated how much effort they expended in generating their arguments on three 9-point scales, e.g. “How much energy did you put into generating your arguments?", taken from (Briñol et al., 2012). Responses were averaged to create a composite measure of perceived effort (M = 5.58, SD = 1.41, Cronbach’s α = .84).

Perceived argument quality — Participants rated how strong they perceived they own arguments to be on three 9-point scales (e.g. “How strong are the arguments that you generated?”; adapted from Briñol et al., 2012). Responses were averaged to create a reliable composite score (M = 5.20, SD = 1.67, Cronbach’s α = .88).

Self-perceived knowledgeability — Participant knowledgeability on the carbon emissions issue was assessed using four 9-point scales (“How knowledgeable do you feel about energy policies?”, “How much information do you feel you have about energy policies?”, “To what extent do you feel you have expertise on energy policies?”, “To what extent do you feel you know the most relevant facts about energy policies?” where 1 = *not at all*, 9 = *extremely*; adapted from Rios et al., 2018). Responses were averaged to create a composite score (M = 3.49, SD = 1.66, Cronbach’s α = .92).

Political orientation — Finally, participants responded to a demographics questionnaire assessing political orientation on the continuous measure (M = 3.55, SD = 1.22), and the categorical measure (Australian Liberal party [equivalent to Republicans in the US] = 36%, Australian Labor party [equivalent to Democrats] = 24%, Greens = 22%, other = 18%).

Results and Discussion

Differences Between Conditions

Similar to Study 1, a principal components analysis revealed that all three moral emotions (anger, contempt, disgust) loaded onto a single moral outrage component explaining 93% of the variance (component loadings: anger = .97, contempt = .96, disgust = .96), such that higher scores reflect greater moral outrage following one’s advocacy attempt.

Next, we proceeded to examine any differences in meta-cognitions between the two conditions. As with previous research (see Briñol et al., 2012), the meta-cognitive variables tested (perceived effort, argument quality, and

source credibility) were moderately correlated (.32 < *r* < .50) and thus we performed a principal components analysis to first determine the separability of these measures. The analysis revealed that all three variables loaded onto a single component explaining 62% of the variance. This component represents general *meta-cognitive confidence* in one's advocacy attempt (component loadings: argument quality = .46, effort = .42, source credibility = .40), such that higher scores reflect greater meta-cognitive confidence in one's advocacy attempt. Component scores (regression method) from the principal components analysis were subjected to an ANCOVA as above.

Similar to Study 1 between-subjects ANCOVAs (Condition: Moral vs Practical) were conducted with initial stance and political orientation as covariates. Replicating the results of Study 1, moral expressiveness significantly differed between the two conditions (see Table 3), indicating that our manipulation is generalizable across attitude issues and sample populations. Similar to Study 1, the two conditions also differed on other moral content, and also on moral outrage. Interestingly, meta-cognitive confidence did not vary between the two advocacy conditions. This suggests that meta-cognitive confidence is unlikely to drive between-condition differences on attitude polarization. Correlations between all variables are indicated in Table 4 below.

Table 3

Means and SEs for All Dictionary Scores as a Function of Condition (Moral Versus Practical), Controlling for Initial Stance and Political Orientation in Study 2

	Moral		Practical		*F*	
Moral foundation	***M***	***SE***	***M***	***SE***	**(*df* = 214)**	η_p^2
Individualizing	0.31	0.07	-0.23	0.07	29.40**	.12
Binding	0.34	0.07	-0.25	0.07	38.36**	.15
Deontology	0.23	0.09	-0.14	0.09	9.33*	.04
Consequentialism	-0.01	0.09	0.07	0.09	0.40	.00
Meta-cognitive confidence	0.01	0.10	-0.01	0.10	0.03	.00
Moral outrage	0.34	0.09	-0.33	0.09	28.31**	.12

*p < .05. **p < .001. Two-tailed tests.

Table 4

Pearson's Correlations Between Continuous Variables in Study 2

Variable	**POL**	**IS**	**PO**	**MO**	**MCC**	**CONS**	**DNT**	**BIND**	**IND**
IND	.12	-.11	-.01	.61**	.08	.61**	.86**	.88**	–
BIND	.17*	-.12	-.04	.45**	.11	.61**	.85**	–	
DNT	.16*	-.14*	-.09	.34**	.05	.76**	–		
CONS	.16*	-.02	-.14*	-.05	.06	–			
MCC	.12	.06	.00	.07	–				
MO	.03	-.13	.16*	–					
PO	-.14*	-.04	–						
IS	.25**	–							
POL	–								

Note. Shorthand notation: Individualizing foundations = IND; Binding foundations = BIND; Deontology = DNT; Consequentialism = CONS; Meta-cognitive confidence = MCC; Moral outrage = MO; Political orientation = PO; Initial stance = IS; Polarization = POL.
*p < .05. **p < .01. Two-tailed tests.

Predictors of Attitude Polarization

Polarization scores for each participant were calculated by multiplying the direction of change (Time 2 attitude in the direction of Time 1 attitude = 1, Time 2 attitude in the opposition direction to Time 1 = -1, Time 1 and Time 2 attitudes unchanged = 0) by the number of scale points moved from Time 1 to Time 2 (maximum 6 scale points). Polarization scores were created based on seminal research investigating (de)polarization in self-persuasion paradigms (see Tesser & Leone, 1977). This method is advantageous compared to pre-post difference scores, because it takes into account both the direction and extremity of attitude change, and thus provides a more complete picture of attitude change which may occur following advocacy interventions. For example, someone who initially indicated *slightly unfavorable* towards the policy at Time 1 (scale point = 3), and indicated *extremely unfavorable* towards the policy at Time 2 (scale point = 1) would have a polarization score of +2 because they changed two scale points and became more extreme in their initial position (i.e. polarized). On the other hand, someone who indicated extremely unfavorable towards the policy at Time 1 (scale point = 1), and indicated slightly unfavorable towards the policy at Time 2 (scale point = 3), would have a polarization score of -2, because they changed by two scale points but became less extreme in their initial position (i.e. depolarized).

Interestingly, the two conditions (moral and practical) did not significantly differ on average polarization scores, $t(216) = -1.41$, $p = .16$. Despite the absence of a direct effect we tested the indirect effect following recommendations by previous researchers (O'Rourke & MacKinnon, 2018; Rucker, Preacher, Tormala, & Petty, 2011). In order to test our prediction that moral versus practical advocacy is consequential, we created four separate mediation models to test which of the following best accounts for variation in polarization: moral expressiveness, moral systems, moral emotions, or meta-cognitive confidence. Each model contained Condition as the independent variable, and two covariates: initial stance and political orientation.

To test the effects of moral expressiveness on polarization, we ran a mediation model using 5000 bootstrap samples and 95% confidence intervals (Hayes, 2013), with individualizing and binding foundation scores as simultaneous mediators. Results revealed a significant total indirect effect (IE) via both individualizing and binding foundations as indicated by partially standardized regression coefficients[iii] (IE = .09, *SE* = .04, 95% CI [.02, .16]; see Figure 1 below)[iv]. Corroborating the regression analysis, the total indirect effect was mainly driven by the significant indirect effect via the binding foundations (IE = .14, *SE* = .07, 95% CI [.01, .27])[vvi].

Next, we tested the second proposed explanation for moral versus practical advocacy effects: moral systems framing (deontological and consequentialist reasoning). Results revealed no significant total indirect effect via both deontology and consequentialism on polarization, IE = .03, *SE* = .03, 95% CI [-.04, .10]. The indirect effects via deontology only IE = .03, *SE* = .03, 95% CI [-.02, .10] and consequentialism only were also not significant, IE = .00, *SE* = .01, 95% CI [-.03, .02]. We then tested the third proposed explanation for our effects: moral outrage. Despite between-condition differences in moral outrage, results revealed no significant indirect effect via moral outrage on polarization, IE = .02, *SE* = .03, 95% CI [-.04, .09]. Given that meta-cognitive confidence did not vary between conditions, we did not formally test this as a mechanism in a mediation model.

Taken together, the findings of Study 2 suggest that relying on moral as opposed to practical advocacy leads to increased attitude polarization via increased moral expressiveness only, and not other advocacy-relevant constructs, such as moral systems, moral outrage, or meta-cognitive confidence. Conversely, the findings suggest that using practical advocacy appeals attenuates attitude polarization by decreasing moral expressiveness.

In Study 3 we sought to explore other downstream consequences of moral versus practical appeals by a) extending our dependent variables to include intentions to proselytize, and b) testing our predictions on the original attitude issue and sample population.

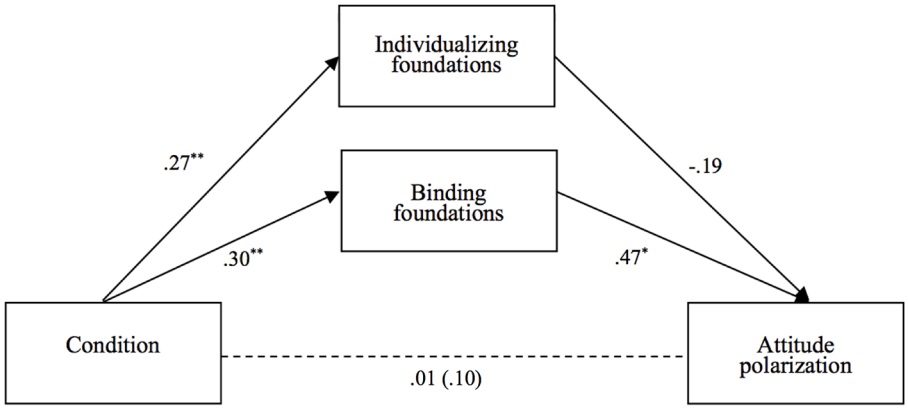

Figure 1. Indirect effect of Condition (moral vs practical) on attitude polarization via both indices of moral expressiveness (coding: Moral = 1, Practical = -1).

Note. The standardized coefficient within brackets represents the total effect of Condition on attitude polarization with covariates included in the model, but without the mediators. The model controls for initial stance and political orientation but these are omitted from the diagram for clarity of presentation.

*$p < .05$. **$p < .01$.

Study 3

Method

Participants and Design

A hundred and sixty-five participants were recruited from Amazon Mechanical Turk ($M = 35.19$, $SD = 10.62$, female = 82). This sample size was sufficient to detect the smallest effect size found in Study 1 (equivalent to $f = .30$) in One-Way ANCOVA's including two groups (Moral vs Practical) and two covariates (initial stance and political orientation) with 80% power.

Materials and Procedure

Participants followed a similar procedure to Study 1, with a few exceptions detailed below. Participants were presented with the attitude issue in Study 1. They then indicated their attitude on the same 7-point scale: "Compared to the number of migrants the government usually takes, how many migrants do you think the US government should take?"; 1 = *much fewer*, 7 = *a lot more; M* = 4.27, *SD* = 1.79). Those who were neutral were excluded from the study ($n = 50$) leaving a final sample of $N = 115$ (Anti = 38, Pro = 77). Participants indicated the extent to which their attitude was based on moral concerns ($M = 5.63$, $SD = 1.56$), and practical concerns ($M = 5.29$, $SD = 1.59$).

Participants were then randomly allocated to advocate for their position on the issue in practical terms ($n = 55$) or moral terms ($n = 60$) using the same task instructions used in Study 1. Participants then completed a thought-

listing task prior to indicating their Time 2 attitude (*M* = 4.33, *SD* = 2.24). Next, participants completed the same meta-cognitive measures as in Study 2: perceived argument quality (*M* = 7.12, *SD* = 1.64, Cronbach's α = .89), perceived effort (*M* = 7.99, *SD* = 1.11, Cronbach's α = .82), and perceived source credibility (*M* = 5.37, *SD* = 1.85, Cronbach's α = .93). Participants then completed the following measures in a randomized fashion: intentions to persuade others, perception of audience persuadability.

Intentions to persuade others — Participant intentions to persuade others was assessed using five 9-point scales. Three items were taken from previous advocacy research (e.g. "How likely would you be to persuade your [family, friends, stranger] of your own position on this topic?", taken from Cheatham and Tormala (2015); *M* = 4.96, *SD* = 2.22, Cronbach's α = .87). We added two more items to measure intentions to persuade the opposition, specifically ("How likely would you be to persuade someone who disagrees with your position on this topic?", "How likely would you be to persuade someone who holds the opposite opinion to you on the issue of migrant intake?"; *M* = 4.93, *SD* = 2.41, Cronbach's α = .90).

Perceptions of audience persuadability — Participant's perception of their advocacy appeal being effective in persuading others was assessed using a 7-point bipolar scale, "In relation to your own attitude toward the issue of migrant intake, to what extent do you think that the arguments you generated will change your target audience's attitude?" (anchored at -3 = move as far away as possible from your possible, 0 = unpersuaded/retain original position, +3 = move as close as possible to your position). This was recoded into a unipolar scale (scale points 1 to 7; *M* = 4.82, *SD* = 0.91).

Finally, participants completed the continuous (*M* = 3.50, *SD* = 1.95) and categorical measures of political orientation (Republican party = 25%, Democratic party = 57%, Libertarian party = 4%, other = 14%).

Results and Discussion

Differences Between Conditions

Corroborating the results of Study 1 and Study 2, a principal components analysis revealed that all three moral emotions (anger, contempt, disgust) loaded onto a single component explaining 92% of the variance (component loadings: anger = .95, contempt = .94, disgust = .98). Similar to Study 2, a principal components analysis revealed that all three meta-cognitive variables (argument quality, effort, source credibility) loaded onto a single component explaining 57% of the variance (component loadings: argument quality = .92, effort = .40, source credibility = .85).

Again replicating our results in Study 1 and Study 2, the two conditions differed on moral expressiveness, such that the moral condition generated higher scores on the individualizing and the binding moral foundations compared to the practical condition, controlling for initial stance and political orientation (see Table 5). Interestingly, unlike in Study 2, moral outrage did not differ between the moral and practical conditions, while unlike in Study 2, meta-cognitive confidence significantly differed between the two advocacy conditions. Corroborating the results of Study 1 and Study 2, the two conditions varied on deontology, but not on consequentialism, although this difference remains small across all studies. Correlations between all variables can be found in Table 6 below.

Table 5

Means and SEs for All Dictionary Scores as a Function of Condition (Moral Versus Practical), Controlling for Initial Stance and Political Orientation in Study 3

	Moral		Practical		*F*	
Moral foundation	***M***	***SE***	***M***	***SE***	**(*df* = 111)**	η_p^2
Individualizing	0.31	0.08	-0.28	0.08	27.56**	.20
Binding	0.43	0.09	-0.50	0.09	56.47**	.34
Deontology	0.17	0.13	-0.23	0.12	5.64*	.05
Consequentialism	-0.12	0.17	0.17	0.13	2.50	.02
Meta-cognitive confidence	0.19	0.13	-0.21	0.13	4.59*	.04
Moral outrage	0.01	0.13	-0.01	0.13	0.00	.00

*p < .05. **p < .001. Two-tailed tests.

Table 6

Pearson's Correlations Between Continuous Variables in Study 3

Variable	**PRO**	**POL**	**IS**	**PO**	**MO**	**MCC**	**CONS**	**DNT**	**BIND**	**IND**
IND	.32**	.04	.06	.15	-.00	.20*	.48**	.71**	.77**	–
BIND	.31**	.01	.02	.11	-.03	.19*	.28**	.68**	–	
DNT	.34**	.08	-.19*	.32**	-.16	.21*	.59**	–		
CONS	.20*	.17	-.08	.17	-.12	.00	–			
MCC	.49**	.11	-.14	.18*	.07	–				
MO	.03	.08	.24*	-.15	–					
PO	.09	.26**	-.59**	–						
IS	.06	.06	–							
POL	.10	–								
PRO	–									

Note. Shorthand notation: Individualizing foundations = IND; Binding foundations = BIND; Deontology = DNT; Consequentialism = CONS; Meta-cognitive confidence = MCC; Moral outrage = MO; Political orientation = PO; Initial stance = IS; Polarization = POL; Proselytization = PRO.

*p < .05. **p < .001. Two-tailed tests.

Note, unlike in Study 1, although an overall regression model containing moral expressiveness, condition, initial stance, and political orientation was significant, R^2 = .13, $F(5, 109) = 3.35$, p = .007, neither of the moral expressiveness indices predicted polarization (individualizing: β = -.02, p = .87; binding: β = -.05, p = .75). We consider explanations for this null effect in the general discussion.

Predictors of Proselytization

We now turn to our primary dependent variable in Study 3: proselytization intentions. Given a) the conceptual similarity between intentions to persuade others, and perceptions of audience persuadability, and b) our desire to create a more parsimonious measure of intentions to proselytize, we conducted a principal components analysis to determine if these items were indeed tapping into one construct. The analysis revealed a single component explaining 74% of the variance (.76 < component loadings < .93). Component scores (regression method) were used in the analyses that follow. This component collectively represents the desire to persuade others of one's

position, and captures the belief that others will be persuaded (hereafter, known as proselytization intentions). The two advocacy conditions significantly differed on average proselytization scores, $t(113) = -2.15$, $p = .03$, such that those in the moral condition were more likely to express intentions to proselytize their opinions ($M = 0.19$, $SD = 1.06$), compared to the practical advocacy condition ($M = -0.21$, $SD = 0.90$).

In order to test our prediction that advocacy condition indirectly influences intentions to proselytize via moral expressiveness, we conducted a mediation analysis on Process (Hayes, 2013) using 5000 bootstrap samples and 95% confidence intervals. Results revealed that the total indirect effect via both binding and individualizing foundations (combined effect) was significant, IE = .16, $SE = .07$, 95% CI [.03, .30]; see Figure 2[vii]. The individual indirect effects via individualizing foundations only, IE = .07, $SE = .07$, 95% CI [-.06, .21] and the binding foundations only were not significant, IE = .08, $SE = .09$, 95% CI [-.10, .28].

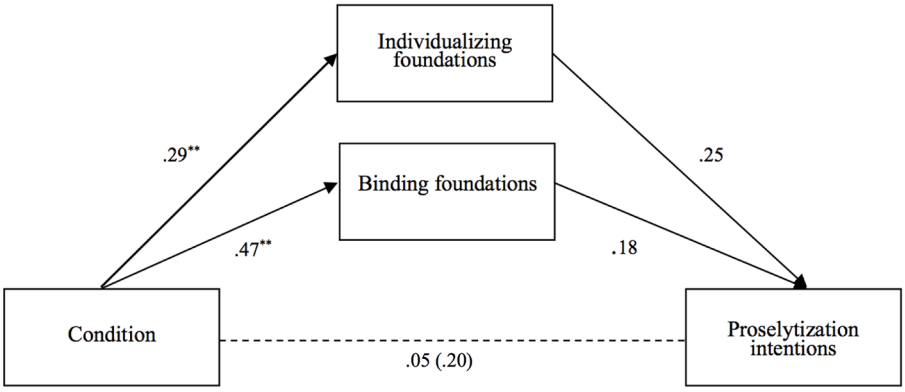

Figure 2. Indirect effect of Condition (moral vs practical) on proselytization intentions via both indices of moral expressiveness. *$p < .05$. **$p < .01$.

In terms of the moral systems account, results revealed no significant total indirect effect via both deontology and consequentialism on proselytization, IE = .05, $SE = .04$, 95% CI [-.03, .14]. The indirect effect via deontology only was marginally significant, IE = .06, $SE = .03$, 95% CI [.01, .13] while the indirect effect via consequentialism only was not significant, IE = -.01, $SE = .02$, 95% CI [-.05, .03]. We did not test moral outrage as an explanation, given the absence of between-condition differences. Finally, corroborating condition-level differences in meta-cognitive confidence, there was a marginally significant indirect mediation via meta-cognitive confidence on proselytization, IE = .09, $SE = .05$, 95% CI [.01, .19].

Taken together, the results of Study 3 suggest that moral advocacy may increase proselytization either via moral expressiveness, deontological framing, or meta-cognitive confidence, and not via consequentialism, or moral outrage. Across all three studies, moral expressiveness tends to emerge as the most reliable mediator of the effects of moral advocacy on polarization and proselytization.

Mini Meta-Analysis

A post-hoc "mini meta-analysis" was conducted across Study 2 and Study 3 following recommendations by Goh, Hall, and Rosenthal (2016), to examine the relationships between each of the two foundation clusters and polar-

ization. Given the similarity in study design, two fixed (versus random) effect meta-analyses were conducted. The first mini meta-analysis was to assess the correlation between the individualizing foundations and polarization (Study 2 r = .12; Study 3 r = .04) and the second was to assess the correlation between binding foundations and polarization (Study 2 r = .17; Study 3 r = .01). Correlations were transformed to Fisher's z, weighted by sample size (Study 2 N = 218; Study 3 N = 115, weight formula = N – 3), and then converted back to correlations for ease of interpretation.

Consistent with the results of the mediation analyses, the relationship between the binding foundations and polarization was significant, overall, Mean r = .12, 95% CI [.01, .22], while the relationship between the individualizing foundations and polarization was not, Mean r = .09, 95% CI [-.02, .20].

General Discussion

Across three studies, we successfully validated a novel advocacy paradigm (moral versus practical advocacy), aimed at understanding how moral language and confidence in advocacy influences post-advocacy attitudes and communicative intentions. While research in standard persuasion suggests that moral framing increases polarization and proselytization, little work has considered why this occurs. Thus, the current research examined the mechanisms through which moral rhetoric may change one's attitudes or communicative intentions following pro-attitudinal advocacy. We tested four possible mechanisms – moral expressiveness (reliance on the five moral foundations), moral systems framing (reliance on deontology vs consequentialist reasoning), increased emotional language use (moral outrage), and confidence in one's advocacy attempt (meta-cognitive confidence). The findings suggest that moral expressiveness is the key mechanism through which morally-framed advocacy attempts lead to polarization and proselytization. In particular, the binding foundations appears to account for the relationship between advocacy condition and polarization across both studies.

To our knowledge, this is the first study to examine the self-persuasive effects of moral versus practical messages in the context of pro-attitudinal advocacy. Our findings support previous work showing that moral rhetoric is persuasive, leads to increased proselytization, decreases willingness to compromise (Clifford, 2019; Ryan, 2017; Valenzuela et al., 2017), leads to the formation of echo chambers (Brady et al., 2017), and even leads to blatant discrimination against those holding opposing views (Pew Research Center, 2016). Our findings suggest a potential alternative to reduce polarization and proselytization: advocacy grounded in practical reasoning. We show that advocacy appeals grounded in practical argumentation leads to lower levels of moral expressiveness (as captured by both individualizing and binding moral foundations), and thus less polarization and lower intentions to proselytize.

In Study 1, we show that the practical advocacy condition generates lower levels of moral expressiveness, on average, compared to the moral condition. The practical condition also generates terms which are less semantically similar to words related to deontology, and moral outrage. In Study 2, we show that only the difference in moral expressiveness is consequential for polarization, such that, the practical condition is less likely to polarize post-advocacy, via lower levels of moral expressiveness. In Study 3, we additionally show that moral expressiveness has implications for behavioral intentions, such as, intentions to proselytize one's opinions. Importantly, those in the practical condition were less likely to proselytize, via lower levels of moral expressiveness. Overall, the findings suggest that using practical compared moral arguments during pro-attitudinal advocacy may lead to more open communicative practices between opposing groups. Further, although previous work in standard persuasion

demonstrates effects of moral values, moral systems, and emotions, in the context of self-generated persuasion (see Briñol et al., 2012; Hovland et al., 1953; Wilson, 1990), we show that the expression of moral foundations, specially, the binding foundations, is most likely to predict polarization and proselytization post-advocacy.

Moral Psychology Theory in Self-Persuasion Research

Construct Distinctiveness

Although the current research treated the different moral theories (e.g. moral expressiveness and moral systems) as separate candidate explanations for our effects, the results suggest potential overlap both within and between distinct theoretical frameworks. For example, we found high correlations between the individualizing and binding moral foundations (moral expressiveness) and also between deontology and consequentialism (moral systems). In addition, we found high correlations between deontology and both moral foundations, suggesting some overlap between the different frameworks as well. This suggests that moral psychology theory may manifest differently in naturalistic expressions of moral language, compared to moral judgement tasks.

In contrast to the current work, the limited previous research using text-analysis to explore themes in moral justifications, shows no correlation between deontology and consequentialism, for example (Wheeler & Laham, 2016). This difference may be attributed to differences in linguistic processes used to justify responses to moral vignettes (which are encountered less often in daily life), compared to justifying attitudes towards controversial issues such as migration and climate-change, which is more likely to occur during conversation. Alternatively, using raw word counts as opposed to a semantic similarity dictionary method may produce differing results, given that the latter is more sensitive to moral content.

On the hand, corroborating the high correlations observed in the current work, some previous research has shown moderate correlations between particular individualizing and binding foundations (e.g. r = .40 between care and loyalty on a climate change issue; see Jansson & Dorrepaal, 2015). In addition, other work theorizing that harm encapsulates all the various moral foundational content (see Schein & Gray, 2015, 2018 for reviews), shows high correlations between theoretically distinct moral foundations (e.g. r = .60 between harm and loyalty; see Schein & Gray, 2015). However, the theoretical rationale for treating the individualizing and binding foundations as separate in the current research is justified given the suite of independent research evidencing the two-factor model of moral foundational content (see Davies, Sibley, & Liu, 2014; Graham et al., 2011; Nilsson & Erlandsson, 2015; Silver & Silver, 2017; van Leeuwen & Park, 2009). Whether an underlying harm foundation encapsulates all moral foundational content is beyond the scope of the current research. Nevertheless, more research is required in the context of self-persuasion to fully understand the implications of moral psychology theory in the context of self-persuasion. This is imperative, given that to our knowledge, the current research is the first to use the dictionary method to explore moral content in the context of pro-attitudinal advocacy.

Importance of the Binding Moral Foundations

Essays created by researchers in standard persuasion typically contain harm related appeals (Luttrell et al., 2019), and appeals to harm are the most common in the real-world (Clifford & Jerit, 2013). Although harm has been proposed to be the most crucial moral foundation (Schein & Gray, 2018), our studies suggest an alternative possibility. Specifically, it appears that the individualizing foundations (containing harm, care, fairness) are less important in driving polarization and in the context of self-persuasion. Rather, advocacy condition exerted indirect effects on polarization primarily via the binding moral foundations (loyalty, authority, purity).

This suggests that targeting the binding foundations may be more effective in reducing polarization during pro-attitudinal advocacy. This notion is supported by previous work in moral reframing, which shows that presenting messages grounded in the binding foundations were more effective in changing attitudes towards recycling, than presenting messages grounded in individualizing/harm foundations (Feinberg & Willer, 2013). The authors suggest that this is because most pro-environmental messages are already framed in terms of the harm/care foundations; people already possess well-established opinions on whether something is harmful and so are unlikely to change opinions when presented with evidence in-line with harm foundations. It is possible that the binding foundations are more effective in producing depolarization because they represent less considered reasons for one's position on an issue. Indeed previous work in self-persuasion shows that people are more likely to depolarize when presented with *novel* arguments in support of the opposition, as opposed to traditional arguments (Burnstein & Vinokur, 1977). Note, however, that the unique effect of the binding foundations did not replicate in accounting for the effect of condition on proselytization. Future research may more fully investigate effects of the moral foundations in self-persuasion contexts.

Null Effects of Political Orientation?

Interestingly, we do not find any moderation by political orientation. Previous work in moral reframing (standard persuasion) consistently shows that liberals are more persuaded when presented with messages containing individualizing foundations, while conservatives are more persuaded by those containing the binding foundations (Feinberg & Willer, 2013, 2015; Voelkel & Feinberg, 2018). However, in the current study, we did not find a) political orientation to predict average levels of moral expressiveness (liberals did not generate higher levels of individualizing foundation terms, and vice versa for conservatives), and b) political orientation to moderate the link between moral expressiveness and polarization (liberals did not polarize more when they expressed greater levels of individualizing foundation terms, and vice versa for conservatives).

Although previous work in standard persuasion suggests "fundamental moral differences separating liberals and conservatives" (Voelkel & Feinberg, 2018) we do not find these differences in the context of self-generated persuasion, suggesting differences in standard versus self-persuasive processes. Indeed, recent work has shown that liberals and conservatives rely on similar moral foundations in making moral judgements of influential figures, suggesting that differences between political partisans may be exaggerated in moral foundational theory (Frimer, Biesanz, Walker, & MacKinlay, 2013). It is possible that political orientation plays a greater role during standard persuasion, but its effects are attenuated in the context of self-generated persuasion, whereby the content of the message generated is more important than demographic characteristics.

However, it is important to note that political orientation significantly correlated with particular moral content in the current research, although these effects were not consistent between and studies, nor large enough to moderate the effects of content on polarization or proselytization. For example, political orientation significantly negatively correlated with both moral foundations and deontology in Study 1, suggesting a link between conservatism and a lower tendency to express content related to these two moral domains. To consider another example, political orientation significantly positively correlated with moral outrage in Study 2, suggesting a link between conservatism and expression of moral outrage. These findings suggest that political orientation may play a more subtle role in the context of self-persuasion, compared to standard persuasion, and presents a interesting direction of research which is yet to be explored in the context of advocacy.

Potential differences between standard and self-persuasive processes is further highlighted by the lack of "moral matching" effects in the current study. Previous work shows that presenting counter-attitudinal moral messages (compared to practical messages) tend to be more effective in persuading those who have a moral attitude basis (see Luttrell et al., 2019). However, this effect does not occur in the current research, again suggesting that demographic characteristics and attitude properties as less important in the context of self-generated persuasion, compared to characteristics of the message itself, such as, moral expressiveness. Future research is required to affirm this claim.

In sum, we show that practical reframing of advocacy appeals may depolarize attitudes and decrease proselytization via self-persuasion. Our findings suggest that moral reframing and moral matching effects may be less effective in reducing close mindedness in the context of pro-attitudinal advocacy. Importantly, the difficulties encountered with encouraging collaboration between liberals and conservatives (Brandt, Reyna, Chambers, Crawford, & Wetherell, 2014) may be combated by encouraging advocacy grounded in practical as opposed to moral reasoning, regardless of one's initial stance or political orientation. These findings carry implications for re-designing online discussion forums (e.g. Reddit), and informing policy recommendations around important socio-political issues such as migration and climate change.

Limitations and Future Directions

Failure to Replicate Polarization-Moral Expressiveness Link

A limitation of the current research is that the indirect effect of advocacy type (moral vs practical) on attitude polarization via moral expressiveness failed to replicate in Study 3. One possibility is that the effect of moral expressiveness on polarization was detected in Study 2 but not in Study 3 due to a smaller sample size. The post-hoc mini meta-analysis revealed that the relationship between the binding foundations and polarization was significant overall, while the relationship between the individualizing foundations and polarization was not. Although the relationship between moral advocacy and polarization can be explained by the expression of binding moral values, this is less clear in explaining the effects of advocacy on proselytization, as neither foundation type predicted proselytization on their own. Future work is required to disentangle the effects of advocacy on the various dependent variables using larger samples from diverse populations.

Another explanation for the diluted effect of polarization in Study 3, is differences in the type of attitude issue tested, as issues vary in their controversy and propensity to evoke emotional reactions (Reeves, Yeykelis, & Cummings, 2016). It is likely that the issue tested in Study 3 (migration) is more affectively-laden and more morally convicted compared to the issue tested in Study 2 (carbon emissions policy, relatively unfamiliar issue). Indeed, a post-hoc *t*-test shows that attitudes towards migration in Study 3 were on average, significantly more based on moral concerns, compared to attitudes towards the carbon emissions policy ($p < .001$, Cohen's $d = 0.78$ medium effect size). This suggests that attitudes towards migration may be more morally ingrained compared to issues related to climate-change (at least in the sample tested), making them more difficult to shift (Krosnick & Petty, 1995).

This is congruent with early research in self-persuasion, showing that depolarization of attitudes tends to be difficult to achieve, in general, (Tesser, 1978), and even more so for attitudes which are highly morally convicted (e.g. capital punishment; Vinokur & Burnstein, 1978). Nevertheless, even on a highly convicted issue, we were able to observe shifts in proselytization following our manipulation, which is quite promising. Encouraging the use of practical reasoning compared to moral reasoning resulted in lower intentions to proselytize, via lower levels of

moral expressiveness, even if attitudes did not change. Future research may examine other downstream consequences of practical versus moral advocacy on communicative intentions such as, willingness to interact with those holding opposing views.

Lack of "Control" Condition

It may be argued that the lack of a true control condition limits the inferences which can be drawn about the self-persuasive effects of advocacy on the outcome variables (depolarization and proselytization). The primary aim of this research was to consider differences specifically between moral and practical advocacy, as opposed to deviation from a control (for similar study designs in pro-attitudinal advocacy which employ two or more experimental conditions without a control condition, see Briñol et al., 2012; Gordijn et al., 2001). Nevertheless, given the lack of a control condition, our inferences are limited to the effects of moral compared to practical advocacy only. Future research may replicate this research with a control condition, in order to more fully draw out the implications of different types of pro-attitudinal advocacy. Further, future research may also investigate whether the effects of moral versus practical advocacy are limited to self-persuasion only, or whether they may apply to standard persuasion as well (i.e. when arguments are presented to participants rather than written by them). We believe that this is a useful direction of research opened up by the current work.

Inability to Draw Causal Inferences

We note that the correlations observed in the current work do not provide strong evidence for causal mediation. This is an inherent limitation of using cross-sectional mediation analysis as a test of causal process (see Kline, 2015). Although we find that advocacy indirectly predicts polarization and proselytization via moral expressiveness, future research is required to demonstrate causal links between these variables. Similar to current work, even though past research typically manipulates an advocacy context, it tends to investigate the relationships between dependent variables in a correlational manner (see Briñol et al., 2012, for work on pro-attitudinal advocacy and meta-cognitive confidence). Future work manipulating the mediator (e.g. manipulating individualizing and binding foundation levels) is required to establish moral expressiveness as the key driver of our effects. Nevertheless, this research highlights important initial evidence that moral expressiveness is associated with polarization and proselytization, specifically in the context of pro-attitudinal advocacy.

Conclusions

We find evidence that moral advocacy may lead to increased social division, via greater expression of moral foundations content, compared to other types of moral content, such as deontology, or moral emotions. We provide evidence for a novel pro-attitudinal advocacy manipulation, i.e. practical advocacy, which may promote depolarization and reduce proselytization via self-persuasion. Our findings suggest that encouraging advocates to generate practical reasons for their attitude leads to less polarization and lower intentions to proselytize one's attitude. The findings provide evidence for a novel and effective method of encouraging free speech, while simultaneously minimizing social conflict.

It is important to note that moral discourse, especially about charged political topics, may have its own value in promoting social cohesion and enabling corrective action, depending on the context. For example, recent research finds that moral language is more persuasive than economic language (i.e. monetary arguments) to influence company management on important social issues faced by employees (Mayer, Ong, Sonenshein, & Ashford, 2019). That is, presenting arguments framed in moral terms as opposed to economic terms is more effective in

mobilizing decision-makers on social issues such as employee health, gender equality, and wage policies. Thus, while practical self-persuasion may be a path to greater harmony, the inverse may also be true: moral arguments may mobilize society in ways that could be practically important. That is, there may be contexts in which each outcome (harmony vs. mobilization) is desirable. Future research may more fully examine the potential of both moral and practical advocacy to bridge ideological divides, by enabling open communication between fragmented social groups.

Notes

i) Dictionaries were constructed using the *embeddingtools* package for R (Crone, 2018), using a pre-trained GloVe model (glove.42B.300d; Pennington, Socher, & Manning, 2014), limiting the model vocabulary to the 250,000 most frequently occurring words.

ii) We conducted a few additional analyses to test if our effects were driven by demographic characteristics, such as, political orientation. A linear regression containing initial stance, political orientation, Condition, and the individualizing (binding) foundations as independent variables revealed that political orientation was not a significant predictor of binding (individualizing) foundations (β = -.04, p = .31; β = -.02, p = .46). Given previous work on moral reframing in the standard persuasion literature, we might have expected political orientation to predict moral expressiveness, such that increased political conservativeness predicts increased expression of the binding foundations, while increased liberalism predicts greater expression of the individualizing foundations. However, this was not the case in the context of self-persuasion. There was also no difference in foundation scores between those who endorsed the different political parties, as captured by our secondary measure of political orientation.

iii) Partially standardized regression coefficient is the default measurement unit in PROCESS, and controls for all variables in the mediation, but does not completely remove variation incurred by the covariates, thus taking into account natural variation in the outcome variable due to confounders (Schielzeth, 2010).

iv) The total indirect effect of condition on polarization via moral expressiveness was significant in a mediation model without the covariates, IE = .14, *SE* = .06, 95% CI [.03, .27]. Again, the effect appears to be driven primarily by the binding foundations, IE = .15, *SE* = .09, 95% CI [.00, .34].

v) Post-hoc moderated mediation analyses on PROCESS (Hayes, 2013) using models 7 and 14 revealed that unlike previous work on moral framing, political orientation was not a significant moderator of the relationship between Condition and moral expressiveness or the relationship between moral expressiveness and polarization. Similarly, unlike previous work on moral matching effects, perceived moral attitude basis did not moderate the relationship between Condition and moral expressiveness or the relationship between moral expressiveness and polarization. Thus, the findings of Study 1 suggest that between-condition differences in polarization is driven by moral expressiveness, and not by other characteristics, such as political orientation, or attitude basis.

vi) One might argue whether this effect is robust controlling for other potential mediators, such as moral outrage and deontology, which also significantly differed between advocacy conditions. Thus, we conducted a secondary simultaneous mediation analysis with the individualizing foundations, binding foundations, deontology and moral outrage as mediators, and political orientation and initial stance as covariates. Results revealed that the binding foundations continued to significantly predict polarization on its own, even after accounting for all other mediators, IE = .32, *SE* = .13, 95% CI [.11, .62].

vii) The total indirect effect of condition on proselytization via moral expressiveness was significant in a mediation model without the covariates as well, IE = .16, *SE* = .07, 95% CI [.04, .30].

Funding

The authors have no funding to report for research design, execution, analysis, interpretation and reporting. An Australian Government Research Training Program Scholarship was bestowed to the corresponding author for the completion of their doctoral dissertation.

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgments

We thank Manuel José Rengifo and Shaheed Azaad from the Melbourne School of Psychological Sciences, for their useful suggestions regarding the mini-meta analysis.

Data Availability

For this article, a dataset is freely available (Abeywickrama, Rhee, Crone, & Laham, 2020a).

Supplementary Materials

The data files used for analysis have been made available on the OSF repository. The Supplementary Materials (Appendices) contain descriptions of the attitude issues examined, advocacy task instructions, and further detail on how the dictionaries were constructed (for unrestricted access, see Index of Supplementary Materials below).

Index of Supplementary Materials

Abeywickrama, R. S., Rhee, J. J., Crone, D. L., & Laham, S. M. (2020a). *Supplementary materials to "Why moral advocacy leads to polarization and proselytization: The role of self-persuasion"* [Research data]. OSF. https://osf.io/nd6vs/

Abeywickrama, R. S., Rhee, J. J., Crone, D. L., & Laham, S. M. (2020b). *Supplementary materials to "Why moral advocacy leads to polarization and proselytization: The role of self-persuasion"* [Appendices]. PsychOpen. https://doi.org/10.23668/psycharchives.3358

References

Anderson, S. E., Potoski, M., DeGolia, A., Gromet, D., Sherman, D., & Van Boven, L. (2014, September). *Mobilization, polarization, and compromise: The effect of political moralizing on climate change politics.* Paper presented at the APSA 2014 Annual Meeting, Sydney, Australia.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., Van Bavel, J. J., & Fiske, S. T. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences of the United States of America, 114*(28), 7313-7318. https://doi.org/10.1073/pnas.1618923114

Brandt, M. J., Reyna, C., Chambers, J. R., Crawford, J. T., & Wetherell, G. (2014). The ideological-conflict hypothesis: Intolerance among both liberals and conservatives. *Current Directions in Psychological Science, 23*(1), 27-34. https://doi.org/10.1177/0963721413510932

Briñol, P., McCaslin, M. J., & Petty, R. E. (2012). Self-generated persuasion: Effects of the target and direction of arguments. *Journal of Personality and Social Psychology, 102*(5), 925-940. https://doi.org/10.1037/a0027231

Burnstein, E., & Vinokur, A. (1977). Persuasive argumentation and social comparison as determinants of attitude polarization. *Journal of Experimental Social Psychology, 13*(4), 315-332. https://doi.org/10.1016/0022-1031(77)90002-6

Cheatham, L., & Tormala, Z. (2015). Attitude certainty and attitudinal advocacy: The unique roles of clarity and correctness. *Personality and Social Psychology Bulletin, 41*(11), 1537-1550. https://doi.org/10.1177/0146167215601406

Clark, J. K., Wegener, D. T., Sawicki, V., Petty, R. E., & Briñol, P. (2013). Evaluating the message or the messenger? Implications for self-validation in persuasion. *Personality and Social Psychology Bulletin, 39*(12), 1571-1584. https://doi.org/10.1177/0146167213499238

Clarkson, J. J., Tormala, Z. L., & Leone, C. (2011). A self-validation perspective on the mere thought effect. *Journal of Experimental Social Psychology, 47*(2), 449-454. https://doi.org/10.1016/j.jesp.2010.12.003

Clarkson, J. J., Valente, M. J., Leone, C., & Tormala, Z. L. (2013). Motivated reflection on attitude-inconsistent information: An exploration of the role of fear of invalidity in self-persuasion. *Personality and Social Psychology Bulletin, 39*(12), 1559-1570. https://doi.org/10.1177/0146167213497983

Clifford, S. (2019). How emotional frames moralize and polarize political attitudes. *Political Psychology, 40*(1), 75-91. https://doi.org/10.1111/pops.12507

Clifford, S., & Jerit, J. (2013). How words do the work of politics: Moral foundations theory and the debate over stem cell research. *The Journal of Politics, 75*(3), 659-671. https://doi.org/10.1017/S0022381613000492

Clifford, S., Jerit, J., Rainey, C., & Motyl, M. (2015). Moral concerns and policy attitudes: Investigating the influence of elite rhetoric. *Political Communication, 32*(2), 229-248. https://doi.org/10.1080/10584609.2014.944320

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY, USA: Lawrence Erlbaum Associates.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour, 1*(11), 769-771. https://doi.org/10.1038/s41562-017-0213-3

Crone, D. (2018). An R package for annotating text data with pre-trained word embedding models [Computer software]. Retrieved from https://github.com/damiencrone/embeddingtools

Davies, C. L., Sibley, C. G., & Liu, J. H. (2014). Confirmatory factor analysis of the Moral Foundations Questionnaire. *Social Psychology, 45*(6), 431-436. https://doi.org/10.1027/1864-9335/a000201

Delton, A. W., DeScioli, P., & Ryan, T. J. (2020). Moral obstinacy in political negotiations. *Political Psychology, 41*(1), 3-20. https://doi.org/10.1111/pops.12612

Department of Environment and Energy. (2018). Thermostats – Energy Saver. Retrieved from https://www.energy.gov/energysaver/thermostats

Ehret, P., Van Boven, L., & Sherman, D. K. (2018). Partisan barriers to bipartisanship understanding climate policy polarization. *Social Psychological & Personality Science, 9*(3), 308-318. https://doi.org/10.1177/1948550618758709

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3.1 manual. *Behavior Research Methods, 39*(2), 175-191. https://doi.org/10.3758/BF03193146

Feinberg, M., & Willer, R. (2013). The moral roots of environmental attitudes. *Psychological Science, 24*, 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., Biesanz, J. C., Walker, L. J., & MacKinlay, C. W. (2013). Liberals and conservatives rely on common moral foundations when making moral judgments about influential people. *Journal of Personality and Social Psychology, 104*(6), 1040-1059. https://doi.org/10.1037/a0032277

Frimer, J. A., Brandt, M. J., Melton, Z., & Motyl, M. (2019). Extremists on the left and right use angry, negative language. *Personality and Social Psychology Bulletin, 45*(8), 1216-1231. https://doi.org/10.1177/0146167218809705

Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., & Dehghani, M. (2016). Morality between the lines: Detecting moral sentiment in text. In *Proceedings of IJCAI 2016 Workshop on Computational Modeling of Attitudes*. Retrieved from http://morteza-dehghani.net/wp-content/uploads/morality-lines-detecting.pdf

Garten, J., Hoover, J., Johnson, K., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior Research Methods, 50*(1), 344-361. https://doi.org/10.3758/s13428-017-0875-9

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass, 10*(10), 535-549. https://doi.org/10.1111/spc3.12267

Gordijn, E. H., Postmes, T., & de Vries, N. K. (2001). Devil's advocate or advocate of oneself: Effects of numerical support on proand counterattitudinal self-persuasion. *Personality and Social Psychology Bulletin, 27*(4), 395-407. https://doi.org/10.1177/0146167201274002

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

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology, 101*(2), 366-385. https://doi.org/10.1037/a0021847

Greene, J. D. (2015). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Law and Ethics of Human Rights, 9*(2), 141-172. https://doi.org/10.1515/lehr-2015-0011

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). Oxford, United Kingdom: Oxford University Press.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998-1002. https://doi.org/10.1126/science.1137651

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

Hayes, A. (2013). *Introduction to mediation, moderation, and conditional process analysis*. New York, NY, USA: Guilford Press.

Hoover, J., Johnson, K., Boghrati, R., Graham, J., & Dehghani, M. (2018). Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation. *Collabra: Psychology, 4*(1), 9-27. https://doi.org/10.1525/collabra.129

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion: Psychological studies of opinion change*. New Haven, CT, USA: Yale University Press.

Jansson, J., & Dorrepaal, E. (2015). Personal norms for dealing with climate change: Results from a survey using moral foundations theory. *Sustainable Development, 23*(6), 381-395. https://doi.org/10.1002/sd.1598

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*(7591), 473-476. https://doi.org/10.1038/nature16981

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-292. https://doi.org/10.2307/1914185

Kline, R. B. (2015). The mediation myth. *Basic and Applied Social Psychology, 37*(4), 202-213. https://doi.org/10.1080/01973533.2015.1049349

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

Kovacheff, C., Schwartz, S., Inbar, Y., & Feinberg, M. (2018). The problem with morality: Impeding progress and increasing divides. *Social Issues and Policy Review, 12*(1), 218-257. https://doi.org/10.1111/sipr.12045

Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 1–24). Mahwah, NJ, USA: Lawrence Erlbaum Associates.

Leidner, B., Kardos, P., & Castano, E. (2018). The effects of moral and pragmatic arguments against torture on demands for judicial reform. *Political Psychology, 39*(1), 143-162. https://doi.org/10.1111/pops.12386

Luttrell, A., Petty, R. E., Briñol, P., & Wagner, B. C. (2016). Making it moral: Merely labeling an attitude as moral increases its strength. *Journal of Experimental Social Psychology, 65*, 82-93. https://doi.org/10.1016/j.jesp.2016.04.003

Luttrell, A., Phillip-Muller, A., & Petty, R. (2019). Challenging moral attitudes with moral messages. *Psychological Science, 30*(8), 1136-1150. https://doi.org/10.1177/0956797619854706

Mayer, D. M., Ong, M., Sonenshein, S., & Ashford, S. J. (2019). The money or the morals? When moral language is more effective for selling social issues. *The Journal of Applied Psychology, 104*(8), 1058-1076. https://doi.org/10.1037/apl0000388

Mayer, N. D., & Tormala, Z. L. (2010). “Think” versus “feel” framing effects in persuasion. *Personality and Social Psychology Bulletin, 36*(4), 443-454. https://doi.org/10.1177/0146167210362981

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour, 2*(6), 389-396. https://doi.org/10.1038/s41562-018-0353-0

Nilsson, A., & Erlandsson, A. (2015). The moral foundations taxonomy: Structural validity and relation to political ideology in Sweden. *Personality and Individual Differences, 76*, 28-32. https://doi.org/10.1016/j.paid.2014.11.049

O’Rourke, H. P., & MacKinnon, D. P. (2018). Reasons for testing mediation in the absence of an intervention effect: A research imperative in prevention and intervention research. *Journal of Studies on Alcohol and Drugs, 79*(2), 171-181. https://doi.org/10.15288/jsad.2018.79.171

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Stroudsburg, PA, USA: Association for Computational Linguistics.

Petty, R. E., & Brinol, P. (2014). The elaboration likelihood and metacognitive models of attitudes. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 172–187). New York, NY, USA: Guilford Press.

Petty, R. E., & Brinol, P. (2015). Emotion and persuasion: Cognitive and meta-cognitive processes impact attitudes. *Cognition and Emotion, 29*(1), 1-26. https://doi.org/10.1080/02699931.2014.967183

Pew Research Center. (2014). *Political polarization in the American public: Growing partisan antipathy.* Retrieved from https://www.people-press.org/2014/06/12/section-2-growing-partisan-antipathy/

Pew Research Center. (2016). *Views of candidate 'insults,' criticism and political divisions.* Retrieved from http://www.people-press.org/2016/10/27/3-views-of-candidate-insults-criticism-and-political-divisions/

Piazza, J., & Sousa, P. (2014). Religiosity, political orientation, and consequentialist moral thinking. *Social Psychological & Personality Science, 5*(3), 334-342. https://doi.org/10.1177/1948550613492826

Qualtrics. (2016). Qualtrics. Provo, UT, USA.

Rapp, C. (2016). Moral opinion polarization and the erosion of trust. *Social Science Research, 58*, 34-45. https://doi.org/10.1016/j.ssresearch.2016.02.008

Reeves, B., Yeykelis, L., & Cummings, J. J. (2016). The use of media in media psychology. *Media Psychology, 19*(1), 49-71. https://doi.org/10.1080/15213269.2015.1030083

Rhee, J. J., Schein, C., & Bastian, B. (2019). The what, how, and why of moralization: A review of current definitions, methods, and evidence in moralization research. *Social and Personality Psychology Compass, 13*(12), Article e12511. https://doi.org/10.1111/spc3.12511

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331-363. https://doi.org/10.1037/1089-2680.7.4.331

Rios, K., Goldberg, M. H., & Totton, R. R. (2018). An informational influence perspective on (non) conformity: Perceived knowledgeability increases expression of minority opinions. *Communication Research, 45*(2), 241-260. https://doi.org/10.1177/0093650217699935

Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass, 5*(6), 359-371. https://doi.org/10.1111/j.1751-9004.2011.00355.x

Ryan, T. J. (2017). No compromise: Political consequences of moralized attitudes. *American Journal of Political Science, 61*(2), 409-423. https://doi.org/10.1111/ajps.12248

Sagi, E., & Dehghani, M. (2014a). Measuring moral rhetoric in text. *Social Science Computer Review, 32*(2), 132-144. https://doi.org/10.1177/0894439313506837

Sagi, E., & Dehghani, M. (2014b). Moral rhetoric in Twitter: A case study of the U.S. federal shutdown of 2013. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014), 36*. Retrieved from https://escholarship.org/uc/item/9sw937kk

Salerno, J. M., & Peter-Hagene, L. C. (2013). The interactive effect of anger and disgust on moral outrage and judgments. *Psychological Science, 24*(10), 2069-2078. https://doi.org/10.1177/0956797613486988

Schein, C., & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin, 41*(8), 1147-1163. https://doi.org/10.1177/0146167215591501

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32-70. https://doi.org/10.1177/1088868317698288

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution, 1*(2), 103-113. https://doi.org/10.1111/j.2041-210X.2010.00012.x

Silver, J. R., & Silver, E. (2017). Why are conservatives more punitive than liberals? A moral foundations approach. *Law and Human Behavior, 41*(3), 258-272. https://doi.org/10.1037/lhb0000232

Simon, H. A. (1979). Rational decision-making in business organizations. *The American Economic Review, 69*(4), 493-513.

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass, 4*(4), 267-281. https://doi.org/10.1111/j.1751-9004.2010.00254.x

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology, 88*(6), 895-917. https://doi.org/10.1037/0022-3514.88.6.895

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology, 58*, 345-372. https://doi.org/10.1146/annurev.psych.56.091103.070145

Tappin, B. M., & McKay, R. T. (2019). Moral polarization and out-party hostility in the US political context. *Journal of Social and Political Psychology, 7*(1), 213-245. https://doi.org/10.5964/jspp.v7i1.1090

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24-54. https://doi.org/10.1177/0261927X09351676

Teeny, J. D., & Petty, R. E. (2018). The role of perceived attitudinal bases on spontaneous and requested advocacy. *Journal of Experimental Social Psychology, 76*, 175-185. https://doi.org/10.1016/j.jesp.2018.02.003

Tesser, A. (1978). Self-generated attitude change. *Advances in Experimental Social Psychology, 11*, 289-338. https://doi.org/10.1016/S0065-2601(08)60010-6

Tesser, A., & Leone, C. (1977). Cognitive schemas and thought as determinants of attitude change. *Journal of Experimental Social Psychology, 13*(4), 340-356. https://doi.org/10.1016/0022-1031(77)90004-X

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences, 7*(7), 320-324. https://doi.org/10.1016/S1364-6613(03)00135-9

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458. https://doi.org/10.1126/science.7455683

Valenzuela, S., Piña, M., & Ramírez, J. (2017). Behavioral effects of framing on social media users: How conflict, economic, human interest, and morality frames drive news sharing. *Journal of Communication, 67*(5), 803-826. https://doi.org/10.1111/jcom.12325

van Bavel, J. J., Packer, D. J., Haas, I. J., & Cunningham, W. A. (2012). The importance of moral construal: Moral versus non-moral construal elicits faster, more extreme, universal evaluations of the same actions. *PLoS One, 7*(11), Article e48693. https://doi.org/10.1371/journal.pone.0048693

van Leeuwen, F., & Park, J. (2009). Perceptions of social dangers, moral foundations, and political orientation. *Personality and Individual Differences, 47*(3), 169-173. https://doi.org/10.1016/j.paid.2009.02.017

van Zomeren, M. (2013). Four core social-psychological motivations to undertake collective action. *Social and Personality Psychology Compass, 7*(6), 378-388. https://doi.org/10.1111/spc3.12031

van Zomeren, M., Postmes, T., & Spears, R. (2012). On conviction's collective consequences: Integrating moral conviction with the social identity model of collective action. *British Journal of Social Psychology, 51*(1), 52-71. https://doi.org/10.1111/j.2044-8309.2010.02000.x

van Zomeren, M., Postmes, T., Spears, R., & Bettache, K. (2011). Can moral convictions motivate the advantaged to challenge social inequality? Extending the social identity model of collective action. *Group Processes & Intergroup Relations, 14*(5), 735-753. https://doi.org/10.1177/1368430210395637

Vinokur, A., & Burnstein, E. (1978). Depolarization of attitudes in groups. *Journal of Personality and Social Psychology, 36*(8), 872-885. https://doi.org/10.1037/0022-3514.36.8.872

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

Wagner, B. C., Briñol, P., & Petty, R. E. (2012). Dimensions of metacognitive judgment implications for attitude change. In P. Briñol & K. G. DeMarree (Eds.), *Social metacognition* (pp. 43–61). New York, NY, USA: Psychology Press.

Wheeler, M. A., & Laham, S. M. (2016). What we talk about when we talk about morality: Deontological, consequentialist, and emotive language use in justifications across foundation-specific moral violations. *Personality and Social Psychology Bulletin, 42*(9), 1206-1216. https://doi.org/10.1177/0146167216653374

Wilson, T. D. (1990). Self-persuasion via self-reflection. In J. M. Olson, M. P. Zanna, & C. P. Herman (Eds.), *Self-inference processes: The Ontario symposium* (Vol. 6, pp. 43–67). Hillsdale, NJ, USA: Lawrence Erlbaum Associates.

Wright, J. C., Cullum, J., & Schwab, N. (2008). The cognitive and affective dimensions of moral conviction: Implications for attitudinal and behavioral measures of interpersonal tolerance. *Personality and Social Psychology Bulletin, 34*(11), 1461-1476. https://doi.org/10.1177/0146167208322557

Yan, H., & Ford, D. (2015, April 28). Baltimore riots: Looting, fires engulf city after Freddie Gray's funeral. *CNN.* Retrieved from https://edition.cnn.com/2015/04/27/us/baltimore-unrest/index.html