Running head: OVERLAP IN MORALS, FACTS, AND PREFERENCES

Examining overlap in behavioral and neural representations of morals, facts, and preferences.

Jordan Theriault[1], Adam Waytz[2], Larisa Heiphetz[3], Liane Young[1]

Word Count: 10,688

[1]Boston College
Department of Psychology
Chestnut Hill, MA, 02467
USA

[2]Northwestern University
Kellogg School of Management
Evanston, IL, 60208
USA

[3]Columbia University
Department of Psychology
New York, NY, 10027
USA

Corresponding Author:
Jordan Theriault
jordan.theriault@bc.edu

Metaethical judgments refer to judgments about the information expressed by moral claims. Moral objectivists generally believe that moral claims are akin to facts, whereas moral subjectivists generally believe that moral claims are more akin to preferences. Evidence from developmental and social psychology has generally favored an objectivist view; however, this work has typically relied on few examples, and analyses have disallowed statistical generalizations beyond these few stimuli. The present work addresses whether morals are represented as fact-like or preference-like, using behavioral and neuroimaging methods, in combination with statistical techniques that can a) generalize beyond our sample stimuli, and b) test whether particular item features are associated with neural activity. Behaviorally, and contrary to prior work, morals were perceived as more preference-like than fact-like. Neurally, morals and preferences elicited common magnitudes and spatial patterns of activity, particularly within dorsal-medial prefrontal cortex (DMPFC), a critical region for social cognition. This common DMPFC activity for morals and preferences was present across whole-brain conjunctions, and in individually localized functional regions of interest (targeting the Theory of Mind network). By contrast, morals and facts did not elicit any neural activity in common. Follow-up item analyses suggested that the activity elicited in common by morals and preferences was explained by their shared tendency to evoke representations of mental states. We conclude that morals are represented as far more subjective than prior work has suggested. This conclusion is consistent with recent theoretical research, which has argued that morality is fundamentally about regulating social relationships.

Keywords: metaethics, morality, social cognition, fMRI, theory of mind

Examining overlap in behavioral and neural representations of morals, facts, and preferences.

Moral claims (e.g. “eating meat is wrong”) can be evaluated on multiple levels. One may agree or disagree with a given claim (a first-order judgment); however, independent of this, one may make a second-order (i.e. *metaethical*) judgment—regardless of whether you agree or disagree, what information does the claim express? Moral objectivists generally believe that moral claims are either true or false, and that this truthfulness is independent of anyone’s personal beliefs (i.e. moral claims are akin to facts). By contrast, moral subjectivists believe that personal beliefs govern whether moral claims are true—or that moral claims cannot be true or false at all (i.e. moral claims are akin to preferences; Sayre-McCord, 1986; for review, see Goodwin & Darley, 2010). Metaethical questions are the subject of intense philosophical debate, yet they are highly relevant to cognitive, social, and moral psychology. Metaethical questions ask how moral information is represented. It is possible that morals are represented as distinct from other sorts of social and non-social information, such as facts and preferences; however, morals, facts, and preferences may also draw on common cognitive processes. Moral objectivists might predict that this common processing should occur between morals and facts, whereas moral subjectivists might predict the same for morals and preferences. In the present work, we address this question of cognitive representation, using a combination of behavioral and neural methods to determine whether morals are represented as more similar to facts or to preferences.

Metaethics and mental state representations

Subjective claims are mind-dependent—their truth depends on the speaker’s mental states (e.g., “chocolate ice cream is better than vanilla” is true *for the speaker* if they believe this). By contrast, objective claims are mind-independent (e.g. “2 + 2 = 4” is true, regardless of what anyone believes; Goodwin & Darley, 2010; Sayre-McCord, 1986). It follows that subjective

claims should evoke mental state representations, because mental state representations are necessary to evaluate the claim[1]. By contrast, objective claims should not necessarily evoke mental state representations, since in this case the mental states are not a precursor for evaluation.

What this all means for moral claims, is that if morals are represented as subjective, then they should elicit greater activity in brain regions responsible for mental state representation. This hypothesis is made testable by recent work in social neuroscience: a set of brain regions—the Theory of Mind (ToM) network—has been consistently implicated in mental state representation (Amodio & Frith, 2006; Decety & Cacioppo, 2012; Saxe & Kanwisher, 2003; Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010; Young & Saxe, 2009; for reviews see Schurz et al., 2014; Van Overwalle, 2009). Within this network, some regions of interest (ROIs) are more active during tasks that involve general forms of social cognition, such as trait inference, or assessing the similarity of others to the self (dorsal/ventral-medial prefrontal cortex; DMPFC, VMPFC; Amodio & Frith, 2006; Decety & Cacioppo, 2012; Harris, Todorov, & Fiske, 2005; Jenkins & Mitchell, 2010; Ma, Vandekerckhove, Van Hoeck, & Van Overwalle, 2012; Mitchell, Banaji, & Macrae, 2005; Ochsner et al., 2005; Schurz et al., 2014; Van Overwalle, 2009; Young & Saxe, 2009). Other ROIs are more active during tasks where participants represent mental states, such as beliefs or intentions (precuneus, and right/left temporoparietal junction; PC, RTPJ, LTPJ; Ciaramidaro et al., 2007; Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011; Fletcher et al., 1995; Gallagher et al., 2000; Gobbini, Koralek, Bryan, Montgomery, &

[1] In the present study, "evaluate" refers to participants rating their agreement with a given claim. Agreement ratings have an advantage over true/false categorization in that they are easy to understand, and critically, they translate well across examples of facts, morals, and preferences. To agree with subjective claims, it is assumed that participants must hold on to some mental state representation (either their own or others', see Saxe, 2009).

Haxby, 2007; Ruby & Decety, 2003; Saxe & Kanwisher 2003; Saxe & Powell, 2006; Vogeley et al., 2001; Young et al., 2010; Young, Cushman, Hauser, & Saxe, 2007; Young, Scholz & Saxe, 2011; Young & Saxe, 2008; 2009). Some ToM ROIs, such as RTPJ, have been shown to play a critical role in moral judgment (Young et al., 2010; Young & Saxe, 2009); however, researchers have hypothesized that these regions are critical to processes underlying moral judgment (e.g. representing intention), rather than being intrinsically "moral areas" (Young & Dungan, 2012), and prior neuroimaging work has generally compared subtypes of moral dilemmas (e.g. intentional vs. accidental violations), as opposed to contrasting moral and non-moral claims. To our knowledge, no prior work has examined neural activity in response to simple moral claims, presented outside of the context of any moral dilemma or judgment.

Given that the ToM network is involved in representing subjective mental states (Saxe & Kanwisher, 2003), we expected that ROIs within this network would be more active as participants read and evaluated preferences and less active for facts. If moral claims require processing subjective mental states (i.e. if morals are represented as subjective), then they too should elicit neural activity in the ToM network, and the extent that these regions overlap with those activated by preferences can act as one metric of their shared cognitive processes (likewise with common activity between morals and facts throughout the brain). Analyses of item features in these regions (described in detail below) can fine-tune inferences about common representations even further.

Metaethics and moral psychology

It is important to situate the present work within the large body of prior research in moral psychology. A great deal of this research might be roughly split into two categories: a) the study of moral judgment and behavior—e.g. moral judgment in response to dilemmas, (e.g. Cushman,

Young, & Hauser, 2006; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Patil, Melsbach, Henning-Fast, & Silani, 2016); the development of moral-based social preferences (e.g. Hamlin, Wynn, & Bloom, 2007); cooperation and behavioral economics (e.g. Rand, Greene, & Nowak, 2012)—and b) the moralization of distinct behaviors (e.g. Schein & Gray, 2015; Graham, Nosek, Haidt, Iyer, Koleva, & Ditto, 2011; Gray, Young & Waytz, 2012; Iyer, Koleva, Graham, Ditto, & Haidt, 2012). For instance, within this latter category, Moral Foundations theorists have proposed that moral violations can be classified into five domains, of which political liberals are primarily concerned with harm and fairness and political conservatives are additionally concerned with loyalty, authority, and purity (Graham et al., 2011). By contrast, other theories have proposed that all of these domains are reducible to harm (Schein & Gray, 2015), or a dyad involving an intentional agent and a suffering victim (Gray et al., 2012). The present work does not fit neatly into either category. We do not ask for moral judgments, and we are not concerned with what makes a claim moral or not. Instead, we simply validated that claims were perceived as moral, then observed what behavioral and neural overlap with facts and preferences this entailed. That said, between the two dominant approaches our method and our findings may be more relevant to the latter—we expand on this in the general discussion.

Our central question concerns how people represent moral information (regardless of what makes it moral), and the work that bears the most direct relevance may lead one to predict that people represent morals as objective (i.e. fact-like). The distinction between morals as either fact-like or preference-like has a parallel in developmental psychology, where research has demonstrated that children and adults draw a distinction between moral and conventional violations (Nichols & Folds-Bennett, 2003; Smetana, 1981; Tisak & Turiel 1988; Turiel, 1978;

Wainryb et al., 2004). In this case, moral violations refer to actions that are universally wrong (e.g. hitting another child is wrong, not just here, but everywhere). Conventional violations refer to actions that are only locally disallowed (e.g. you may not wear pajamas to class, but there may be other schools where you may). Thus, under this paradigm, morals are definitionally objective claims. Recent work in social psychology and experimental philosophy adds some nuance to this moral-conventional distinction: although morals are largely perceived as fact-like, some moral claims are perceived as more objective than others (Beebe, 2014; Goodwin & Darley, 2008; 2012; Heiphetz & Young, in press; Sarkissian et al., 2011; Wright et al., 2013). The present work uses several methodological advances to better characterize the cognitive representation of the moral domain. First, we use a novel approach to measure metaethical judgments that avoids constraining participants' responses. Second, we use an analytical approach that can generalize beyond our set of example stimuli, and can account for item features that may coincide with domain differences (such as intrinsic differences in valence between morals, facts and preferences). We describe each advance in turn below.

Measuring metaethics

Measuring metaethical judgment requires that researchers create questions that are interpretable to an audience without philosophical training. This has been a methodological concern throughout prior work; for instance, researchers have argued that it would be "a somewhat pointless exercise to ask naïve participants to produce fine distinctions between sophisticated meta-ethical views. .... [Instead, researchers] need ways to pose questions about the topic that are understandable to human participants without philosophical training" (Goodwin & Darley, 2010, p. 165). To solve this problem, it has been proposed that researchers ask "whether people take their [moral] beliefs to be objectively true statements of fact, or alternatively,

subjective preferences or attitudes" (Goodwin & Darley, 2010, p. 165). For instance, participants may read moral propositions—alongside propositions about social conventions, aesthetic tastes, and scientific facts—and categorize each as true, false, or an opinion/attitude (Goodwin & Darley, 2008). The present work builds on this approach.

We wanted to test participants' intuitions about metaethics without unnecessarily constraining their responses. Prior work has typically imposed a zero-sum relationship between judgments of morals as objective or subjective (e.g. Goodwin & Darley, 2008; 2012), and, while this may reflect the philosophical distinction, it also constrains how participants are allowed to express their intuitions. It is possible that participants see morals as both fact-like and preference-like to some extent, and a categorical (or one-dimensional) approach rules out this outcome before testing it. To address this, we had participants read moral claims (among facts and preferences) and make a comparison. Rather than categorizing claims (e.g. "eating meat is wrong") as either objective or subjective, participants rated each claim on three scales, presented simultaneously (Figure 1): "To what degree is this statement about [facts, morals, preferences]?" We expected that all morals would be perceived as moral-like; however, the question of interest was which secondary feature would dominate. Are morals, overall, perceived as more fact-like or more preference-like?

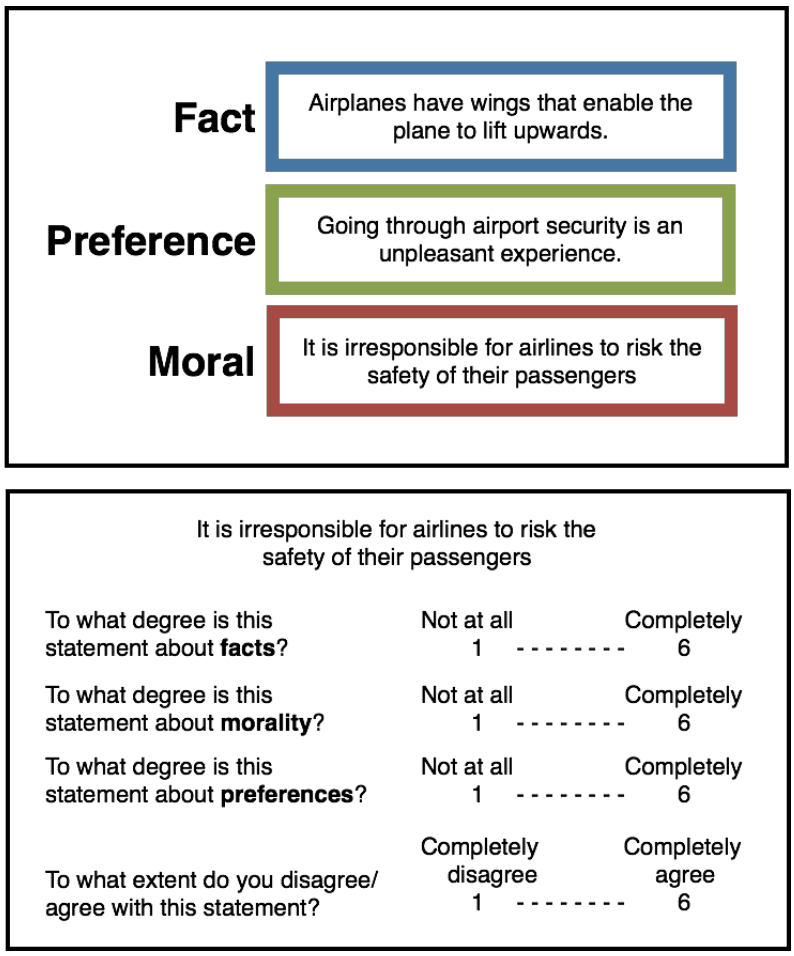

Figure 1. Sample Stimuli and Behavioral Task. Participants read 72 claims in total, evenly divided between morals, facts, and preferences. For each claim, all rating prompts were presented simultaneously, and there was no explicit indication as to whether any claim was a moral, fact, or preference. See Appendix A for the full text of all stimuli.

Analytic approach

Our analytical approach differed from prior work in that it allowed for statistical generalizations beyond the sampled set of stimuli. Researchers face a particular set of statistical hurdles when comparing domains (e.g. morals, facts, and preferences), where those domains are comprised of sets of example stimuli. Although one can never test every possible moral claim (e.g. "eating meat is wrong" is one of countless possible moral claims), with enough examples one might hope that the results are generalizable past the specific set. Unfortunately, this hope is not statistically supported (Clark, 1973; Cornfield & Tukey, 1956; Judd, Westfall & Kenny,

2012). To generalize beyond a sample of stimuli, one must treat those stimuli as random effects (while at the same time treating subjects as random effects). This "crossed random effects" design is not possible in many traditional analyses (e.g. ANOVA). For instance, averaging across stimuli in each domain and then performing traditional analyses across subjects is not sufficient. In this case, one is only licensed to conclude that the result would replicate in another group of subjects *with the exact same set of stimuli*—domains and their exemplars are perfectly confounded, and, under normal circumstances, Type I error rates for conclusions about domain differences can exceed 50% (Judd et al., 2012; Westfall, Kenny, & Judd, 2014). In the present work, we used linear mixed effects analyses, modeling crossed random effects for subjects and stimuli (Baayen, Davidson, & Bates, 2008; Judd, Westfall, & Kenny, 2012; Westfall, Kenny, & Judd, 2014). This analytic technique allowed us to statistically account for the heterogeneity of stimuli in each domain, meaning that our conclusions are generalizable beyond our specific examples, applying instead to sampled populations of morals, facts, and preferences.

Of course, morals, facts, and preferences also differ in intrinsic ways, and these intrinsic differences will be confounded with domain differences. This is particularly concerning for neural analyses; some brain regions may be active for both morals and facts (or for both morals and preferences), but presumably this activity is related to some more basic feature of the stimuli (e.g. valence, reading ease), rather than the socially constructed domain (Young & Dungan, 2012). Item analyses allowed us to turn this confound to our advantage (e.g. Bruneau, Dufour, & Saxe, 2013; Dodell-Feder et al., 2011): given that domains (and the stimuli that comprise them) differ in intrinsic ways, which features of these stimuli are related to neural activity? We can determine the item features responsible for domain differences by first identifying domain differences in neural activity (e.g. within a ROI, morals and preferences may elicit greater

activity than facts) and then adding item features (e.g. valence) as covariates. If particular item features can reduce the initial domain difference to non-significance, then they may explain *why* morals elicit common activity with facts or preferences. This analysis is directly related to our central aim: we want to know if morals are represented as similar to facts or preferences, and item analyses license more specific inferences about the dimensions responsible for this similarity.

Present work

The present work is concerned with the cognitive representation of moral claims: do people represent morals as more similar to objective facts or to subjective preferences? Study 1 probed this question behaviorally, simultaneously asking participants to rate the extent that morals (presented among facts and preferences) were "about [morals, facts, and preferences]." This method was selected to make metaethical questions interpretable without constraining participants' responses. Study 2 examined neural activity as participants evaluated claims about morals, facts, and preferences (rating their agreement with each). First, we performed a whole-brain random effects analysis to identify brain regions where morals and facts (or morals and preferences) elicited activity in common. Next, we examined activity within ToM ROIs (regions implicated in social cognition) and used item analyses to examine the relationship between ROI activity and item features, collected from independent online samples and from text analysis software (Coh-Metrix 3.0; Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Louwerse, Cai, & Graesser, 2014). These item analyses allowed us to identify which underlying features were responsible for observed differences in neural activity between morals, facts, and preferences.

Study 1

Method

Participants. We recruited participants online using Amazon Mechanical Turk (AMT) at an approximate rate of $5/hour, in line with standard AMT compensation rates. Our final sample consisted of 68 adults (36 female; M_{Age} = 34.0 years, SD_{Age} = 11.1 years), after excluding 11 participants for failing a simple attention check that asked them to describe any claim they had read. Using standard assumptions about variance components among random effects (Westfall, Judd, & Kenny, 2014), our subjects and stimuli should allow us to detect effects sizes as small as .303 at 80% power. The Boston College Institutional Review Board approved Studies 1 and 2, and each participant provided consent before beginning.

Procedure. Participants were instructed that they would read a series of claims, and, for each, rate their agreement and the extent to which it was about facts, about morals, and about preferences (*Dimension*: fact-like/moral-like/preference-like). Agreement was measured with a single question: "To what extent do you disagree/agree with this statement?" (1 – "Completely disagree"; 6 – "Completely agree"). Dimension ratings were presented as a set of three questions: "To what degree is this statement about [facts, morals, preferences]" (1 – "not at all"; 6 – "completely")? These questions were presented simultaneously, and their order was counterbalanced across participants. Claims were presented one at a time, at the top of the page, and participants were given no indication that any claim was designed to be a fact, moral, or preference.

At the end of the survey, participants answered two brief questionnaires (not discussed in this paper) about their general stance toward moral objectivity (Forsyth, 1980) and consequences of that stance. Following these questionnaires, participants provided demographic information. Participants were generally socially liberal (M = 5.3, SD = 1.6, 7-point scale anchored at 1,

"Socially Conservative", and 7, "Socially Liberal"), as indicated by a one-sample t-test against the scale mid-point, $t(67) = 6.82$, $p < .001$, $d = .83$.

Stimuli. Participants read 72 claims in total, divided evenly between content categories (24 facts, 24 morals, and 24 preferences; see Appendix A for the full text of all stimuli). Claims did not contain any mental state markers (e.g., "She thinks," "He believes"), which might have explicitly engaged ToM. Claims within each content category were refined across a series of pilot studies to ensure that the moral claims we generated were not perceived as more fact-like or preference-like than they were moral-like. The present study used the final set of stimuli generated from this process. Content categories also contained consensus sub-categories (i.e. sub-categories were designed to elicit either agreement, disagreement, or no consensus across individuals), which are explored in greater detail elsewhere (Theriault, Waytz, Heiphetz, & Young, under review).

Statistical methods. We use mixed effects analyses throughout this paper, following recommendations to model crossed by-subject and by-item random effects (Baayen, Davidson, & Bates, 2008; Judd et al., 2012; Westfall et al., 2014). This analysis allows for generalizations beyond a sample of participants (as is the case for standard statistical analyses, such as ANOVA), but also beyond a sample of stimuli (which is not the case for most standard statistical analyses). We performed analyses using R (R Core Team, 2016) and the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015), and obtained *p* values for fixed effects using the Kenward-Roger approximation of degrees of freedom, implemented in *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2015) and *pbkrtest* packages (Halekoh & Højsgaard, 2014). We followed the recommendation of Barr, Levy, Scheepers, and Tily (2013) in using the maximal random-intercepts structure justified by the design: we modeled by-subject and by-item random

intercepts, as well as all by-subject and by-item random slopes justified by the design. Random slopes were removed from the model only when the model failed to converge (Baayen et al., 2008).

Results

First, we validated our *a priori* content categories (facts, morals, and preferences), using paired *t*-tests to compare mean ratings on our three dimensions (fact-like, moral-like, preference-like; Figure 2). Consistent with our design, facts were perceived as more fact-like, $M_{Fact\text{: Fact-like}} = 5.46$, than moral-like, $M_{Fact\text{: Moral-like}} = 1.13$, or preference-like, $M_{Fact\text{: Preference-like}} = 1.29$, ts > 33, ps < .001, ds > 4.1. Preferences were perceived as more preference-like, $M_{\text{Preference: Preference-like}} = 5.68$, than fact-like, $M_{\text{Preference: Fact-like}} = 1.57$, or moral-like, $M_{\text{Preference: Moral-like}} = 1.21$, ts > 36, ps < .001, ds > 3.9. And morals were perceived as more moral-like, $M_{\text{Moral: Moral-like}} = 4.82$, than fact-like, $M_{\text{Moral: Fact-like}} = 2.11$, $t(67) = 19.6$, $p < .001$, $d = 2.95$ or preference-like, $M_{\text{Moral: Preference-like}} = 4.21$, $t(67) = 3.8$, $p < .001$, $d = 1.64$.

In the analysis above, morals emerged as principally moral-like, but also as largely preference-like. Indeed, when repeating the analysis using a maximal mixed effects model, morals just barely remained significantly more moral-like than preference-like, $z = 2.02$, $p = .043$ (for full mixed effects analysis, see Table S1 of the online supplemental materials). By contrast, even in this more stringent mixed effects model, morals were robustly more preference-like than fact-like, $z = 7.45$, $p < .001$.

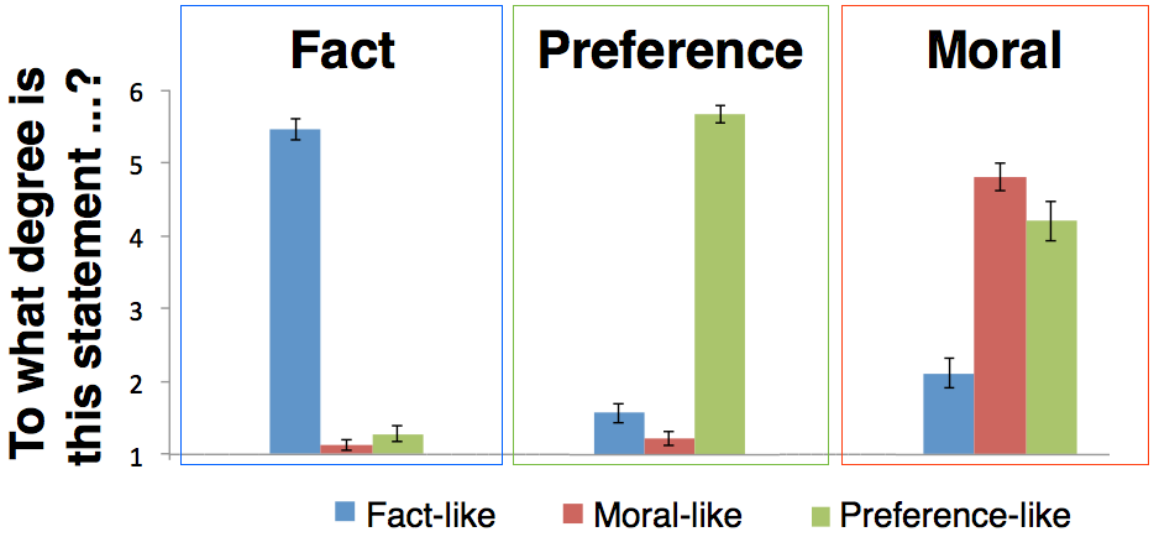

Figure 2. Behavioral Ratings. Claims were rated highest on their content-consistent dimension (e.g., facts were rated as fact-like), but morals were also rated as more preference-like than fact-like. Error bars indicate 95% confidence intervals. For estimates derived from the mixed effects analysis, see Table S1 of the online supplemental materials.

Discussion

According to Study 1, if participants are allowed the flexibility to rate moral claims as any combination of moral-like, fact-like, and preference-like, then moral claims are perceived as highly preference-like. This result is surprising, as prior work has suggested that morals are largely seen as objective (Nichols & Folds-Bennett, 2003; Smetana, 1981; Tisak & Turiel 1988; Turiel, 1978; Wainryb et al., 2004). Although recent work has demonstrated that this objectivity is variable—some moral claims are more objectivity than others (Goodwin & Darley, 2008; 2012)—the conclusion remained the same: morals are perceived as highly objective. It is possible that our sample of stimuli were exceptional in some way, and that in another sample morals would be perceived as more fact-like and less preference-like; however, this is unlikely, as we were able to replicate the effect in an independent sample of stimuli, derived from items used in the Moral Foundations Questionnaire (Graham et al., 2011; Iyer et al., 2012; see supplemental study in the online supplemental materials). Still, based on this behavioral result

alone, it is difficult to answer *why* exactly morals and preferences are perceived as similar, i.e. what are the underlying features that are responsible for their perceived similarity? We aimed to address this question in Study 2, performing a neural analysis, paired with an analysis of item features[2].

Study 2

Behaviorally, morals were perceived as highly preference-like. Morals and preferences may also elicit neural activity in common, and the brain regions in which this common activity occurs can help us better understand the basis of their similarity; however, reverse inferences such as these are also extremely limited in their explanatory power (Poldrack, 2006). Thus, we also use item analyses to supplement our interpretation. Most likely, morals and preferences are intrinsically different from facts along many dimensions (e.g. emotional valence, social relevance). Of these dimensions, some may explain common neural activity better than others. In our item analysis, we tested several item features (using stimuli ratings collected from independent online samples), asking whether any particular feature could explain common activity elicited by morals and preferences, relative to facts. We were particularly interested in the ToM network, given its role in representing subjective mental states; thus, we used an established independent functional localizer to identify regions of interest (ROIs) in this network (Dodell-Feder et al., 2011; Koster-Hale et al., 2013; Saxe & Kanwisher, 2003; Young et al., 2007; 2010; 2011).

[2] The fMRI data used in Study 2 is also analyzed in a separate study (Theriault et al., under review). Analyses are not repeated between the two studies: the present study focuses on domain-level similarity between morals, facts, and preferences, and attempts to explain similarity on the basis of item features. The separate study focuses on the relationship between neural activity and within-domain variability in metaethical judgment (i.e. why are some moral claims seen as more objective than others?).

Method

Participants. Our final sample consisted of 25 right-handed participants (12 female, 12 male, 1 unspecified; M_{age} = 27.1 years, SD_{age} = 5.4 years), recruited through an online posting for a $65 cash payment (two additional participants were recruited but were not analyzed due to excessive movement, which was identified during spatial preprocessing, before any analysis was performed). Of these 25 participants, two completed only a subset of the scan session runs: one completed only the first five runs due to experimenter error, and in another, a movement artifact during run 4 rendered only the first three runs useable. These partial cases were included in all analyses except for multi-voxel pattern analysis (MVPA), a technique that used iterative combinations across the full set of runs to compute correlations, such that any data loss would drastically reduce the number of combinations. For another one of the 25 participants, we were unable to collect post-scan ratings. Participants were a community sample of native English speakers with no reported history of learning disabilities, previous psychiatric or neurological disorders, or a history of drug or alcohol abuse.

Procedure. Participants completed the study during a single session. Twenty were run at the Center for Brain Science Neuroimaging Facility at Harvard University, and an additional five were run at the Martinos Imaging Center at the Massachusetts Institute of Technology. Scanning parameters and equipment were identical between sites (see below). Inside the scanner, participants underwent a structural scan and then performed the experimental task. Participants read each claim and reported their agreement (1 – “strongly agree”; 4 – “strongly disagree”; scores were reverse coded for convenience). Participants were also allowed to use their thumb to

indicate "don't know," which was coded as an empty cell[3]. We presented stimuli across six runs (12 claims per run, evenly divided between facts, morals and preferences). Each trial began with the presentation of a claim (6 s), followed by an agreement rating (+4 s), followed by fixation (+12 s). Each experimental run was 4 min 52 s long, totaling 29 min 12 s across 6 runs; the total scan time was 68 min 8 s due to the inclusion of a structural scan (6 min 3 s), a functional localizer (two 4 min 46 s runs), and a second study not reported here (involving responses to moral dilemmas; 29 min 12 s). Stimuli were presented in white text on a black background on a projector, viewable through a mirror mounted on the headcoil. The experimental protocol was run on an Apple Macbook Pro using Matlab 7.7.0 (R2008b) with Psychophysics Toolbox.

In a post-scan behavioral session, participants re-read all claims on an Apple Macbook Pro and provided dimension ratings for each—"To what degree is this statement about [facts, morals, preferences]" (1 – "not at all"; 7 – "completely")? At the end of the post-scan session they provided additional demographic information. As in Study 1, participants were generally socially liberal (M = 5.3, SD = 2.0, 7-point scale, anchored at 1, "Socially Conservative", and 7, "Socially Liberal"), as indicated by a one-sample t-test against the scale mid-point, $t(22) = 3.18$, $p = .004$, $d = .66$.

Stimuli. Stimuli were the same as those described in Study 1 (see Appendix A for the full text of all stimuli). As in Study 1, content categories also contained consensus sub-categories. ROI activity in response to these subcategories is explored in greater detail elsewhere (Theriault et al., under review).

[3] This "don't know" option was provided to avoid confusion, as a subset of facts was designed to be generally unknown to participants, making agreement responses ambiguous. The majority (71.6%) of "don't know" responses were within this sub-group category, and the next highest occurrence was 7.3% for an equivalent group of preferences, designed to not elicit strong agreement or disagreement.

fMRI imaging and analysis. Scanning was performed using a 3.0 T Siemens Tim Trio MRI scanner (Siemens Medical Solutions, Erlangen, Germany) and a 12-channel head coil at both the Center for Brain Science Neuroimaging Facility at Harvard University, and the Martinos Imaging Center at the Massachusetts Institute of Technology. Thirty-six slices with 3mm isotropic voxels, with a 0.54mm gap between slices to allow for full brain coverage, were collected using gradient-echo planar imaging (TR = 2000 ms, TE = 30 ms, flip angle = 90°, FOV = 216 x 216 mm; interleaved acquisition). Anatomical data were collected with T1-weighted multi-echo magnetization prepared rapid acquisition gradient echo image (MEMPRAGE) sequences (TR = 2530 ms, TE = 1.64 ms, FA = 7°, 1mm isotropic voxels, 0.5mm gap between slices, FOV = 256 x 256 mm). Data processing and analysis were performed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) and custom software. The data were motion-corrected, realigned, normalized onto a common brain space (Montreal Neurological Institute, MNI), and spatially smoothed using a Gaussian filter (full-width half-maximum = 8 mm kernel), and high-pass filtered (128 Hz). Whole-brain conjunction analyses and MVPA were performed using a GLM with three regressors of interest: fact, moral, and preference categories. Analyses within functional ROIs are described in detail below.

Whole-brain conjunction analysis. Whole-brain conjunction analyses compared two whole-brain random effects contrasts, examining activity elicited in common between two content categories compared to the one remaining content category—e.g. (Moral > Fact) + (Preference > Fact). Contrasts were first modeled for each participant, then entered into a second level random effects analysis across all participants. Conjunction analyses compared two of these contrasts at a time, providing a visualization of the voxels that were significant for both contrasts. Following recent recommendations (Eklund, Nichols, & Knutson, 2016), we

performed permutation tests (5000 samples) to achieve a cluster-corrected familywise error rate of $\alpha = .05$ in each contrast, while thresholding voxels at $p < .001$ (uncorrected; recommended by Woo, Krishnan & Wager, 2014). Permutation tests were performed using SnPM 13 (http://warwick.ac.uk/snpm; Nichols & Holmes, 2001),

ToM localizer task. We used an independent functional localizer to identify ROIs associated with ToM (Dodell-Feder et al., 2011). The task consisted of 20 scenarios presented across two 4 min 46 s scans: 10 stories about mental states (false-belief condition) and 10 stories about physical representations (false-photograph condition). Stimuli were matched in complexity; see http://saxelab.mit.edu/superloc.php for the complete set. Each story was presented for 10 s and was followed by a statement about the story that was judged as true or false (4 s). A boxcar for the full duration (14 s) was used to model stories in both conditions. Activity was estimated in each voxel for both conditions, and a simple contrast was performed to estimate voxels showing significantly greater activity for mental stories than physical stories ($p < .001$, $k > 10$). ROIs were defined as contiguous voxels in a 9mm-radius of the peak voxel that passed the contrast threshold (for peak coordinates, see Table S2 of the online supplemental materials).

It was possible that the cluster extent threshold chosen for our functional localizer was too liberal, as it was derived from an arbitrary 10 voxel threshold (with voxels thresholded at $p < .001$). We used this arbitrary threshold was so that our results could be easily compared with prior work, which has used the same parameters (Dodell-Feder et al., 2011; Koster-Hale et al., 2013; Saxe & Kanwisher, 2003; Young et al., 2007; 2010; 2011); however, we also wanted to ensure that our findings were not dependent on it. How best to balance Type I and Type II error when selecting functional ROIs is an open question (Degryse et al., 2017), so we selected ROIs

based on the peak coordinates from a whole brain random effects contrast (belief > photograph) across all participants, and replicated the central analyses below (see supplemental analyses in the online supplemental materials; for peak coordinates, see Table S3 of the online supplemental materials). The results of this analysis are identical to the ROI analyses reported below (Figure S2 of the online supplemental materials).

Functional ROI response magnitude analysis. For our experimental task, we used a slow event-related design to model blood oxygen level dependent (BOLD) activity in each functional ROI. Events were defined as beginning when text first appeared and continuing for the length of the claim and agreement response (10 s). The time-window was adjusted for hemodynamic lag so that data were collected at 4–14 seconds from onset (Dodell-Feder et al., 2011). To model neural activity in each ROI, we transformed BOLD activity at each time point of the experimental task into percent signal change (PSC = raw BOLD magnitude for (condition – fixation)/fixation). The data at each time point were centered at the mean PSC of the run. Given that we center PSC for each run, there is no simple interpretation of our ROI findings with respect to the x-axis; this is not a concern, as the comparisons of interest are between conditions. Averaging run-centered PSC across the duration of the scenario provided a single PSC value for each ROI, for each participant, for each condition.

ROI multi-voxel pattern analysis. For each functional ROI, MVPA compared spatial patterns of activity between two conditions. We used the Haxby split half method (Haxby et al., 2001), splitting each participant's unsmoothed BOLD activity into two equal sets of runs (partitions). A vector of βs represented the voxels in each ROI, and this vector was averaged separately in each partition. MVPA compared correlations *within* and *between* conditions. *Within correlations* correlated vectors across partitions within one condition, while *between correlations*

correlated vectors across partitions between the two conditions being compared. Correlations were Fisher transformed and calculated across all possible iterations of partitions (e.g. 1, 2, 3 vs. 4, 5, 6; 1, 2, 4 vs. 3, 5, 6; etc.). Subject-wise classification accuracy within a contrast was calculated across iterations by summing cases in which the within correlation exceeded the between correlation and dividing by the total number of comparisons. A contrast was significant if, across participants, classification accuracy exceeded chance (50%) in a one-tailed, one sample t-test. Note that our approach to MVPA relied on correlational distance (as opposed to Euclidean distance, Mahalanobis distance, etc.), meaning that any observed differences are independent of condition differences in the ROI response magnitude analyses described above (Norman, Polyn, Detre, & Haxby, 2006).

Item analyses. We performed mixed effects analyses using R (R Core Team, 2016), the *lme4* package (Bates et al., 2015), the Kenward-Roger approximation of degrees of freedom (*lmerTest*, Kuznetsova et al., 2015; *pbkrtest*, Halekoh & Højsgaard, 2014), and the maximal justified random-intercepts structure (Baayen et al., 2008; Barr, et al., 2013). Several item features were used as covariates, which might rule out alternative hypotheses. These included features explored in prior work (Dodell-Feder et al., 2011): arousal/valence ($N_{\text{Subjects}} = 17$), ratings ($N_{\text{Subjects}} = 18$), the presence of a person ($N_{\text{Subjects}} = 20$), and arousal/valence ($N_{\text{Subjects}} = 17$; note that arousal and valence were measured using two unipolar positivity and negativity scales—based on prior work, arousal was the sum of these scales and valence was the difference; Kron, Goldstein, Lee, & Gardhouse, 2013). These data were collected from independent online samples in which participants read the complete set of stimuli from Study 1. We also examined mean Study 1 item-wise agreement ratings (as opposed to in-scanner ratings from Study 2, where the range of response was restricted to a 4-point scale). Additional covariates measured syntactic

and semantic features of claims—i.e. word count, reading ease, anaphor reference, intention verb incidence, causal verb incidence, causal verb ratio, noun concreteness, noun familiarity, noun imageability, negation density, number of modifiers, and left embeddedness (see Table S8 for covariate summary statistics; see Appendix B for complete descriptions of covariates). Syntactic and semantic covariates were collected using *Coh Metrix 3.0* (http://cohmetrix.com), an online linguistic analysis tool (Graesser, McNamara, Louwerse, & Cai, 2004). Finally, we collected reaction times in response to the in-scanner rating task, and this was included as a nuisance parameter in all final models.

Results

Behavioral results. We collected fact-like, moral-like, and preference-like ratings for each claim in a post-scan behavioral session. These ratings were consistent with the patterns observed in Study 1. In a maximal mixed effects analysis, people perceived morals as more preference-like than fact-like, $z = 4.4$, $p < .001$ (for full results, see Table S4 of the online supplemental materials).

Neural results. Study 1 and the behavioral results from Study 2 suggest that morals are generally perceived as more preference-like than fact-like. Here, we asked whether morals and preferences, relative to facts, also elicit neural activity in common. First, we performed a series of whole-brain conjunction analyses, mapping common activity across two contrasts. Of these, the conjunction of (Moral > Fact) + (Preference > Fact) revealed the most activity in common (Figure 3a), with overlap in both DMPFC (peak coordinates: moral > fact [-4, 56, 30], preference > fact [-2, 54, 24]) and VMPFC (peak coordinates: moral > fact [2, 48, -12], preference > fact [4, 40, -20]). By contrast, the conjunction of (Moral > Preference) + (Fact > Preference) revealed no activity in common (Figure 3b). Notably, although less relevant to our key questions, we found

that preferences and facts, relative to morals, elicited common activity in left middle frontal gyrus, and bilateral superior parietal lobule (Figure S1 of the online supplemental materials); this was notable because, in terms of whole-brain neural activity, facts appeared to have more in common with preferences than with morals (for peak cortical coordinates of each contrast, see Table S5 of the online supplemental materials). Thus, morals and preferences, relative to facts, appear to elicit neural activity in common, particularly within medial prefrontal cortex.

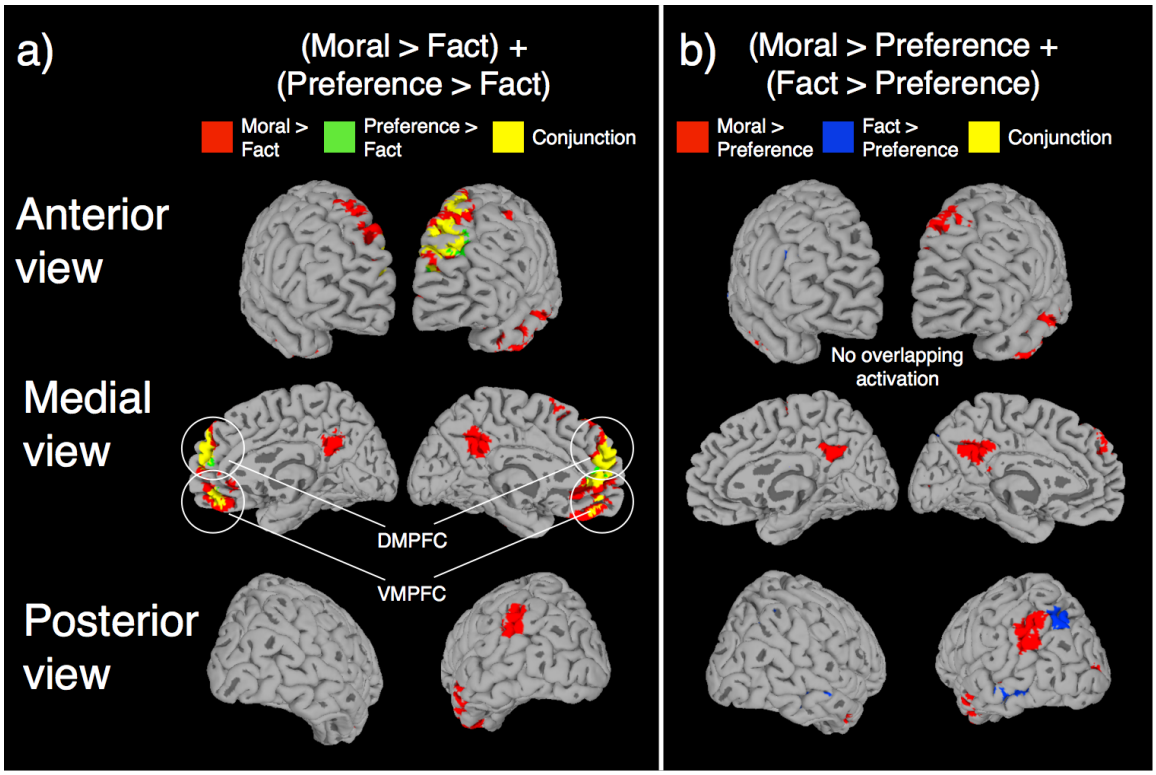

Figure 3. Whole-brain Conjunction Analyses. (a) Morals and preferences, relative to facts, elicited common activity in DMPFC and VMPFC. (b) Morals and facts, relative to preferences, did not elicit any activity in common. Permutation tests (5000 samples) were used to achieve a cluster-corrected familywise error rate of $\alpha = .05$ in each contrast, while thresholding voxels at $p < .001$ (uncorrected). Permutation testing was performed using SnPM 13 (http://warwick.ac.uk/snpm; Nichols & Holmes, 2001). Peak coordinates for each contrast are reported in Table S5 of the online supplemental materials.

To more directly probe neural activity related to ToM, we performed analyses within ToM ROIs (DMPFC, VMPFC, PC, RTPJ, LTPJ) identified for each individual in an independent functional localizer task. This analysis depended on observing a significant contrast between

localizer conditions for each ROI, meaning that *N* for each ROI varied based on successful localization (N_{DMPFC} = 20/25; N_{VMPFC} = 20/25; N_{PC} = 23/25; N_{RTPJ} = 25/25; N_{LTPJ} = 24/25). For each ROI, we performed a repeated measures ANOVA comparing neural activity for morals, facts, and preferences, followed by condition contrasts. Contrast *p* values are corrected for three comparisons to achieve a familywise α of .05 within each ROI ($p_{corrected}$ = .0167). In *Item Analysis*, we also present linear mixed effects analyses, which are capable of generalizing beyond our sample of stimuli.

ROI analyses were consistent with the whole-brain analyses. Morals and preferences, relative to facts, both elicited greater activity in DMPFC and VMPFC (Figure 4). One-way ANOVAs revealed a main effect of content in DMPFC, $F(2, 38) = 33.41$, $p < .001$, $\eta_p^2 = .31$, where both morals, $z = 7.65$, $p < .001$, $d = 1.71$, and preferences, $z = 6.32$, $p < .001$, $d = 1.41$, elicited greater activity than facts. Likewise, in VMPFC, $F(2, 38) = 12.11$, $p < .001$, $\eta_p^2 = .15$, both morals, $z = 3.87$, $p < .001$, $d = 1.09$, and preferences, $z = 3.01$, $p = .006$, $d = 0.68$, elicited greater activity than facts. In both DMPFC and VMPFC, there was no significant difference in neural activity elicited by morals and preferences: DMPFC, $z = 1.33$, $p = .377$, $d = 0.30$; VMPFC, $z = 1.81$, $p = .166$, $d = 0.40$. Thus, in DMPFC and VMPFC, morals and preferences appear to elicit common neural activity.

In PC, RTPJ, and LTPJ, morals elicited greater activity than both facts and preferences. In LTPJ preferences also elicited greater activity than facts; this contrast was marginal in PC and non-significant in RTPJ (Figure 4). One-way ANOVAs revealed a main effects of content in: (a) PC, $F(2, 44) = 25.66$, $p < .001$, $\eta_p^2 = .20$, such that morals elicited greater activity than both facts, $z = 6.99$, $p < .001$, $d = 1.46$, and preferences, $z = 4.84$, $p < .001$, $d = 1.01$, while preferences elicited marginally more activity than facts, $z = 2.15$, $p = .080$, $d = 0.45$; (b) RTPJ, $F(2, 48) =$

10.85, $p < .001$, $\eta_p^2 = .09$, such that morals elicited greater activity than both facts, $z = 4.54$, $p < .001$, $d = 0.91$, and preferences, $z = 3.17$, $p = .004$, $d = 0.63$, while preferences and facts did not differ, $z = 1.37$, $p = .355$, $d = 0.27$; and (c) LTPJ, $F(2, 46) = 32.2$, $p < .001$, $\eta_p^2 = .18$, such that morals elicited greater activity than both facts, $z = 8.01$, $p < .001$, $d = 1.63$, and preferences, $z = 4.43$, $p < .001$, $d = 0.90$, while preferences also elicited greater activity than facts, $z = 3.58$, $p = .001$, $d = 0.73$. Thus, in PC, RTPJ, and LTPJ, morals appear to elicit greater activity than both facts and preferences.

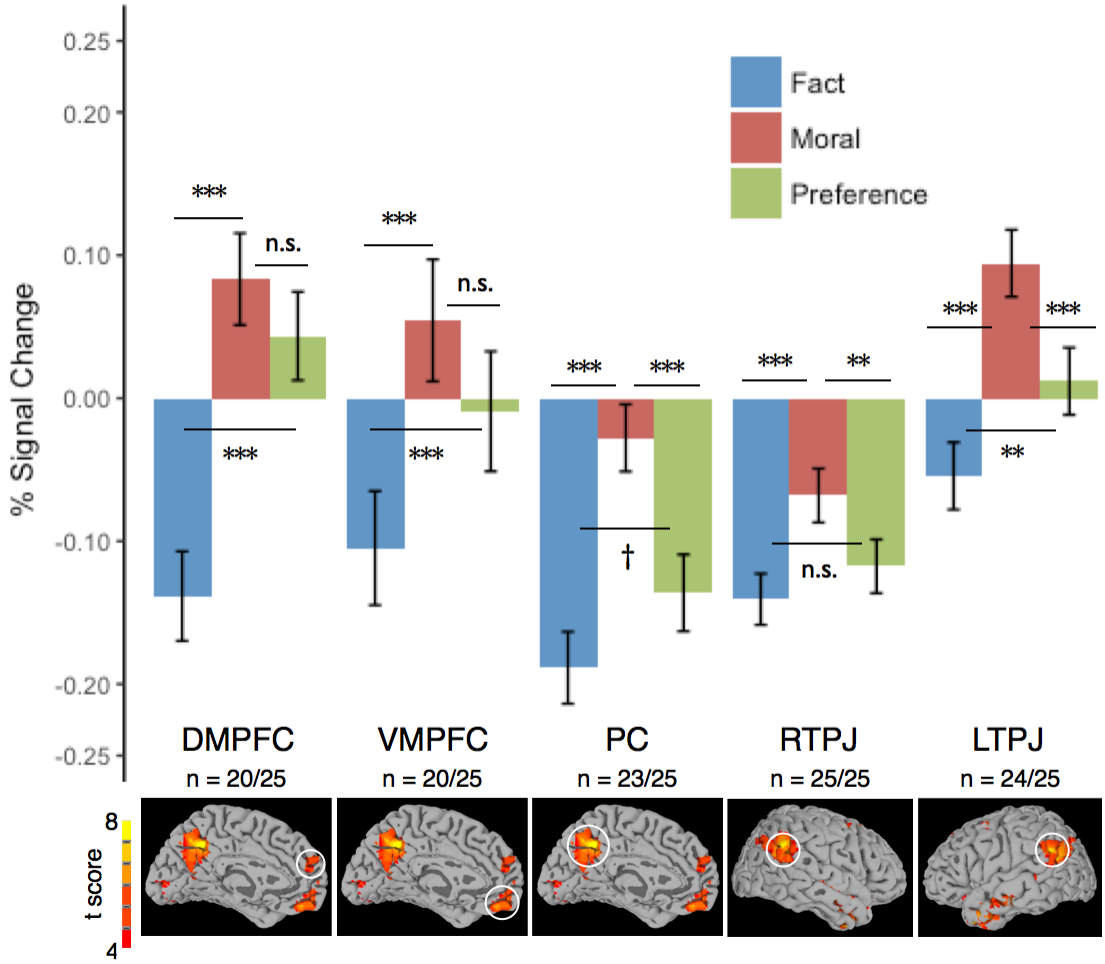

Figure 4. Response Magnitude Across Content (fact/moral/preference) and ROIs. Morals and preferences both elicit greater activity than facts in DMPFC and VMPFC, whereas morals elicit greater activity than both facts and preferences in PC, RTPJ, and LTPJ. ROIs were identified for each individual using an independent functional localizer (Dodell-Feder et al., 2011), meaning that *N* for each ROI varies based on successful localization. Error bars indicate 95% confidence

intervals of condition means. *** $p < .001$; ** $p < .01$; * $p < .05$; † $p < .10$. For mixed effects regression analysis coefficients, see Table 1 and Table S6 of the online supplemental materials.

To test the specificity of the effects we observed in the ToM ROIs, we also explored a set of ROIs hypothesized to have no unique relation to social cognition. It was possible that morals could elicit activity more similar to facts in these non-social brain regions. We defined seven ROIs using peak coordinates from the reverse inference map for the term "working memory" at neurosynth.org (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011; for peak coordinates, see Table S7 of the online supplemental materials). ROIs were: left/right anterior middle frontal gyrus; left/right posterior middle frontal gyrus; left/right supramarginal gyrus; and medial superior frontal gyrus. For each, we defined a 9mm spheres around the peak coordinate. PSC was extracted using the same method as for functional ROIs. Across these ROIs, there was no evidence that moral claims were processed as more similar to facts, compared to preferences (see supplemental analyses and Figure S3 of the online supplemental materials).

MVPA provided us with an additional method of comparing neural representations of morals to facts and preferences: it allowed us to examine how easily categories could be distinguished by spatial correlations between their voxel-wise activity. We tested whether MVPA could more easily distinguish between morals and facts, or between morals and preferences. Importantly, we conducted MVPA using a correlational distance metric, meaning that the analysis was independent of overall mean differences (i.e. independent of the ANOVA analyses above). For each ROI within each participant, we used iterative split-half correlations (Haxby et al., 2001) to generate discrimination accuracy scores for the two contrasts (Moral-versus-Fact, Moral-versus-Preference). In each ROI, paired sample *t*-tests compared contrast discrimination accuracy (Figure 5). *P* values reflect significance after Bonferroni correction for multiple comparisons to achieve familywise $\alpha = .05$ across five comparisons ($p_{corrected} = .01$). The

classifier was significantly more accurate at discriminating between morals and facts, compared to morals and preferences in three of our five ROIs: DMPFC, $t(18) = 4.30$, $p = .002$, $d = 0.99$, PC, $t(20) = 5.81$, $p < .001$, $d = 1.27$, and LTPJ, $t(21) = 2.99$, $p = .035$, $d = 0.64$; the effect was marginal in RTPJ, $t(22) = 2.04$, $p = .266$, $d = 0.43$, and VMPFC, $t(17) = 1.90$, $p = .370$, $d = 0.45$. Thus, independent of mean differences in the magnitude of neural activity, morals are represented as more similar to preferences than to facts in DMPFC, PC, and LTPJ.

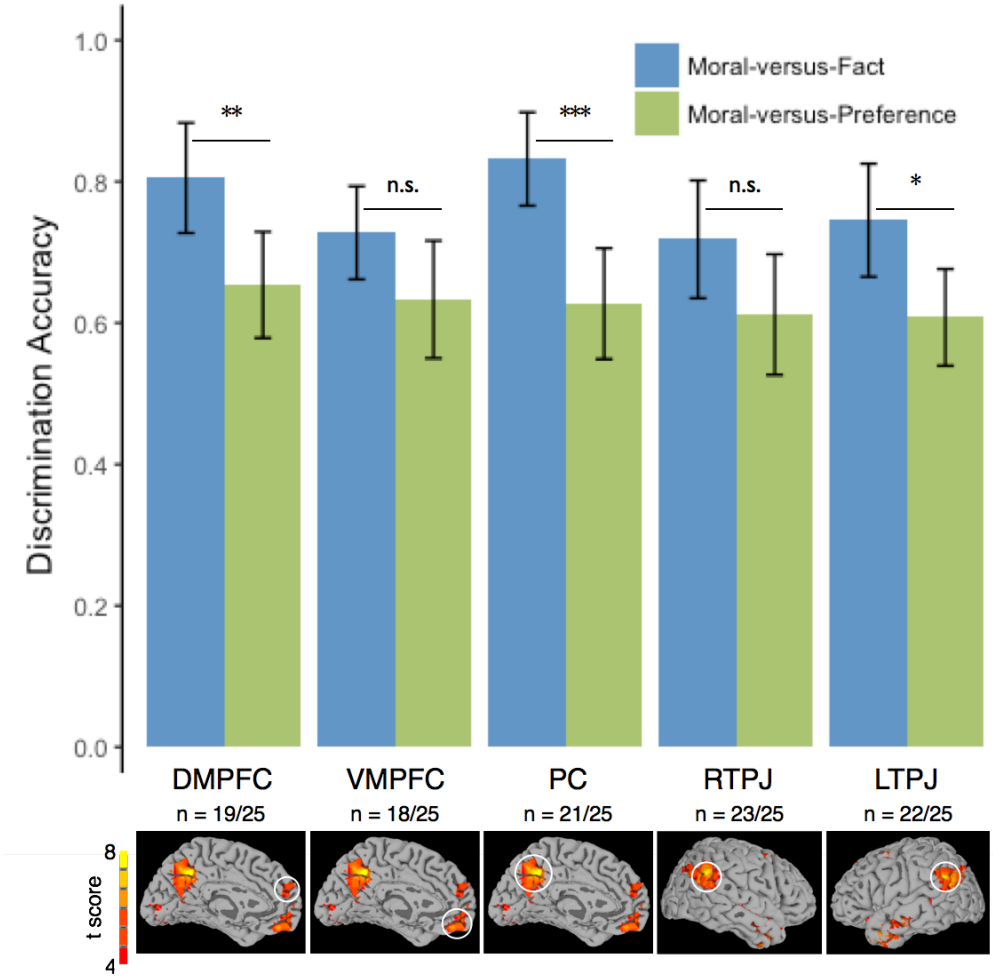

Figure 5. MVPA Discrimination Accuracy for Morals Versus Facts and Preferences. In DMPFC, PC, and LTPJ, morals and facts are more accurately discriminated than morals and preferences, based on the spatial correlation of voxel-wise activity (i.e. independent of mean differences in neural activity, presented in Figure 4). ROIs were identified for each individual using an independent functional localizer (Dodell-Feder et al., 2011), meaning that *N* for each ROI varies based on successful localization. Two participants had partial data and were excluded from this

analysis. Error bars indicate 95% confidence intervals. *** $p < .001$; ** $p < .01$; * $p < .05$; † $p < .10$.

Item Analysis

Our analyses above demonstrated that morals, facts, and preferences elicit different magnitudes and patterns of activity in ToM ROIs. Item analyses using linear mixed effects models allowed us to improve on these analyses in two ways: a) by modeling by-item random effects, allowing us to generalize beyond our specific sample of items—a step that has rarely been taken in prior work (*c.f.* Judd et al., 2012), and b) by including covariates measuring item features (collected in independent samples; see Appendix B), allowing us to address *why* morals and preferences elicited activity in common. For each ROI, mixed effects models were built in three steps (Table 1 & Table S6 of the online supplemental materials). First, we replicated the ROI analyses reported above: dummy coding morals and preferences against facts while controlling for the maximal by-subject and by-item random effects structure. Next, we identified which item features were viable covariates: we dropped the dummy coded categories from our model and modeled each covariate as a single fixed effect predicting ROI activity, controlling for by-subject and by-item random intercepts. (Alternatively, we could choose covariates by identifying which item features differ across domains; for this analysis, see Tables S8 and S9 of the online supplemental materials). Finally, significant covariates were entered as fixed effects (Barr et al., 2013) one at a time to the initial model, in the order of their significance (noting if and when categorical effects of morals and preferences became marginal or non-significant). Reaction time was considered a nuisance parameter and was always controlled for after accounting for significant covariates.

Mixed effects analyses within ROIs were consistent with the ANOVAs reported above (Table 1 & Table S6 of the online supplemental materials). Both morals and preferences elicited

greater activity than facts in DMPFC (morals, b = 0.222, $t(35.1) = 5.94$, $p < .001$; preferences, b = 0.182, $t(40.1) = 5.14$, $p < .001$), VMPFC (morals, b = 0.159, $t(32.8) = 3.91$, $p < .001$; preferences, b = 0.098, t(33.8) = 2.15, $p = 0.039$), and LTPJ (morals, b = 0.148, $t(49.5) = 5.00$, $p < .001$; preferences, b = 0.066, $t(56.3) = 2.40$, $p = .020$). Morals, but not preferences, elicited greater activity than facts in PC (morals, b = 0.158, $t(58.9) = 4.70$, $p < .001$; preferences, b = 0.051, $t(58.9) = 1.53$, $p = .132$), and in RTPJ (morals, b = 0.072, $t(31.9) = 3.55$, $p = .001$; preferences, b = 0.023, $t(34.6) = 1.35$, $p = .187$).

Table 1. Mixed effects analysis for DMPFC across all claims, examining ROI percent signal change (PSC) for morals and preferences relative to facts.

ROI	Step	Model: R Syntax	Coefficients
DMPFC	***Hypothesis testing***	lmer(PSC ~ Moral + Preference + (1|Item) + (Moral+Preference|ID))	******Moral:*** $\beta = 0.222$, $t(35.1) = 5.94$, $p = 9.1 \times 10^{-7}$ ******Preference:*** $\beta = 0.182$, $t(40.1) = 5.14$, $p = 7.5 \times 10^{-6}$
	Identify potential covariates	lmer(PSC ~ MentalState + (1|Item) + (1|ID))	******Mental States:*** $\beta = 0.078$, $t(70.0) = 8.74$, $p = 7.9 \times 10^{-13}$
		lmer(PSC ~ Arousal + (1|Item) + (1|ID))	******Arousal:*** $\beta = 0.069$, $t(70.1) = 4.33$, $p = 4.9 \times 10^{-5}$
		lmer(PSC ~ NounFamiliarity + (1|Item) + (1|ID))	****Noun Familiarity:*** $\beta = 0.002$, $t(70.1) = 2.38$, $p = .020$
		lmer(PSC ~ NounConcreteness + (1|Item) + (1|ID))	****Noun Concreteness:*** $\beta = -0.0005$, $t(69.8) = 2.27$, $p = .026$
		lmer(PSC ~ PersonPresent + (1|Item) + (1|ID))	****Person Present:*** $\beta = 0.077$, $t(70.0) = 2.19$, $p = .032$
		lmer(PSC ~ NounImageability + (1|Item) + (1|ID))	****Noun Imageability:*** $\beta = -0.0005$, $t(69.8) = 2.05$, $p = .044$
	Attempt to disprove hypothesis	***Marginal/non-significant model:*** lmer(PSC ~ MentalState + Moral + Preference + (1|Item) + (Moral+Preference|ID))	***Moral:*** $\beta = 0.119$, $t(74.6) = 1.58$, $p = .118$ ***Preference:*** $\beta = 0.098$, $t(71.2) = 1.54$, $p = .129$ ***Mental States:*** $\beta = 0.039$, $t(68.0) = 1.57$, $p = .120$
		Full model: lmer(PSC ~ RT + NounImageability + PersonPresent + NounConcreteness + NounFamiliarity + Arousal + MentalState + Moral + Preference + (1|Item) + (Moral+Preference|ID))	***Moral:*** $\beta = 0.118$, $t(67.4) = 1.52$, $p = .132$ ***Preference:*** $\beta = 0.097$, $t(64.8) = 1.43$, $p = .157$ ***Mental States:*** $\beta = 0.037$, $t(62.2) = 1.30$, $p = .200$ ***Arousal:*** $\beta = 0.004$, $t(62.5) = 0.19$, $p = .849$ *****Noun Familiarity:*** $\beta = 0.002$, $t(60.8) = 2.70$, $p = .008$ ***Noun Concreteness:*** $\beta = -0.00006$, $t(60.5) = 0.12$, $p = .904$ ***Person Present:***

	β = 0.003, t(60.8) = 1.13, p = .262
	Noun Imageability:
	β = -0.00007, t(61.0) = 0.01, p = .990
	Reaction Time:
	β = -0.0009, t(1325.0) = 0.08, p = .940

Remaining ROIs are presented in Table S6 of the online supplemental materials. Analyses were performed using R (R Core Team, 2016), and the *lme4* package (Bates et al., 2015), using the Kenward-Roger approximation of degrees of freedom (*lmerTest*, Kuznetsova et al., 2015; *pbkrtest*, Halekoh & Højsgaard, 2014). *** $p < .001$; ** $p < .01$; * $p < .05$; † $p < .10$. β represent standardized regression coefficients.

Covariate analyses in DMPFC, VMPFC, and LTPJ all indicated that the neural activity elicited by morals and preferences was almost entirely accounted for by their common tendency to evoke thoughts about an agent's mental states—i.e. beliefs, desires, thoughts, experiences (Dodell-Feder, 2011; see Appendix B for complete descriptions of covariates). All covariates were individually entered as fixed effects predicting neural activity, and significant covariates were noted: (a) in DMPFC these were mental state ratings, $p < 1.0 \times 10^{-12}$, arousal, $p < 1.0 \times 10^{-4}$, noun familiarity, $p < .05$, noun concreteness, $p < .05$, the presence of a person, $p < .05$, and noun imageability, $p < .05$; (b) in VMPFC these were mental state ratings, $p < 1.0 \times 10^{-4}$, arousal, $p < .001$, the presence of a person, $p < .05$, and reaction time, $p < .05$; (c) in LTPJ these were mental state ratings, $p < 1.0 \times 10^{-6}$, the presence of a person, $p < 1.0 \times 10^{-4}$, arousal, $p < .01$, intentional verb incidence, $p < .05$, negation density, $p < .05$, reading ease, $p < .05$, and number of modifiers, $p < .05$. We added these potential covariates to our initial model, testing if and when effects of content became marginal or non-significant (Table S6 of the online supplemental materials). (a) In DMPFC, after controlling for mental state ratings, both morals, b = 0.119, t(74.6) = 1.58, p = .118, and preferences, b = 0.098, t(71.2) = 1.54, p = .129, dropped to marginal significance, and remained marginal after controlling for arousal, noun familiarity, noun concreteness, the presence of a person, noun imageability, and reaction time (Table 1). (b) In VMPFC, after controlling for mental state ratings, both morals, b = 0.091, t(70.6) = 0.91, p = .368, and preferences, b = 0.043, t(70.9) = 0.48, p = .629, dropped to non-significance. (c) In

LTPJ, after controlling for mental state ratings, both morals, b = 0.051, $t(74.3) = 0.77$, $p = .443$, and preferences, b = -0.013, $t(74.5) = 0.23$, $p = .819$, dropped to non-significance. Thus, the common neural activity that morals and preferences elicit appears to stem from their tendency to evoke mental state representations.

Covariate analyses revealed that PC and RTPJ activity, elicited by morals, could be explained to some extent by item features related to social cognition. Potential covariates were identified as described above: (a) in PC these were mental state ratings, $p < 1.0 \times 10^{-4}$, the presence of a person, $p < 1.0 \times 10^{-4}$, intentional verb incidence $p < .01$, arousal, $p < .05$, and reading ease, $p < .05$; (b) in RTPJ, these were mental state ratings, reaction time, $p < .001$, mental state ratings, $p < .001$, and noun familiarity, $p < .05$. We added these potential covariates to a model for each ROI, testing if and when the coefficient for morals, dummy coded against facts and preferences, became marginal or non-significant (Table S6 of the online supplemental materials). (a) In PC, morals only dropped to marginal significance after controlling for mental state ratings, the presence of a person, intention verb incidence, and arousal, b = 0.067, $t(63.6) = 1.95$, $p = .055$, and PC remained marginal after adding reading ease and reaction time to the model, b = 0.064, $t(61.0) = 1.64$, $p = .068$. (b) In RTPJ, after controlling for mental state ratings and reaction time, morals dropped to non-significance, b = 0.030, $t(43.6) = 1.63$, $p = .110$. Thus, the activity elicited by moral claims in PC and RTPJ can be explained to some extent by their tendency to evoke mental state representations, although in PC this may not completely explain the observed effect.

Discussion

Study 2 examined (a) whether perceived behavioral similarities among morals, facts, and preferences, initially observed in Study 1, were also reflected in brain regions associated with

ToM, and (b) if they were reflected, what underlying processes might be responsible for that similarity. Generally, across the ToM network morals were represented as more similar to preferences than to facts. This was particularly true in medial prefrontal cortex: in both whole-brain and ROI analyses, morals and preferences elicited common activity in DMPFC and VMPFC. Furthermore, in DMPFC, morals were more easily distinguished from facts than from preferences, based on the voxel-wise patterns of activity (an independent metric from overall BOLD differences). In DMPFC, VMPFC, and LTPJ, common activity elicited by morals and preferences, relative to facts, stemmed from both morals and preferences eliciting mental state inferences (i.e. inferences about agents' beliefs, thoughts, and desires). Note that it was *exclusively* this difference in mental state content that accounted for DMPFC and VMPFC activity in response to morals and preferences, as opposed to any other intrinsic differences between categories that we tested for (e.g. valence, arousal; Tables S8–S9 of the online supplemental materials). Surprisingly, we also observed greater activity in PC, RTPJ, and LTPJ for moral claims relative to both facts and preferences. We speculate on the meaning of this finding in the General Discussion, below.

General Discussion

Two studies examined metaethical judgment, testing whether morals are represented as objective or subjective. If people represents morals as subjective, then they should perceive morals as relatively more preference-like and morals should elicit more neural activity in common with preferences, particularly within brain regions associated with mental state representation. This is what we observed. In Study 1, participants read claims about morals, facts, and preferences, and rated each claim on the extent that it was about morals, about facts, and about preferences (Figures 1–2). Morals were perceived as relatively more preference-like

than fact-like across our sample of moral claims (and in an independent set of moral claims, adapted from the Moral Foundations Questionnaire—Graham et al., 2011; Iyer et al., 2012; see Figure S4 of the online supplemental materials). In Study 2, participants read the original set of claims while undergoing fMRI, allowing us to compare neural activity elicited by morals, facts, and preferences. Here too, morals were represented as more similar to preferences than to facts—morals and preferences elicited overlapping activity (and voxel-wise patterns of activity) across ROIs in the ToM network, and particularly within DMPFC (Figures 3–5). In a subsequent item analysis, we observed that the activity elicited in common by morals and preferences could be almost entirely explained by their shared tendency to evoke representations of mental states (e.g. experiences, beliefs, thoughts, & desires; Dodell-Feder et al, 2011). Initially we had anticipated that preferences would act as a high water mark for activity in brain regions for social processing, and that activity for moral claims would fall somewhere between activity for facts and preferences. However, we were surprised to find that moral claims actually elicited *greater* activity than both facts and preferences in PC, RTPJ, and LTPJ, all critical nodes in the ToM network; that is, based on neural activity, moral claims were processed as more social than preferences, a category selected for its social relevance. Taken together, Studies 1 and 2 suggest that (a) people represent morals as largely similar to preferences, and (b) this common representation stems from both morals' and preferences' tendency to evoke mental state representations; that is, morals are seen as social information.

Morals are represented as preference-like

In the present work, people reported that morals are more similar to preferences than prior work has emphasized (Beebe, 2014; Goodwin & Darley, 2008; 2012; Nichols & Folds-Bennett, 2003; Smetana, 1981; Tisak & Turiel 1988; Turiel, 1978; Wainryb et al., 2004; Wright

et al., 2013). The present work also seems to contradict a position expressed by some philosophers; namely, that the majority of non-philosophers are moral objectivists; that they believe morals are fact-like; that "… moral questions have correct answers; that the correct answers are made correct by objective moral facts … [and that] we can discover what these objective moral facts determined by circumstance are" (Smith, 1994, p. 6). Non-philosophers may be moral objectivists—our research cannot rule this out—however, our results should also give some pause to those who claim that similarity to facts is a central feature of the moral domain. In the present work, neural and behavioral evidence consistently demonstrates that morals share more in common with preferences than with facts.

Several methodological advances may explain the discrepancy between prior work and our own findings. First, our behavioral analyses avoid imposing categorical or one-dimensional distinctions (i.e. we do not require that fact-like morals necessarily be less preference-like). This approach avoids constraining comparisons, which could exaggerate categorical differences. If prior work correctly concluded that people represent morals as preeminently fact-like, then this saliency should emerge naturally in our method; however, the similarity between morals and preferences emerged instead. Second, our work can make statistical generalizations in a way that prior work could not. We used a large sample of stimuli, but critically, we analyzed these stimuli using mixed effects analyses, modeling by-item random effects (Baayen et al., 2008; Barr et al., 2013; Judd et al., 2012; Westfall et al., 2014), a method that allows for statistical generalizations beyond our specific items. Prior work targeting morality as a separable domain from other sorts of information (e.g. conventional norms) has been criticized for its selection of examples (Nichols & Folds-Bennett, 2003; Smetana, 1981; Tisak & Turiel 1988; Turiel, 1978; Wainryb et al., 2004; for criticism, see Gabennmesch, 1990; Kelly et al., 2007; Machery, 2012), but this

criticism has typically been made on the grounds of conceptual generalizability: critics charge that the work has focused on "prototypical" moral issues—e.g., inflicting harm—with only an occasional nod to "non-prototypical" moral issues—e.g. abortion (Turiel et al., 1991). These conceptual criticisms, valid as they may be, put the cart before the horse: conceptual criticisms are typically applied to conclusions with statistical support (Cornfield & Tukey, 1956), and if items are not treated as random effects then researchers are not licensed to make *any* generalization beyond the examples they have tested (Judd et al., 2012). Note that conceptual criticisms could be applied to our own results (as they could be applied to any statistical inference). We intentionally omitted controversial moral issues, and our results cannot directly speak to their properties (e.g. abortion, same-sex marriage; for a more thorough treatment of these topics see Skitka, Bauman & Sargis, 2005). It is possible that the general trends we have identified will carry over into this domain (and other sub-domains of morality, see supplemental study in the online supplemental materials), but additional work is necessary to confirm our supposition.

Morals are socially informative

So far we have shown that moral claims are not represented as objective to the extent that prior work has asserted, but the positive case is equally important: Can the present work say anything about what morals are? Behaviorally, morals are perceived as far more similar to preferences than has been previously suggested, but neurally, morals actually outstripped preferences, eliciting greater activity across several social brain regions, such as PC, RTPJ, and LTPJ. Note that the present study is not equipped to speak to the function of specific brain regions, but by drawing on prior work we can speculate on what the observed activity implies about the nature of moral content. The DMPFC, where the greatest overlap in activity between

morals and preferences emerged, is a key region implicated in social cognition (Amodio & Frith, 2006; Mitchell et al., 2005; Ochsner et al., 2005) and has been implicated in processing stable personal traits (Harris et al., 2005; Jenkins & Mitchell, 2010), even in the absence of explicit instruction (Ma et al., 2012). That is, DMPFC activity has been associated with learning something about a person. Morals and preferences may be perceived as similar on account of their both being rich sources of social information.

Brain regions where morals elicited more activity than preferences have been implicated in processing beliefs and intentions (e.g. innocent intentions following an accident; Young & Saxe, 2009). However, recent accounts have moved to consider these findings in a more general framework of hierarchical predictive coding (Koster-Hale & Saxe, 2013). In this hierarchical predictive coding framework, it is presumed that the brain works to build a stable model of the world, issuing predictions about incoming sensory information (Friston, 2010; Hohwy, 2013; Rao & Ballard, 1999). Social predictions, processed in the ToM network, are abstracted from sensory information and situated near the top of this hierarchy (Koster-Hale & Saxe, 2013). When a prediction is violated, the model must be updated to account for this prediction error (for review, see Clark, 2013). Consistent with this, the same regions that we have examined (i.e. the ToM network) also support impression updating, showing increased activity when inconsistent information about a known social agent is presented (Mende-Siedlecki, Baron, & Todorov, 2013). If activity in these regions roughly reflects the magnitude of prediction error (Koster-Hale & Saxe, 2013), then moral claims may elicit greater activity (compared to preferences) in PC, RTPJ, and LTPJ because morals license stronger social predictions. That is, when participants received moral information they are able to make a stronger prediction about the anonymous speaker (e.g. what other moral beliefs they may have, whether the participant would like or

dislike this person). Consistent with this, recent work has shown that people perceive moral beliefs (compared to preferences) as more central to identity—e.g. a brain injury that alters one's moral beliefs changes one's identity more than a brain injury altering preferences (Strohminger & Nichols, 2014; 2015). According to this hypothesis—which we are testing in ongoing work—the observed discrepancy between morals and preferences does not reflect a difference in kind, but rather a difference in degree: both morals and preferences can provide social information (violating social predictions about an anonymous speaker), but morals are more informative—in part because there are certain moral beliefs that we expect everyone to endorse (e.g. slavery is wrong). In sum, moral beliefs appear to be distinguished (from facts, but possibly even from preferences) by their salience as social information.

Future Directions

"Which" actions people moralize is an area of heated debate within moral psychology (e.g. Fiske & Rai, 2014; Graham et al., 2011; Gray et al., 2012; Janoff-Bulman & Carnes, 2013), and while the present work cannot directly address the controversy, it may help to contextualize it. Behaviorally, we allowed participants to rate the extent that moral claims were fact-like, moral-like, and/or preference-like. In Study 1, people rated moral claims as more moral-like than preference-like, but this difference was slight (just past the threshold of significance in a full mixed effects analysis). It was possible that people would view more prototypical moral claims as more moral-like, more fact-like, and less preference-like. However, in a supplemental study (Figure S4 & Table S10 of the online supplemental materials), using claims adapted from the Moral Foundations Questionnaire (Graham et al., 2011; Iyer et al., 2012), we found that this was not the case: across all domains (e.g. harm, fairness, purity, authority, loyalty), people (regardless of political ideology) viewed moral claims as more preference-like than fact-like. Furthermore,

and surprisingly, moral claims were only rated as more moral-like than preference-like in the harm domain. Based on this, one might conclude that harm is the most prototypical moral domain, and that other domains are only moralized to the extent they involve harm (Gray et al., 2012; Schein & Gray, 2015). However, an alternative is also possible. All domains were moralized to some extent, and to focus on relative moral-like and preference-like ratings would overlook that fact that *all* moral claims were perceived as highly preference-like. This, combined with our neuroimaging finding that morals are salient sources of social information, suggests that morality may be best understood as rooted in predictions about social relationships. Fortunately, several theories have advanced the argument that morality is embedded in social contexts and relationships (Carnes, Lickel, & Janoff-Bulman, 2015; Fiske & Rai, 2014; Janoff-Bulman & Carnes, 2013; Rai & Fiske, 2011; Heiphetz, Strohminger, & Young, 2016; Strohminger & Nichols, 2015; 2016). Future work could apply our method to a broader sample of stimuli to test the relative prominence of features in a given claim (e.g. "To what degree is this statement about… [morality/social relationships/etc.]".

Separately, there remains the interesting question of why moral claims have been thought to be objective in such a wide range of prior work. Moral conviction researchers have emphasized that people can be motivated to avoid compromise for their most strongly held moral beliefs (e.g. Skitka et al., 2005). Likewise, communities enshrine certain moral beliefs as laws or ethical codes, making them a social reality. If people are pushed to defend their moral beliefs, then they may express that they are more fact-like then they would under other circumstances (Fisher, Knobe, Strickland, & Keil, 2017). For this reason, future work might benefit from distinguishing moral processing from the defense of moral beliefs. The former may address the

representation of moral information, while the latter is more relevant to motivated cognition and communication.

Conclusion

Questions about the metaethical status of moral claims are questions about how moral information is represented. Moral objectivists have argued that people represent morals as fact-like (Railton, 1986; Shafer-Landau, 2003; Smith, 1994) and prior work in psychology and experimental philosophy has generally favored this objectivist view (Turiel, 1978; Wainryb et al., 2004) with the recent caveat that some moral claims may be more objective than others (Beebe, 2014; Goodwin & Darley, 2008; 2012; Heiphetz & Young, in press; Sarkissian et al., 2011; Wright et al., 2013). Evidence from the present work favors the alternative, subjectivist view: that behaviorally and neurally, people represent moral claims as largely preference-like. This evidence speaks to philosophical debates about the metaethical status of moral claims, and while it certainly cannot conclude them, it demonstrates that the social relevance of moral claims is more salient than their objectivity—specifically, across a wide range of stimuli, morals and preferences both elicit activity in brain regions associated with social cognition and mental state representations. The social nature of moral claims is consistent with recent theoretical work, which has argued that morals are fundamentally about regulating social relationships (Fiske & Rai, 2014; Heiphetz et al., 2016; Janoff-Bulman & Carnes, 2013; Rai & Fiske, 2011). Taken together, our findings help to situate the moral domain within the broader constellation of social and non-social information, bringing into focus the underlying cognitive processes that support moral cognition.

Acknowledgements

We thank Fiery Cushman, Drew Linsley, Sean MacEvoy, Jonathan Phillips, James Russell, and members of the Morality Lab for feedback. This work was supported by the Alfred P. Sloan Foundation BR2012-004 (L.Y.), the Dana Foundation (L.Y.), the National Science Foundation SMA-1408989 (L.H.), and Natural Sciences and Engineering Research Council of Canada PGSD3-420445 (J.T.).

References

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*, 268–277. http://dx.doi.org/10.1038/nrn1884

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. http://dx.doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed effects models using lme4. *Journal of Statistical Software, 67,* 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Beebe, J. R. (2014). How different kinds of disagreement impact folk metaethical judgments. In J. C. Wright & H. Sarkissian (Eds.), *Advances in experimental moral psychology: Affect, character, and commitments* (pp. 167–187). New York, NY: Bloomsbury.

Bruneau, E., Dufour, N., & Saxe, R. (2013). How we know it hurts: Item analysis of written narratives reveals distinct neural responses to others' physical pain and emotional suffering. *PLoS One, 8,* e63085. http://dx.doi.org/10.1371/journal.pone.0063085

Carnes, N. C., Lickel, B., & Janoff-Bulman, R. (2015). Shared Perceptions Morality Is Embedded in Social Contexts. *Personality and Social Psychology Bulletin*, *41(3)*, 351–362. http://dx.doi.org/10.1177/0146167214566187

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181-204. http://dx.doi.org/10.1017/S0140525X12000477

Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359. http://dx.doi.org/10.1016/s0022-5371(73)80014-3

Ciaramidaro, A., Adenzato, M., Enrici, I., Erk, S., Pia, L., Bara, B. G., & Walter, H. (2007). The intentional network: How the brain reads varieties of intentions. *Neuropsychologia, 45*, 3105–3113. http://dx.doi.org/10.1016/j.neuropsychologia.2007.05.011

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *The Annals of Mathematical Statistics, 27,* 907–949. http://dx.doi.org/10.1214/aoms/1177728067

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological science*, *17*(12), 1082-1089. http://dx.doi.org/10.1111/j.1467-9280.2006.01834.x

Decety, J., & Cacioppo, S. (2012). The speed of morality: A high-density electrical neuroimaging study. *Journal of Neurophysiology, 108*, 3068–3072. http://dx.doi.org/10.1152/jn.00473.2012

Degryse, J., Seurinck, R., Durnez, J., Gonzalez-Castillo, J., Bandettini, P. A., & Moerkerke, B. (2017). Introducing alternative-based thresholding for defining functional regions of interest in fMRI. *Frontiers in Neuroscience*, *11*: 222. http://dx.doi.org/10.3389/fnins.2017.00222

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *NeuroImage, 55*, 705–712. http://dx.doi.org/10.1016/j.neuroimage.2010.12.040

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113(28)*, 7900–7905. http://dx.doi.org/10.1073/pnas.1602413113

Fiske, A. P., & Rai, T. S. (2014). *Virtuous violence: Hurting and killing to create, sustain, end, and honor social relationships*. UK: Cambridge University Press.

Fisher, M., Knobe, J., Strickland, B., & Keil, F. C. (2017). The influence of social interaction on intuitions of objectivity and subjectivity. *Cognitive science*, *41(4),* 1119–1134. http://dx.doi.org/10.1111/cogs.12380

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., & Frith, C. D. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition, 57*, 109–128. http://dx.doi.org/10.1016/0010-0277(95)00692-r

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature Reviews Neuroscience*, *11*(2), 127-138. http://dx.doi.org/10.1038/nrn2787

Forsyth, D. R. (1980). A taxonomy of ethical ideologies. *Journal of Personality and Social Psychology, 39,* 175–184. http://dx.doi.org/10.1037//0022-3514.39.1.175

Gabennesch, H. (1990). The perception of social conventionality by children and adults. *Child Development, 61,* 2047–2059. http://dx.doi.org/10.2307/1130858

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of 'theory of mind' in verbal

and nonverbal tasks. *Neuropsychologia, 38*, 11–21. http://dx.doi.org/10.1016/s0028-3932(99)00053-6

Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal of Cognitive Neuroscience, 19,* 1803–1814. http://dx.doi.org/10.1162/jocn.2007.19.11.1803

Goodwin, G. P., & Darley, J. M. (2008). The psychology of meta-ethics: Exploring objectivism. *Cognition, 106*, 1339–1366. http://dx.doi.org/10.1016/j.cognition.2007.06.007

Goodwin, G. P., & Darley, J. M. (2010). The perceived objectivity of ethical beliefs: Psychological findings and implications for public policy. *Review of Philosophy and Psychology, 1*, 161–188. http://dx.doi.org/10.1007/s13164-009-0013-4

Goodwin, G. P., & Darley, J. M. (2012). Why are some moral beliefs perceived to be more objective than others? *Journal of Experimental Social Psychology, 48*, 250–256. http://dx.doi.org/10.1016/j.jesp.2011.08.006

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*, 193–202. http://dx.doi.org/10.3758/bf03195564

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. http://dx.doi.org/10.1037/a0021847

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101-124. http://dx.doi.org/10.1080/1047840X.2012.651387

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105-2108. http://dx.doi.org/10.1126/science.1062872

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed Models — The R package pbkrtest. *Journal of Statistical Software, 59,* 1–32. http://dx.doi.org/10.18637/jss.v059.i09

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557-559. http://dx.doi.org/10.1038/nature06288

Harris, L. T., Todorov, A., & Fiske, S. T. (2005). Attributions on the brain: Neuro-imaging dispositional inferences, beyond theory of mind. *NeuroImage, 28*, 763–769. http://dx.doi.org/10.1016/j.neuroimage.2005.05.021

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*, 2425–2430. http://dx.doi.org/10.1126/science.1063736

Heiphetz, L., & Young, L. L. (in press). Can only one person be right? The development of objectivism and social preferences regarding widely shared and controversial moral beliefs. *Cognition*. http://dx.doi.org/10.1016/j.cognition.2016.05.014

Hohwy, J. (2013). *The predictive mind*. New York, NY: Oxford University Press.

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PloS one*, *7*(8), e42366. http://dx.doi.org/10.1371/journal.pone.0042366

Janoff-Bulman, R., & Carnes, N. C. (2013). Surveying the moral landscape moral motives and group-based moralities. *Personality and Social Psychology Review*, *17*(3), 219-236. http://dx.doi.org/10.1177/1088868313480274

Jenkins, A. C., & Mitchell, J. P. (2010). Mentalizing under uncertainty: Dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex, 20*, 404–410. http://dx.doi.org/10.1093/cercor/bhp109

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103,* 54–69. http://dx.doi.org/10.1037/a0028347

Kelly, D., Stich, S., Haley, K. J., Eng, S. J., & Fessler, D. M. T. (2007). Harm, affect, and the moral/conventional distinction. *Mind & Language, 22*, 117–131. http://dx.doi.org/10.1093/acprof:oso/9780199733477.003.0013

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 5648–5653. http://dx.doi.org/10.1073/pnas.1207992110

Kron, A., Goldstein, A., Lee, D. H-J., & Gardhouse, K. (2013). How are you feeling? Revisiting the quantification of emotional qualia. *Psychological Science, 24*, 1503–1511. http://dx.doi.org/10.1177/0956797613475456

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests in linear mixed effects models [Computer software manual]. http://CRAN.R-project.org/package=lmerTest. (R Package version 2.0-25).

Ma, N., Vandekerckhove, M., Van Hoeck, N., & Van Overwalle, F. (2012). Distinct recruitment of temporo-parietal junction and medial prefrontal cortex in behavior understanding and trait identification. *Social Neuroscience, 7*, 591–605. http://dx.doi.org/10.1080/17470919.2012.686925

Machery, E. (2012). Delineating the moral domain. *The Baltic International Yearbook of Cognition, Logic and Communication, 7,* 1–14. http://dx.doi.org/10.4148/biyclc.v7i0.1777

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (7 September, 2014). Coh-Metrix version 3.0. *http://cohmetrix.com*.

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *Journal of Neuroscience*, *33*(50), 19406-19415. http://dx.doi.org/10.1523/JNEUROSCI.2334-13.2013

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). General and specific contributions of the medial prefrontal cortex to knowledge about mental states. *NeuroImage, 28*, 757–762. http://dx.doi.org/10.1016/j.neuroimage.2005.03.011

Nichols, S., & Folds-Bennett, T. (2003). Are children moral objectivists? Children's judgments about moral and response-dependent properties. *Cognition, 90*, B23–B32. http://dx.doi.org/10.1016/s0010-0277(03)00160-4

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Science, 10*, 424–430. http://dx.doi.org/10.1016/j.tics.2006.07.005

Ochsner, K. N., Beer, J. S., Robertson, E. R., Cooper, J. C., Gabrieli, J. D. E., Kihsltrom, J. F., & D'Esposito, M. (2005). The neural correlates of direct and reflected self-knowledge. *NeuroImage,* 28, 797–814. http://dx.doi.org/10.1016/j.neuroimage.2005.06.069

Patil, I., Melsbach, J., Hennig-Fast, K., & Silani, G. (2016). Divergent roles of autistic and alexithymic traits in utilitarian moral judgments in adults with autism. *Scientific reports*, *6*, 23637. http://dx.doi.org/10.1038/srep23637

Poldrack, R. A. (2006). Can cognitive processing be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10,* 59–63. http://dx.doi.org/10.1016/j.tics.2005.12.004

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychological review*, *118*(1), 57–75. http://dx.doi.org/10.1037/a0021867

Railton, P. (1986). Moral realism. *Philosophical Review, 95*, 163–207. http://dx.doi.org/10.2307/2185589

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427-430. http://dx.doi.org/10.1038/nature11467

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79-87. http://dx.doi.org/10.1038/4580

Ruby, P., & Decety, J. (2003). What you believe versus what you think they believe: A neuroimaging study of conceptual perspective-taking. *European Journal of Neuroscience, 11,* 2475–2480. http://dx.doi.org/10.1046/j.1460-9568.2003.02673.x

Sarkissian, H., Park, J., Tien, D., Wright, J.C., & Knobe, J. (2011). Folk moral relativism. *Mind & Language, 26*, 482–505. http://dx.doi.org/10.1111/j.1468-0017.2011.01428.x

Saxe, R. (2009). The happiness of the fish: Evidence for a common theory of one's own and others' actions. *The handbook of imagination and mental simulation*, 257-266.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage, 19*, 1835–1842. http://dx.doi.org/10.1016/s1053-8119(03)00230-1

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science, 17*, 692–699. http://dx.doi.org/10.1111/j.1467-9280.2006.01768.x

Sayre-McCord, G. (1986). The many moral realisms. *The Southern Journal of Philosophy, 24 (Supplement)*, 1–22. http://dx.doi.org/10.1111/j.2041-6962.1986.tb01593.x

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. http://dx.doi.org/10.1177/0146167215591501

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews, 42,* 9–34. http://dx.doi.org/10.1016/j.neubiorev.2014.01.009

Shafer-Landau, R. (2003). *Moral realism: A defense*. Oxford, UK: Oxford University Press.

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*, 895–917. http://dx.doi.org/10.1037/0022-3514.88.6.895

Smetana, J. G. (1981). Preschool children's conceptions of moral and social rules. *Child Development, 52,* 1333–1336. http://dx.doi.org/10.2307/1129527

Smith, M. (1994). *The moral problem*. Oxford, UK: Blackwell

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159-171. http://dx.doi.org/10.1016/j.cognition.2013.12.005

Strohminger, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science*, *26*(9), 1469-1479. http://dx.doi.org/10.1177/0956797615592381

Theriault, J., Waytz, A., Heiphetz, L., & Young, L. (under review). Metaethical judgment relies on activity in right temporoparietal junction: Evidence from neuroimaging and transcranial magnetic stimulation. *Neuroimage*.

Tisak, M. S. & Turiel, E. (1988). Variation in seriousness of transgressions and children's moral and conventional concepts. *Developmental Psychology, 24,* 352–357. http://dx.doi.org/10.1037/0012-1649.24.3.352I

Turiel, E. (1978). Social regulations and domains of social concepts. In W. Damon (Ed.), *New directions for child development. Vol. 1* (pp. 45–74). San Francisco, CA: Jossey-Bass.

Turiel, E., Hildebrandt, C., Wainryb, C., & Saltzstein, H. D. (1991). Judging social issues: Difficulties, inconsistencies, and consistencies. *Monograms of the Society for Research in Child Development, 56*, 1–116. http://dx.doi.org/10.2307/1166056

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*, 829–858. http://dx.doi.org/10.1002/hbm.20547

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P. … Zilles, K. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage, 14,* 170–181. http://dx.doi.org/10.1006/nimg.2001.0789

Wainryb, C., Shaw, L. S., Langley, M., Cottam, K., & Lewis, R. (2004). Children's thinking about diversity of belief in the early school years: Judgments of relativism, tolerance, and disagreeing persons. *Child Development, 75,* 287–703. http://dx.doi.org/10.1111/j.1467-8624.2004.00701.x

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participant respond to samples of stimuli. *Journal of Experimental Psychology: General, 143,* 2020–2045. http://dx.doi.org/10.1037/xge0000014

Woo, C. W., Krishnan, A., Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage, 91,* 412–419. http://dx.doi.org/10.1016/j.neuroimage.2013.12.058

Wright, J. C., Grandjean, P. T., & McWhite, C. B. (2013). The meta-ethical grounding of our moral beliefs: Evidence for meta-ethical pluralism. *Philosophical Psychology, 26*, 336–361. http://dx.doi.org/10.1080/09515089.2011.633751

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, *8*(8), 665-670. http://dx.doi.org/10.1038/nmeth.1635

Young, L., Camprodon, J., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 6753–6758. http://dx.doi.org/10.1073/pnas.0914826107

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of*

Sciences of the United States of America, *104,* 8235–8240. http://dx.doi.org/10.1073/pnas.0701408104

Young, L., & Dungan, J. (2012). Where in the brain is morality? Everywhere and maybe nowhere. *Social Neuroscience, 7,* 1–10. http://dx.doi.org/10.1080/17470919.2011.569146

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage, 40,* 1912–1920. http://dx.doi.org/10.1016/j.neuroimage.2008.01.057

Young, L., & Saxe, R. (2009). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience, 21,* 1396–1405. http://dx.doi.org/10.1162/jocn.2009.21137

Young, L., Scholz, J., & Saxe, R. (2011). Neural evidence for "intuitive prosecution": The use of mental state information for negative moral verdicts. *Social Neuroscience, 6,* 302–315. http://dx.doi.org/10.1080/17470919.2010.529712

Supplemental Materials

Supplemental Study

In Study 1, we observed that moral claims were perceived as more preference-like than fact-like. However, it is critical to replicate this finding with a separate set of moral stimuli. Below we analyze a secondary set of moral claims, adapted from the Moral Foundations Questionnaire (Graham, Nosek, Haidt, Iyer, Koleva, & Ditto, 2011; Iyer, Koleva, Graham, Ditto, & Haidt, 2012), which provide a taxonomy of the moral space, with domains related to harm, fairness, loyalty, authority, purity, and economic / lifestyle liberty. In every domain, moral claims were perceived as more preference-like than they were fact-like, by both liberals and conservatives in our sample, suggesting that the observations in the main paper are not simply a consequence of our stimuli sample.

Method

Participants. We recruited participants online using Amazon Mechanical Turk (AMT) at an approximate rate of $5/hour, in line with standard AMT compensation rates. Our final sample consisted of 100 adults (49 female, 50 male, 1 unspecified; M_{Age} = 36.4 years, SD_{Age} = 12.5 years), after excluding 2 participants for failing a simple attention check that asked them to describe any claim they had read. The Boston College Institutional Review Board approved this study, and each participant provided consent before beginning.

Procedure. The procedure was identical to that described for Study 1. Participants rated their agreement with claims and the extent that each was about facts, about morals, and about preferences: “To what degree is this statement about [facts, morals, preferences]” (1 – “not at all"; 6 – “completely”)? Following these questionnaires, participants provided demographic information.

As a group, participants were neither liberal or conservative (M = 3.8, SD = 2.0, 7-point scale anchored at 1, “Socially Conservative”, and 7, “Socially Liberal”), as indicated by a one-sample t-test against the scale mid-point, $t(98) = 0.76$, $p = .451$. To achieve a politically neutral sample, we recruited in two batches, approximately one month apart. First, we collected a sample of 50 participants, which leaned liberal (as in Study 1); then, for our second sample of 50 participants, we advertised that we were specifically interested in people with conservative political views. We were explicit that we would not screen participants by ideology in either sample (doing so would incentivize lying about political beliefs to qualify).

Stimuli. As in Study 1, participants read claims about facts, morals and preferences. Fact and preference claims were the same as in Study 1(24 facts, 24 preferences; see Appendix A). Study 1 moral claims were replaced with claims drawn from the Moral Foundations Questionnaire (MFQ; Graham et al., 2011), and additional items related to economic and personal liberty (Iyer et al., 2012). These questionnaires break up moral concerns into distinct domains: e.g. harm, fairness, loyalty, authority, and purity. Harm and fairness are endorsed most strongly by political liberals, whereas political conservatives endorse a combination of all domains. The economic and personal liberty domains were added to explore libertarian morality.

From the MFQ, we selected items where participants were asked to rate their agreement. These items were either used verbatim or minimally edited to remove the first-person perspective, making them consistent with our stimuli (e.g. “I think it’s morally wrong that rich children inherit a lot of money while poor children inherit nothing.” became “It’s wrong that rich children inherit a lot of money while poor children inherit nothing.”, whereas “It is more important to be a team player than to express oneself.” was unchanged; see Appendix C for all stimuli and alterations). We used 22 moral claims in total: 3 harm, 3 fairness, 3, loyalty, 3

authority, 3 purity, 4 economic liberty, 2 lifestyle liberty, and 1 control item ("It is better to do good than to do bad.").

Statistical methods. Our primary motivation for this study was to confirm that the high preference-like ratings we observed in Study 1 were not simply due to our selection of moral claims. Thus, for our purposes it was enough to examine contrasts within the sample of stimuli, rather than to generalize beyond it (as in Study 1). For this reason, we performed repeated measures ANOVAs within each moral domain, using a by-participant average across stimuli in each (as opposed to a full mixed effects analysis, crossing by-subject and by-item random effects (Baayen et al., 2008; Judd et al., 2012; Westfall et al., 2014). We followed up ANOVAs with condition contrasts, comparing the extent to which people perceived examples in each domain as fact-like, moral-like, and preference-like (p values are corrected for three comparisons, $p_{\text{corrected}} = .0167$).

Results and Discussion

Within morals, repeated measures ANOVAs identified a significant main effect of dimension (fact-like/moral-like/preference-like) within each domain: Harm, $F(2, 198) = 122.16$, $p < .001$, $\eta_p^2 = .458$, Fairness, $F(2, 198) = 67.20$, $p < .001$, $\eta_p^2 = .325$, Purity, $F(2, 198) = 148.43$, $p < .001$, $\eta_p^2 = .501$, Loyalty, $F(2, 198) = 192.02$, $p < .001$, $\eta_p^2 = .558$, Authority, $F(2, 198) = 47.23$, $p < .001$, $\eta_p^2 = .236$, Economic Liberty, $F(2, 198) = 133.48$, $p < .001$, $\eta_p^2 = .482$, Lifestyle Liberty, $F(2, 198) = 135.97$, $< .001$, $\eta_p^2 = .481$, Control, $F(2, 198) = 85.41$, $p < .001$, $\eta_p^2 = .371$. We followed up these main effects with contrasts, comparing the extent that examples in each domain were relatively perceived as fact-like, moral-like and preference-like. Contrasts are presented in Figure S4 and Table S10. In every domain, people perceived moral claims as more preference-like than fact-like. Given that endorsement of moral domains differs between liberals

and conservatives, we explored whether relative dimension ratings differed in politically defined subgroups. For the question "Please indicate your political orientation relating to social issues" [1 – Very Conservative; 7 – Very Liberal], we categorized participants answering above the midpoint as liberal (n = 37), and those answering below as conservative (n = 46). For both groups, all domains were perceived as more preference-like than fact-like (Figure S4; Table S10).

These findings are intriguing, and leave open many questions for future research. For instance, among the moral domains, the only examples that were perceived as more moral-like than preference-like were within the harm domain (and the control statement). How this relates to ongoing debates regarding how claims are moralized (e.g. Graham et al., 2011; Gray, Young & Waytz, 2012) is beyond the scope of the present paper. For present purposes, these findings demonstrate that people generally perceive moral claims, both the sample used in the main paper and the independent sample used here, as more preference-like than fact-like.

Supplemental Analysis For fMRI Study

Working memory ROI analysis. In all working memory ROIs where we observed a difference between conditions, morals elicited less activity than facts, and in several cases morals also elicited less activity than preferences (Figure S3 of the online supplemental materials). We performed a repeated measures ANOVA within each ROI, observing a main effect of content in five of seven ROIs: (a) left anterior middle frontal gyrus, $F(2, 48) = 5.80$, $p = .006$, $\eta_p^2 = .015$; (b) right anterior middle frontal gyrus, $F(2, 48) = 9.82$, $p < .001$, $\eta_p^2 = .060$; (c) left supramarginal gyrus, $F(2, 48) = 3.52$, $p = .037$, $\eta_p^2 = .029$; (d) right supramarginal gyrus, $F(2, 48) = 6.34$, $p = .004$, $\eta_p^2 = .062$; and (e) medial superior frontal gyrus, $F(2, 48) = 3.61$, $p = .035$,

$\eta_p^2 = .009$. There was no significant main effects in either left, $F(2, 48) = 0.06$, $p = .940$, $\eta_p^2 = .0002$, or right posterior middle frontal gyrus, $F(2, 48) = 0.46$, $p = .633$, $\eta_p^2 = .001$.

We followed up significant main effects with contrast analyses (contrast p values are corrected for multiple comparisons within each ROI). In (a) left anterior middle frontal gyrus, morals elicited less activity than both facts, $z = 3.03$, $p = .007$, $d = 0.61$, and preferences, $z = 2.86$, $p = .012$, $d = 0.57$, while preferences and facts did not differ, $z = 0.17$, $p = 0.984$, $d = 0.03$. In (b) right anterior middle frontal gyrus, morals also elicited less activity than both facts, $z = 4.42$, $p < .001$, $d = 1.01$, and preferences, $z = 2.51$, $p = .032$, $d = 0.57$, while preferences and facts did not differ, $z = 1.91$, $p = .136$, $d = 0.44$. In (c) left supramarginal gyrus, morals elicited less activity than facts, $z = 2.58$, $p = .027$, $d = .78$, but there was no difference in activity between morals and preferences, $z = 0.71$, $p = .755$, $d = .22$, or between preferences and facts, $z = 1.86$, $p = .151$, $d = .56$. In (d) right supramarginal gyrus, facts elicited greater activity than both morals, $z = 3.09$, $p = .006$, $d = .818$, and preferences, $z = 3.08$, $p = .006$, $d = .816$, while morals and preferences did not differ, $z = 0.01$, $p = .999$, $d = .003$. Finally, in (e) medial superior frontal gyrus, morals elicited less activity than preferences, $z = 2.67$, $p = .021$, $d = .60$, but there was no difference in activity between morals and facts, $z = 1.59$, $p = .248$, $d = .36$, or between facts and preferences, $z = 1.08$, $p = .528$, $d = .24$.

Anatomically defined ToM ROI analysis. Anatomically defined ToM ROIs were each defined as a 9mm sphere surrounding a peak coordinate identified in a whole brain random effects analysis of the functional localizer contrast (false belief > false photograph) across all participants. Peak coordinates are reported in Table S3 of the online supplemental materials. For each ROI, we performed a repeated measures ANOVA comparing neural activity for morals,

facts, and preferences, followed by condition contrasts. Contrast p values are corrected for three comparisons to achieve a familywise α of .05 within each ROI ($p_{corrected} = .0167$).

As in the ROI analysis reported in the main paper, morals and preferences, relative to facts, both elicited greater in DMPFC and VMPFC (Figure S2 of the online supplemental materials). Main effects of content were significant in both ROIs: DMPFC, $F(2, 48) = 40.67$, $p < .001$, $\eta_p^2 = .326$, VMPFC, $F(2, 48) = 9.48$, $p < .001$, $\eta_p^2 = .108$. Within DMPFC, both morals, $z = 7.95$, $p < .001$, $d = .1.59$, and preferences, $z = 7.67$, $p < .001$, $d = .1.53$, elicited greater activity than facts, but were not distinguishable from each other, $z = 0.28$, $p = .957$, $d = .0.06$. Likewise, within VMPFC, both morals, $z = 6.01$, $p < .001$, $d = 1.20$, and preferences, $z = 4.04$, $p < .001$, $d = 0.81$, elicited greater activity than facts, but were not distinguishable from each other, $z = 1.97$, $p = .120$, $d = 0.39$.

Results for PC, RTPJ, and LTPJ were also identical to the results for the individually localized functional ROIs reported in the main paper. We observed main effects of content within PC, $F(2, 48) = 22.36$, $p < .001$, $\eta_p^2 = .197$, RTPJ, $F(2, 48) = 10.78$, $p < .001$, $\eta_p^2 = .058$, and LTPJ, $F(2, 48) = 16.80$, $p < .001$, $\eta_p^2 = .132$. Within PC, morals elicited greater activity than both facts, $z = 6.56$, $p < .001$, $d = 1.31$, and preferences, $z = 4.40$, $p < .001$, $d = 0.88$, while preferences elicited marginally more activity than facts, $z = 2.17$, $p = .077$, $d = 0.43$. In RTPJ, morals elicited greater activity than both facts, $z = 4.34$, $p < .001$, $d = .0.87$, and preferences, $z = 3.61$, $p < .001$, $d = .72$, while preferences and facts were indistinguishable, $z = 0.73$, $p = .747$, $d = .15$. Finally, in LTPJ, morals elicited greater activity than both facts, $z = 5.80$, $p < .001$, $d = .1.16$, and preferences, $z = 2.93$, $p = .010$, $d = .59$, while preferences also elicited greater activity than facts, $z = 2.87$, $p = .011$, $d = .57$. Thus, our findings from anatomically defined ToM ROIs (based on the peak coordinates of the localizer contrast, within our sample) are identical to those

reported using individually, functionally defined ROIs. Morals and preferences both elicited greater activity relative to facts across the medial prefrontal cortex, while morals elicited greater activity than both facts and preferences in the posterior ToM ROIs: precuneus, and bilateral temporoparietal junction.

Alternative Item Analysis. In the main body, we performed an item analysis to better characterize ROI activity in response to morals, facts, and preferences. In the second step of this item analysis, we identified potential covariates, by testing which item features were associated with ROI activity, in the absence of fixed effects of category.

Here, we take an alternative approach; for each item feature, we tested for differences across categories to identify what intrinsic differences existed between morals, facts, and preferences. Table S8 of the online supplemental materials displays descriptive statistics for each feature and category, along with the results of a one-way ANOVA across categories. Among syntactic and semantic covariates, there were significant differences across categories in noun concreteness, $F(2, 69) = 4.90$, $p = .010$ and noun imageability, $F(2, 69) = 3.94$, $p = .024$, and a marginal difference in left embeddedness, $F(2, 69) = 2.76$, $p = .070$. Among online norming measures, there were significant differences across categories in valence, $F(2, 69) = 6.88$, $p = .002$, arousal, $F(2, 69) = 19.52$, $p < .001$, whether a person was present, $F(2, 69) = 4.46$, $p = .015$, whether claims evoked mental states, $F(2, 69) = 196.4$, $p < .001$, and agreement $F(2, 69) = 3.20$, $p = .047$.

We flagged these item features as potential covariates and used in the subsequent analysis of ROI activity. As in the main body, we added each covariate as a fixed effect, one at a time in the order of their significance. However, as mental state ratings were entered first in all prior ROI analyses, adding them first here would simply replicate the analysis presented in Table S6;

thus, we held out mental state ratings and entered them as the final covariate. This allowed us to test whether the categorical effects for morals and preferences could survive correction for all other intrinsic differences. If they could, then the inclusion of mental states as a covariate would be necessary to explain their effect (rather than only sufficient).

We focus here on DMPFC for demonstration purposes, (but see Table S9 for analyses of other ROIs). The analysis confirmed that mental state ratings were responsible for the neural activity elicited by morals, as opposed to other intrinsic differences between categories. In DMPFC, morals and preferences remained significant when controlling for all intrinsic differences *but* mental states: Morals, $\beta = 0.205$, $t(43.1) = 4.65$, $p < .001$, Preferences, $\beta = 0.159$, $t(46.4) = 3.66$, $p < .001$; when mental states were added to the model these effects were reduced to non-significance: Morals, $\beta = 0.100$, $t(69.2) = 1.25$, $p = .214$, Preferences, $\beta = 0.068$, $t(67.7) = 0.94$, $p = .350$. Thus, our conclusion that morals and preferences elicit common activity given that both elicit mental state representations is supported by this alternative method of covariate selection, in addition to the method used in the main body.

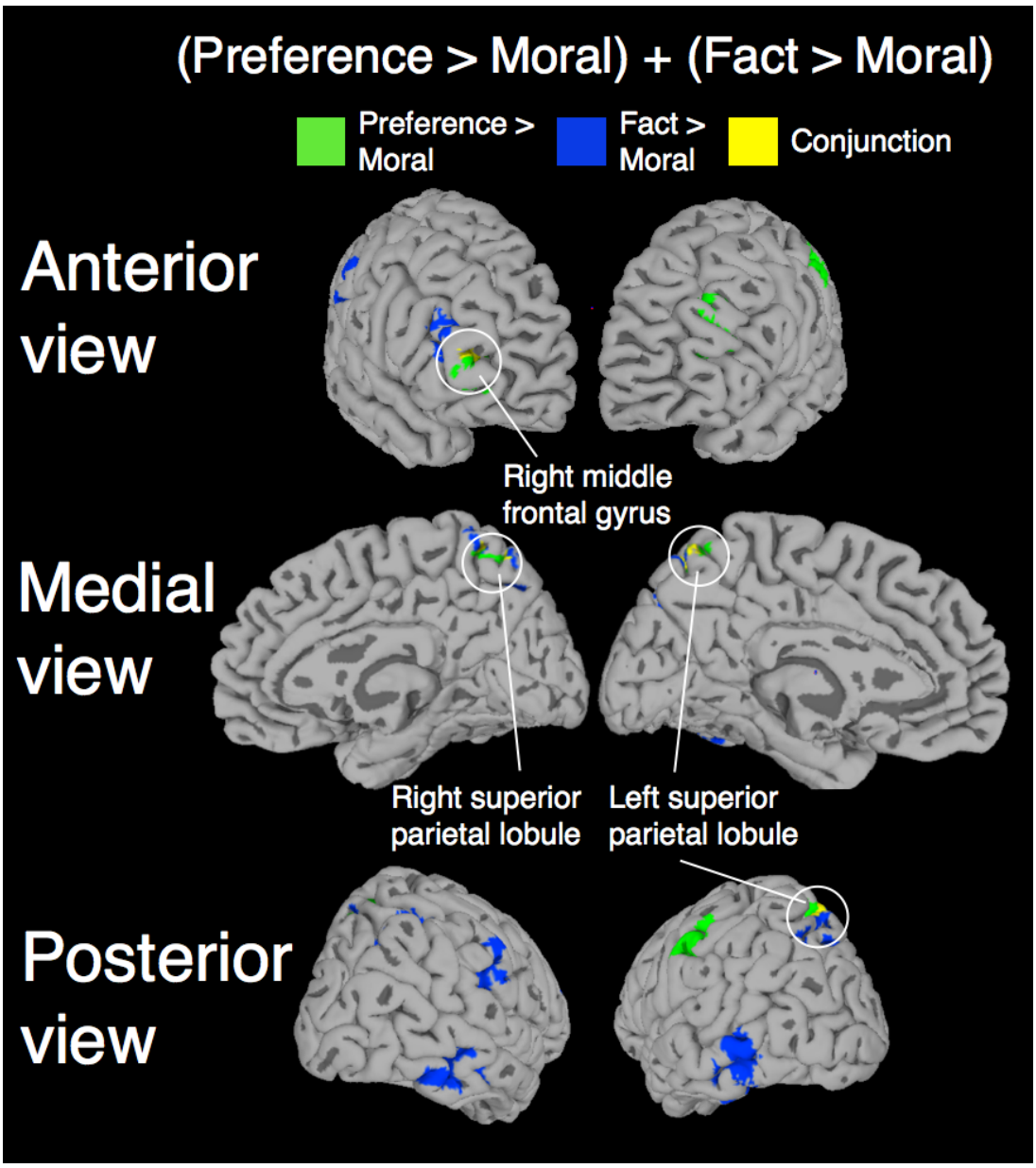

Figure S1. Whole-brain Conjunction Analysis for Preferences and Facts Relative to Morals. Preferences and facts, relative to morals, elicited common activity in left middle frontal gyrus, peak coordinates: preference >moral [38, 44, 8], fact > moral [38, 42, 10]; left superior parietal lobule, peak coordinates: preference >moral [-10, -66, 58], fact > moral [-8, -68, 60]; and right superior parietal lobule, peak coordinates: preference >moral [8, -58, 66], fact > moral [10, -68, 60]. Permutation tests (5000 samples) were used to achieve a cluster-corrected familywise error rate of $\alpha = .05$ in each contrast, while thresholding voxels at $p < .001$ (uncorrected). Permutation testing was performed using SnPM 13 (http://warwick.ac.uk/snpm; Nichols & Holmes, 2001). Coordinates are reported in MNI space. Peak coordinates for each contrast are reported in Table S5.

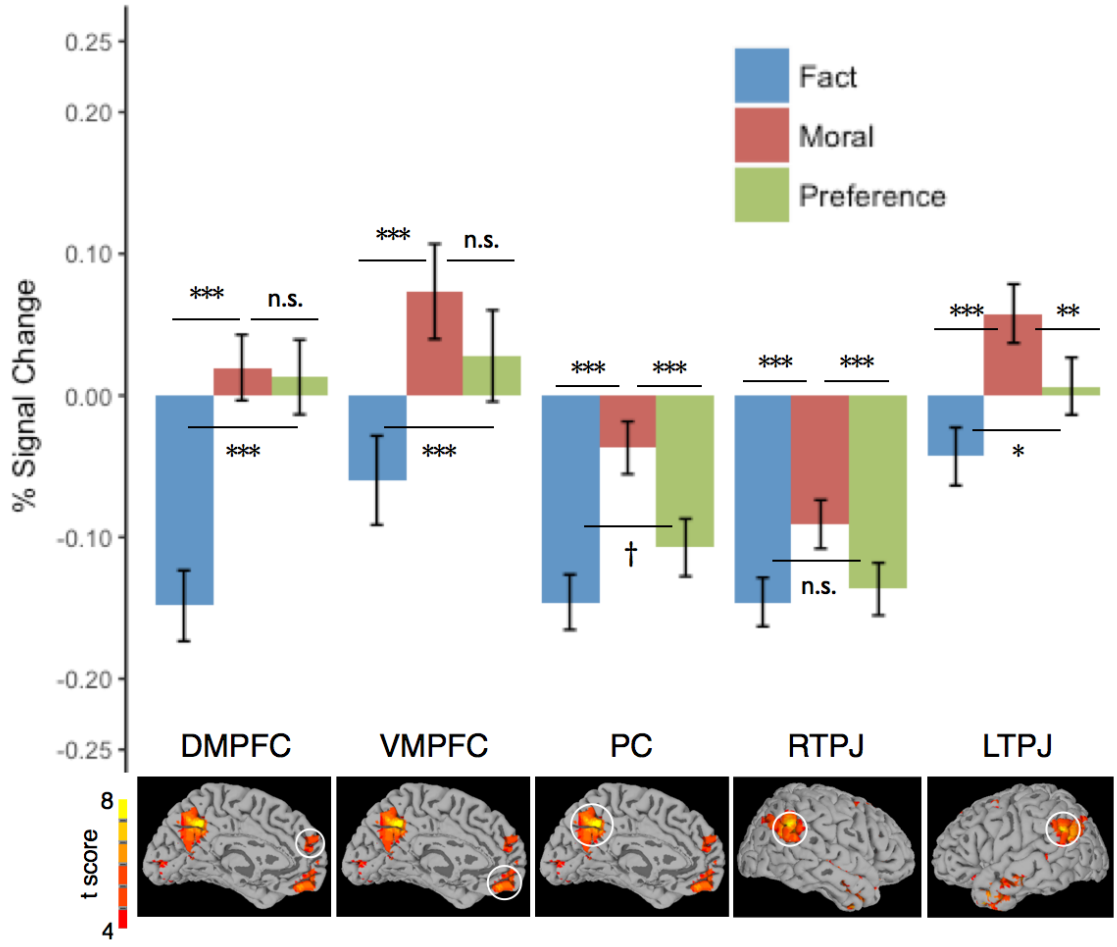

Figure S2. Response Magnitude Across Content (fact/moral/preference) and anatomically defined ToM ROIs. ROIs were identified using the peak coordinates of a whole brain contrast of the localizer contrast (false belief > false photograph) across all participants. Each ROI is defined as a 9mm sphere around these peak coordinates. Coordinates are reported in Table S3. Error bars indicate 95% confidence intervals of condition means. *** $p < .001$; ** $p < .01$; * $p < .05$; † $p < .10$.

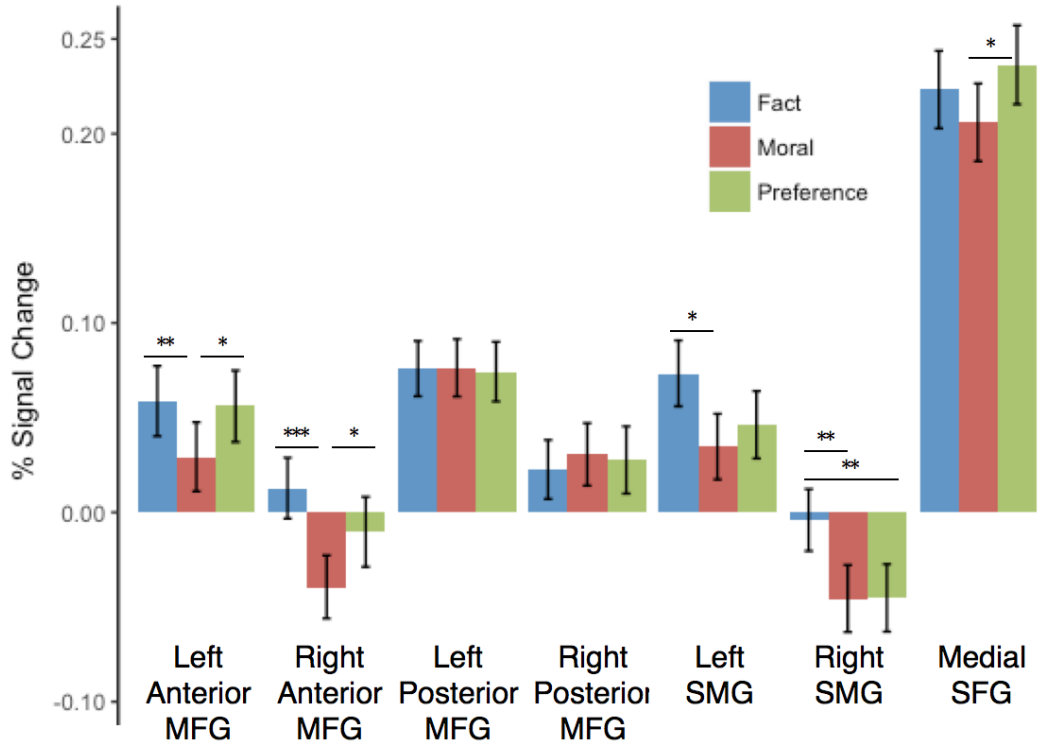

Figure S3. Response Magnitude Across Content (fact/moral/preference) and working memory ROIs. ROIs were identified using the reverse inference map for "working memory" at neurosynth.org (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). MFG = middle frontal gyrus; SMG = supramarginal Gyrus; SFG = superior frontal gyrus. Error bars indicate 95% confidence intervals of condition means. *** $p < .001$; ** $p < .01$; * $p < .05$; † $p < .10$. Any contrasts not marked are non-significant.

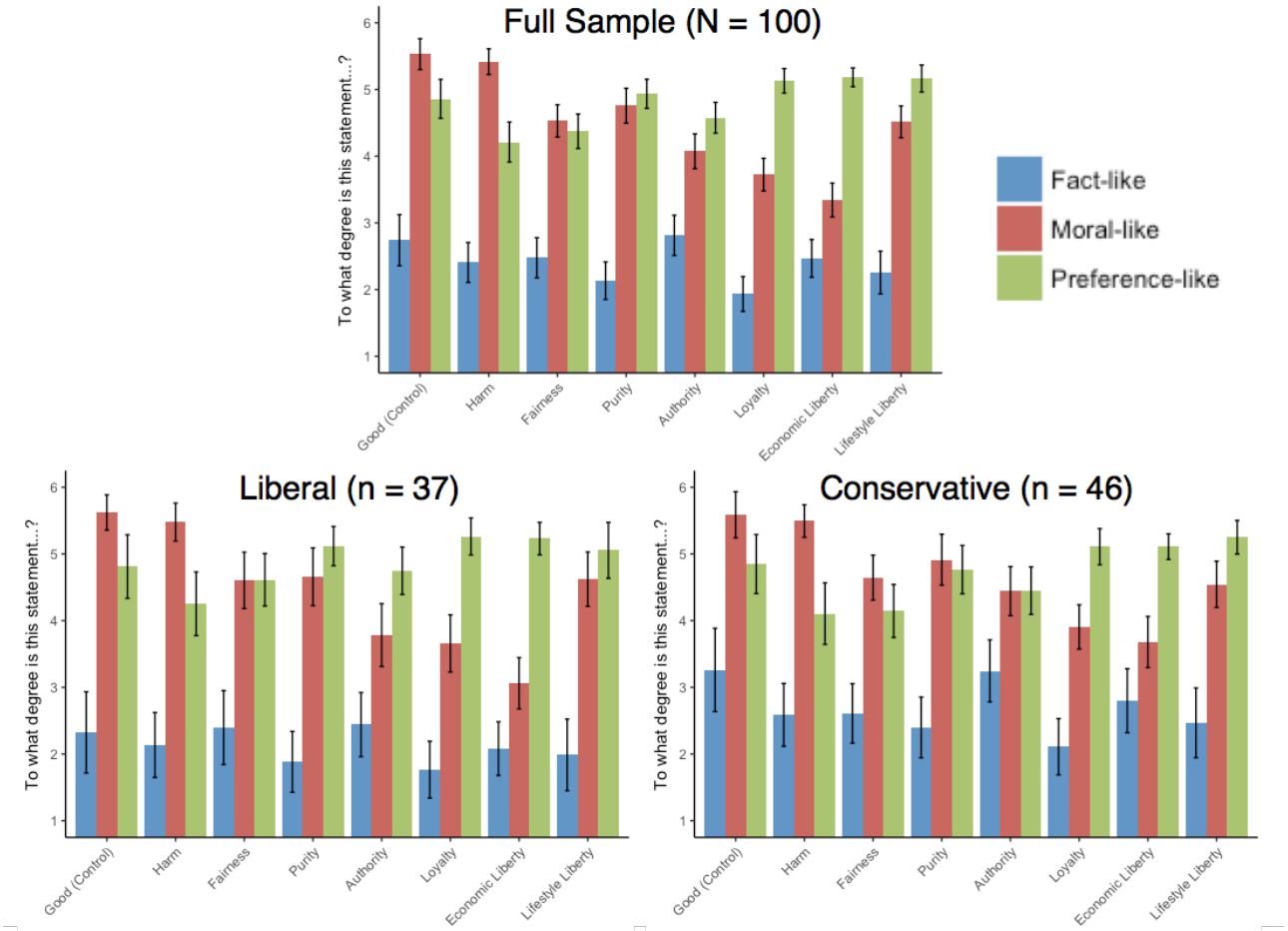

Figure S4. Supplemental Study MFQ Behavioral Ratings. Across all domains, morals were rated as more preference-like than fact-like. This pattern held even when splitting the sample based on political orientation. Participants were grouped as liberal or conservative based on their response to the question: "Please indicate your political orientation relating to social issues" [1 – Very Conservative; 7 – Very Liberal]. Liberals answered above the midpoint (> 4) and conservatives answered below the midpoint (< 4); 16 participants answered at the midpoint and were not grouped, while 1 participant gave no answer. Error bars indicate 95% confidence intervals. For contrast values and associated significance, see Table S10.

Table S1. Study 1 behavioral results.

Category: Morals			
Model: lmer(DV ~ Dimension + (Dimension | ID) + (Dimension | Item)			
	F statistic	***p***	
Dimension (main effect)	$F(2, 48.4)$† = 114.1	<.001	
Post-hoc paired t-tests	***z ratio***	***p***	***Mean Diff (SE)***
Moral-like > Fact-like	$z = 14.5$	<.001*	2.70 (0.19)
Moral-like > Preference-like	$z = 2.02$	.043*	0.61 (0.30)
Preference-like > Fact-like	$z = 7.45$	<.001***	2.10 (0.28)
Morals are perceived as more moral-like and preference-like than fact-like.			
Category: Preferences			
Model: lmer(DV ~ Dimension + (Dimension | ID) + (Dimension | Item)			
	F statistic	***p***	
Dimension (main effect)	$F(2, 67.9)$† = 817.1	<.001	
Post-hoc paired t-tests	***z ratio***	***p***	***Mean Diff (SE)***
Preference-like > Fact-like	$z = 29.7$	$p < .001$	4.11 (0.14)
Preference-like > Moral-like	$z = 39.8$	$p < .001$	4.47 (0.11)
Fact-like ~> Moral-like	$z = 4.8$	$p < .001$	0.36 (0.08)
Preferences are perceived as more preference-like than they are moral-like or fact-like.			
Category: Facts			
Model: lmer(DV ~ Dimension + (Dimension | ID) + (Dimension | Item)			
	F statistic	***p***	
Dimension (main effect)	$F(2, 34.3)$† = 350.5	<.001	
Post-hoc paired t-tests	***z ratio***	***p***	***Mean Diff (SE)***
Fact-like > Moral-like	$z = 26.3$	<.001***	4.33 (0.16)
Fact-like > Preference-like	$z = 23.7$	<.001***	4.18 (0.18)
Preference-like > Moral-like	$z = 2.7$	.007**	0.15 (0.06)
Facts are perceived as more fact-like than they are moral-like or preference-like.			

Post-hoc tests are uncorrected for multiple comparisons. †In these analyses we used the Satterthwaite approximation of degrees of freedom for reasons of computational expense. *** $p < .001$; ** $p < .01$; * $p < .05$.

Table S2. Study 2 peak individual ROI coordinates for ToM functional localizer.

Region	N	x	y	z	k
RTPJ	25/25	51 +/- 6	-53 +/- 4	25 +/- 5	263 +/- 84
LTPJ	24/25	-49 +/- 7	-57 +/- 4	26 +/- 5	214 +/- 80
PC	23/25	1 +/- 4	-56 +/- 5	37 +/- 5	262 +/- 81
DMPFC	20/25	2 +/- 4	56 +/- 5	26 +/-7	128 +/- 86
VMPFC	20/25	1 +/- 4	56 +/- 4	-8 +/- 6	120 +/- 72

Mean and standard deviation, across participants, of peak coordinates for false belief > false photo contrast (Dodell-Feder et al., 2011). All coordinates reported in MNI space.

Table S3. Theory of Mind network peak coordinates – group analysis

Region	x	y	z	T score
DMPFC	0	58	22	5.62
VMPFC	0	44	-20	7.69
PC	0	-52	40	10.81
RTPJ	52	-60	24	10.55
LTPJ	-56	-56	28	9.69

ROIs were a 9mm sphere around the reported coordinates. T scores represent difference scores in the false belief > false photograph contrast, in a random effects analysis across all subjects (df = 24). Permutation testing (5000 samples) ensured that this analysis was cluster-corrected to achieve a familywise error rate of $\alpha = .05$, holding voxels at $p < .001$ (uncorrected). All coordinates are reported in MNI space.

Table S4. Study 2 behavioral results.

Category: Morals			
Model: lmer(DV ~ Dimension + (Dimension | ID) + (Dimension | Item)			
	F statistic	***p***	
Dimension (main effect)	$F(2, 34.2) = 46.7$	<.001	
Post-hoc paired t-tests	***z ratio***	***p***	***Mean Diff (SE)***
Moral-like > Fact-like	$z = 9.7$	<.001	3.56 (0.37)
Moral-like > Preference-like	$z = 4.0$	<.001	1.70 (0.43)
Preference-like > Fact-like	$z = 4.4$	<.001	1.86 (0.42)
Morals are perceived as more moral-like and preference-like than fact-like.			
Category: Preferences			
Model: lmer(DV ~ Dimension + (Dimension | ID) + (Dimension | Item)			
	F statistic	***p***	
Dimension (main effect)	$F(2, 27.6) = 73.8$	<.001	
Post-hoc paired t-tests	***z ratio***	***p***	***Mean Diff (SE)***
Preference-like > Fact-like	$z = 12.0$	$p < .001$	4.34 (0.36)
Preference-like > Moral-like	$z = 11.6$	$p < .001$	4.60 (0.40)
Fact-like ~> Moral-like	$z = 1.4$	$p = .307$	0.27 (0.19)
Preferences are perceived as more preference-like than they are moral-like or fact-like.			
Category: Facts			
Model: lmer(DV ~ Dimension + (Dimension | ID) + (0 + Moral-like + Preference-like | Item)			
	F statistic	***p***	
Dimension (main effect)	$F(2, 21.9) = 107.4$	<.001	
Post-hoc paired t-tests	***z ratio***	***p***	***Mean Diff (SE)***
Fact-like > Moral-like	$z = 14.4$	<.001***	5.14 (0.36)
Fact-like > Preference-like	$z = 13.6$	<.001***	5.00 (0.37)
Preference-like > Moral-like	$z = 2.0$	.094	0.15 (0.07)
Facts are perceived as more fact-like than they are moral-like or preference-like.			

Table S5. Study 2 whole-brain random effects contrasts: peak coordinates.

Contrast	Name	Cluster Size	Peak T	x	y	z
(Moral > Fact)	L Superior Frontal Gyrus	4020	10.14	-4	56	30
			9.01	-16	38	48
			8.48	-20	48	34
	L Precentral Gyrus	1142	8.39	-48	2	36
			7.02	-64	-18	-14
			6.72	-58	-16	-24
	M Precuneus	828	7.05	-8	-48	32
			5.95	-6	-56	36
			5.36	6	-50	24
	L Superior Temporal Gyrus	666	6.67	-40	-56	28
			5.68	-50	-60	38
			5.48	-52	-64	26
	R Cerebellum	486	8.99	26	-78	-34
	L Inferior Frontal Gyrus	341	5.67	-32	24	-18
			5.21	-50	18	4
			4.28	-52	22	12
	R Inferior Temporal Gyrus	158	5.64	36	14	-44
			5.58	50	0	-34
			4.75	44	10	-36
	R Superior Frontal Gyrus	143	7.33	16	24	62
			6.02	14	36	54
	L Cerebellum	139	5.23	-24	-80	-34
			3.82	-34	-84	-30
	L Middle Frontal Gyrus	121	5.3	-40	10	50
			4.4	-26	2	44
	L Pons	121	4.61	-12	-40	-26
			4.32	-10	-30	-36
			3.65	-4	-24	-38
(Moral > Preference)	M Precuneus	1055	5.7	-4	-54	30
			5.64	14	-44	26
			5.49	-10	-44	32
	L Medial Temporal Gyrus	1047	7.05	-52	-4	-30
			6.38	-62	-18	-12
			6.16	-62	-14	-20
	L Superior Temporal Gyrus	850	6.51	-44	-60	32
			5.42	-54	-68	30
			5.34	-40	-68	42
	L Occipital Gyri	331	5.11	-20	-92	2
			4.9	-12	-94	6
			4.85	-18	-88	-10
	L Superior Frontal Gyrus	198	5.74	-18	36	46
			4.61	-4	48	44
			4.49	-10	42	52
	R Medial Temporal Gyrus	172	5.97	50	-2	-34
			5.02	58	-6	-34
			4.13	62	-6	-26
	L Medial Temporal Gyrus	159	5.79	-52	-44	10
			4.42	-52	-34	-4
			4.02	-54	-48	0

	R Precentral Gyrus	109	4.99	12	-14	66
			4.49	16	-18	56
			4.47	24	-20	54
	L Superior Frontal Gyrus	100	5.3	-8	32	54
			4.98	-12	24	52
			4.05	-12	18	44
(Preference > Fact)	M Superior Frontal Gyrus	1762	9.61	-2	54	24
			8.4	2	52	16
			8.17	-22	48	34
	R Cerebellum	565	6.97	36	-84	-32
			6.59	28	-80	36
			5.06	46	-66	-34
	M Straight Gyrus	157	4.74	4	40	-20
			4.65	-2	48	-14
(Preference > Moral)	R/L Superior Parietal Lobule	337	5.94	-10	-66	58
			4.83	12	-58	62
			4.71	-4	-60	62
	L Supramarginal Gyrus	244	6.75	-58	-28	38
			4.85	-52	-30	52
			4.66	-56	-36	48
	R Middle Frontal Gyrus	199	6.24	38	44	8
			3.93	42	52	-4
			3.72	52	36	2
	L Middle Frontal Gyrus	184	5.75	-40	48	14
			4.86	-38	42	30
			4.73	-44	42	22
(Fact > Preference)	L Angular Gyrus	364	5.37	-32	-76	42
			5.11	-28	-74	50
			5.02	-30	-66	48
	L Inferior Temporal Gyrus	287	5.76	-56	-60	-8
			5.25	-62	-52	-10
			4.71	-46	-54	-18
	R Angular Gyrus	218	5.43	30	-50	36
			5.04	36	-66	42
			4.28	34	-74	38
	R Intraparietal Sulcus	155	5.74	18	-58	26
			4.18	20	-66	38
	R Middle Temporal Gyrus	154	5.01	62	-42	-8
			4.97	58	-48	-14
			4.56	60	-40	16
	R Middle Frontal Gyrus	123	6.25	44	30	24
(Fact > Moral)	L Inferior Temporal Gyrus	618	6.52	-58	-56	-20
			5.89	-50	-48	-26
			5.65	-58	-62	2
	R Inferior Temporal Gyrus	552	6.35	56	-50	-20
			6.21	64	-38	-16
			5.91	60	-56	-16
	R Parietooccipital Transition Zone	455	6.2	12	-68	42
			5.84	16	-62	24
			5.47	8	-56	70
	L Parietooccipital Transition	455	5.45	-10	-74	46

Zone		5.35	-8	-68	60
		4.95	-30	-74	42
R Middle Frontal Gyrus	318	5.28	38	42	10
		5.08	46	38	12
		4.9	50	32	24
R Angular Gyrus	282	5.5	38	-56	54
		4.76	38	-50	44
		4.43	36	-70	44
R Parietal Operculum	187	5.12	62	-30	30
		4.75	62	-30	42

Contrasts were first modeled for each participant, and entered into a random effects analysis across all participants. Permutation tests (5000 samples) were used to achieve a cluster-corrected familywise error rate of $\alpha = .05$ in each contrast, while thresholding voxels at $p < .001$ (uncorrected). Permutation testing was performed using SnPM 13 (http://warwick.ac.uk/snpm; Nichols & Holmes, 2001). All coordinates reported in MNI space.

Table S6. Study 2 mixed effects analysis across all claims, examining ROI percent signal change (PSC) for morals and preferences relative to facts.

ROI	Step	Model: R Syntax	Coefficients
DMPFC	***Hypothesis testing***	lmer(PSC ~ Moral + Preference + (1|Item) + (Moral+Preference|ID))	******Moral:*** $\beta = 0.222$, $t(35.1) = 5.94$, $p = 9.1 \times 10^{-7}$ ******Preference:*** $\beta = 0.182$, $t(40.1) = 5.14$, $p = 7.5 \times 10^{-6}$
	Identify potential covariates	lmer(PSC ~ MentalState + (1|Item) + (1|ID))	******Mental States:*** $\beta = 0.078$, $t(70.0) = 8.74$, $p = 7.9 \times 10^{-13}$
		lmer(PSC ~ Arousal + (1|Item) + (1|ID))	******Arousal:*** $\beta = 0.069$, $t(70.1) = 4.33$, $p = 4.9 \times 10^{-5}$
		lmer(PSC ~ NounFamiliarity + (1|Item) + (1|ID))	****Noun Familiarity:*** $\beta = 0.002$, $t(70.1) = 2.38$, $p = .020$
		lmer(PSC ~ NounConcreteness + (1|Item) + (1|ID))	****Noun Concreteness:*** $\beta = -0.0005$, $t(69.8) = 2.27$, $p = .026$
		lmer(PSC ~ PersonPresent + (1|Item) + (1|ID))	****Person Present:*** $\beta = 0.077$, $t(70.0) = 2.19$, $p = .032$
		lmer(PSC ~ NounImageability + (1|Item) + (1|ID))	****Noun Imageability:*** $\beta = -0.0005$, $t(69.8) = 2.05$, $p = .044$
	Attempt to disprove hypothesis	***Marginal/non-significant model:*** lmer(PSC ~ MentalState + Moral + Preference + (1|Item) + (Moral+Preference|ID))	***†Moral:*** $\beta = 0.119$, $t(74.6) = 1.58$, $p = .118$ ***†Preference:*** $\beta = 0.098$, $t(71.2) = 1.54$, $p = .129$ ***Mental States:*** $\beta = 0.039$, $t(68.0) = 1.57$, $p = .120$
		Full model: lmer(PSC ~ RT + NounImageability + PersonPresent + NounConcreteness + NounFamiliarity + Arousal + MentalState + Moral + Preference + (1|Item) + (Moral+Preference|ID))	***Moral:*** $\beta = 0.118$, $t(67.4) = 1.52$, $p = .132$ ***Preference:*** $\beta = 0.097$, $t(64.8) = 1.43$, $p = .157$ ***Mental States:*** $\beta = 0.037$, $t(62.2) = 1.30$, $p = .200$ ***Arousal:*** $\beta = 0.004$, $t(62.5) = 0.19$, $p = .849$ *****Noun Familiarity:*** $\beta = 0.002$, $t(60.8) = 2.70$, $p = .008$ ***Noun Concreteness:*** $\beta = -0.00006$, $t(60.5) = 0.12$, $p = .904$ ***Person Present:*** $\beta = 0.003$, $t(60.8) = 1.13$, $p = .262$ ***Noun Imageability:*** $\beta = -0.00007$, $t(61.0) = 0.01$, $p = .990$ ***Reaction Time:*** $\beta = -0.0009$, $t(1325.0) = 0.08$, $p = .940$
VMPFC	***Hypothesis testing***	lmer(PSC ~ Moral + Preference + (1|Item) + (Moral+Preference|ID))	******Moral:*** $\beta = 0.159$, $t(32.8) = 3.91$, $p = 4.3 \times 10^{-4}$ ****Preference:*** $\beta = 0.098$, $t(33.8) = 2.15$, $p = .039$
	Identify potential covariates	lmer(PSC ~ MentalState + (1|Item) + (1|ID))	******Mental States:*** $\beta = 0.050$, $t(70.1) = 4.15$, $p = 9.1 \times 10^{-5}$
		lmer(PSC ~ Arousal + (1|Item) + (1|ID))	*****Arousal:*** $\beta = 0.054$, $t(70.0) = 2.99$, $p = .004$
		lmer(PSC ~ PersonPresent + (1|Item) + (1|ID))	****Person Present:*** $\beta = 0.087$, $t(69.9) = 2.30$, $p = .024$

		lmer(PSC ~ RT + (1|Item) + (1|ID))	***Reaction Time:*** $\beta = 0.049$, $t(1260.4) = 2.24$, $p = .026$
	Attempt to disprove hypothesis	***Marginal/non-significant model:*** lmer(PSC ~ MentalState + Moral + Preference + (1|Item) + (Moral+Preference|ID))	***Moral:*** $\beta = 0.091$, $t(70.6) = 0.91$, $p = .368$ ***Preference:*** $\beta = 0.043$, $t(70.9) = 0.48$, $p = .629$ ***Mental States:*** $\beta = 0.025$, $t(68.4) = 0.74$, $p = .464$
		Full model: lmer(PSC ~ RT + PersonPresent + Arousal + MentalState + Moral + Preference + (1|Item) + (Moral+Preference|ID))	***Moral:*** $\beta = 0.095$, $t(68.3) = 0.90$, $p = .372$ ***Preference:*** $\beta = 0.066$, $t(69.8) = 0.70$, $p = .489$ ***Mental States:*** $\beta = 0.014$, $t(67.9) = 0.34$, $p = .739$ ***Arousal:*** $\beta = 0.013$, $t(67.8) = 0.52$, $p = .606$ ***Person Present:*** $\beta = 0.055$, $t(66.4) = 1.37$, $p = .176$ ***Reaction Time:*** $\beta = 0.036$, $t(1222.7) = 2.11$, $p = .035$
LTPJ	***Hypothesis testing***	lmer(PSC ~ Moral + Preference + (1|Item) + (Moral+Preference|ID))	****Moral:*** $\beta = 0.148$, $t(49.5) = 5.00$, $p = 7.5 \times 10^{-6}$ ***Preference:*** $\beta = 0.066$, $t(56.3) = 2.40$, $p = .020$
	Identify potential covariates	lmer(PSC ~ MentalState + (1|Item) + (1|ID))	****Mental States:*** $\beta = 0.047$, $t(70.0) = 5.63$, $p = 3.4 \times 10^{-7}$
		lmer(PSC ~ PersonPresent + (1|Item) + (1|ID))	****Person Present:*** $\beta = 0.113$, $t(69.9) = 4.53$, $p = 2.4 \times 10^{-5}$
		lmer(PSC ~ Arousal + (1|Item) + (1|ID))	***Arousal:*** $\beta = 0.040$, $t(70.0) = 3.03$, $p = .003$
		lmer(PSC ~ IntentionVerb + (1|Item) + (1|ID))	***Intentional Verb Incidence:*** $\beta = 0.001$, $t(70.1) = 2.58$, $p = .012$
		lmer(PSC ~ NegationDense + (1|Item) + (1|ID))	***Negation Density:*** $\beta = 0.001$, $t(69.9) = 2.57$, $p = .012$
		lmer(PSC ~ ReadingEase + (1|Item) + (1|ID))	***Flesch Reading Ease:*** $\beta = -0.001$, $t(69.9) = 2.55$, $p = .013$
		lmer(PSC ~ NumModifiers + (1|Item) + (1|ID))	***Number of Modifiers:*** $\beta = -0.050$, $t(69.8) = 2.28$, $p = .026$
	Attempt to disprove hypothesis	***Marginal/non-significant model:*** lmer(PSC ~ MentalState + Moral + Preference + (1|Item) + (Moral+Preference|ID))	***Moral:*** $\beta = 0.051$, $t(74.3) = 0.77$, $p = .443$ ***Preference:*** $\beta = -0.013$, $t(74.5) = 0.23$, $p = .819$ ***Mental States:*** $\beta = 0.036$, $t(68.5) = 1.61$, $p = .111$
		Full model: lmer(PSC ~ RT + NumModifiers + ReadingEase + NegationDense + IntentionVerb + Arousal + PersonPresent + MentalState + Moral + Preference + (1|Item) + (Moral+Preference|ID))	***Moral:*** $\beta = 0.092$, $t(62.5) = 1.47$, $p = .146$ ***Preference:*** $\beta = 0.052$, $t(53.0) = 0.98$, $p = .330$ ***Mental States:*** $\beta = 0.015$, $t(62.8) = 0.65$, $p = .518$ ***Person Present:*** $\beta = 0.063$, $t(61.7) = 2.56$, $p = .013$ ***Arousal:*** $\beta = -0.009$, $t(62.5) = 0.66$, $p = .511$ ***Intentional Verb Incidence:***

			β = -0.00005, t(60.5) = 0.14, p = .887 ****Negation Density:*** β = 0.001, t(64.1) = 2.59, p = .012 ***Flesch Reading Ease*** β = -0.0006, t(60.9) = 1.17, p = .245 ***Number of Modifiers:*** β = -0.023, t(61.5) = 1.35, p = .181 ***Reaction Time*** β = 0.006, t(1578.0) = 0.50, p = .495
PC	***Hypothesis testing***	lmer(PSC ~ Moral + Preference + (1|Item) + (Moral+Preference|ID))	******Moral:*** β = 0.158, t(58.9) = 4.70, p = 1.63 x 10^{-5} ***Preference:*** β = 0.051, t(58.9) = 1.53, p = .132
		lmer(PSC ~ Moral + (1|Item) + (Moral|ID))	******Moral:*** β = 0.133, t(61.8) = 4.60, p = 2.1 x 10^{-5}
	Identify potential covariates	lmer(PSC ~ MentalState + (1|Item) + (1|ID))	******Mental States:*** β = 0.047, t(70.0) = 4.47, p = 2.9 x 10^{-5}
		lmer(PSC ~ PersonPresent + (1|Item) + (1|ID))	******Person Present:*** β = 0.125, t(70.0) = 4.14, p = 9.7 x 10^{-5}
		lmer(PSC ~ IntentionVerb + (1|Item) + (1|ID))	*****Intentional Verb Incidence:*** β = 0.001, t(70.1) = 2.79, p = .007
		lmer(PSC ~ Arousal + (1|Item) + (1|ID))	****Arousal:*** β = 0.038, t(70.0) = 2.44, p = .017
		lmer(PSC ~ ReadingEase + (1|Item) + (1|ID))	****Flesch Reading Ease:*** β = -0.002, t(69.9) = -2.20, p = .031
	Attempt to disprove hypothesis	***Marginal/non-significant model:*** lmer(PSC ~Arousal + IntentionVerb + PersonPresent + MentalState + Moral + (1|Item) + (Moral|ID))	***†Moral:*** β = 0.067, t(63.6) = 1.95, p = .055 ***†Mental States:*** β = 0.029, t(66.2) = 1.87, p = .066 ****Person Present:*** β = 0.069, t(66.0) = 2.19, p = .032 **Intention Verb Incidence:** β = 0.0004, t(66.2) = 1.18, p = .241 **Arousal:** β = -0.010, t(66.2) = 0.55, p = .584
		Full model: lmer(PSC ~ RT + ReadingEase + Arousal + IntentionVerb + PersonPresent + MentalState + Moral + (1|Item) + (Moral|ID))	***†Moral:*** β = 0.064, t(61.0) = 1.64, p = .068 ****Mental States:*** β = 0.035, t(66.7) = 1.89, p = .027 ***†Person Present:*** β = 0.056, t(64.5) = 1.82, p = .081 ***Intention Verb Incidence:*** β = 0.0004, t(63.6) = 1.24, p = .249 ***Arousal:*** β = -0.012, t(65.5) = 0.63, p = .502 ***Flesch Reading Ease:*** β = -0.0006, t(64.1) = 0.79, p = .354 ***Reaction Time:*** β = -0.003, t(1489.0) = 0.48, p = .796
RTPJ	***Hypothesis testing***	lmer(PSC ~ Moral + Preference + (1|Item) + (Moral+Preference|ID))	*****Moral:*** β = 0.072, t(31.9) = 3.55, p = .001 ***Preference:*** β = 0.023, t(34.6) = 1.35, p = .187

	lmer(PSC ~ Moral + (1|Item) + (Moral|ID))	***Moral:* β = 0.060, t(32.9) = 3.61, p = .001
Identify potential covariates	lmer(PSC ~ RT + (1|Item) + (1|ID))	****Reaction Time:* β = 0.028, t(1633.5) = 3.82, p = 1.3 x 10^{-4}
	lmer(PSC ~ MentalState + (1|Item) + (1|ID))	****Mental States:* β = 0.021, t(70.1) = 4.02, p = 1.4 x 10^{-4}
	lmer(PSC ~ NounFamiliarity + (1|Item) + (1|ID))	**Noun Familiarity:* β = 0.001, t(70.0) = 2.06, p = .043
Attempt to disprove hypothesis	***Marginal/non-significant model:*** lmer(PSC ~ RT + MentalState + Moral + (1|Item) + (Moral|ID))	*Moral:* β = 0.030, t(43.6) = 1.63, p = .110 ***Mental States:* β = 0.017, t(63.3) = 7.92, p = .008 ****Reaction Time:* β = 0.027, t(1615.0) = 3.69, p = 2.3 x 10^{-4}
	Full model: lmer(PSC ~ NounFamiliarity + RT + MentalState + Moral + (1|Item) + (Moral|ID	†*Moral:* β = 0.035, t(41.4) = 1.91, p = .063 **Mental States:* β = 0.016, t(79.4) = 2.52, p = .014 ****Reaction Time:* β = 0.027, t(1604.0) = 3.71, p = 2.2 x 10^{-4} **Noun Familiarity:* β = 0.0008, t(69.4) = 2.53, p = .014

Analyses were performed using R (R Core Team, 2016), and the *lme4* package (Bates et al., 2015), using the Kenward-Roger approximation of degrees of freedom (*lmerTest*, Kuznetsova et al., 2015; *pbkrtest*, Halekoh & Højsgaard, 2014). *** p < .001; ** p < .01; * p < .05; † p < .1. β represent standardized regression coefficients.

Table S7. Working memory network ROI coordinates.

Region	x	y	z	Z ratio
L Anterior MFG	-42	30	26	9.81
R Anterior MFG	42	34	28	10.24
L Posterior MFG	-28	0	54	8.84
R Posterior MFG	32	4	52	10.37
L SMG	-36	-50	44	11.14
R SMG	40	-50	46	9.50
Medial SFG	0	16	48	8.16

ROIs were a 9mm sphere around the reported coordinates. Z ratios correspond to the reverse inference map for "working memory" at neurosynth.org (Yarkoni et al., 2011). The Z ratio represents the extent that this voxel is *preferentially* related to the term "working memory". All coordinates are reported in MNI space. MFG = middle frontal gyrus; SMG = supramarginal Gyrus; SFG = superior frontal gyrus.

Table S8 Item-wise covariates: descriptive statistics and ANOVAs.

Question name	Descriptive Statistics: *M* (*S.D.*)			ANOVA
	Facts N_{Items} = 24	Morals N_{Items} = 24	Preferences N_{Items} = 24	
Coh Metrix 3.0 Measures				
Word count	12.1 (2.3)	12.0 (2.4)	11.8 (2.4)	$F(2, 69) = 0.09$, $p = .912$
Flesch reading ease	62.0 (21.6)	55.8 (17.0)	61.2 (23.4)	$F(2, 69) = 0.62$, $p = .542$
Anaphor reference	65.7 (9.8)	69.2 (10.7)	68.8 (13.0)	$F(2, 69) = 0.70$, $p = .502$
Intentional verb incidence	14.8 (35.5)	25.8 (46.2)	8.80 (29.8)	$F(2, 69) = 1.26$, $p = .292$
Causal verb incidence	38.3 (43.8)	23.8 (42.6)	18.8 (38.0)	$F(2, 69) = 1.43$, $p = .246$
Causal verb ratio	0.10 (0.29)	0.19 (0.38)	0.12 (0.30)	$F(2, 69) = 0.41$, $p = .663$
Noun concreteness	438.2 (62.8)	406.2 (62.3)	379.1 (71.1)	$F(2, 69) = 4.90$, $p = .010$ **
Noun familiarity	574.0 (18.8)	573.2 (15.8)	578.2 (21.5)	$F(2, 69) = 0.49$, $p = .615$
Noun imageability	466.8 (56.4)	439.2 (57.6)	420.4 (58.8)	$F(2, 69) = 3.94$, $p = .024$ *
Negation density	7.1 (24.6)	8.4 (28.7)	3.2 (15.7)	$F(2, 69) = 0.32$, $p = .729$
Number of modifiers	1.01 (0.52)	0.71 (0.58)	0.86 (0.52)	$F(2, 69) = 1.94$, $p = .151$
Left embeddedness	3.54 (2.06)	2.50 (2.13)	3.96 (2.44)	$F(2, 69) = 2.76$, $p = .070$ †
Online Norming Measures				
Agreement	4.11 (1.36)	3.96 (1.47)	3.96 (1.29)	$F(2, 69) = 3.20$, $p = .047$ *
Valence	0.91 (1.75)	-1.47 (2.24)	0.38 (2.86)	$F(2, 69) = 6.89$, $p = .002$ **
Arousal	5.42 (0.75)	6.60 (0.78)	6.50 (0.66)	$F(2, 69) = 19.57$, $p < .001$ ***
(Positive Rating)	3.16 (0.88)	2.57 (1.04)	3.44 (1.47)	$F(2, 69) = 3.57$, $p = .033$ *
(Negative Rating)	2.26 (1.02)	4.04 (1.32)	3.06 (1.47)	$F(2, 69) = 11.60$, $p < .001$ ***
Mental imagery	4.18 (0.74)	4.20 (0.62)	4.36 (0.71)	$F(2, 69) = 0.47$, $p = .629$
Mental state	2.14 (0.54)	4.70 (0.50)	4.24 (0.38)	$F(2, 69) = 196.4$, $p < .001$ ***
Person present	0.31 (0.44)	0.57 (0.43)	0.23 (0.39)	$F(2, 69) = 4.46$, $p = .015$ *
In-scanner				
Reaction time	1.26 (0.17)	1.38 (0.25)	1.27 (0.22)	$F(2, 69) = 2.06$, $p = .135$

Coh Metrix ratings are calculated using an online tool at http://cohmetrix.com (Graesser et al., 2004; McNamara et al., 2014). Online samples were collected using Amazon Mechanical Turk. All measures are described in detail in appendix B.

Table S9. Supplemental analysis mixed effects analysis across all claims, examining ROI percent signal change (PSC) for morals and preferences relative to facts, and intrinsic differences between categories.

ROI	Step	Model: R Syntax	Coefficients
DMPFC	***Hypothesis testing***	lmer(PSC ~ Moral + Preference + (1|Item) + (Moral+Preference|ID))	******Moral:*** β = 0.222, t(35.1) = 5.94, p = 9.1 x 10^{-7} ******Preference:*** β = 0.182, t(40.1) = 5.14, p = 7.5 x 10^{-6}
	All intrinsic differences except for mental states	lmer(PSC ~ Moral + Preference + Arousal + Valence + NounConcreteness + PersonPresent + NounImageability + Agreement + (1|Item) + (Moral+Preference|ID)	*****Moral:** β = 0.205, t(43.1) = 4.65, p = 3.2 x 10^{-5} *****Preference:** β = 0.159, t(46.4) = 3.66, p = 6.5 x 10^{-4} **Arousal:** β = 0.005, t(63.7) = 0.31, p = .759 **Valence:** β =0.005, t(63.0) = 0.74, p = .461 **Noun Concreteness:** β =-0.0004, t(62.9) = 0.84, p = .406 **†Person Present:** β = 0.049, t(63.2) = 1.77, p = .082 **Noun Imageability:** β = 0.0002, t(62.8) = 0.44, p = .663 **Agreement:** β = -0.011, t(62.9) = 1.10, p = .277
	All intrinsic differences	lmer(PSC ~ Moral + Preference + Arousal + Valence + NounConcreteness + PersonPresent + NounImageability + Agreement + MentalStates (1|Item) + (Moral+Preference|ID)	**Moral:** β = 0.100, t(69.2) = 1.25, p = .214 **Preference:** β = 0.068, t(67.7) = 0.94, p = .350 **Arousal:** β = -0.009, t(62.4) = 0.46, p = .647 **Valence:** β = 0.005, t(62.0) = 0.85, p = .400 **Noun Concreteness:** β = -0.0005, t(61.9) = 0.99, p = .328 **Person Present:** β = 0.041, t(62.2) = 1.47, p = .147 **Noun Imageability:** β = 0.0003, t(61.8) = 0.62, p = .540 **Agreement:** β = -0.016, t(61.9) = 1.48, p = ..145 **Mental States:** β = 0.047, t(62.0) = 1.56, p = .124
VMPFC	***Hypothesis testing***	lmer(PSC ~ Moral + Preference + (1|Item) + (Moral+Preference|ID))	******Moral:*** β = 0.159, t(32.8) = 3.91, p = 4.3 x 10^{-4} ****Preference:*** β = 0.098, t(33.8) = 2.15, p = .039

	All intrinsic differences except for mental states	lmer(PSC ~ Moral + Preference + Arousal + Valence + NounConcreteness + PersonPresent + NounImageability + Agreement + (1|Item) + (Moral+Preference|ID)	****Moral:** β = 0.141, t(56.7) = 2.70, p = .009 **Preference:** β = 0.067, t(49.8) = 1.17, p = .247 **Arousal:** β = 0.017, t(63.5) = 0.76, p = .448 **Valence:** β = 0.009, t(62.9) = 1.07, p = .287 **Noun Concreteness:** β = -0.0006, t(62.3) = 0.92, p = .360 **Person Present:** β = 0.055, t(62.8) = 1.42, p = .160 **Noun Imageability:** β = 0.0007, t(62.1) = 0.86, p = .394 **Agreement:** β = -0.015, t(62.7) = 1.04, p = .305
	All intrinsic differences	lmer(PSC ~ Moral + Preference + Arousal + Valence + NounConcreteness + PersonPresent + NounImageability + Agreement + MentalStates + (1|Item) + (Moral+Preference|ID)	**Moral:** β = 0.105, t(63.8) = 0.97, p = .337 **Preference:** β = 0.036, t(66.2) = 0.36, p = .719 **Arousal:** β = 0.013, t(62.0) = 0.49, p = .624 **Valence:** β = 0.009, t(61.9) = 1.09, p = .281 **Noun Concreteness:** β = -0.0007, t(61.3) = 0.95, p = .348 **Person Present:** β = 0.052, t(62.0) = 1.32, p = .192 **Noun Imageability:** β = 0.0007, t(61.1) = 0.89, p = .378 **Agreement:** β = -0.016, t(61.7) = 1.09, p = .280 **Mental States:** β = 0.016, t(62.0) = 0.38, p = .708
LTPJ	***Hypothesis testing***	lmer(PSC ~ Moral + Preference + (1|Item) + (Moral+Preference|ID))	******Moral:*** β = 0.148, t(49.5) = 5.00, p = 7.5 x 10^{-6} ****Preference:*** β = 0.066, t(56.3) = 2.40, p = .020
	All intrinsic differences except for mental states	lmer(PSC ~ Moral + Preference + Arousal + Valence + NounConcreteness + PersonPresent + NounImageability + Agreement + (1|Item) + (Moral+Preference|ID)	*****Moral:** β = 0.128, t(60.2) = 3.70, p = 4.7 x 10^{-4} **†Preference:** β = 0.006, t(64.0) = 1.75, p = .085 **Arousal:** β = 0.002, t(63.4) = 0.14, p = .891 **Valence:** β = 0.005, t(62.7) = 0.90, p = .374 **Noun Concreteness:** β = -0.0001, t(62.7) = 0.29, p = .774 *****Person Present:** β = 0.086, t(63.0) = 3.59, p = 6.4 x 10^{-4} **Noun Imageability:** β = 0.00003, t(62.5) = 0.06, p = .949 **Agreement:** β = -0.005, t(62.9) = 0.62, p = .540

	All intrinsic differences	lmer(PSC ~ Moral + Preference + Arousal + Valence + NounConcreteness + PersonPresent + NounImageability + Agreement + MentalStates (1|Item) + (Moral+Preference|ID)	**Moral:** β = 0.066, t(67.7) = 0.98, p = .331 **Preference:** β = 0.051, t(65.9) = 0.09, p = .933 **Arousal:** β = -0.005, t(62.2) = 0.39, p = .701 **Valence:** β = 0.005, t(62.1) = 0.96, p = .339 **Noun Concreteness:** β = -0.0002, t(61.7) = 0.38, p = .702 ****Person Present:** β = 0.081, t(62.1) = 3.33, p = .001 **Noun Imageability:** β = 0.00003, t(62.0) = 0.06, p = .956 **Agreement:** β = -0.008, t(62.0) = 0.87, p = .385 **Mental States:** β = 0.028, t(62.2) = 1.08, p = .287
PC	***Hypothesis testing***	lmer(PSC ~ Moral + Preference + (1|Item) + (Moral+Preference|ID))	*****Moral:*** β = 0.158, t(58.9) = 4.70, p = 1.63 x 10^{-5} ***Preference:*** β = 0.051, t(58.9) = 1.53, p = .132
		lmer(PSC ~ Moral + (1|Item) + (Moral|ID))	*****Moral:*** β = 0.133, t(61.8) = 4.60, p = 2.1 x 10^{-5}
	All intrinsic differences except for mental states	lmer(PSC ~ Moral + Arousal + Valence + NounConcreteness + PersonPresent + NounImageability + Agreement + (1|Item) + (Moral |ID)	****Moral:** β = .106, t(61.8) = 3.12, p = .003 **Arousal:** β = .011, t(64.0) = 0.76, p = .448 **Valence:** β = .003, t(63.8) = 0.46, p = .648 ***Noun Concreteness:** β = -0.001, t(63.5) = 2.05, p = .045 ****Person Present:** β = 0.090, t(63.8) = 3.07, p = .003 **Noun Imageability:** β = 0.001, t(63.5) = 1.61, p = .113 **Agreement:** β = .000008, t(63.7) = 0.001, p = .999
	All intrinsic differences	lmer(PSC ~ Moral + Arousal + Valence + NounConcreteness + PersonPresent + NounImageability + Agreement + MentalStates + (1|Item) + (Moral |ID)	†**Moral:** β = 0.076, t(64.6) = 1.94, p = .057 **Arousal:** β = -0.007, t(63.0) = 0.37, p = .714 **Valence:** β = 0.002, t(62.9) = 0.25, p = .805 †**Noun Concreteness:** β = -0.001, t(62.6) = 1.92, p = .059 ****Person Present:** β = 0.090, t(62.8) = 3.09, p = .003 **Noun Imageability:** β = 0.001, t(62.5) = 1.68, p = .097 **Agreement:** β = 0.002, t(62.8) = 0.19, p = .851 **Mental States:** β = 0.024, t(63.2) = 1.43, p = .157

RTPJ	***Hypothesis testing***	lmer(PSC ~ Moral + Preference + (1|Item) + (Moral+Preference|ID))	****Moral:*** β = 0.072, t(31.9) = 3.55, p = .001 ***Preference:*** β = 0.023, t(34.6) = 1.35, p = .187
		lmer(PSC ~ Moral + (1|Item) + (Moral|ID))	****Moral:*** β = 0.060, t(32.9) = 3.61, p = .001
	All intrinsic differences except for mental states	lmer(PSC ~ Moral + Arousal + Valence + NounConcreteness + PersonPresent + NounImageability + Agreement + (1|Item) + (Moral |ID)	****Moral:** β = 0.067, t(47.4) = 3.40, p = .001 **Arousal:** β = 0.004, t(64.3) = 0.50, p = .619 **Valence:** β = 0.004, t(64.1) = 1.03, p = .306 ***Noun Concreteness:** β = -0.0002, t(63.4) = 2.01, p = .048 **Person Present:** β = 0.001, t(64.0) = 0.08, p = .927 **Noun Imageability:** β = 0.0005, t(63.4) = 1.58, p = .119 **Agreement:** β = -0.005, t(63.9) = 0.88, p = .380
	All intrinsic differences	lmer(PSC ~ Moral + Arousal + Valence + NounConcreteness + PersonPresent + NounImageability + Agreement + MentalStates + (1|Item) + (Moral |ID)	***Moral:** β = 0.054, t(57.1) = 2.38, p = .020 **Arousal:** β = -0.004, t(64.5) = 0.42, p = .673 **Valence:** β = 0.003, t(63.2) = 0.85, p = .400 **†Noun Concreteness:** β = -0.0005, t(62.6) = 1.89, p = .063 **Person Present:** β = 0.001, t(63.1) = 0.07, p = .942 **Noun Imageability:** β = 0.0005, t(63.0) = 1.63, p = .106 **Agreement:** β = -0.004, t(63.0) = 0.72, p = .475 **Mental States:** β = 0.011, t(63.7) = 1.21, p = .230

Analyses were performed using R (R Core Team, 2016), and the *lme4* package (Bates et al., 2015), using the Kenward-Roger approximation of degrees of freedom (*lmerTest*, Kuznetsova et al., 2015; *pbkrtest*, Halekoh & Højsgaard, 2014). *** p < .001; ** p < .01; * p < .05; † p < .1. β represent standardized regression coefficients.

Table S10. Dimension Rating Contrasts Across MFQ Domains.

		Full Sample (N = 100)		Liberal (n = 37)		Conservative (n = 46)	
Contrast	**MFQ Domain**	**Diff (SE)**	**z ratio**	**Diff (SE)**	**z ratio**	**Diff (SE)**	**z ratio**
(Moral-like – Fact-like)	**Good (Control)**	2.79 (0.22)	12.67 ***	3.30 (0.33)	10.04 ***	2.33 (0.34)	6.83 ***
	Harm	3.01 (0.19)	15.73 ***	3.34 (0.30)	11.33 ***	2.91 (0.28)	10.20 ***
	Fairness	2.05 (0.19)	10.74 ***	2.21 (0.32)	6.85 ***	2.04 (0.28)	7.33 ***
	Purity	2.62 (0.18)	14.43 ***	2.77 (0.28)	9.93 ***	2.51 (0.27)	9.19 ***
	Authority	1.26 (0.19)	6.74 ***	1.34 (0.30)	4.52 ***	1.20 (0.28)	4.26 ***
	Loyalty	1.79 (0.16)	10.95 ***	1.89 (0.25)	7.66 ***	1.80 (0.24)	7.37 ***
	Economic Liberty	0.88 (0.17)	5.25 ***	0.98 (0.24)	4.00 ***	0.88 (0.26)	3.38 **
	Lifestyle Liberty	2.26 (0.18)	12.27 ***	2.64 (0.32)	8.25 ***	2.08 (0.27)	7.57 ***
(Preference-like – Fact-like)	**Good (Control)**	2.12 (0.22)	9.63 ***	2.49 (0.33)	7.57 ***	1.59 (0.34)	4.66 ***
	Harm	1.80 (0.19)	9.43 ***	2.12 (0.30)	7.18 ***	1.52 (0.28)	5.33 ***
	Fairness	1.90 (0.19)	9.92 ***	2.22 (0.32)	6.88 ***	1.54 (0.28)	5.53 ***
	Purity	2.80 (0.18)	15.41 ***	3.23 (0.28)	11.58 ***	2.37 (0.27)	8.65 ***
	Authority	1.76 (0.19)	9.43 ***	2.31 (0.30)	7.77 ***	1.20 (0.28)	4.29 ***
	Loyalty	3.20 (0.16)	19.55 ***	3.50 (0.25)	14.16 ***	3.00 (0.24)	12.30 ***
	Economic Liberty	2.72 (0.17)	16.28 ***	3.15 (0.24)	12.87 ***	2.31 (0.26)	8.85 ***
	Lifestyle Liberty	2.91 (0.18)	15.80 ***	3.06 (0.32)	9.60 ***	2.78 (0.27)	10.15 ***
(Moral-like – Preference-like)	**Good (Control)**	0.67 (0.22)	3.04 **	0.81 (0.81)	2.47 *	0.74 (0.34)	2.17 †
	Harm	1.20 (0.19)	6.30 ***	1.23 (0.30)	4.15 ***	1.39 (0.28)	4.87 ***
	Fairness	-0.16 (0.19)	0.82	-0.01 (0.32)	0.03	0.50 (0.28)	1.80
	Purity	-0.18 (0.18)	0.98	0.46 (0.28)	1.64	0.15 (0.27)	0.54
	Authority	-0.50 (0.19)	2.69 *	-0.96 (0.30)	3.25 **	-0.01 (0.28)	0.03
	Loyalty	-1.41 (0.16)	8.60	-1.60 (0.25)	6.50 ***	-1.20 (0.24)	4.93 ***
	Economic Liberty	-1.84 (0.17)	11.04 ***	-2.17 (0.24)	8.86 ***	-1.43 (0.26)	5.48 ***
	Lifestyle Liberty	-0.65 (0.18)	3.53 **	-0.43 (0.32)	1.35	-0.71 (0.27)	2.58 *

Participants were grouped as liberal or conservative based on their response to the question: "Please indicate your political orientation relating to social issues" [1 – Very Conservative; 7 – Very Liberal]. Liberals answered above the midpoint (> 4) and conservatives answered below the midpoint (< 4); 16 participants answered at the midpoint and were not grouped, while 1 participant gave no answer. All p values are corrected for three multiple comparisons (contrasts within each domain and sample grouping; $p_{corrected} = .0167$). $p_{family\text{-}wise} = .05$, $p_{corrected} = .00208$. *** $p < .001$; ** $p < .01$; * $p < .05$; † $p < .1$.

Supplemental References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed effects models using lme4. *Journal of Statistical Software, 67,* 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33*, 497–505. http://dx.doi.org/10.1080/14640748108400805

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *NeuroImage, 55*, 705–712. http://dx.doi.org/10.1016/j.neuroimage.2010.12.040

Fellbaum, C. (1998). *Wordnet: An electronic lexical database.* Cambridge, MA: MIT press.

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. http://dx.doi.org/10.1037/a0021847

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101-124. http://dx.doi.org/10.1080/1047840X.2012.651387

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed Models — The R package pbkrtest. *Journal of Statistical Software, 59,* 1–32. http://dx.doi.org/10.18637/jss.v059.i09

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PloS one*, *7*(8), e42366. http://dx.doi.org/10.1371/journal.pone.0042366

Kron, A., Goldstein, A., Lee, D. H-J., & Gardhouse, K. (2013). How are you feeling? Revisiting the quantification of emotional qualia. *Psychological Science, 24*, 1503–1511. http://dx.doi.org/10.1177/0956797613475456

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests in linear mixed effects models [Computer software manual]. http://CRAN.R-project.org/package=lmerTest. (R Package version 2.0-25).

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K.J. (1990). Introduction to wordnet: an on-line lexical database* *International Journal of Lexicography 3*, 235–244. http://dx.doi.org/10.1093/ijl/3.4.235

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, *8*(8), 665-670. http://dx.doi.org/10.1038/nmeth.1635

Appendix A. Experimental Stimuli.

Studies 1-2.

Fact	Moral	Preference
High-agreement		
In sports-based afterschool programs children participate in sports such as baseball or basketball to name a few.	The goal of sports should be to teach children that respect for others is more important than winning.	Afterschool programs involving sports are more fun than most of the alternatives available to children.
In a full-term human pregnancy, babies spend nine months in a woman's womb.	Parents should be willing to make sacrifices for the benefit of their baby.	Babies that are temperamental are aggravating to spend time around.
Airplanes have wings that enable the plane to lift upwards.	It is irresponsible for airlines to risk the safety of their passengers.	Going through airport security is an unpleasant experience.
University professors teach classes but also conduct research.	Professors should not tolerate students cheating on their exams.	Professors who play videos make their classes more entertaining.
A breathalyzer is used to determine whether a driver is intoxicated.	Driving after drinking heavily is a stupid and selfish way to behave.	Having a drink every now and then is a good way to relax.
Touchscreens are used in a variety of electronics, including smartphones.	The deplorable conditions of Chinese electronics workers should not be ignored.	Using touchscreens is a much more satisfying way to interact with computers.
Low-agreement		
Medical students at hospitals are able to perform surgeries with little to no training.	It is fine for doctors to accidentally kill a small number of patients per year.	Having a doctor listen attentively to your medical concerns is awful.
Coffee beans grow particularly well in freezing cold climates, such as Alaska and Russia.	Child labor in coffee bean farming is acceptable because it lowers the market price.	Drinking coffee is a miserable experience when you are tired and need energy.
The sand on beaches is usually transported there from nearby deserts.	Private beaches are immoral, as everyone should be able to share the space.	While at a hot beach, it is agonizing to dip your toes in the cool water.
Fish are able to live outside of water for an extended time.	Sport fishing to kill and eat fish is barbaric and evil.	Nothing is more appealing than the smell of rotting fish.
In humans, the liver pumps blood throughout the body.	Universal donors should be obligated to donate their blood.	Having blood drawn is a pleasurable experience.
Cockroaches are a type of cold-blooded reptiles related to snakes.	It is wrong to harm cockroaches just because humans find them disgusting.	Cockroaches are delicious to eat because of their hard and crunchy shell.

Mid-agreement		
The very first waffle cone was invented in Chicago, Illinois, at a state fair.	It is unethical for businesses to promote sugary products to children.	Any ice cream flavor tastes better when served in a crunchy waffle cone.
Monopoly pieces were made from wood, not metal, during WWI.	It is wrong to cheat when playing games such as Monopoly.	Many games are better than Monopoly, which is incredibly boring.
The author J.K. Rowling has two younger siblings, one brother and one sister.	Harry Potter should be banned from school libraries for idolizing witchcraft.	The Harry Potter books are engaging and delightful to read, even for adults.
A town in North Dakota holds the world record for the tallest snowman.	People should help their elderly neighbors clear snow from their driveway.	In the wintertime, it is fun to catch snowflakes on the tip of your tongue.
The oldest sandals in the world were found in Oregon's Paisley Caves.	It is wrong to knowingly buy sandals made using sweatshop labor.	Because sandals have fewer styles, they are less fun to go shopping for.
Hummer trucks were first marketed to civilians in 1990.	Good Americans buy American cars, such as Hummers.	Nothing is more awesome than driving in a Hummer.
There are more fish species in the Amazon River than in the Atlantic Ocean.	Eating fish is acceptable if they were treated humanely when caught or raised.	Sitting in a boat and fishing all day long is boring and a waste of time.
The first CD made for commercial release was the rock CD: "Born in the USA".	Music stores should prevent children from buying CDs with violent or sexist lyrics.	Rock music is pleasing to the ear, and much more agreeable than rap music.
Newtown Pippin was the first apple variety exported from the US.	It is unjust for businesses to allow apples to rot rather than giving them to the needy.	Green apples are too sour to be an enjoyable lunchtime snack.
Of all types of birds, owls are the ones that can see the color blue.	Destroying the habitats of owls through deforestation is deplorable.	The "hoots" of owls in the woods make camping more enjoyable.
The dog breed, Basenji, is the world's only barkless dog breed.	Dog racing is harmful and exploitative to the dogs being raced.	Dogs are not worth the stress and aggravation it takes to own them.
Saturn’s moon, Titan, is the only moon known to have clouds.	It is wrong to use animals as disposable space shuttle test pilots.	Gazing at planets through a telescope is a satisfying activity.

Supplemental Study.

Facts and preferences are unchanged from Studies 1 & 2.

Moral	
Modifications from Moral Foundations Questionnaire (Graham et al., 2011; Iyer et al.,2012) are bolded	
Original	**Adapted**
Control (Good)	
It is better to do good than to do bad.	It is better to do good than to do bad.
Harm	
Compassion for those who are suffering is the most crucial virtue.	Compassion for those who are suffering is the most crucial virtue.
One of the worst things a person could do is hurt a defenseless animal.	One of the worst things a person could do is hurt a defenseless animal.
It can never be right to kill a human being.	It can never be right to kill a human being.
Fairness	
When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.	When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.
Justice is the most important requirement for a society.	Justice is the most important requirement for a society.
I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing.	**It's** wrong that rich children inherit a lot of money while poor children inherit nothing.
Purity	
People should not do things that are disgusting, even if no one is harmed.	People should not do things that are disgusting, even if no one is harmed.
I would call some acts wrong on the grounds that they are unnatural.	Some acts **are** wrong on the grounds that they are unnatural.
Chastity is an important and valuable virtue.	Chastity is an important and valuable virtue.
Authority	
Respect for authority is something all children need to learn.	Respect for authority is something all children need to learn.
Men and women each have different roles to play in society.	Men and women each have different roles to play in society.
If **I were** a soldier **and** disagreed with **my** commanding officer's orders, **I would** obey anyway because that is **my** duty.	If a soldier disagreed with **his** commanding officer's orders, **he should** obey anyway because that is **his** duty.
Loyalty	
I am proud of **my** country's history.	**Citizens should be** proud of **their** country's history.
People should be loyal to their family members, even when they have done something wrong.	People should be loyal to their family members, even when they have done something wrong.
It is more important to be a team player than to	It is more important to be a team player than to

express oneself.	express oneself.
Economic Liberty	
People who are successful in business have a right to enjoy their wealth as they see fit.	People who are successful in business have a right to enjoy their wealth as they see fit.
Society works best when it lets individuals take responsibility for their own lives without telling them what to do.	Society works best when it lets individuals take responsibility for their own lives without telling them what to do.
The government interferes far too much in our everyday lives.	The government interferes far too much in our everyday lives.
Property owners should be allowed to develop their land or build their homes in any way they choose, as long as they don't endanger their neighbors.	Property owners should be allowed to develop their land or build their homes in any way they choose, as long as they don't endanger their neighbors.
Lifestyle Liberty	
I think everyone should be free to do as they choose, so long as they don't infringe upon the equal freedom of others.	Everyone should be free to do as they choose, so long as they don't infringe upon the equal freedom of others.
People should be free to decide what group norms or traditions they themselves want to follow.	People should be free to decide what group norms or traditions they themselves want to follow.

Appendix B. List of covariates with descriptions.

Question name	Source	Description
Word count	Coh Metrix 3.0	Number of words in statement.
Flesch reading ease	Coh Metrix 3.0	Measures reading difficult through the average sentence length and number of syllables per word. Higher scores indicate more difficulty.
Anaphor reference	Coh Metrix 3.0	Measures the number of times a single idea is referenced by counting the use of anaphors (e.g. pronouns: he, she, it; ellipsis markers: did, was).
Intentional verb incidence	Coh Metrix 3.0	Measures intentional information by counting verbs categorized as intentional by Wordnet ratings (Fellbaum, 1998; Miller et al., 1990).
Causal verb incidence	Coh Metrix 3.0	Measures causal information by counting verbs categorized as causal by WordNet ratings.
Causal verb ratio	Coh Metrix 3.0	Measures the cohesion of causal events to actors through the ratio of causal particles (e.g. because, if) to causal verbs. Higher scores indicate increased cohesion and easier readability.
Noun concreteness	Coh Metrix 3.0	Measures concreteness of content words (e.g. chair is high in concreteness, democracy is low) using the mean concreteness ratings of content words, taken from human ratings in the MRC Psycholinguistics Database (Coltheart, 1981).
Noun familiarity	Coh Metrix 3.0	Measures the familiarity of content words using the mean familiarity ratings of all content words, taken from human ratings in the MRC Psycholinguistic Database.
Noun imageability	Coh Metrix 3.0	Measures the imageability of content words using the mean familiarity ratings of all content words, taken from human ratings in the MRC Psycholinguistic Database.
Negation density	Coh Metrix 3.0	Provides a measure of syntactic complexity (i.e. working memory load) through the count of negative expressions in the text (e.g. not, un-).
Number of modifiers	Coh Metrix 3.0	Provides a measure of syntactic complexity (i.e. working memory load) through the mean number of modifiers per noun phrase.
Left embeddedness	Coh Metrix 3.0	Provides a measure of syntactic complexity (i.e. working memory load) through the mean number of words before the main verb in a sentence.
Agreement	Scores taken from Study 1 N = 68	"To what extent do you agree / disagree with this statement?" (1-7; "strongly disagree"-"strongly agree").
Valence	Scores taken from	Valence was the difference between unipolar positive

	online norming study. N = 17	and negative ratings (Kron et al., 2013), described below: *Instructions:* "Please rate your feelings regarding this statement using the following two scales. An extreme unpleasant rating means you feel completely unpleasant, unhappy, annoyed, unsatisfied, melancholic, or despaired. An extreme pleasant rating means you feel completely pleased, happy, satisfied, content or hopeful." *Ratings*: Negative valence (1-8; "no unpleasant feelings"-"strong unpleasant feelings") and positive valence (1-8; "no pleasant feelings"-"strong pleasant feelings").
Arousal	Scores taken from online norming study. N = 17	Arousal was the sum of unipolar positive and negative ratings, described above. Recent work has demonstrated that summed unipolar valence ratings are highly correlated with physiological measures of arousal, and may be superior to separately measuring arousal (Kron et al., 2013).
Mental imagery	Scores taken from online norming study. N = 20	"To what extent did you picture or imagine what the statements described as you read?" (1-7; "very little"-"very much"; Dodell-Feder et al., 2011).
Mental state	Scores taken from online norming study. N = 18	"To what extent did this statement make you think about someone's experiences, thoughts, beliefs and/or desires?" (1-7; "very little"-"very much"; Dodell-Feder et al., 2011).
Person present	Scores taken from online norming study. N = 20	"Does this statement mention people or a person?" ("Yes" / "No").
Reaction time	In-scanner N = 20	The time from the appearance of the in-scanner agreement rating prompt to the input of a response by the participant.

Coh Metrix ratings are calculated using an online tool at http://cohmetrix.com (Graesser et al., 2004; McNamara et al., 2014). In online samples, participants who did not correctly answer a catch question (asking them to describe any of the 72 statements they had read) were excluded from analysis. This caused some variability in N across covariates.