Developing Best Practices for the Implicit Measurement of Moral Foundation Salience

Jacob T. Fisher[1]

Frederic R. Hopp[1]

Sujay Prabhu[2]

Ron Tamborini[2]

Rene Weber[1*]

[1]UC Santa Barbara Media Neuroscience Lab - Department of Communication

[2]Michigan State University - College of Communication Arts and Sciences

* Correspondence should be addressed to René Weber (renew@comm.ucsb.edu);

University of California, Department of Communication - Media Neuroscience Lab,

Santa Barbara, CA 93106-4020 (renew@comm.ucsb.edu)

Abstract

Individual moral intuitions influence numerous media processes and effects. In light of this, recent efforts aim to assess these moral intuitions via implicit measurement paradigms, such as the Moral Foundations Affect Misattribution Procedure (MF-AMP; Tamborini, Prabhu, Lewis, Grizzard, & Eden, 2016), and the Lexical Decision Task (LDT; Gantman & Van Bavel, 2014). Despite the promise of these measures for indexing individual moral salience, broader implementation of these procedures is limited by a lack of easy administration and extensibility, as well as underdeveloped best practices for stimulus selection, randomization, and presentation times. We herein introduce an emerging approach to addressing these shortcomings with an open-source tool that is: a) easy and quick to administer, b) based on normed and weighted moral words, and c) modifiable in light of currently evolving best-practices. To demonstrate the usability of this tool, we present data from a pilot study of 79 participants. In this study, we find that: a) moral words are more correctly identified than non-words in the LDT, b) the valence of moral word primes influences the affect attributed to non-word targets, c) this influence is associated with a word's weight in the Extended Moral Foundations Dictionary (E-MFD), and d) moral affect misattribution is associated with political orientation.

Keywords: moral intuition, moral foundations theory, MIME, AMP, LDT, open-source, open-science

Developing Best Practices for the Implicit Measurement of Moral Foundation Salience

Moral intuitions of media users play an instrumental role in a wide range of media processes and effects (Tamborini, 2011). The Model of Intuitive Morality and Exemplars (MIME, Tamborini, 2013) serves as a theoretical framework predicting and explicating how these moral intuitions influence multimedia processing, how individual differences in moral intuitions may predict media preferences, and how these media preferences may influence content production and media systems as a whole. The MIME argues that individuals selectively expose themselves to media content that aligns with their moral intuitions (Bowman, Jöckel, & Dogruel, 2012) and more favorably evaluate narratives that adhere to their moral concerns (Tamborini, Eden, et al., 2013). Furthermore, the MIME argues that long-term exposure to morally-laden media content may shift audiences' moral sensibilities toward those that are made salient during media exposure (Tamborini, Weber, Eden, Bowman, & Grizzard, 2010) and that subsequently— via marketing campaigns and consumer research—content producers become aware of audiences' moral tendencies and adjust content production efforts to maximize appeal within certain audiences (Bowman, Lewis, & Tamborini, 2014; Mastro, Enriquez, & Bowman, 2013).

Further advancements in our understanding of the reciprocal influence processes between audiences' moral intuitions, media consumption, and production hinges on two primary challenges: 1) Reliable and valid extraction of moral information from media texts (Weber et al., 2018) and, 2) Reliable and valid assessment of the moral sensibilities of individuals and audiences (Tamborini et al., 2016).

Progress toward the first goal has recently been found in the development of computational tools for large-scale extraction of moral information from media texts, such as the Extended Moral Foundations Dictionary (E-MFD; Hopp et al., 2018). Assessment of individual differences in moral foundation salience has proven more elusive. In order to measure individuals' moral sensibilities, the majority of research in this area has relied on self-report scales, such as the Moral Foundations Questionnaire (MFQ; Graham et al., 2011). Yet, individuals completing the MFQ are more likely to engage in a slow and rational response process while deliberating their moral concerns. Hence, the MFQ is limited in its ability to assess fast and intuitive responses to moral stimuli that comprise the majority of moral judgment and decision-making processes (Haidt, 2001). Furthermore, self-report scales do not allow a researcher to investigate how moral judgment processes take place in real time. This is especially problematic in biophysiological paradigms like psychophysiology or functional magnetic resonance imaging (fMRI).

These limitations have motivated growing efforts to assess individuals' moral sensibilities using more intuitive measurement procedures. For example, Tamborini and colleagues (2016) have employed a modified version of the Affect Misattribution Procedure that gauges the salience of moral intuitions within an individual or group (MF-AMP; Tamborini, Prabhu, Lewis, Grizzard, & Eden, 2016). In this task, the quickness and valence of individuals' reactions to moral words are used as an indicator of the relative salience of moral foundations of interest. Others have relied on Lexical Decision Tasks (LDT; Gantman & Van Bavel, 2014), in which moral words have been shown to elicit quicker and more accurate responses than non-moral words. In assessing individuals' moral sensibilities in an intuitive manner, these approaches allow for measurement strategies that are closer to the theoretical tenets of moral foundations theory.

Despite their promise for developing understanding of individual differences in intuitive moral salience, broad implementation of these implicit measures is limited by several key factors: First, their setup and analysis often requires painstakingly designing the procedure in stimulus presentation software (such as MediaLab[1] or DirectRT[2]). These software tools are often expensive, proprietary, and platform-specific, thus limiting replication efforts. Second, these emerging approaches rely on moral stimuli selected using manual, ad-hoc strategies. This limits the informativeness of the conclusions that may be drawn regarding individual differences in moral salience measured using these tasks, as they leave unclear whether observed variation is due to the moral nature of the words or due to other factors (e.g. semantic grouping or phonetic features; (Firestone & Scholl, 2016). Third, as these procedures are quite new, best practices do not yet exist for experimental design and analysis. This is especially salient in that minor differences in a protocol (e.g. presenting a word for 40 vs. 80 msec) can greatly influence dependent variables of interest in these tasks (Gantman & Van Bavel, 2014).

In an attempt to mitigate these shortcomings, we herein introduce an experimental protocol for the measurement of intuitive moral foundations that is: 1)

[1] http://www.empirisoft.com/medialab.aspx

[2] http://www.empirisoft.com/directrt.aspx

easy to use and administer, 2) fully open-source and thus replicable and extendable, 3) constructed using validated, weighted moral words, non-moral words, and non-words. In this manuscript, we briefly review the theoretical foundations of this work, focusing on Moral Foundations Theory (MFT; Graham et al., 2013) which provides the theoretical basis for our measurement protocols, and the Model of Intuitive Morality and Exemplars (MIME; Tamborini, 2011) which introduces the framework through which moral intuitions are predicted to influence media selection and processing. Next, we discuss previous implementations of the AMP and LDT, highlighting their respective contributions and limitations. Finally, we outline the design and rationale of the current approach, presenting preliminary analyses from a pilot study. We conclude with a discussion of the benefits, limitations, and future directions of implicit measurement of moral intuitions.

Literature Review

Moral Foundations Theory

MFT argues that humans are equipped with five innate, universal moral intuitions that exist among individuals across cultures and societies. These *moral foundations* include care/harm (involving intuitions of sympathy, compassion, and nurturance), fairness/cheating (including notions of rights and justice), loyalty/betrayal (supporting moral obligations of patriotism and “us vs. them” thinking), authority/subversion (including concerns about traditions and maintaining social order), and sanctity/degradation (including moral disgust and spiritual concerns related to the body). MFT argues that humans are born with cognitive modules that are attuned to recognizing morally-relevant information, but that the relative salience of each moral foundation is modulated over the lifespan through learning and acculturation processes. In this approach, nature provides only a “first draft” (Graham et al., 2013) of the moral building blocks that subsequently are edited by personal experience. In support of this notion, empirical evidence demonstrates that environmental pressures (e.g., pathogen prevalence; van Leeuwen, Koenig, Graham, & Park, 2014) and (political) socialization (Graham, Nosek, & Haidt, 2012) can shape the importance that an individual, culture, or society assigns to each of the foundations (Graham et al., 2011).

Critically, recent work suggests that exposure to morally-laden media content is an important factor in shaping individuals’ moral intuition salience (Tamborini, 2011). Moral cues are salient in a wide range of media formats spanning news (Bowman et al., 2014; Weber et al., 2018), television shows (Hahn et al., 2017; Lewis & Mitchell, 2014; Weber et al., 2008), movie subtitles (Lewis et al., 2017), and song lyrics (Hahn et al., 2018; Long & Eveland, 2018). The MIME proposes that long-term exposure to moral media may shift audiences’ moral judgment towards the social conventions made salient during media exposure (Tamborini, 2011). Accordingly, media narratives are often attributed a powerful cultural role for moral education (cite Haidt & Joseph, 2007).

Furthermore, as postulated by the MIME, the relationship between morally-laden media content and audiences’ moral intuitions is not one-directional, but cyclical. As media users selectively expose themselves to content that aligns with their moral sensibilities, producers become aware of audiences' preferences and adjust productions to maximize the appeal—and thus the success—of their content (Bowman et al., 2014; Mastro et al., 2013). These tailored productions are then reintroduced into individuals' media diets, thus completing the cycle. While emerging literature provides evidence for the existence of these dynamic transactions (Hopp, Fisher, & Weber, 2019), further validations and refinements of the MIME's propositions require methodological innovation in at least two core areas: 1) Reliable and valid extraction of moral information from media texts (Weber et al., 2018) and, 2) Reliable and valid assessment of the moral sensibilities of individuals and audiences (Tamborini & Weber, in press). Much progress has recently been made regarding the extraction of moral rhetoric from media texts (see Hopp et al., 2018; Weber et al., 2018) but the assessment of individual differences in moral intuitions has received comparatively little academic consideration. As such, glaringly few methods or best practices exist in this area. In view of addressing this critical gap, we briefly review previous paradigms for measuring moral foundation salience, highlighting their respective contributions and limitations.

Measuring Moral Intuitions

Moral foundations questionnaire. The majority of previous research measuring individual differences in moral foundation salience has relied on self-report, questionnaire-based assessments. By far the most commonly administered scale is the Moral Foundations Questionnaire (MFQ; Graham et al., 2011). The MFQ consists of 30 items that aim to assess individual differences in ascribed importance of the five moral foundations for reaching moral judgments (Graham et al., 2011). For example, the MFQ captures the salience of the moral foundations when considering “whether or not someone acted unfairly”, “whether or not someone showed a lack of loyalty”, or “whether or not someone was cruel.” In relation to media use, the MFQ has been administered in a variety of contexts. For example, studies have used the MFQ to measure how individuals’ moral foundation salience predicts media choice (Bowman et al., 2012) and media appeal (Tamborini, 2013). Likewise, the MFQ has been used to measure shifts in individuals’ moral saliences following prolonged media exposure (Tamborini et al., 2010). Lastly, the MFQ has also been used to assess the salience of moral foundations in coders in content-analytic paradigms (for an overview, see Weber et al., 2018), with the aim of creating coder pairs that are similar in their moral foundation salience to maximize intercoder reliability.

As these studies emphasize, the MFQ has contributed to notable empirical advancements. Despite this, it is clear that its self-report nature may have several side-effects that compromise its theoretical validity (Tamborini, 2011). First, asking individuals to evaluate their moral sensibilities likely triggers a slow and rational deliberation process. However, as the social-intuitionist model of morality (Haidt, 2001) postulates, the majority of moral judgments are quick and intuitive gut-reactions, rather than careful deliberations. Furthermore, when being asked to explicate why certain actions as morally wrong, people often experience *moral dumbfounding* (Haidt, Bjorklund, & Murphy, 2000) in which they cannot find a rational answer to an intuitively-felt moral concern. In light of these findings, it is clear that a self-report measure such as the MFQ is limited in its ability to access these subconsciously held moral sensibilities. In addition, although the MFQ is fairly easy to administer, its 30-item battery may be a concern in longer paradigms or in which study duration is costly—such as in fMRI.

With these limitations in mind, recent efforts have aimed to develop measurement paradigms that aim to assess moral intuitions in a more intuitive fashion. Most prominently, these approaches draw either on adaptations of either affect misattribution procedures (AMP; Payne, Cheng, Govorun, & Stewart, 2005) or lexical decision tasks (Gantman & Van Bavel, 2014).

Moral Foundations Affect Misattribution Procedure (MF-AMP). In the AMP, a participant is exposed to a series of images or words, followed by unrelated (usually neutral or uninterpretable) words or images. Participants are asked to rate how the latter stimulus made them feel, typically using a binary (pleasant or unpleasant). Consistently, individuals attribute valence to the second stimulus in a manner associated with the first stimulus. This occurs even when participants are explicitly instructed not to base their judgments on the first stimulus (Payne & Lundberg, 2014). Recent work by Tamborini and colleagues (2016) has re-purposed the AMP to measure the salience of different moral foundations following media exposure.

Generally speaking, the MF-AMP rests on the assumption that moral intuitions can attach positive or negative affect to entities or concepts that are morally relevant. In this framework, some morally relevant concepts might be associated with positive or negative affect and stored in a person's long-term memory with their affective connotations intact. For example, hearing a story about a father slapping his disobeying child might trigger a strongly negative affective reaction in a person with a high care/harm salience. Yet, another person might experience less negative affect in response to this situation as they might perceive the parenting action to be upholding the authority/subversion foundation. Accordingly, the MF-AMP aims to examine the degree to which an individual attributes importance to a moral foundation by implicitly measuring the affective evaluation of a concept or behavior related to that foundation. In doing so, it is argued that the MF-AMP provides a less intrusive and more intuitive assessment of individuals' moral foundation salience.

To operationalize these affective evaluations, previous studies drawing on the MF-AMP presented subjects with moral words that were believed to reflect either the upholding (e.g., "help" for care) or violation (e.g., "kill" for harm) of certain moral foundations. These moral prime words are presented for very short time intervals (e.g., 75-100 msec) and are then followed by an evaluation target (e.g., a Chinese character) for a slightly longer time (~ 300 ms). The participants are asked to rate the target stimulus as pleasant or unpleasant, and are reminded to not base their judgments on the previously presented word. The rationale behind this procedure is that participants will "misattribute" the positive (negative) affect triggered by the moral word to the target stimulus. Individual differences in moral foundation salience are then assessed by measuring (a) the valence of a person's rating of the target stimulus in light of the vice/virtue nature of the preceding moral word and, (b) the reaction time with which people rated the target stimulus. For example, individuals that are high in the care/harm foundation would be expected to quickly rate target stimuli as pleasant (unpleasant) if they were preceded by a moral word that indicates the upholding (violation) of care. Several recent studies suggest that the MF-AMP is capable of identifying shifts in moral foundation salience and has been shown capable of identifying even small increases in the accessibility of intuitions (Tamborini, Prabhu, Hahn, Idzik, & Wang, 2014; Tamborini et al., 2016; Tamborini, Prabhu, Wang, & Grizzard, 2013).

Lexical Decision Task. In the LDT, participants are briefly flashed a string of letters composing a word or a non-word and are asked to identify whether the presented string of letters was a word or non-word. Importantly, presentation time is extremely brief (typically around 40 msec), making correct identification of words or non-words quite difficult. As work by Van Bavel and colleagues (Gantman & Van Bavel, 2014) has demonstrated, as presentation time inches nearer to the threshold of perceptual awareness, moral words are more likely to be correctly classified as words compared to non-moral words. This finding was termed the *moral pop-out* effect. Although the LDT was initially designed to examine whether morally relevant stimuli are more likely to reach perceptual awareness, we propose that it may well be extended to measure individual differences in moral intuitions. For example, individuals that strongly endorse the care/harm foundation may be more likely to correctly classify moral words as words if they relate to the care foundation. Interestingly, though, Van Bavel and colleagues (2014) were not able to demonstrate a correlation between participants' moral foundation salience as indexed by the MFQ and lexical decision ratings in both accuracy and response time. This finding may have resulted from either a poor selection of moral stimuli (i.e., words were chosen ad-hoc and were not classified into individual foundations) or it may be due to a lack of correlation between the rational deliberation triggered by the MFQ and the intuitive decisions indexed by the LDT.

Toward Best Practices for MF-AMP and LDT

While both the MF-AMP and the LDT have opened valuable research avenues to assess moral intuitions in a non-invasive, implicit, and intuitive fashion, we argue that more work is needed to develop best practices for their implementation. This is due to three primary factors. First, these protocols are often implemented in software packages that are expensive, proprietary, and platform-specific, limiting collaborative efforts. Second, thus far the MF-AMP and the LDT have relied on ad-hoc word selection procedures that limit the generalizability of findings. As (Firestone & Scholl, 2016) have highlighted, implicit biases in word perception could be due to many factors other than a word's status as a moral word, especially when word lists are small and non-normed. The incorporation of validated and normed word lists will allow for more clarity in this area. Finally, both the MF-AMP and the LDT are sensitive to even minor protocol changes. For example, recent work reports that presenting a word for 80 msec as opposed to 40 msec almost entirely erases the difference between moral and non-moral words in the LDT (Gantman & Van Bavel, 2014). As such, replicability of these findings is contingent upon developing and sharing stimuli that closely hew to the standards of previous work.

As a next step toward the development of these best-practices, we introduce recent work designing and implementing the MF-AMP and the LDT in an open-source presentation software using previously validated and normed lists of moral words, neutral words, and non-words. In addition, this protocol provides full randomization between prime and target word pairs in the MF-AMP. To provide first steps toward validating this protocol, we provide data from a pilot study. Based on extant literature in moral intuition salience, we predict that participants would exhibit more correct responses to moral words in the LDT as opposed to neutral words. Furthermore, we predict that participants will report more negative evaluations of a target word when it is preceded by a moral vice word as opposed to a moral virtue word. While not predicted a-priori, based on previous work, we also show that participant's responses to moral vice and moral virtue words are associated with the weight of these words in the E-MFD, which provides additional evidence for the suggested procedure's validity. Finally, we predict that these responses would differ between naturally-occurring moral groups, such as between liberals and conservatives.

Method

A pilot experiment was conducted in early 2019 at a large university in the western United States. In this experiment, participants completed the Moral Foundations Affect Misattribution Procedure (MF-AMP) as well as a Lexical Decision Making (LDT) task. Moral words in each task were chosen based on their categories and weights in the Extended Moral Foundations Dictionary (E-MFD; Hopp et al., 2018).

Stimuli for this experiment were coded in PsychoPy, an open-source stimulus creation and presentation program based on the Python programming language (Peirce, 2007).

Procedures

Participants (n = 79, M_{age} = 19.87) completed both tasks in a dimly lit computer lab with ten computers. Each computer was positioned in a cubicle, limiting distraction from other participants. Upon entering the lab, each participant was greeted, given a consent form, and asked not to begin until prompted. After all participants in the session were seated, a researcher read a brief description of each of the tasks. Upon completion of the oral instructions, participants were instructed to place their right hand on the arrow pad of the keyboard in front of them, with their index finger positioned on the left arrow key and their ring finger positioned on the right arrow key.

Half of the participants completed the MF-AMP first and the other half completed the LDT first. A brief series of textual instructions was presented before each task, reminding participants of the goal of the task and of proper hand placement. Stimuli were presented on Dell 1600 x 900 monitors with a 60 Hz refresh rate. In both tasks, words were presented as black text on a light gray background. All text was positioned in the center of the screen and subtended .05 degrees of vertical visual angle. Upon completing both tasks, participants responded to a short questionnaire.

Lexical Decision Making Task. During the LDT, participants are told that they will be completing a task that will measure their ability to quickly recognize words and non-words. No information in the oral or textual instructions cued participants as to the moral nature of the words. In each trial, a fixation cross was presented on the screen for a random duration between 100 and 300 msec, followed by a 60 msec presentation of a word or non-word, followed by a backward mask of ampersands with a length corresponding to the previously presented word or non-word (See Figure 1). The mask remained on the screen until the participants made their choice. Participants were instructed to press the left arrow key if they believe they saw a non-word, or the right arrow key if they believed they saw a word. After viewing a series of instruction screens but before beginning the task, participants completed a practice round of four trials (two words and two nonwords). Participants completed 200 rounds of the LDT, viewing 50 moral words (ten from each foundation), 50 neutral words, and 100 non-words. In total, the LDT took about 4 minutes to complete. At the conclusion of the LDT, participants were presented with a 7-point scale asking to what extent they responded randomly in the task (1 = *not at all*, 7 = *completely, M* = 4.17).

Moral Foundations Affect Misattribution Procedure (MF-AMP). During the MF-AMP, participants were told that they would be completing a task designed to better understand the contribution of linguistic cues to word perception. In the task, participants reported whether a nonsense word (from here on referred to as the *target*) presented briefly on screen made them feel pleasant or unpleasant. Before the presentation of the target, another word is flashed on the screen. This word (from here on referred to as the *prime*), was either a moral word, a neutral word, or a non-word. Participants were told that this word was a cue that the nonsense word they should rate was about to appear on screen, and were instructed to try their best not to let the prime words bias their judgments of the target words (Payne et al., 2005; Tamborini et al., 2016). No indication was given as to the nature of the prime words.

In each trial, participants were presented with a fixation cross (100-300 msec), followed by the prime word (100 msec), a blank screen (125 msec), the target word (300 msec) and a mask (see Figure 2).The mask remained on screen until participants responded. A reminder of the correct key to press in order to report feeling pleasant or unpleasant was presented along with the mask. Presentation times for each of these screens were adapted from (Tamborini et al., 2016). The fixation cross was adapted from (Gantman & Van Bavel, 2014) in order to to reduce participants' tendency to get lulled into a rhythmic response pattern.

In addition to the fixation cross, the protocol outlined here differs from that of Tamborini and colleagues (2016) in two primary ways: first, rather than presenting a Chinese character as the target, this study employs a non-word target using English phonemes. This change was initiated due to the fact that the subject pool at the university in which this study was conducted contains a relatively high proportion of Chinese participants. This introduced the risk that the Chinese character would be interpretable by some of our participants but not others, and that these interpretations may introduce unaccounted for variance in the attribution procedure. In order to account for variation in valence responses that may be due to phonetic features of the target words, ten non-words were included as prime words for in each participant. As an additional control, the words "pleasant" and "unpleasant" were also included as primes. Participants completed four rounds of 82 target/prime pairs (70 moral primes, 10 non-word primes, and 2 control primes) for a total of 328 trials. In total, the MF-AMP took about 8 minutes to complete. At the conclusion of the task, participants were presented with three 7-point scales (1 = *not at all*, 7 = *completely*) asking to what extent they: 1) based their responses on the first word they were presented in each trial (*M* = 4.18), 2) based their responses on their evaluations of the non-words (*M* = 4.07) and, 3) responded randomly in the task (*M* = 3.71).

Word Selection and Randomization

Moral words were selected for the LDT and the MF-AMP based on their weighting in the Extended Moral Foundations Dictionary (Hopp et al., 2018). The 24 highest-weighted words per foundation (ten virtue and ten vice) were chosen for inclusion in the full word list, resulting in 140 moral words. The words “pleasant” and “unpleasant” were added to the list as controls. Neutral words for the experiment were pulled at random from the list from stop-words from the Stanford natural language processing toolkit (NLTK; Bird, Klein, & Loper, 2009). 192 non-words were selected from the English Lexicon Project (Balota et al., 2007) and matched to the moral and neutral words along multiple dimensions, including word length, lexical frequency, and number of phonemes. This resulted in a final word list containing 384 words.

A randomization algorithm was created in Python in order to appropriately distribute words from each category (care, fairness, authority, loyalty, sanctity, neutral, control, nonword) into each task on a per-participant basis (see Figure 3). A unique word list was generated for each participant from this general word list. The list for the LDT consisted of the 50 neutral words, 50 randomly selected moral words (5 vice and 5 virtue per foundation), and 100 randomly selected non-words. For the MF-AMP, the list consisted of the two control words, 70 moral words (7 virtue and 7 vice words per foundation), 10 randomly selected non-word primes, and 82 randomly selected non-word targets. All random selection was conducted without replacement to ensure that no participant sees the same word in both the LDT and the MF-AMP. For the MF-AMP, matches between the primes and the targets were randomized for each of the four prime repetitions to ensure that prime-target pairs were not duplicated.

Survey Items

All survey items were completed in Qualtrics. The main measures of interest in this experiment are moral foundation salience and political orientation. Moral foundation salience was assessed using the Moral Foundations Questionnaire (MFQ; Graham et al., 2011). This measure consists of 30 7-item scales. The first 15 items concern the importance of each of the moral foundations when judging whether something is right or wrong. In the last 15 items, participants report their agreement or disagreement with 15 morally-laden statements (e.g. "Respect for authority is something all children need to learn"). Participants' overall moral salience is calculated as the mean of all items, and the salience of each particular domain is calculated as the mean of each domain subscale. Political orientation was assessed using a one-item scale in which participants reported their views on a 5-point scale (1 = *strongly liberal,* 5 = *strongly conservative*). Two "catch" items are also presented in the questionnaire to isolate participants who did not pay attention during the questionnaire. Two participants were excluded due to inappropriate responses on these items. Due to a technical error in the survey instrument, survey responses from an additional 10 participants were unavailable. As such, these participants are excluded from analyses involving survey data.

Dependent Measures

In the LDT, dependent measures of interest are correctness and reaction times (RTs). All incorrect guesses are coded as 0 and all correct guesses are coded as 1. RTs were measured in milliseconds. Following (Whelan, 2008), RTs more than 3 standard deviations away from the participant's mean RT were removed, and reaction times below 100 msec were replaced with 100msec.

In the AMP, dependent measures of interest are RT, response valence, and a composite measure of response valence and reaction time (here called ϕ). This composite measure was constructed following the method introduced by Tamborini and colleagues (2016). To construct this measure, we divided a participant's reaction time within a particular trial by their mean reaction time for all trials. The inverse of this standardized reaction time was then multiplied by the participants response valence in that trial (either -1 or 1) to arrive at an overall index of the valence and salience of the word presented in that trial. As each word was repeated across four trials, the ϕ score for each word was created by taking the mean of all four trials in which the participant saw that word as a prime. As each participant was presented 82 unique prime words, each participant will have 82 unique ϕ values.

Pilot Results

In this study, we hypothesized that participants would respond more correctly overall to moral words in the LDT as opposed to neutral words, that participants would respond more negatively in the MF-AMP when the prime was a moral vice word as opposed to a moral virtue word, and that participant's responses to moral vice and moral virtue words would be associated with the weight of these words in the E-MFD. Furthermore, we predicted that these responses would differ between naturally-occurring moral groups, such as between liberals and conservatives. Initial compliance checks revealed that 7 participants did not adequately complete the tasks (e.g. self-reported completely random responses, or pressed "left" or "right" for over 90% of trials). Data from these seven participants was excluded from further analyses.

A t-test was conducted to test the hypothesis that participants would respond more correctly to moral words as opposed to neutral words in the LDT. The difference between moral and neutral words was significant for response correctness (t(2, 136) = 1.90, p = .029, one-tailed) such that participants were more accurate for moral words (M = .78) than for neutral words (M = .72). Non-words elicited the least accurate responses overall (M = .64; See Figure 4). Thus, this hypothesis was supported.

A t-test was conducted to test the hypothesis that participants would rate a non-word as more unpleasant when it followed a moral vice word prime than when it followed a moral virtue word prime in the MF-AMP. The difference between moral virtue and moral vice primes was significant (t(2, 140) = -9.78, p < .001) such that participants rated non-words as more unpleasant following moral vice primes (M =

-.41) than following moral virtue primes (*M* = .22). Interestingly, participants also reported feeling unpleasant following non-word primes (*M* = -.31). Participants' unpleasant or pleasant ratings following moral vice or virtue primes was also associated with the word's weight in the E-MFD, such that higher-weighted vice words in the dictionary primed more unpleasant feelings from participants (*r*(60) = -0.31, *p* = .017) and that higher-rated virtue words in the dictionary primed more pleasant feelings from participants (*r*(60) = 0.35, *p* = .006; see Figure 5).

Finally, it was predicted that political orientation would be associated with responses to moral words in the MF-AMP. As predicted, political orientation was significantly associated with responses (*r*(60) = .21, p < .001). As an exploratory analysis, we sought to understand how these responses differed between foundations. Based on moral foundations theory, it could be expected that liberals and conservatives would not differ in their reactions to care and fairness words, but that they may differ in their responses to authority, loyalty, and sanctity words such that conservatives would be more sensitive to these words than liberals (Graham et al., 2012). This analysis was probed using a series of within-domain correlations. As predicted by MFT, the association between political orientation and word responses was significant within authority (*r*(60) = .21, p = .046) and loyalty (*r*(60) = .39, p = .003) domains, and approached significance within the sanctity domain *r*(60) = .22, p = .103). These associations were not significant for care (*r*(60) = .12, p = .381) or fairness (*r*(60) = .18, p = .193; see Figure 6).

Discussion

MFT and the MIME predict that individual differences in media exposure and message processing can be attributed to differences in the salience of moral intuitions (Graham et al., 2011; Tamborini, 2011, 2013). Although differences in moral intuitions can be measured using self-report scales such as the Moral Foundations Questionnaire (MFQ), the use of self-report questionnaires carry with them numerous limitations that circumscribe their utility for understanding implicit processes like moral judgements. Progress toward implicit measurement of moral foundation salience has been found in emerging work investigating the *moral pop-out* effect observed by Gantman and Van Bavel (2014) and in work by Tamborini and colleagues tying differences in affect misattribution to differences in media exposure (2013; 2014; 2016).

To ensure the broader utility of these measures, here we introduced an open-source version of the MF-AMP and the LDT that relies on normed and validated word banks and that allows for flexible implementation and replication. In order to demonstrate the utility of this protocol, we provided data from a pilot study. In this study, we hypothesized that moral words would elicit more accurate responses in the LDT (replicating the findings of Gantman and Van Bavel, 2014), and that moral virtue (vice) words would elicit positive (negative) affect misattributions toward non-words in the MF-AMP. Furthermore, we proposed two exploratory hypotheses regarding the associations between affect attributions and word weights in the E-MFD (Hopp et al., 2018), and that affect responses would vary as a function of political orientation. Finally, we proposed an additional exploratory hypothesis that differences in moral attributions between liberals and conservatives would be driven by the authority, loyalty, and sanctity domains. All of these hypotheses were supported.

The approach presented here, although a step forward from extant approaches, does carry two primary limitations. Most notably, it remains uninvestigated how minute differences in the protocol for the MF-AMP and the LDT may lead to differences in the utility of implicit measurements conducted using these measures. For example, Tamborini and colleagues have primarily employed a Chinese character as the misattribution target whereas here we employed a non-word with phonemes characteristic of the English language. Future work should systematically vary these responses. Second, the responses collected following the MF-AMP and the LDT point to a need for theoretical clarity regarding the extent to which these processes reflect a true misattribution of moral salience. For example, participants reported basing their judgments on the primes at a higher rate than the targets (*M* = 4.18 vs. *M* = 4.06), although explicitly instructed to not base their judgments on the primes. Participants seeming awareness of this "misattribution" raises important questions regarding the truly implicit nature of this task. Future work should seek to understand the extent to which experimental parameters (e.g. presentation times) influence attributions, but also participants' awareness of those attributions.

Conclusion and Future Directions

The work presented here provides initial support for the usability of the proposed paradigm for measurement of implicit moral attitudes. Participants more correctly identified moral words from the E-MFD as opposed to neutral words, more often attributed positive valence to a non-word following a virtue word, and more often attributed negative valence to a non-word following a vice word. These responses differed as a function of the words' weight in the E-MFD. Finally, these responses varied in relation to political orientation—a naturally-occurring moral group. Although these findings are promising, more work is needed to replicate and verify these findings, as well as solidify best practices for the MF-AMP and the LDT. These improvements will contribute to increased clarity regarding individual differences in moral salience as well as how these differences may predict media-related outcomes of interest like media choice, character evaluations, and other processes.

References

Mastro, D., Enriquez, M., Bowman, N. D., Prabhu, S. & Tamborini, R. (2013) Morality subcultures and media production: How Hollywood minds the morals of its audience. In R. Tamborini (Ed), *Media and the moral mind* (75-92). London, UK: Routledge.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Bowman, N. D., Jöckel, S., & Dogruel, L. (2012). A question of morality? The influence of moral salience and nationality on media preferences. *Communications: The European Journal of Communication Research*, *37*(4), 345–369.

Bowman, N. D., Lewis, R. J., & Tamborini, R. (2014). The morality of May 2, 2011: A content analysis of U.S. headlines regarding the death of Osama Bin Laden. *Mass Communication and Society*, *17*(5), 639–664. https://doi.org/10.1080/15205436.2013.822518

Firestone, C., & Scholl, B. J. (2016). 'Moral Perception' Reflects Neither Morality Nor Perception. *Trends in Cognitive Sciences*, *20*(2), 75–76. https://doi.org/10.1016/j.tics.2015.10.006

Gantman, A. P., & Van Bavel, J. J. (2014). The moral pop-out effect: Enhanced perceptual awareness of morally relevant stimuli. *Cognition*, *132*(1), 22–29. https://doi.org/10.1016/j.cognition.2014.02.007

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013).

Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In P. Devine & A. Plant (Eds.), *Advances in Experimental Social Psychology* (Vol. 47, pp. 55–130). https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Graham, J., Nosek, B. A., & Haidt, J. (2012). The Moral Stereotypes of Liberals and Conservatives: Exaggeration of Differences across the Political Spectrum. *PLoS ONE, 7*(12), e50092. https://doi.org/10.1371/journal.pone.0050092

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

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Hopp, F. R., Fisher, J. T., & Weber, R. (2019). The dynamic relationship between news frames and real-world events: A hidden Markov model approach. *Paper Presented at the 69th Annual Conference of the International Communication Association*. Presented at the Washington DC, USA. Washington DC, USA.

Hopp, F. R., Mangus, J. M., Swanson, R., Gordon, A., Khooshabeh, P., & Weber, R. (2018). Developing and validating the moral foundations dictionary for news narratives: a crowd-sourced approach. *Paper Presented at the 69th Annual Conference of the International Communication Association*. Presented at the Prague, CZ. Prague, CZ.

Mastro, D., Enriquez, M., & Bowman, N. D. (2013, October 2). Morality subcultures and media production: How Hollywood minds the morals of its audience. https://doi.org/10.4324/9780203127070-12

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*(3), 277–293. http://dx.doi.org/10.1037/0022-3514.89.3.277

Payne, B. K., & Lundberg, K. (2014). The affect misattribution procedure: ten years of evidence on reliability, validity, and mechanisms. *Social & Personality Psychology Compass, 8*(12), 672–686. https://doi.org/10.1111/spc3.12148

Peirce, J. W. (2007). PsychoPy—psychophysics software in python. *Journal of Neuroscience Methods, 162*(1), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Tamborini, R. (2011). Moral Intuition and Media Entertainment. *Journal of Media Psychology, 23*(1), 39–45. https://doi.org/10.1027/1864-1105/a000031

Tamborini, R. (2013). Model of intuitive morality and exemplars. In R. Tamborini (Ed.), *Media and the Moral Mind.* London, UK: Routledge.

Tamborini, R., Eden, A., Bowman, N. D., Grizzard, M., Weber, R., & Lewis, R. J. (2013). Predicting media appeal from instinctive moral values. *Mass Communication and Society, 16*(3), 325–346. https://doi.org/10.1080/15205436.2012.703285

Tamborini, R., Prabhu, S., Hahn, L., Idzik, P., & Wang, L. (2014). News exposure's influence on the salience of moral intuitions: Testing the reliability of the Intuitive Motivation–Affect Misattribution Procedure (IM–AMP). *Paper*

Presented at the 64th Annual Meeting of the International Communication Association. Presented at the Seattle, WA. Seattle, WA.

Tamborini, R., Prabhu, S., Lewis, R. J., Grizzard, M., & Eden, A. (2016). The influence of media exposure on the accessibility of moral intuitions and associated affect. *Journal of Media Psychology, 30*(2), 79–90. https://doi.org/10.1027/1864-1105/a000183

Tamborini, R., Prabhu, S., Wang, L., & Grizzard, M. (2013). Setting the moral agenda: News exposure's influence on the salience of moral intuitions. *Paper Presented at the 63rd Annual Meeting of the International Communication Association*. Presented at the London, UK. London, UK.

Tamborini, R., & Weber, R. (in press). Advancing the model of intuitive morality and exemplars. In *The Routledge Handbook of Communication Science and Biology*. New York, NY: Routledge.

Tamborini, R., Weber, R., Eden, A., Bowman, N. D., & Grizzard, M. (2010). Repeated Exposure to Daytime Soap Opera and Shifts in Moral Judgment Toward Social Convention. *Journal of Broadcasting & Electronic Media, 54*(4), 621–640. https://doi.org/10.1080/08838151.2010.519806

van Leeuwen, F., Koenig, B. L., Graham, J., & Park, J. H. (2014). Moral concerns across the United States: associations with life-history variables, pathogen prevalence, urbanization, cognitive ability, and social class. *Evolution and Human Behavior, 35*(6), 464–471. https://doi.org/10.1016/j.evolhumbehav.2014.06.005

Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., …

Tamborini, R. (2018). Extracting latent moral information from text narratives: relevance, challenges, and solutions. *Communication Methods and Measures, 12*(2–3), 119–139. https://doi.org/10.1080/19312458.2018.1447656

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record, 58*(3), 475–482. https://doi.org/10.1007/BF03395630

Figures

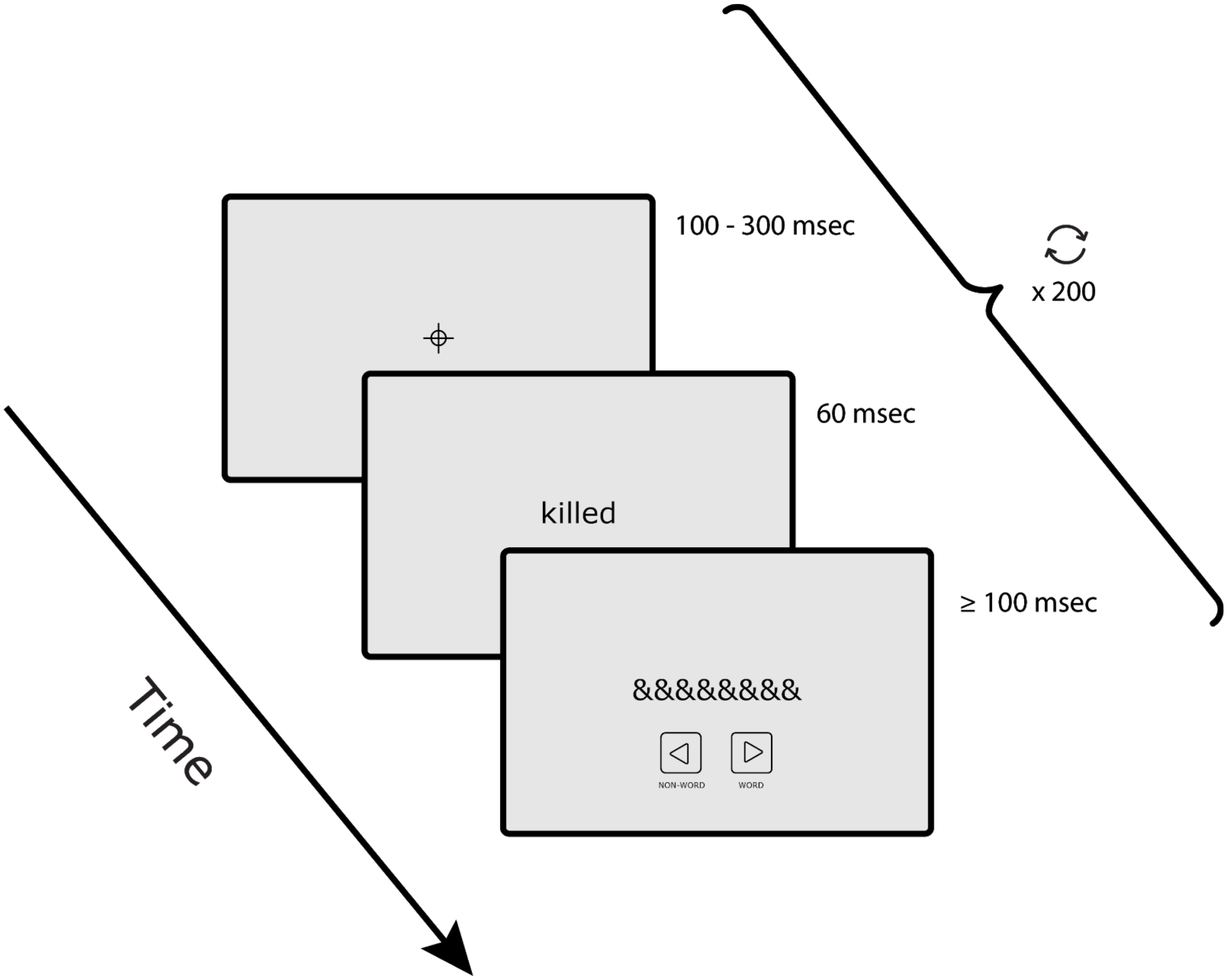

Figure 1: Schematic of lexical decision task (LDT). Participants saw a fixation cross for a random period of time between 100 and 300 msec, followed by a word or non-word (60 msec), followed by a mask (≥ 100 msec). The mask remained on screen until participants pressed “left” or “right” to indicate their guess.

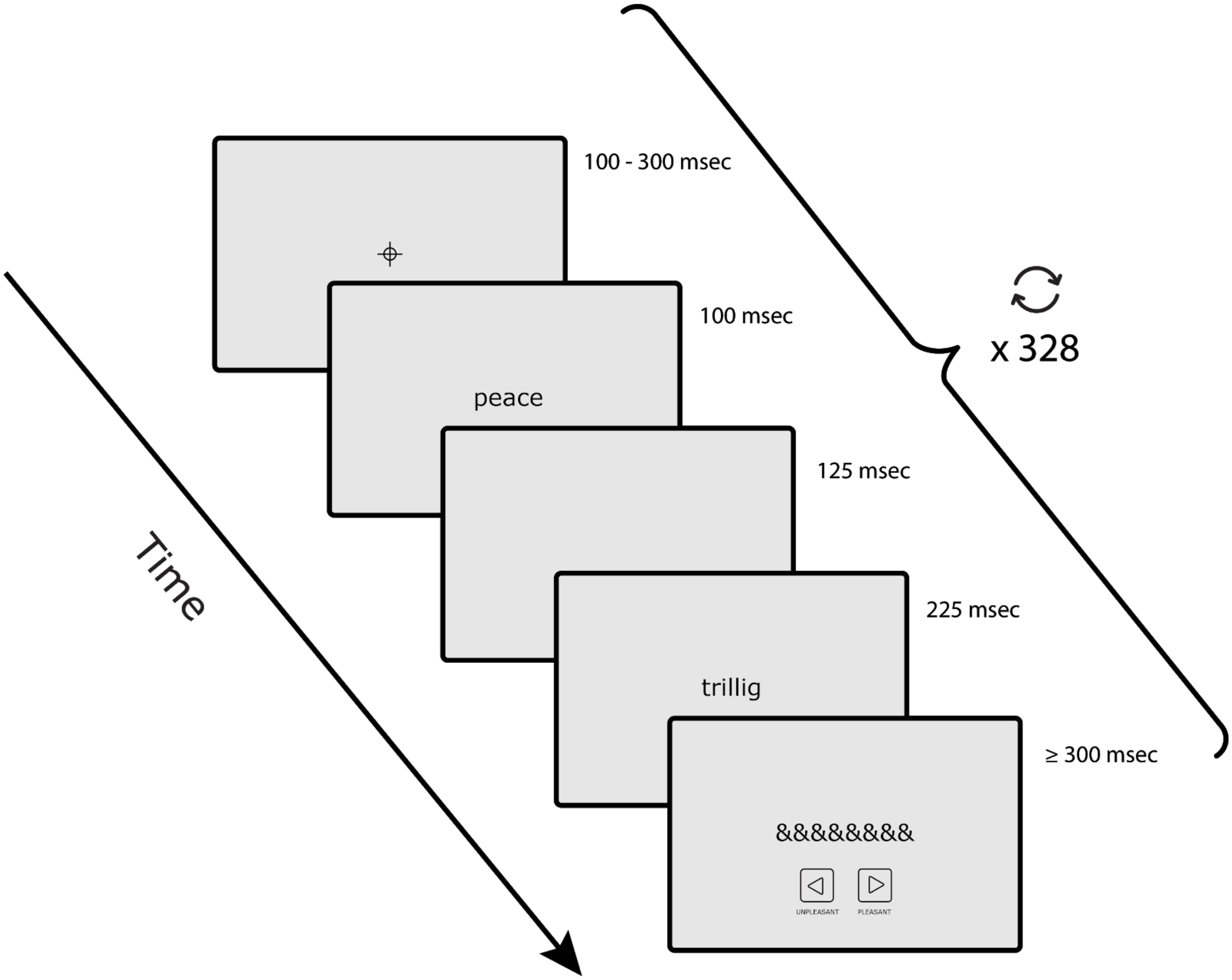

Figure 2: Schematic of Affect Misattribution Procedure (AMP). Participants saw a fixation cross for a random period of time between 100 and 300 msec, followed by a word (100 msec), followed by a blank screen (125 msec), a non-word (225 msec), and a mask (≥ 300 msec). The mask remained on screen until participants pressed “left” or “right” to indicate their guess.

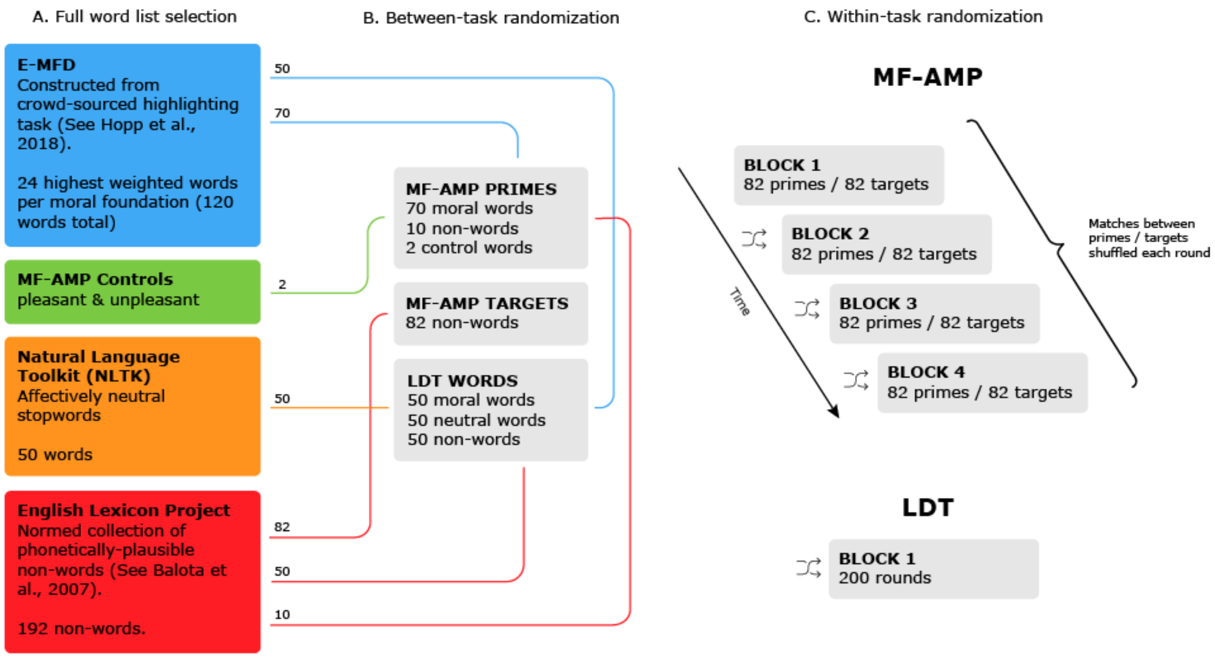
A. Full word list selection
B. Between-task randomization
C. Within-task randomization
E-MFD
Constructed from crowd-sourced highlighting task (See Hopp et al., 2018).
24 highest weighted words per moral foundation (120 words total)
MF-AMP Controls
pleasant & unpleasant
Natural Language Toolkit (NLTK)
Affectively neutral stopwords
50 words
English Lexicon Project
Normed collection of phonetically-plausible non-words (See Balota et al., 2007).
192 non-words.
50
70
2
50
82
50
10
MF-AMP PRIMES
70 moral words
10 non-words
2 control words
MF-AMP TARGETS
82 non-words
LDT WORDS
50 moral words
50 neutral words
50 non-words
MF-AMP
BLOCK 1
82 primes / 82 targets
BLOCK 2
82 primes / 82 targets
BLOCK 3
82 primes / 82 targets
BLOCK 4
82 primes / 82 targets
Time
Matches between primes / targets shuffled each round
LDT
BLOCK 1
200 rounds

Figure 3: Schematic of word selection and randomization procedure. 384 words were selected in total for the experiment (120 moral words, 2 control words, 50 neutral words, and 192 non-words). A randomization algorithm created a unique word list per participant, selecting 82 words from this list as MF-AMP primes (70 moral words, 10 non-words, and 2 control words), 82 words as MF-AMP targets (all non-words), and 200 words for the LDT (50 moral words, 50 neutral words, and 50 primes). Within the MF-AMP, matches between primes and targets were randomized each round.

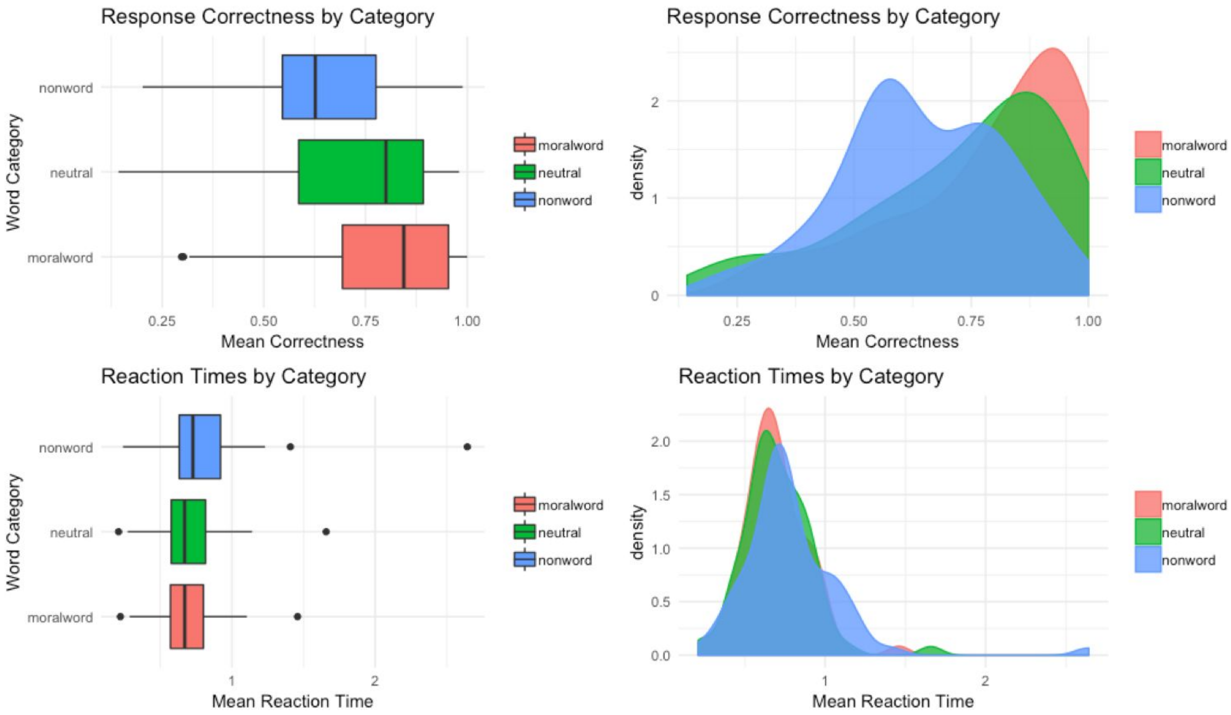

Figure 4: Plots of response correctness and reaction times per category. In the LDT, differences between word categories were significant for response correctness such that participants were the most accurate for moral words (t = 1.90, *M* = .78, *p* = .03), followed by neutral words (*M* = .72), and non-words (*M* = .64).

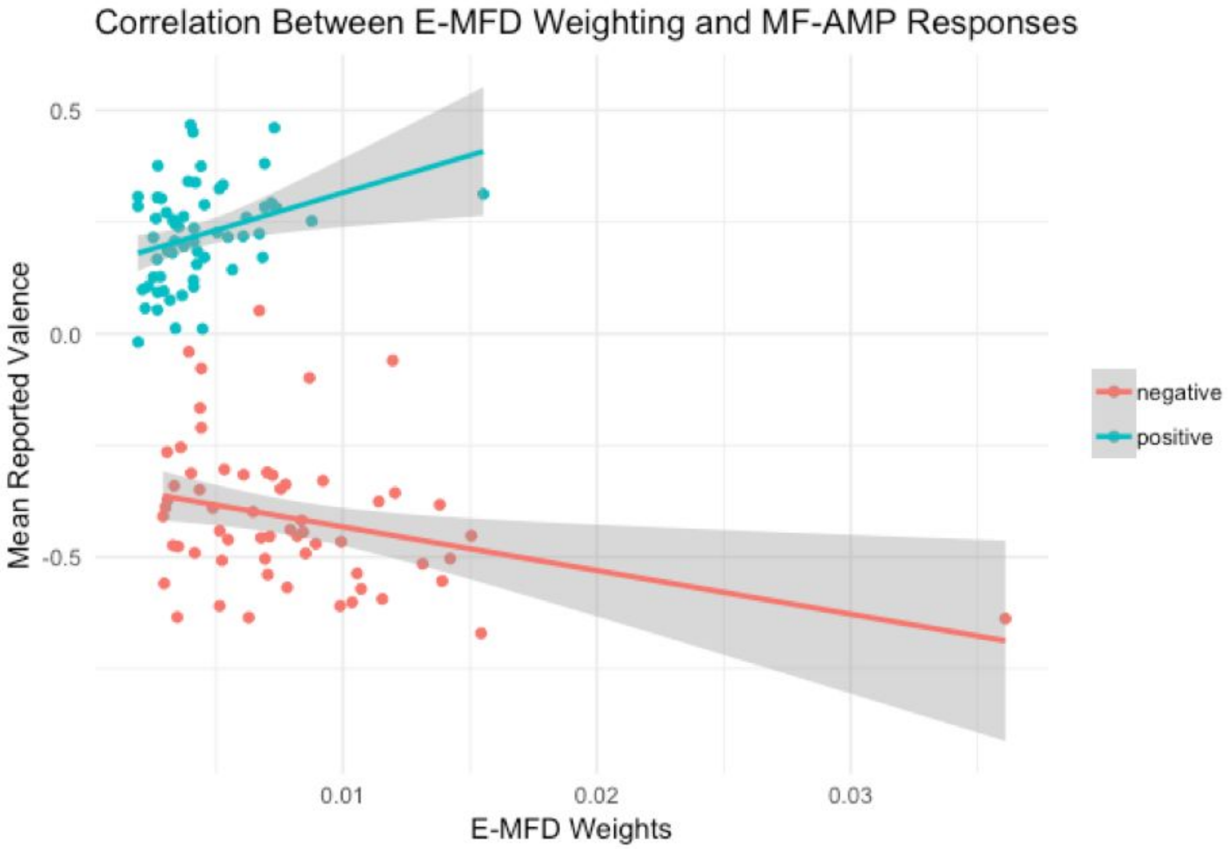

Figure 5: Correlation between E-MFD Weights and Responses in the MF-AMP. Participants' feeling unpleasant or pleasant following moral vice or virtue primes was associated with the word's weight in the E-MFD, such that higher-weighted vice words in the dictionary primed more unpleasant feelings from participants (*r*(60) = -0.31, *p* = .017) and that higher-rated virtue words in the dictionary primed more pleasant feelings from participants (*r*(60) = 0.35, *p* = .006).

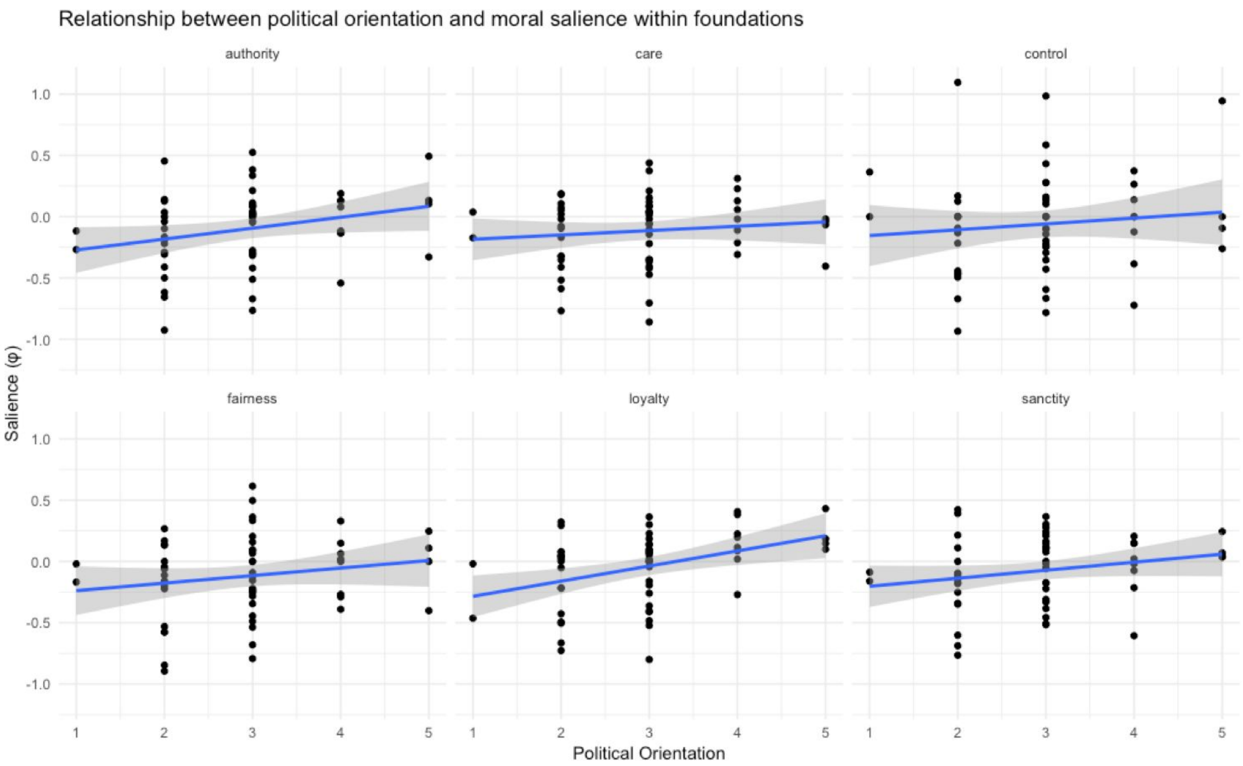

Figure 6: Relationship between political orientation and moral foundation salience within foundations. Based on moral foundations theory, it could be expected that liberals and conservatives would not differ in their reactions to care and fairness words, but that they may differ in their responses to authority, loyalty, and sanctity words such that conservatives would be more sensitive to these words than liberals. This analysis was probed using a series of within-domain correlations. As predicted by MFT the association between political orientation and word responses was significant within authority (*r*(60) = .21, p = .046) and loyalty (*r*(60) = .39, p = .003) domains, and approached significance within the sanctity domain *r*(60) = .22, p = .103). These associations were not significant for care (*r*(60) = .12, p = .381) or fairness (*r*(60) = .18, p = .193).