Morality as a Scaffold for Social Prediction

Author: Jordan Eugene Theriault

MORALITY AS A SCAFFOLD FOR SOCIAL PREDICTION

JORDAN E. THERIAULT

A dissertation

submitted to the Faculty of

the department of Psychology

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Boston College
Morrissey College of Arts and Sciences
Graduate School

July, 2017

MORALITY AS A SCAFFOLD FOR SOCIAL PREDICTION

Jordan E. Theriault

Advisor: Liane L. Young, Ph.D

Theory of mind refers to the process of representing others' mental states. This process consistently elicits activity in a network of brain regions: the theory of mind network (ToMN). Typically, theory of mind has been understood in terms of content, i.e. representing the semantic content of someone's beliefs. However, recent work has proposed that ToMN activity could be better understood in the context of social prediction; or, more specifically, prediction error—the difference between observed and predicted information. Social predictions can be represented in multiple forms—e.g. dispositional predictions about who a person is, prescriptive norms about what people should do, and descriptive norms about what people frequently do. **Part 1** examined the relationship between social prediction error and ToMN activity, finding that the activity in the ToMN was related to both dispositional, and prescriptive predictions. **Part 2** examined the semantic content represented by moral claims. Prior work has suggested that morals are generally represented and understood as objective, i.e. akin to facts. Instead, we found that moral claims are represented as far more social than prior work had anticipated, eliciting a great deal of activity across the ToMN. **Part 3** examined the relationship between ToMN activity and metaethical status, i.e. the extent that morals were perceived as objective or subjective. Objective moral claims elicited less ToMN activity, whereas subjective moral claimed elicited more. We argue that this relationship is best understood in the context of prediction, where objective moral claims represent strong social priors about what most people will believe. Finally, I expand on this finding and argue that a theoretical approach incorporating social prediction has serious implications for morality, or more specifically, for the motivations underlying normative compliance. People may be compelled to observe moral rules because doing so maintains a predictable social environment.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

First of all, I want to thank Amelia Brown for feeding the cat and for keeping me sane as I finished this final writing binge. I want to thank Liane Young for all of the advice and guidance along the way, and for (occasionally) challenging me to push beyond the limits of what I thought was possible. I want to thank Elizabeth Kensinger, Sean MacEvoy, and Jim Russell for their feedback and for being willing to read this entire dissertation. I want to thank Lisa Feldman Barrett for welcoming me into her lab, and for being willing to speak with a graduate student that she had never met on the eve of a book tour. Thank you Alek Chakroff, James Dungan, Brendan Gaesser, Larisa Heiphetz, Laura Niemi, Josh Rottman, Lily Tsoi, and Emily Wasserman for making the lab an exciting place to work, relax, and field test ideas. I would also like to thank Hiram Brownell, Andrea Herberlein, and Barry Schneider for giving candid advice and mentorship. Thank you, Nadine Weidman, for teaching my favorite course at Boston College, and for giving me a radically different perspective on how to think about psychology and my work. Finally, I would like to thank my parents for their love and support, and for tolerating my long-term absence from home.

Introduction

Theory of mind refers to the ability to represent the mental states of others (Premack & Woodruff, 1978). Their thoughts, beliefs, desires, and intentions cannot be directly observed, yet, to navigate the social world, it is critical that they be understood. To solve this problem, it has been argued that we make an inference: we theorize that people's behaviors are driven by internal mental states (Baron-Cohen, 2001; Gopnik, 2005; Onishi & Baillargeon, 2005; Saxe, 2009), e.g. if Jane left her shoes under the couch and they were later moved, an observer may infer that Jane has the belief: "My shoes are under the couch". Theory of mind almost certainly subsumes a number of sub processes (Schaafsma et al., 2014), and recent theoretical work has argued that neural signatures of theory of mind may represent *social prediction error*, i.e. the discrepancy between expected and observed social behavior (Joiner et al., 2017; Koster-Hale & Saxe, 2013). In this dissertation, I will present evidence for this claim, then argue that this theoretical perspective may have particularly powerful implications for our understanding of morality.

Social Prediction and Theory of Mind

Considering theory of mind in the context of social prediction does not change the fundamental problem. The goal remains to infer, based on observable evidence, an underlying causal structure that is responsible for an agent's behavior (Koster-Hale & Saxe, 2013). A predictive approach contextualizes this goal within a broader framework; e.g. social prediction is a generalized, abstract case of the fundamental problem that the brain evolved to solve: predicting all incoming sensory information (Clark, 2013; Friston, 2010). Models using this theoretical approach are referred to as *predictive coding models*, and are based around the assumption that new sensory information is only encoded in terms of prediction error—the difference between observed and predicted input. Prediction error runs on the same logic as image compression (as in .gif, or .jpeg files; Figure 1): most information is redundant, and it would be inefficient to represent each pixel of an image, so instead, images can be compressed by encoding differences. (Also, on this line of thought, points of change would be a natural fundamental unit for encoding information, e.g. edges and boundaries; Rao & Ballard, 1999). Predictive coding models argue that the brain implements a hierarchy of these predictions, each level of which is increasingly abstracted from the sensory input (Clark, 2013). The proposal then, is that social predictions underlying theory of mind are driven by the discrepancy between observed and predicted information.

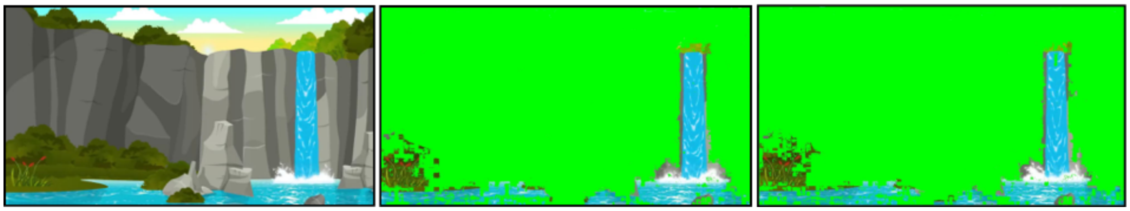


Figure 1. Example of .gif compression. Three frames of an animation are depicted. The waterfall flows, and the flowers in the bottom left corner sway in the wind, but most pixels in the image remain stationary. The image can be compressed by encoding repeated pixels as green space, meaning that only pixels that differ from their prior values need to be updated. In other words, in each frame, pixels only encode prediction error—the difference between the predicted and observed value for that pixel.

Neural Signatures of Theory of Mind

One of the most consistent findings in social neuroscience is that a variety of social tasks elicit activity in the same few brain regions (e.g. Blakemore et al., 2003; Kircher et al., 2009; Ma et al., 2012; Saxe & Kanwisher, 2003; see Schurz et al., 2014 for review). These regions include medial prefrontal cortex, precuneus (or posterior cingulate cortex), and bilateral temporoparietal junction, and have been collectively referred to as the "theory of mind network" (ToMN; although the same network has gone by many names in other lines of research; Barrett, 2017).

It has recently been suggested that ToMN activity could be interpreted in terms of social prediction error, i.e. social tasks elicit ToMN activity because they license (and update, via prediction error) social predictions (Koster-Hale et al., 2013). **Part 1** examines this connection between ToMN activity and social prediction error directly, using a detailed set of moral narratives and examining multiple forms of social prediction—e.g. *dispositional predictions*, about who a person is (Gilbert & Malone, 1995), *prescriptive norms*, about what a person should do (Cialdini et al., 1990), and *descriptive norms*, about what people frequently do (Cialdini et al., 1990). ToMN activity was related to both dispositional and prescriptive prediction error, but not to descriptive prediction error. That is, across a number of stories, ToMN is most active when dispositional and normative social predictions are violated.

The Content of Moral Claims

Putting aside prediction for a moment, there is an ongoing philosophical debate regarding what semantic content is expressed in moral claims. This debate refers to the *metaethical* status of morality: Regardless of whether you agree or disagree with a given claim (e.g. "eating meat is wrong"), what semantic information has the claim expressed? Facts generally express *objective* information about the nature of the world. Preferences generally express *subjective* information about a particular person's beliefs. Where morals fall with respect to those two extremes is unresolved.

Prior work, in both philosophy and developmental psychology has generally favored a moral objectivist position. For instance, philosophers have argued that most non-philosophers believe 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). In developmental psychology, researchers have demonstrated that children distinguish *moral* and *conventional* violations (Turiel, 1978; Wainryb et al., 2004). Moral violations (e.g. hitting another child) are universally wrong, and cannot be made right by the endorsement of an authority (e.g. a teacher says that it is ok). By contrast, conventional violations (e.g. wearing pajamas to school) are only locally wrong, and can be licensed with permission (e.g. a teacher declares pajama day).

A great deal of this prior work, however, suffers from serious methodological flaws. For instance, forcing participants to classify morals as true, false, or an opinion/preference (Goodwin & Darley, 2008), could force participants to make classifications that have no analogue in the brain. Even more seriously, prior work has made statistical distinctions between categories (e.g. morals, vs. facts, vs. preferences), using methods that cannot generalize beyond the *exact* stimuli that were used. In one well-cited example, facts, morals, and preferences were each represented by two examples, and conclusions were generalized to the entire moral domain (Wainryb et al., 2004). **Part 2** examines the metaethical status of morality, while attempting to address these problems: We used a large sample of stimuli, combined with mixed effects analyses (which can generalize beyond samples of both subjects and stimuli) and discovered that behaviorally and neurally, moral claims are represented as more similar to social preferences than to objective facts. Both morals and preferences elicited activity in the medial prefrontal cortex. Surprisingly, however, moral claims actually elicited greater

ToMN activity than preferences, a control category that had been designed as a benchmark of social content.

Metaethical Variance and Moral Predictions

That moral claims elicited widespread activity in the ToMN (**Part 2)** could be interpreted to mean that the moral domain is distinguished by strong underlying representations of social content. However, if ToMN activity is associated with social prediction error (as the findings in **Part 1** suggest), then a second possibility is raised: moral claims may be related to social prediction. **Part 3** examines variance in ToMN activity in detail, providing evidence consistent with this account.

Recent work has established that non-philosophers do not hold a unified metaethical stance (e.g. Beebe, 2014; Goodwin & Darley, 2012; Wright et al., 2013)—people see some moral claims as more objective (e.g. "slavery is wrong") and others as more subjective (e.g. "eating meat is wrong"). We observed that ToMN activity was correlated with this metaethical variance. Moral claims that were perceived as subjective elicited greater ToMN activity, and moral claims that were perceived as objective elicited less. Exploratory analyses also suggested that this relationship was not driven by underlying differences in mental state representation.

If subjective moral claims elicit greater ToMN activity, then does the ToMN represent moral subjectivity? This seems unlikely. Even accepting that mental functions can be localized to specific brain regions (and this is controversial; Barrett, 2017; Uttal, 2001), classifying brain regions as *for* metaethical judgment seems like it would be going too far. It has been argued that there are no intrinsically "moral" brain regions (Young &

Dungan, 2012); extending this argument, brain regions for metaethical judgment only seem even more unlikely.

By contrast: Are objective moral claims predictable? I would propose that they are, and that social prediction may underlie the observed relationship between ToMN activity and metaethical judgment. Objective moral claims (as opposed to more subjective morals) might be thought of as strong social priors—default predictions about what other people believe. For instance, without having to ask a stranger, you might predict that they hold certain moral beliefs (e.g. that drinking and driving is bad, that slavery is wrong, etc.). If ToMN activity is associated with social prediction error, then there would be a relatively small discrepancy between predicted and observed information for objective moral claims, explaining the decreased ToMN activity that we observed.

Outline

In three parts, I present evidence that morality is related to social prediction. In **Part 1**, we demonstrate that activity in the theory of mind network (ToMN) is related to prediction error in the context of both dispositional judgments and prescriptive norms. In **Part 2**, we demonstrate that moral claims are characterized by their social relevance, rather than their objectivity; and that simple statements about morality elicit widespread activation across the ToMN. In **Part 3**, we demonstrate that, for moral claims, this ToMN activity is related to metaethical judgment: subjective claims elicit greater ToMN activity, and objective claims elicit less. Based on this, I argue that reconsidering morality in terms of prediction may help to explain why we feel compelled to conform to moral norms.

Part 1

Social Prediction in the Theory of Mind Network

Jordan Theriault[1] & Liane Young[1]

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

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

Abstract

Social tasks reliably elicit activity in the theory of mind network (ToMN), but it has remained unclear how this activity should be interpreted. Recently, it has been suggested that the ToMN may encode prediction error in social contexts. Classic work in social psychology has identified a number of sources that could license social predictions, such as dispositional information, prescriptive norms, and descriptive norms. The present work examined the by-stimuli relationship between these sources of social prediction and ToMN activity. In Study 1, we tested a stimuli set consisting of detailed moral narratives, where an agent made a moral (or immoral) decision, that was later reframed, inducing participants to change their initial moral judgments. These scenarios were normed using an online sample (N = 554), and the dimensional structure underlying these normative ratings was obtained using a principal components analysis. Within our stimuli set, dispositional, prescriptive, and descriptive predictions were distinguished sources of information. In Study 2, we examined the relationship between ToMN activity and dispositional, prescriptive, and descriptive prediction error. Participants (N= 20) read the same moral scenarios inside the scanner, providing their moral judgment. We compared by-stimuli estimates extracted from Study 1 with by-stimuli estimates of neural activity in Study 2, extracted from each ToMN ROI. Across the ToMN, ROI activity was related to descriptive and prescriptive prediction error, but not to descriptive prediction error. This relationship was specific to scenarios where new, morally relevant information was presented (as opposed to morally irrelevant, control information). Thus, the present work provides evidence that the ToMN encodes prediction error in social contexts.

Introduction

The theory of mind network (ToMN) is a network of brain regions that has been consistently implicated in tasks that involve representing mental states (e.g. belief, intentions; Ciaramidaro et al., 2007; Dodell-Feder et al., 2011; Fletcher et al., 1995; Gallagher et al., 2000; Gobbini et al., 2007; Ruby & Decety, 2003; Saxe & Kanwisher 2003; Saxe & Powell, 2006; Vogeley et al., 2001; Young et al., 2007; 2010) and social information more generally (Amodio & Frith, 2006; Harris et al., 2005; Jenkins & Mitchell, 2010; Ma et al., 2012; Mitchell et al., 2005; Yeshurun et al., 2017; for review, see Schurz et al., 2014; Van Overwalle, 2009). Although these regions are reliably active during social tasks, less is known about the computations implemented within them. Recently, it has been proposed that the ToMN may represent prediction error in social contexts (Koster-Hale & Saxe, 2013; also see Joiner et al., 2017). The present work used an item analysis to test this proposal within a set of detailed moral scenarios, in which participants formed and then updated moral judgments. This item analysis allowed us use contextually rich stories to interrogate processes underlying ToMN activity at a finer grain than traditional fMRI contrasts (Westfall et al., 2016), asking whether ToMN activity tracks by-stimuli differences in social prediction error.

Theory of mind refers to the ability to represent internal mental states (Premack & Woodruff, 1978). The ToMN is active for a variety of social tasks (Schurz et al., 2014), such as reading about false beliefs (e.g. Dodell-Feder et al., 2011; Saxe & Kanswisher, 2003), watching social animations (e.g. Blakemore et al., 2003), making strategic decisions in economic games (e.g. Kircher et al., 2009), and impression formation (Baron et al., 2011; Bhanji & Beer, 2013; Cloutier et al., 2011; Ma et al., 2012;

Mende-Siedlecki & Todorov, 2016; Mende-Siedlecki et al., 2013; Schiller et al., 2009). Furthermore, the ToMN overlaps (at least partially; Mars et al., 2012, also see Schurz et al., 2017) with the default mode network, a network of regions active during rest that are thought to be critical for generating an internal model of the world (Barrett, 2016; Buckner, 2012; Hassabis & Maguire, 2010). That such diverse tasks elicit such broad overlapping activation raises the possibility that ToMN activity represents some more general underlying process (or collection of processes)—albeit a process tightly linked with social cognition. Consistent with this, some researchers have argued that theory of mind is most likely not a singular process (Schaafsma et al., 2014; Heyes, 2014): for instance, to represent a mind, one would presumably need to (at the very least) distinguish oneself from others, track goals and intentions, and understand causality (Schaafsma et al., 2014). Others have called attention to the functional overlap in critical regions in the ToMN, e.g. the temporoparietal junction (TPJ) is well positioned for a high-level integrative role, sitting at the nexus of regions implicated in memory, attention, social cognition, and language (Carter & Huettel, 2013). However, the dilemma remains: ToMN is strongly associated with social processing, but it is unclear how best to interpret ToMN activity.

A promising hypothesis comes from recent theoretical work connecting theory of mind with a predictive coding framework (Koster-Hale & Saxe, 2013). Predictive coding models aim at offering a unifying framework of neural computation, proposing that the brain is a "hierarchical prediction machine" (A. Clark, 2013), attempting to match internally generated predictions with incoming sensory information (Barrett, 2017; Friston, 2010). This framework offers a radically different approach than classic representational theories (e.g. Fodor, 1983), where dedicated mechanisms transform sensory information for use in some central cognitive space. Instead, predictive coding models propose that the brain actively predicts sensory information, and updates these predictions on the basis of "prediction error"—the difference between observed and predicted input. In this framework, social predictions could be characterized as high-level, abstracted predictions about incoming sensory information. Koster-Hale & Saxe (2013) proposed that the ToMN may represent such high-level social predictions. Prior research could be interpreted as consistent with this account (e.g. Dungan et al., in press; Mende-Siedlecki et al., 2013); however, to our knowledge, the relationship between ToMN activity and social prediction error has not been directly examined. In the present work, we examine this relationship and attempt to go further, borrowing distinctions from social psychology to examine multiple forms of social prediction.

Social predictions can take several forms. For instance, classic work in social psychology has demonstrated that behavior can be attributed to dispositional or situational sources (Gilbert & Malone, 1995)—e.g. "Dave was short with the waiter because Dave is a jerk" vs. "Dave was short with the waiter because the parking meter was running out". Furthermore, within a situation, we can make predictions on the basis on social norms, i.e. implicit expectations about how people will behave. Social norms are not singular, and prior work has typically made a distinction between prescriptive and descriptive norms (Brauer & Chaurand, 2010; Cialdini et al., 1990). Prescriptive norms refer to expectations based on moral or social values. Descriptive norms refer to expectations based on statistical frequency. For instance, descriptively most people will drive a short distance rather than walk; however, prescriptively this is frowned on (i.e.

people often do it, but they shouldn't; Brauer & Chaurand, 2010). Note that the concept of a prescriptive norm is similar to a moral judgment, but is framed in terms of expectation rather than rules, a feature we return to in the general discussion. Recent work has also considered these classic effects in a predictive framework (Bach & Schenke, 2017). In sum, information about people (i.e. dispositional information) and about descriptive/prescriptive norms could be used to generate social priors. These sources of social prediction were examined in the present work.

Present Work

The present work tested the relationship between multiple forms of social prediction and ToMN activity. Allowing multiple forms of social prediction to coexist in a single experimental design is not trivial; for instance, impression formation has been studied by presenting discrete, contradictory pieces of information about individuals (e.g. Mende-Siedlecki et al., 2013), which targets dispositional information, but omits situational social norms. To address this problem, we designed a series of detailed narratives, each of which elicited an initial moral judgment, and subsequently induced participants to update that judgment. Each scenario described an agent facing a moral dilemma: it described the background and potential outcomes (e.g. a hospital administrator must choose to save one sick child, or create a larger immunization program); it presented the agent's decision (e.g. the administrator creates the immunization program); and then it reframed the dilemma with additional information (e.g. the hospital board has promoted past administrators who began new programs). The additional information reframed the scenario, rather than offering a direct contradiction. Scenarios were also intentionally wide-ranging, eliciting initial and reframed moral judgments of varying strengths. The intention was to create variability, so that by-stimuli differences in dispositional, prescriptive, and descriptive prediction error could be examined.

In Study 1 we characterized the stimuli by collecting online ratings of each scenario on a number of measures and then performing a principal components analysis to identify the underlying dimensions. Within our stimuli set, dispositional, prescriptive, and descriptive sources of social prediction were clearly distinguished. In Study 2, we examined the relationship between Study 1 by-stimuli component scores, and by-stimuli ToMN activity. We used mixed effects analyses in both studies to model by-subject and by-stimuli variance, allowing us to extract by-stimuli estimates in each study (best linear unbiased predictors; BLUPs; Baayen, 2008; Baayen et al., 2008). Across ROIs (regions of interest), ToMN activity was associated with dispositional and prescriptive prediction error.

Study 1

In Study 1 we collected behavioral ratings of all stimuli and examined the relationship among them, performing a principal components analysis to identify the underlying dimensions. Questions were selected to test the extent that reframing information violated predictions about the agent, or the extent that the agent's decision violated prescriptive or descriptive norms (Table 1). Questions also measured more general stimuli features used in prior item analyses of the ToMN (mental state inferences, mental imagery; Dodell-Feder et al., 2011), in addition to valence, arousal and moral judgment.

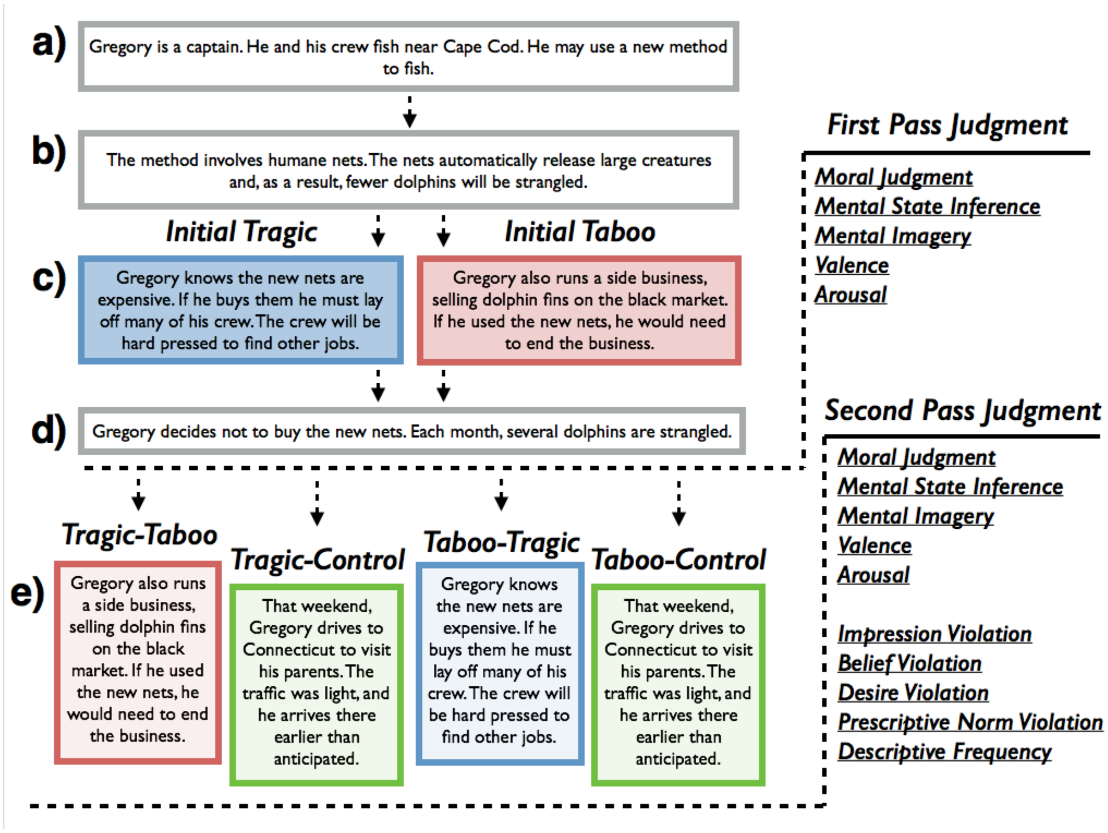
Method

Participants. Participants were recruited online using Amazon Mechanical Turk (AMT) at an approximate rate of $6/hour, in line with standard AMT compensation rates. Participants were recruited in two cohorts. Cohort one consisted of 239 adults (132 female, 1 unspecified; M_{Age} = 36.4 years, SD_{Age} = 12.1 years), after excluding four participants for failing a simple attention check, asking them to briefly describe any of the scenarios they had read. Cohort two consisted of 315 adults (140 female, 1 unspecified; M_{Age} = 33.6 years, SD_{Age} = 9.4 years), after excluding seven participants for failing the same attention check. The Boston College Institutional Review Board approved Studies 1 and 2, and each participant provided consent before beginning.

Stimuli. Stimuli consisted of twenty-four scenarios, adapted or inspired from prior work (Critcher et al., 2012; Lichtenstein et al., 2007; Tetlock et al., 2000; Uhlmann et al., 2013; Appendix A). Scenarios described *tragic* and *taboo* dilemmas. Tragic dilemmas forced agents to choose between two moral outcomes, whereas taboo dilemmas forced agents to choose between a moral and selfish outcome. For instance, in a tragic dilemma, Gregory, a fisherman, could save the jobs of his crew, but at the expense of killing more dolphins; in a taboo dilemma, he could save the dolphins, but at a personal cost (Figure 1; Appendix A).

First, scenarios were described as tragic or taboo, then later, they were reframed. For instance, in a *tragic–taboo* scenario, at first Gregory must choose between saving dolphins or saving his crew's jobs. He chooses to save his crew. Next, participants learn that this decision sustained Gregory's side-business: selling black-market dolphin fins. In the *taboo–tragic* version of this scenario, at first Gregory must choose between his side-business and saving dolphins. He chooses his side business. Next, participants learn that this decision saved Gregory's crew. Thus, participants ultimately read the same content for tragic–taboo and taboo–tragic scenarios, only presented in different orders. These reframing scenarios were intermixed with control scenarios, which appended morally irrelevant information (Figure 1).

For each scenario, participants provided either one or two judgments. *First pass judgments* were made after the initial dilemma (Figure 1d), and *second pass judgments* were made during the reframed dilemma (Figure 1e). In cohort one, all segments were presented sequentially, whereas in cohort two segments a–d were presented simultaneously (along with first pass judgments), followed by segment e. This change was made to decrease page loading times.


a) Gregory is a captain. He and his crew fish near Cape Cod. He may use a new method to fish.

b) The method involves humane nets. The nets automatically release large creatures and, as a result, fewer dolphins will be strangled.

c) Initial Tragic: Gregory knows the new nets are expensive. If he buys them he must lay off many of his crew. The crew will be hard pressed to find other jobs.

Initial Taboo: Gregory also runs a side business, selling dolphin fins on the black market. If he used the new nets, he would need to end the business.

d) Gregory decides not to buy the new nets. Each month, several dolphins are strangled.

e) Tragic-Taboo: Gregory also runs a side business, selling dolphin fins on the black market. If he used the new nets, he would need to end the business.

Tragic-Control: That weekend, Gregory drives to Connecticut to visit his parents. The traffic was light, and he arrives there earlier than anticipated.

Taboo-Tragic: Gregory knows the new nets are expensive. If he buys them he must lay off many of his crew. The crew will be hard pressed to find other jobs.

Taboo-Control: That weekend, Gregory drives to Connecticut to visit his parents. The traffic was light, and he arrives there earlier than anticipated.

First Pass Judgment: Moral Judgment, Mental State Inference, Mental Imagery, Valence, Arousal

Second Pass Judgment: Moral Judgment, Mental State Inference, Mental Imagery, Valence, Arousal, Impression Violation, Belief Violation, Desire Violation, Prescriptive Norm Violation, Descriptive Frequency

Figure 1. Example scenario. Scenarios consisted of segments which could be substituted or rearranged to form four conditions: Tragic–Taboo, Tragic–Control, Taboo–Tragic, and Taboo–Control. For each scenario, participants made either both first and second pass

judgments, or only a second pass judgment. The text above is abbreviated. 24 scenarios were used in total; for full text see Appendix A.

Procedure. Each participant read 24 scenarios (6 tragic–taboo; 6 taboo–tragic; 6 tragic–control; 6 taboo–control), presented in a semi-random order to counterbalance condition–scenario combinations across participants. Ten measures were collected in total. In cohort one, participants provided first and second pass judgments of either *mental state inference*, *mental imagery*, or *valence* and *arousal* (Kron et al., 2013). In cohort two, all participants provided first and second pass moral judgments, and subgroups of participants provided second pass judgments of either *impression violation, belief violation, desire violation, prescriptive norm violation*, or *descriptive frequency* (Figure 1; Table 1). Measures for prescriptive and descriptive norms were adapted from Brauer & Chaurand (2010).

Analysis. We used mixed effects analyses to model each measure. Critically, these models included crossed by-subject and by-stimuli random effects. In traditional models (e.g. ANOVA) these two sources of variance cannot be modeled simultaneously, meaning that we would be forced to average across stimuli (or across participants), and limit our conclusions to the *exact* stimuli (or participants) that were tested (Baayen et al., 2008; H. Clark, 1973; Judd et al, 2012; Westfall et al., 2016). Another reason for this analysis, is that it allowed us to easily extract BLUPs (best linear unbiased predictors; Baayen, 2008; Baayen et al., 2008). BLUPs were by-stimuli estimates of behavioral ratings for each scenario and condition. BLUPs were preferable to simple by-stimuli averages for two reasons: a) by-stimuli BLUPs are independent from by-subject variance, meaning that estimates were specific to the scenarios (and could be compared with by-stimuli estimates of ToMN activity in Study 2); and b) by-stimuli BLUPs incorporate the sample distribution into the estimate (i.e. they are semi-pooled estimates; Gelman, Hill, & Yajima, 2012), meaning that they anticipate regression to the mean and mitigate against outliers.

Data was modeled using R (R Core Team, 2016) and the lme4 package (Bates et al., 2015). All mixed effects models used the maximal random effects structure, including all necessary by-subject and by-stimuli random slopes and intercepts (Barr et al., 2013). For measures where first-pass and second-pass judgments were collected (moral judgment, mental state inference, mental imagery, valence, and arousal), models specified fixed, by-subject, and by-stimuli random effects as: A + B + AxB + AxC + AxBxC, where A = *Time Point* (first pass, 0/second pass, 1), B = *Initial Condition* (tragic, -0.5/taboo, +0.5), C = *Reframing Condition*, (control, -0.5/reframed, +0.5); the main effect for Reframing Condition was omitted because the distinction between control and reframed conditions was only relevant to second pass judgments. For measures where only second-pass judgments were collected (impression violation, belief violation, desire violation, prescriptive norm violation, and descriptive norm violation), models specified fixed, by-subject, and by-stimuli random effects as: B + C + BxC, where B = *Initial Condition* (tragic, -0.5/taboo, +0.5), C = *Reframing Condition*, (control, -0.5/reframed, +0.5).

Any model simplifications to achieve convergence are described in Results below, and convergence was achieved before testing any relationships of interest. Models were simplified by temporarily removing correlations between random effects, inspecting the uncorrelated model and removing any random effects parameters with zero variance, then returning correlations to the model.

Results

Modeling. Across cohorts, we collected ten measures: a) impression violation, b) belief violation, c) desire violation, d) prescriptive norm violation, e) descriptive norm violation, f) moral judgment, g) mental state inference, h) mental imagery, i) valence, and j) arousal. Model simplification was necessary only for moral judgment, where the by-subject Time Point x Initial Condition parameter was dropped, meaning that by-subject random effects could not be separately estimated for tragic–control and taboo–control conditions. Condition mean estimates for each measure are reported in Table 1. For each model, we extracted by-stimuli BLUPs for each scenario and condition combination, providing 96 data points for each measure.

Table 1. Behavioral rating condition means.

		Initial Tragic	**Initial Taboo**	**Tragic–Taboo**	**Tragic–Control**	**Taboo–Tragic**	**Taboo–Control**
Measure	**Text**	***M (SE)***	***M (SE)***	***M (SE)***	***M (SE)***	***M (SE)***	***M (SE)***
Mental State Inference (n = 80)	"To what extent did this story make you think about someone's experiences, thoughts, beliefs, and/or desires?" 1–7; "not at all" – "very much"	5.48 (0.12)	5.40 (0.11)	5.04 (0.15)	4.54 (0.19)	5.24 (0.14)	4.75 (0.18)
Mental Imagery (n = 80)	"To what extent did you picture or imagine the events of the story happening as you read?	5.38 (0.12)	5.33 (0.12)	5.38 (0.12)	3.77 (0.19)	5.41 (0.13)	3.78 (0.19)
Valence (n = 79)	*Difference of positive and negative unipolar scales.* "Please rate your feelings regarding this statement on two scales:" 1–8; "no unpleasant feelings – "strong unpleasant feelings" 1–8; "no unpleasant feelings – "strong unpleasant feelings"	-1.81 (0.41)	-3.60 (0.40)	-2.90 (0.46)	-1.58 (0.40)	-1.27 (0.44)	-3.01 (0.39)
Arousal (n = 79)	*Sum of positive and negative unipolar scales.*	8.18 (0.17)	8.38 (0.13)	8.41 (0.16)	7.87 (0.21)	8.37 (0.18)	8.11 (0.17)
Moral Judgment (n = 315)	"Are **<agent>**'s actions moral?" 1–7; "not at all" – "completely"	4.08 (0.20)	2.62 (0.21)	3.30 (0.19)	4.12 (0.20)	4.01 (0.18)	2.79 (0.21)
Impression Violation (n = 63)	"Is this new information **inconsistent** with your previous impression of <agent>?"	n/a	n/a	3.80 (0.20)	2.13 (0.15)	4.12 (0.20)	2.15 (0.15)
Belief Violation (n = 61)	"Is this new information **inconsistent** with what you previously thought <agent> believed?"	n/a	n/a	4.08 (0.21)	2.21 (0.19)	4.45 (0.19)	1.99 (0.17)
Desire Violation (n = 67)	"Is this new information **inconsistent** with what you previously thought <agent> desired?"	n/a	n/a	3.82 (0.19)	2.36 (0.19)	3.92 (0.21)	2.33 (0.20)
Prescriptive Norm Violation (n = 62)	"**With this new information in mind,** to what extent is <agent>'s decision deviant (i.e. to what extent does it go against the	n/a	n/a	4.24 (0.21)	3.07 (0.21)	3.21 (0.14)	4.09 (0.25)

	norms of our society)?”						
Descriptive Frequency (n = 62)	“**With this new information in mind,** to what extent is <agent>'s decision common (i.e. to what extent is it frequently observed in our society)?”	n/a	n/a	4.15 (0.18)	4.22 (0.17)	4.17 (0.17)	4.12 (0.20)

Principal Components Analysis. A principal components analysis reduced the dimensionality of our measures. We tested 2-factor, 3-factor, 4-factor, 5-factor, and 6-factor varimax rotated solutions (Table 2 & Table S1 of the online supplemental materials). To fit all variables, we wanted to ensure that the communalities—i.e. the proportion of variance explained for each variable—were high in all cases. A 2-factor solution was a poor fit for descriptive frequency ($h2$ = .23). A 3-factor solution offered an improvement, but remained a poor fit for mental imagery ($h2$ = .51). A 4-factor solution was a reasonable fit for the data, with the worst fit being for arousal ($h2$ = .68); however, communalities were high across all variables in the 5-factor solution ($h2_{\mathrm{min}}$ = .89), and a 6-factor solution offered only modest improvement ($h2_{\mathrm{min}}$ = .92). Based on this, we decided to use the 5-factor solution. Factors were: 1) dispositional prediction error (impression violation, belief violation, desire violation, & mental state inference); 2) prescriptive prediction error (moral judgment, valence, prescriptive norm violation; factor loadings were reverse scored); 3) mental imagery; 4) descriptive prediction error (descriptive frequency, reverse scored); and 5) arousal. Factor loadings and fit for all solutions are reported in Table 2, and correlations among measures are visualized in Figure 2.

Table 2. Principal components analysis; 5-factor solution.

Measure	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Communality
Mental Imagery	0.41		0.90			0.99
Mental State Inference	0.82		0.44		0.21	0.91
Valence		0.94				0.89
Arousal	0.50	-0.27			0.80	1
Descriptive Frequency		0.20		0.97		0.99
Prescriptive Norm Violation		-0.87		-0.34		0.89
Belief Violation	0.95					0.97
Desire Violation	0.97					0.96
Impression Violation	0.96					0.96
Moral Judgment		0.96				0.94
Component loading	3.85	2.68	1.11	1.08	0.78	
Proportional variance explained	.39	.27	.11	.11	.08	
Cumulative variance explained	.39	.65	.76	.87	.95	

PCA was performed on by-stimuli BLUPs, extracted from Study 1 models for each measure. Factor loadings with an absolute value < .2 are omitted for ease of interpretation. Principal components were varimax rotated.

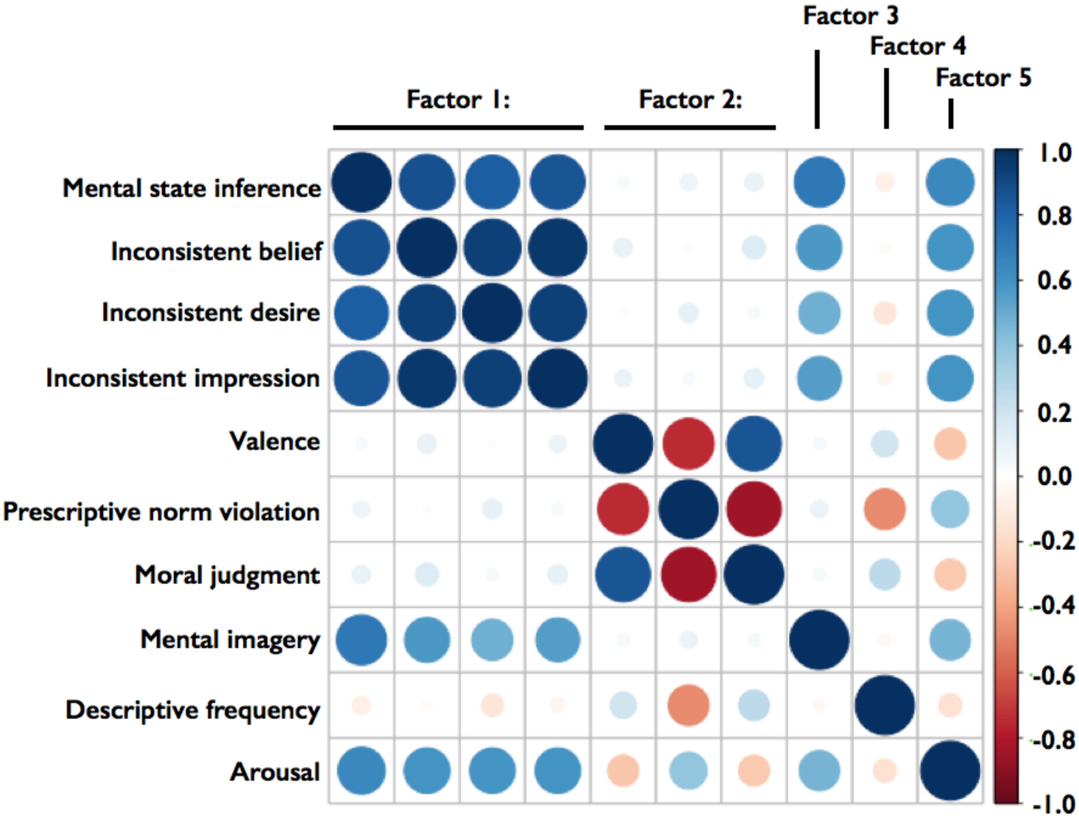


Figure 2. Correlations among by-stimuli BLUPs. Each measure was estimated in a separate model, and all measures were collected from separate participants (with the exception of moral judgment). Dimension reduction was achieved using principal components analysis with varimax rotation. The 5-factor solution fit all variables well, and was clearly interpretable. Factor 1 was interpreted as dispositional prediction error; factor 2 as prescriptive prediction error (reverse coded), factor 3 as mental imagery; factor 4 as descriptive prediction error (reverse coded); and factor 5 as arousal. Note that factors 3 and 5 also partially load on factor 1. Factors were extracted as weighed averages based on all loadings (Table 2).

Discussion

Study 1 identified dimensions underlying our stimuli set. Principal components analysis distinguished between dispositional, prescriptive, and descriptive prediction error (Figure 2; Table 2). Mental imagery and valence comprised additional dimensions. With by-stimuli measures of social prediction error established, we could compare ratings with ToMN activity, estimated in Study 2.

Study 2

Study 2 examined whether by-stimuli differences in social prediction error were related to BOLD activity in the ToMN. Participants read the scenarios analyzed in Study 1 while undergoing fMRI. The ToMN was identified using an independent functional localizer, and BOLD activity was extracted for five ROIs: dorsal-/ventral-medial prefrontal cortex, precuneus, and right/left temporoparietal junction (DMPFC/VMPFC/PC/RTPJ/LTPJ). By-stimuli estimates were extracted from each ROI and compared with Study 1 component scores (testing for overall effects of ToMN activity and interactions across ROIs).

Method

Participants. Our final sample consisted of 20-right handed participants (10 female, 9 male, 1 unspecified; M_{age} = 27.3 years, SD_{age} = 5.0 years), after excluding two participants from analysis due to excessive movement, identified during spatial preprocessing. The sample size was chosen in advance to be consistent with fMRI studies of social cognition (e.g., Fourie et al., 2014; Koster-Hale et al., 2013; Mende-Siedlecki et al., 2013; Ratner et al., 2012; Young & Saxe, 2009). Participants were a community sample, recruited through an online posting and given a $65 cash payment. Participants were native English speakers with no reported history of learning disabilities, previous psychiatric or neurological disorders, or a history of drug or alcohol abuse.

Stimuli and Measures. Stimuli were identical to those used in Study 1. Participants were told that they would read about a character and provide a moral judgment at the end of the story, so they should keep their judgment in mind as they read. At the end of each scenario, participants provided a single moral judgment ("How morally wrong?"; 1 – "Not at all", 4 – "Very"). We collected one second pass judgment (as opposed to both first and second pass judgments) to avoid interrupting participants as they read.

Procedure. Scenarios were projected onto a screen in the scanner, and appeared cumulatively in five parts (+10 s each, 50 s total). Moral judgments were probed on a separate screen (+4 s), followed by fixation (+12 s). Runs were 4.6 mins each, and the total scan time was 64.6 mins, due to the inclusion of a second study reported elsewhere (involving reading and evaluating claims about facts, morals, and preferences; Theriault et al., 2017a, 2017b). Stimuli were presented in white text on a black background, using Matlab 7.7.0 (R2008b), on an Apple Macbook Pro.

As in Study 1, each participant read 24 scenarios (6 tragic–taboo; 6 taboo–tragic; 6 tragic–control; 6 taboo–control), presented in a semi-random order to counterbalance condition–scenario combinations across participants.

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 the Center for Brain Science Neuroimaging Facility at Harvard University. 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), spatially smoothed using a Gaussian filter (full-width half-maximum = 5 mm kernel), and high-pass filtered (128 Hz).

ToMN Localizer Task. ToMN ROIs were identified using an independent functional localizer task (Dodell-Feder et al., 2011), contrasting ten stories about mental states (*false-belief*) and ten stories about physical representations (*false-photograph*). Stories were matched in complexity across conditions; see http://saxelab.mit.edu/superloc.php for the complete set. Each trial presented a story (10 s) and a statement about the story, which participants rated true or false (+4 s). A boxcar for the full duration was used to model BOLD (blood oxygen level dependent) activity. A simple contrast (false belief > false photograph; $p < .001$, $k > 10$) identified significant ToMN voxels for each participant.

ROIs were defined for DMPFC, VMPFC, PC, RTPJ, and LTPJ, and each comprised all significant voxels in a 9mm-radius sphere surrounding the peak voxel in each location (for coordinates, see Table S2 of the online supplemental materials). It is worth noting that the chosen threshold is lenient (Eklund et al., 2016); however, our principal aim was to compare our findings with prior work which has identified the ToMN (e.g. Dodell-Feder et al., 2011; Saxe & Kanwisher, 2003), which required that we apply the localizer using the same method. Note that ROIs were defined for each subject, based on the localizer contrast; if no voxels were active then no ROI was defined, meaning that Ns differ across ROIs (Table S2).

Functional ROI Analysis. Within each ToMN ROI, we transformed BOLD activity at each time point into percent signal change (PSC = raw BOLD magnitude for [condition – fixation]/fixation), offsetting the time course by 4 seconds to account for hemodynamic lag (Dodell-Feder et al., 2011). For each subject, PSC was mean-centered for each run. Across the entire sample, outliers within each ROI were removed using an iterated Grubbs test (Grubbs, 1969), at a threshold of $p < .001$. PSC was then averaged across three epochs: *background epoch* (4–24 s, before scenarios diverged into initial tragic and initial taboo; Figure 1); *first pass epoch* (24–44 s; after scenarios diverged into initial tragic and initial taboo); and *second pass epoch* (44–58 s; after scenarios were reframed and diverged into four conditions, including moral judgment). To simplify analysis, values for the background epoch were dropped from analysis, and we entered first and second pass epoch PSC for each subject and scenario into a mixed effects analysis.

Mixed Effects Analysis. As in Study 1, we used mixed effects analyses to build models with crossed by-subject and by-stimuli random effects (Baayen et al., 2008; Judd et al., 2012; Westfall et al., 2016). From these models, we extracted by-stimuli BLUPs, estimating activity for each scenario and condition, within each ToMN ROI. Data was modeled using R (R Core Team, 2016) and the lme4 package (Bates et al., 2015).

Mixed effects models were first modeled as maximal, including all necessary by-subject and by-stimuli random slopes and intercepts (Barr et al., 2013). All models were initially specified as: A + B + AxB + AxC + AxBxC, where A = *Epoch* (first pass epoch, 0/second pass epoch, 1), B = *Initial Condition* (tragic, -0.5/taboo, +0.5), C = *Reframing Condition*, (control, -0.5/reframed, +0.5); as in Study 1, the main effect for Reframing Condition was omitted because the distinction between control and reframed conditions was only relevant in the second pass epoch. Any model simplification to achieve convergence is described in Results below, and convergence was achieved before testing any relationships of interest. Steps for simplification were the same as described in Study 1.

Results

Modeling. PSC for each ToMN ROI was fit using mixed effects models (see Mixed Effects Analysis in Methods). No simplification was necessary for PC, and VMPFC, RTPJ, and LTPJ were simplified by removing by-stimuli random intercepts, which for our purposes meant that all scenarios in these ROIs varied by condition around a fixed point (see Barr et al., 2013). DMPFC was simplified further by removing by-stimuli random intercepts, and by-stimuli random slopes for Initial Condition and Initial Condition x Epoch, meaning that, within DMPFC, by-stimuli variance was removed in the first pass epoch, and reduced in the second pass epoch. It was promising that so little simplification was necessary, as the design called for the estimation of a fairly large number of random effects. That convergence could be achieved suggests that our sample size met at least the minimum requirements for estimation (granted, more accurate estimates will always be produced in a larger sample). Condition mean estimates for each ROI are reported in Table 3. Model details are reported in Table S3 of the online supplemental materials. As in Study 1, we extracted by-stimuli BLUPs for each scenario and condition combination, providing 96 data points for each ROI.

Table 3. ToMN percent signal change condition means.

	Initial Tragic	Initial Taboo	Tragic–Taboo	Tragic–Control	Taboo–Tragic	Taboo–Control
ROI	***M (SE)***	***M (SE)***	***M (SE)***	***M (SE)***	***M (SE)***	***M (SE)***
DMPFC	0.105 (0.031)	0.089 (0.029)	-0.018 (0.043)	-0.111 (0.038)	-0.048 (0.042)	-0.070 (0.043)
VPMFC	0.101 (0.036)	0.075 (0.052)	0.039 (0.046)	-0.050 (0.049)	-0.020 (0.048)	-0.045 (0.048)
PC	0.129 (0.027)	0.117 (0.027)	0.042 (0.035)	-0.025 (0.041)	0.078 (0.050)	-0.037 (0.026)
RTPJ	0.050 (0.023)	0.061 (0.022)	-0.005 (0.025)	0.005 (0.034)	0.017 (0.027)	0.019 (0.028)
LTPJ	0.141 (0.029)	0.144 (0.027)	0.073 (0.038)	-0.007 (0.031)	0.060 (0.045)	-0.028 (0.034)

ToMN activity–behavioral component score analysis. Across by-stimuli BLUPs, we examined the relationship between component scores, derived in Study 1, and ToMN activity. As component scores were derived only for second pass judgments, we did not analyze first pass ToMN ROI activity (although they were still included in the model above, deriving by-stimuli BLUPs). Each component score was fit with a linear model, first adding Initial Condition, Reframing Condition, and their interaction, then adding ToMN activity and its interaction with all terms, and finally adding ROI interactions (no main effect of ROI was included, as component scores were identical across ROIs; for details, see Table S4 of the online supplemental materials).

For dispositional prediction error (Figure 3a), we observed an interaction between ToMN activity and Reframing Condition, $F(1, 456) = 4.44$, $p = .035$, such that dispositional prediction error was associated with ToMN activity within tragic–taboo/taboo–tragic scenarios, $B = 0.10$, $t(487) = 2.54$, $p = .022$, but not within tragic–control/taboo–control scenarios, $B = -0.01$, $t(487) = 0.18$, $p = .980$ (corrected for two comparisons; $\alpha_{\text{familywise}} = .05$; single-step method; *multcomp* package, Hothorn et al., 2008). The interaction was not qualified by any further interaction with ROI, $F(4, 456) = 1.35$, $p = .250$; however, for visualization the effect is broken down by ROI in Figure 3a. For prescriptive prediction error (Figure 3b), we also observed an interaction between ToMN activity and Reframing Condition, $F(1, 456) = 9.94$, $p = .002$, such that prescriptive prediction error was also associated with ToMN activity within tragic–taboo/taboo–tragic scenarios, $B = 0.18$, $t(487) = 2.46$, $p = .028$, but not within tragic–control/taboo–control scenarios, $B = -0.06$, $t(487) = 0.50$, $p = .852$ (corrected for two comparisons; $\alpha_{\text{familywise}} = .05$; single-step method). This interaction was also not qualified by an ROI interaction, *F*(4, 456) = 0.53, *p* = .710, although as above, we visualize each ROI in Figure 3b for the sake of clarity. ToMN activity was not related to descriptive prediction error (Figure 3c; Table S4), mental imagery (Figure S1, Table S4), or arousal (Figure S1, Table S4). Thus, when participants were required to update their initial impressions and moral judgments, ToMN activity tracked the magnitude of the dispositional and prescriptive prediction error. By contrast, we did not observe this pattern in a component measuring descriptive prediction error .

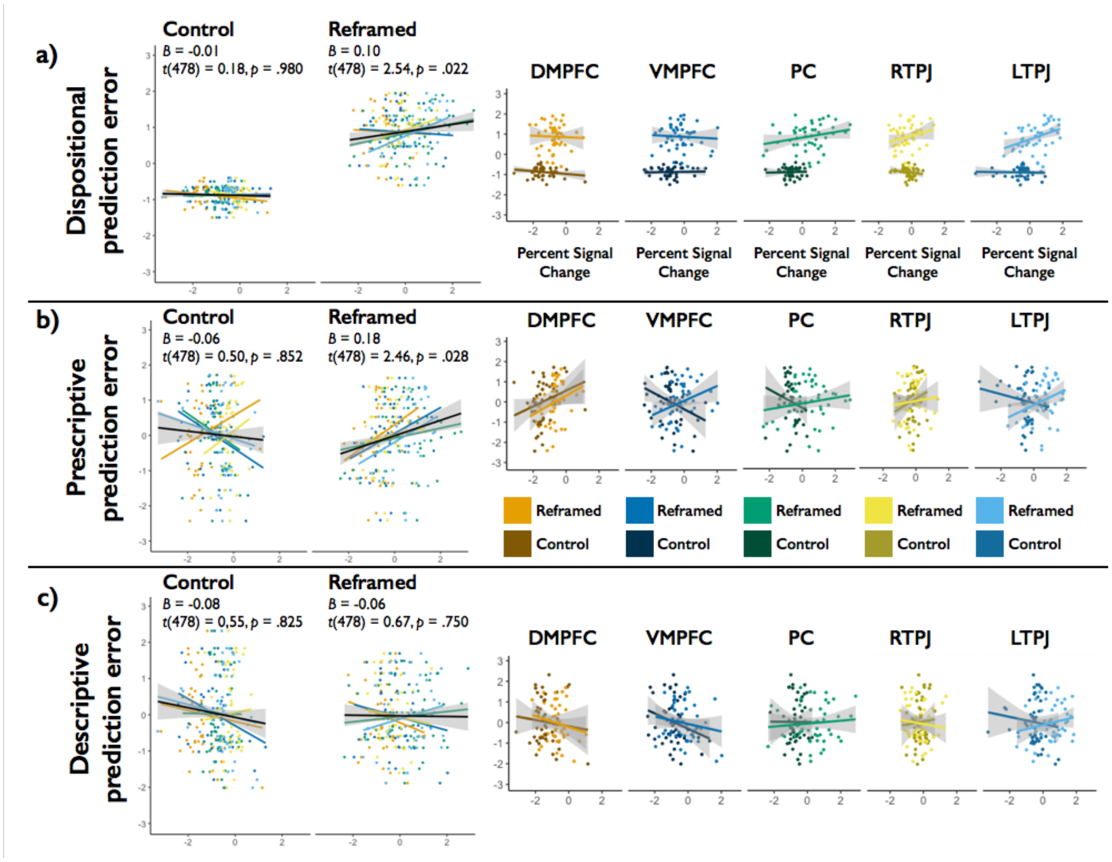


Figure 3. ToMN–behavioral component score relationships. Main effects of ToMN activity (left) and their breakdown by ROI (right). Interactions with ROI were not significant, but separate results for individual ROIs are presented for ease of interpretation. (a) Dispositional prediction error was related to ToMN activity within reframed scenarios (i.e. tragic–taboo/taboo–tragic) but not within control scenarios (i.e. tragic–control/taboo–control). (b) Prescriptive prediction error showed the same pattern, and was related to ToMN activity for reframed, but not control, scenarios. (c) Descriptive prediction error showed no relationship with ToMN activity. Note that component scores were orthogonal, so common patterns cannot stem from correlations among component

scores. Relationships for mental imagery, and arousal component scores were non-significant and are presented in Figure S1, of the online supplemental materials. P values are corrected for two comparisons in each model ($\alpha_{\text{familywise}} = .05$; single-step method).

General Discussion.

In the present work, we tested the relationship between ToMN (theory of mind network) activity and several forms prediction error. Across ToMN ROIs, activity was associated with both dispositional and prescriptive prediction error (Figure 3), but not descriptive prediction error (i.e. violations of descriptive norms). That is, scenarios which contradicted either moral norms or specific impressions of the agent, elicited greater activity across the ToMN. Measures of prediction error were aggregated component scores (extracted from a principal components analysis in Study 1), and were orthogonally rotated to remove correlations among factors—meaning that the observed relationships between dispositional and prescriptive prediction error cannot be explained by a correlation between dependent variables. Thus, our findings are consistent with the proposal that the ToMN is involved in the computation of social prediction error (Koster-Hale & Saxe, 2013; Joiner et al., 2017).

Notably, the relationship between ToMN activity and both dispositional and prescriptive prediction error was significant only during reframed scenarios (tragic–taboo/taboo–tragic), where participants were given additional, morally relevant information. Neither relationship was significant across control scenarios (tragic–control/taboo–control). For dispositional prediction error, this non-significance could plausibly be explained by a lack of variance (Figure 3), i.e. it is possible that a different set of control scenarios, which violated personal impressions (without reframing moral judgment), would be associated with ToMN activity. Thus, one weakness of the present study is that such variance could not be explored. However, prescriptive judgments were variable across control scenarios—as participants formed first pass moral judgments during the initial dilemma (initial tragic/initial taboo) and maintained them in their second pass judgments (Table 1). Thus, ToMN activity was related to prescriptive prediction error only when new, morally relevant information was presented.

If ToMN is responsive to social prediction error, then future work may better account for observed patterns of activity by considering the normative contexts in which stimuli are presented. For instance, prior work in impression formation has contrasted positive and negative impression updating (e.g. Bhanji & Beer, 2013), finding separate neural correlates for each. However, other work suggests that diagnosticity is a common denominator for (at least some) neural activity related to impression updating (Mende-Siedlecki et al., 2013). For instance, putting aside prescriptive norms for the moment, within the domain of moral behaviors, extreme immorality is descriptively uncommon—people behave themselves more often than not. By contrast, in the domain of proficiency, it is descriptively uncommon to encounter people who are extremely skilled. Across both domains, updating based on descriptive frequency elicited activity in regions such as left superior temporal sulcus, and left/right ventral-lateral prefrontal cortex (Mende-Siedlecki et al., 2013), although effects were generally stronger for moral updating. It may be that social predictions based around morality are particularly salient in social situations (see also Theriault et al., 2017b)—our social world is predictable when others behave morally, and when they let us down it is important that we update our social predictions.

The ToMN is broadly involved in social cognition, and based on the present study alone, it is impossible to make conclusions about the precise relationship between mental state representation and social prediction; for instance, does *all* mental state representation boil down to prediction? Or is prediction only one part of the broader construct? Despite this, it is worth pointing out that we included a measure of mental state representation (used in prior work; Dodell-Feder et al., 2011), and it loaded heavily onto our factor tracking dispositional prediction error (Figure 2; Table 2). Intuitively, this makes sense, as people should be most inclined to consider mental states when an agent does not behave as predicted. In other words, mental state inference is a solution to the problem of prediction when a system does not behave according to a transparent system of input and output (Dennett, 1987; Theriault & Young, 2014).

Finally, it is worth emphasizing what we consider to be the benefits of the item analysis approach taken in the present work. Social behavior is complicated, and, in our opinion, social categories do not lend themselves to the level of specificity that simple BOLD contrasts require—there is simply no plausible way to control for all confounds and deliver a clean contrast. At the other end of the spectrum, computational models can cleanly characterize a process, but also require that the task be somewhat removed from naturalistic contexts (e.g. economic games). Item analysis offers a middle path, where researchers can relax constraints on their sample of stimuli but, at the same time, conduct fine-grained analyses of the dimensions that emerge across the it. Furthermore, this approach is advantageous in that stimuli can be normed and reused, meaning that other researchers can utilize our stimuli (and normed estimates) to test their own hypotheses about the underlying dimensions (see Appendix B for the complete stimuli set and paired BLUPs).

Conclusion

ToMN activity is associated with social prediction error, consistent with a predictive coding account of social cognition (Koster-Hale & Saxe, 2013). Although future research is necessary to better characterize the relationship between ToMN activity and social cognition, the present work is consistent with an account that imagines the brain as a "hierarchical prediction machine" (A. Clark, 2013), using information about people, contexts, and statistical frequency to anticipate its social environment.

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. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

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

Bach P, & Schenke K. C. (2017). Predictive social perception: Towards a unifying framework from action observation to person knowledge. *Social and Personality Psychology Compass, 11*, 1751–11, e12312. https://dx.doi.org/10.1111/spc3.12312

Baron, S. G., Gobbini, M. I., Engell, A. D., & Todorov, A. (2010). Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions. *Social Cognitive and Affective Neuroscience*, *6*, 572-581. http://dx.doi.org/10.1093/scan/nsq086

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

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*, 1-23. http://dx.doi.org/10.1093/scan/nsw154

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

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

Bhanji, J. P., & Beer, J. S. (2013). Dissociable neural modulation underlying lasting first impressions, changing your mind for the better, and changing it for the worse. *Journal of Neuroscience*, *33*, 9337-9344. http://dx.doi.org/10.1523/JNEUROSCI.5634-12.2013

Blakemore, S. J., Boyer, P., Pachot-Clouard, M., Meltzoff, A., Segebarth, C., & Decety, J. (2003). The detection of contingency and animacy from simple animations in the human brain. *Cerebral Cortex*, *13*, 837-844. http://dx.doi.org/10.1093/cercor/13.8.837

Brauer, M., & Chaurand, N. (2010). Descriptive norms, prescriptive norms, and social control: An intercultural comparison of people's reactions to uncivil behaviors. *European Journal of Social Psychology*, *40*, 490-499. http://dx.doi.org/10.1002/ejsp.640

Buckner, R. L. (2012). The serendipitous discovery of the brain's default network. *Neuroimage*, *62*, 1137-1145. http://dx.doi.org/10.1016/j.neuroimage.2011.10.035

Carter, R. M., & Huettel, S. A. (2013). A nexus model of the temporal–parietal junction. *Trends in Cognitive Sciences*, *17*, 328-336. http://dx.doi.org/10.1016/j.tics.2013.05.007

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*, 1015–1026. http://dx.doi.org/10.1037/0022-3514.58.6.1015

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

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 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

Cloutier, J., Gabrieli, J. D., O'young, D., & Ambady, N. (2011). An fMRI study of violations of social expectations: when people are not who we expect them to be. *NeuroImage*, *57*, 583-588. http://dx.doi.org/10.1016/j.neuroimage.2011.04.051

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science*, *4*, 308-315. http://dx.doi.org/10.1177/1948550612457688

Dennett, D. C. (1989). *The intentional stance*. MIT press.

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

Dungan, J., Stepanovic, M., & Young, L. (in press). Theory of mind for processing unexpected events across contexts. *Social Cognitive and Affective Neuroscience*. http://dx.doi.org/10.1093/scan/nsw032

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*, 7900–7905. http://dx.doi.org/10.1073/pnas.1602413113

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

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT press.

Fourie, M. M., Thomas, K. G., Amodio, D. M., Warton, C. M., & Meintjes, E. M. (2014). Neural correlates of experienced moral emotion: an fMRI investigation of emotion in response to prejudice feedback. *Social Neuroscience*, *9*, 203-218. http://dx.doi.org/10.1080/17470919.2013.878750

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

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

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*, 189-211. http://dx.doi.org/10.1080/19345747.2011.618213

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*, 21. http://dx.doi.org/10.1037/0033-2909.117.1.21

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

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

Hassabis, D., & Maguire, E. A. (2009). The construction system of the brain. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*, 1263-1271. http://dx.doi.org/10.1098/rstb.2008.0296

Heyes, C. (2014). False belief in infancy: a fresh look. *Developmental Science*, *17*, 647-659. http://dx.doi.org/10.1111/desc.12148

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal 50(3)*, 346--363. http://dx.doi.org/10.1002/bimj.200810425

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

Joiner, J., Piva, M., Turrin, C., & Chang, S. (2017). Social learning through prediction error in the brain. *npj Science of Learning, 2*, 8. http://dx.doi.org/10.1038/s41539-017-0009-2

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

Kircher, T., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J., van Os, J., & Krach, S. (2009). Online mentalising investigated with functional MRI. *Neuroscience Letters*, *454*, 176-181. http://dx.doi.org/10.1016/j.neulet.2009.03.026

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron, 79*, 836–848. http://dx.doi.org/10.1016/j.neuron.2013.08.020

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*, *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).

Lichtenstein, S., Gregory, R., & Irwin, J. (2007). What's bad is easy: Taboo values, affect, and cognition. *Judgment and Decision Making, 2*, 169–188.

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

Mars, R. B., Neubert, F. X., Noonan, M. P., Sallet, J., Toni, I., & Rushworth, M. F. (2012). On the relationship between the "default mode network" and the "social brain". *Frontiers in Human Neuroscience, 6,* 189. http://dx.doi.org/10.3389/fnhum.2012.00189

Mende-Siedlecki, P., & Todorov, A. (2016). Neural dissociations between meaningful and mere inconsistency in impression updating. *Social Cognitive and Affective Neuroscience*, *11*, 1489-1500. http://dx.doi.org/10.1093/scan/nsw058

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*, 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

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and Brain Sciences*, *1*, 515-526. http://dx.doi.org/10.1017/S0140525X00076512

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ratner, K. G., Kaul, C., & Van Bavel, J. J. (2012). Is race erased? Decoding race from patterns of neural activity when skin color is not diagnostic of group boundaries. *Social Cognitive and Affective Neuroscience*, *8*, 750-755. http://dx.doi.org/10.1093/scan/nss063

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

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

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences, 19,* 65–72. http://dx.doi.org/10.1016/j.tics.2014.11.007

Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., & Phelps, E. A. (2009). A neural mechanism of first impressions. *Nature Neuroscience*, *12*, 508-514. http://dx.doi.org/10.1038/nn.2278

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

Schurz, M., Tholen, M. G., Perner, J., Mars, R. B., & Sallet, J. Specifying the brain anatomy underlying temporo-parietal junction activations for theory of mind: A review using probabilistic atlases from different imaging modalities. *Human Brain Mapping*. http://dx.doi.org/10.1002/hbm.23675

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*, 853–870. http://dx.doi.org/10.1037//0022-3514.78.5.853

Theriault, J., & Young, L. (2014). Taking an 'Intentional Stance' on Moral Psychology. In J. Systma (ed.), *Advances in Experimental Philosophy of Mind*. Continuum Press.

Theriault, J., Waytz, A., Heiphetz, L., & Young, L. (in press). Examining overlap in behavioral and neural representations of morals, facts, and preferences. *Journal of Experimental Psychology: General*. http://dx.doi.org/10.1037/xge0000350

Theriault, J., Waytz, A., Heiphetz, L., & Young, L. (under review). Theory of Mind network activity is associated with metaethical judgment: An item analysis.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*, 326-334. http://dx.doi.org/10.1016/j.cognition.2012.10.005

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

Westfall, J., Nichols, T. E., & Yarkoni, T. (2016). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Wellcome Open Research*, 1. http://dx.doi.org/10.12688/wellcomeopenres.10298.2

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

Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., & Hasson, U. (2017). Same story, different story: the neural representation of interpretive frameworks. *Psychological Science*, *28*, 307-319. http://dx.doi.org/10.1177/0956797616682029

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., 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

Appendix A. Scenario Text.

Scenario 1				
a)	Rebecca is in charge of running a yearlong drug trial at McAdam Hospital. The drug was given to the experimental group of patients, and a placebo was given to the control group. At two months, early results suggest that the drug is effective.			
b)	Rebecca has the option to give the control group the medicine early. This could potentially save the lives of patients who would die without immediate access to the treatment.			
	Initial Tragic		**Initial Taboo**	
c)	Holding the trial at the original length would produce more conclusive data. This would help develop better treatments in the long run, and save the lives of patients in the future.		Giving treatment to the control group before a study is complete is frowned on in the medical community. If Rebecca ends the study early, she will have trouble progressing her career.	
d)	Rebecca thinks very carefully and ultimately decides to continue the study at its original length. The drug trial continues for the remainder of the year, but some patients in the control group die during this time.			
	Tragic–Taboo	**Tragic–Control**	**Taboo–Tragic**	**Taboo–Control**
e)	Giving treatment to the control group before a study is complete is frowned on in the medical community. If Rebecca ends the study early, she will have trouble progressing her career.	After Rebecca arrives home from work, she makes herself spaghetti for dinner and watches television. After dinner she washes the dishes and takes a shower before going to bed.	Holding the trial at the original length would produce more conclusive data. This would help develop better treatments in the long run, and save the lives of patients in the future.	After Rebecca arrives home from work, she makes herself spaghetti for dinner and watches television. After dinner she washes the dishes and takes a shower before going to bed.

<table>
<tr><th><b>Scenario 2</b></th><th colspan="4"></th></tr>
<tr><td><b>a)</b></td><td colspan="4">Jessica is in charge of a subcommittee of the Environmental Protection Agency and must break a tie in a vote. The vote is on whether to approve a project proposed by a drug company.</td></tr>
<tr><td><b>b)</b></td><td colspan="4">If Jessica rejects the project then the drug company will be prevented from harvesting old growth forests to develop their drug. This would prevent severe environmental damage which would wipe out many endangered species.</td></tr>
<tr><td></td><td colspan="2"><b>Initial Tragic</b></td><td colspan="2"><b>Initial Taboo</b></td></tr>
<tr><td><b>c)</b></td><td colspan="2">By approving the project, Jessica will be allowing highly effective treatments for multiple sclerosis to be developed. No other treatment is nearly as effective as this drug, and the new drug would help thousands of people across America.</td><td colspan="2">Jessica's boss, the head of the EPA, would personally benefit if this new drug were to be produced. By approving the project Jessica would win his gratitude and he would be more likely to grant the promotion she planned to ask for.</td></tr>
<tr><td><b>d)</b></td><td colspan="4">Jessica thinks very carefully and ultimately decides to approve the project. The drug is developed, and the old growth forest is destroyed.</td></tr>
<tr><td></td><td><b>Tragic–Taboo</b></td><td><b>Tragic–Control</b></td><td><b>Taboo–Tragic</b></td><td><b>Taboo–Control</b></td></tr>
<tr><td><b>e)</b></td><td>Jessica's boss, the head of the EPA, would personally benefit if this new drug were to be produced. By approving the project Jessica would win his gratitude and he would be more likely to grant the promotion she planned to ask for.</td><td>Jessica calls several of her close friends and makes plans to see a movie the following weekend. The movie had received good reviews in the local newspaper and Jessica has seen all of the leading actor's previous movies.</td><td>By approving the project, Jessica will be allowing highly effective treatments for multiple sclerosis to be developed. No other treatment is nearly as effective as this drug, and the new drug would help thousands of people across America.</td><td>Jessica calls several of her close friends and makes plans to see a movie the following weekend. The movie had received good reviews in the local newspaper and Jessica has seen all of the leading actor's previous movies.</td></tr>
</table>

Scenario 3				
a)	Emil owns a small farm in Argentina. Emil is considering expanding his farm, which would allow him to grow more varieties of fruits and vegetables.			
b)	Emil sells his crops to a nearby village. He knows that what he grows does not contain enough nutrients for a healthy diet. By expanding his farm he could save the villagers from malnourishment.			
	Initial Tragic		**Initial Taboo**	
c)	Emil knows that the area of the rainforest that borders on his farm contains an exceptional number of endangered species, and that expanding into it will cause many of them to die out.		Emil regularly brings in tourists who pay for tours of the rainforest near his house, and continuing to give these tours will be more profitable then planting more crops and feeding the village.	
d)	Emil thinks very carefully and ultimately decides not to expand his farm into the rainforest. His farm is not developed any further, and the villagers continue to suffer from malnourishment.			
	Tragic–Taboo	**Tragic–Control**	**Taboo–Tragic**	**Taboo–Control**
e)	Emil regularly brings in tourists who pay for tours of the rainforest near his house, and continuing to give these tours will be more profitable then planting more crops and feeding the village.	Emil received a call the next day from an acquaintance he had lost contact with. They agreed to meet the next week at a café in the city in order to catch up on their lives since high school.	Emil knows that the area of the rainforest that borders on his farm contains an exceptional number of endangered species, and that expanding into it will cause many of them to die out.	Emil received a call the next day from an acquaintance he had lost contact with. They agreed to meet the next week at a café in the city in order to catch up on their lives since high school.

<table>
<tr><th>Scenario 4</th><th colspan="4"></th></tr>
<tr><td>a)</td><td colspan="4">Sanjeev is a government official in India who is in charge of a local wildlife preservation. A family of endangered tigers has been attacking people on the border of the preserve, and she must decide what to do about it.</td></tr>
<tr><td>b)</td><td colspan="4">The World Wildlife Foundation has asked that they be allowed to capture the tigers alive and relocate them. This will leave the villagers in danger for longer, but will save the tigers.</td></tr>
<tr><td>c)</td><td colspan="2"><b>Initial Tragic</b>
The nearby villagers are terrified and afraid to let their children out of their houses. By sending the army to kill the tigers Sanjeev could eliminate the danger almost immediately.</td><td colspan="2"><b>Initial Taboo</b>
Sanjeev knows that the teeth of these tigers are very valuable and can be sold as an aphrodisiac. By sending the army to kill the tigers Sanjeev could take a share of the profit.</td></tr>
<tr><td>d)</td><td colspan="4">Sanjeev thinks very carefully and ultimately decides to send in the army. The army kills the entire family of endangered tigers.</td></tr>
<tr><td>e)</td><td><b>Tragic–Taboo</b>
Sanjeev knows that the teeth of these tigers are very valuable and can be sold as an aphrodisiac. By sending the army to kill the tigers Sanjeev could take a share of the profit.</td><td><b>Tragic–Control</b>
The next afternoon, Sanjeev goes for a walk to get exercise. She forgot to eat breakfast that morning and becomes hungry after 30 minutes. She decides to take the shorter route home.</td><td><b>Taboo–Tragic</b>
The nearby villagers are terrified and afraid to let their children out of their houses. By sending the army to kill the tigers Sanjeev could eliminate the danger almost immediately.</td><td><b>Taboo–Control</b>
The next afternoon, Sanjeev goes for a walk to get exercise. She forgot to eat breakfast that morning and becomes hungry after 30 minutes. She decides to take the shorter route home.</td></tr>
</table>

Scenario 5				
a)	Sarah and her five year old son Jeffery were recently evicted from their house and are living in a homeless shelter. A very rich couple has approached Sarah explaining that they are unable to conceive and that they would be willing to adopt Jeffery.			
b)	The couple has told Sarah that they will be moving to California. They do not plan to return and it is unlikely that Sarah will ever see her son again if they adopt him.			
	Initial Tragic		**Initial Taboo**	
c)	Sarah can barely feed Jeffery, let alone provide a comfortable life for him. She is absolutely certain that Jeffery would be very well off with his new family.		Sarah has been told by the couple that they will buy her an expensive new car to replace her old one. They will purchase the car when she signs over custody of Jeffery.	
d)	Sarah thinks very carefully and ultimately decides to allow the family to adopt Jeffery. Jeffery moves to California with the family, and Sarah never sees him again.			
	Tragic–Taboo	**Tragic–Control**	**Taboo–Tragic**	**Taboo–Control**
e)	Sarah has been told by the couple that they will buy her an expensive new car to replace her old one. They will purchase the car when she signs over custody of Jeffery.	Sarah goes to her former high school's basketball game that weekend. Her high school used to rank poorly, but recently has improved its standing in the league.	Sarah can barely feed Jeffery, let alone provide a comfortable life for him. She is absolutely certain that Jeffery would be very well off with his new family.	Sarah goes to her former high school's basketball game that weekend. Her high school used to rank poorly, but recently has improved its standing in the league.

<table>
<tr><th>Scenario 6</th><th colspan="4"></th></tr>
<tr><td>a)</td><td colspan="4">Candace is the Mayor of a small mid-western American city. The city needs to make cuts to the education budget. The council has brought two proposals to her about programs that could potentially be cut.</td></tr>
<tr><td>b)</td><td colspan="4">Candace could cut after school programs for at-risk youth. This would leave many adolescents without alternatives to getting involved with drugs and crime. It would almost certainly negatively affect their future prospects.</td></tr>
<tr><td>c)</td><td colspan="2">**Initial Tragic**
Cutting day-care programs would put a financial strain on single parents. Parents with jobs would need to work more hours and see their children less. Parents without jobs would have to balance childcare and their time to job hunt.</td><td colspan="2">**Initial Taboo**
Cutting day-care programs would make parents unhappy, and Candace is up for reelection next year. The votes of these parents were essential to her victory in the last election. Losing the support of parents could lose her the next election.</td></tr>
<tr><td>d)</td><td colspan="4">Candace thinks carefully and ultimately decides to cut the after school programs for at-risk youth. As a result, many of these children get involved with drugs and crime.</td></tr>
<tr><td>e)</td><td>**Tragic–Taboo**
Cutting day-care programs would make parents unhappy, and Candace is up for reelection next year. The votes of these parents were essential to her victory in the last election. Losing the support of parents could lose her the next election.</td><td>**Tragic–Control**
Candace takes her dog, Spot, on a walk through the downtown city park. The park allows dogs to go off of their leashes, but only in certain areas. When Candace reaches the area, she lets Spot off of his leash and throws the ball with him.</td><td>**Taboo–Tragic**
Cutting day-care programs would put a financial strain on single parents. Parents with jobs would need to work more hours and see their children less. Parents without jobs would have to balance childcare and their time to job hunt.</td><td>**Taboo–Control**
Candace takes her dog, Spot, on a walk through the downtown city park. The park allows dogs to go off of their leashes, but only in certain areas. When Candace reaches the area, she lets Spot off of his leash and throws the ball with him.</td></tr>
</table>

<table>
<tr><th>Scenario 7</th><th colspan="4"></th></tr>
<tr><td><b>a)</b></td><td colspan="4">Michael is 36 years old and is visiting his mother in the hospital. She is completely paralyzed and unable to speak. She had told Michael that in such situations she wants to be euthanized and the decision is legally Michael’s to make.</td></tr>
<tr><td><b>b)</b></td><td colspan="4">Michael could ask the doctors to euthanize his mother. The procedure would be painless and if it was not performed she would live for years, completely unable to move or speak.</td></tr>
<tr><td></td><td colspan="2"><b>Initial Tragic</b></td><td colspan="2"><b>Initial Taboo</b></td></tr>
<tr><td><b>c)</b></td><td colspan="2">Michael’s wife is adamantly opposed to euthanasia. She understands Michael’s mother’s circumstances, but has told Michael that she will divorce him if he has his mother euthanized.</td><td colspan="2">Michael’s mother is very wealthy and has set up automatic deposits to Michael’s bank account. When she dies, her fortune will be donated to her favorite charity, and Michael will stop receiving money.</td></tr>
<tr><td><b>d)</b></td><td colspan="4">Michael thinks very carefully and ultimately decides to leave his mother to die naturally. His mother remains conscious but unable to move or speak.</td></tr>
<tr><td></td><td><b>Tragic–Taboo</b></td><td><b>Tragic–Control</b></td><td><b>Taboo–Tragic</b></td><td><b>Taboo–Control</b></td></tr>
<tr><td><b>e)</b></td><td>Michael’s mother is very wealthy and has set up automatic deposits to Michael’s bank account. When she dies, her fortune will be donated to her favorite charity, and Michael will stop receiving money.</td><td>On his way home, Michael listens to the radio. He hears that winter is expected to come early this year. As soon as he arrives home he writes himself a note to put the winter tires on the car on Friday.</td><td>Michael’s wife is adamantly opposed to euthanasia. She understands Michael’s mother’s circumstances, but has told Michael that she will divorce him if he has his mother euthanized.</td><td>On his way home, Michael listens to the radio. He hears that winter is expected to come early this year. As soon as he arrives home he writes himself a note to put the winter tires on the car on Friday.</td></tr>
</table>

<table>
<tr><th>Scenario 8</th><th colspan="4"></th></tr>
<tr><td>a)</td><td colspan="4">Erica is a 25-year-old woman who is seven months pregnant and single after her husband’s death in a car accident. After a recent visit to the doctor, she learns that her baby has a rare chronic medical condition.</td></tr>
<tr><td>b)</td><td colspan="4">Erica could choose to have her baby. Although the medical condition will be debilitating, the baby’s life expectancy is expected to be completely normal.</td></tr>
<tr><td></td><td colspan="2"><b>Initial Tragic</b></td><td colspan="2"><b>Initial Taboo</b></td></tr>
<tr><td>c)</td><td colspan="2">Erica knows that her baby would be in extreme pain for his entire life. The condition causes skin to be hypersensitive and painful to any sort of touch.</td><td colspan="2">Erica knows that her baby’s medical condition would require her specialized equipment. She would need to move to a cheap apartment in order to afford this.</td></tr>
<tr><td>d)</td><td colspan="4">Erica thinks carefully and ultimately decides to have an abortion. Her unborn baby is aborted, and Erica suffers no negative consequences from the abortion.</td></tr>
<tr><td></td><td><b>Tragic–Taboo</b></td><td><b>Tragic–Control</b></td><td><b>Taboo–Tragic</b></td><td><b>Taboo–Control</b></td></tr>
<tr><td>e)</td><td>Erica knows that her baby’s medical condition would require her specialized equipment. She would need to move to a cheap apartment in order to afford this.</td><td>When Erica cooks dinner that night she accidentally burns the potatoes. She notices that the smoke detector does not go off and replaces the batteries.</td><td>Erica knows that her baby would be in extreme pain for his entire life. The condition causes skin to be hypersensitive and painful to any sort of touch.</td><td>When Erica cooks dinner that night she accidentally burns the potatoes. She notices that the smoke detector does not go off and replaces the batteries.</td></tr>
</table>

Scenario 9				
a)	Abby is the CEO of Morrison Motors, a large car manufacturing company. Abby must make a decision about whether to issue a recall due to a defect in the Ellipsis line of cars.			
b)	Abby could issue a recall to fix this defect, which would return the thousands of Ellipsis cars to the factory. This would protect customers from the fatal accidents that can occur when the brakes fail.			
	Initial Tragic		**Initial Taboo**	
c)	Abby knows that the finances of the company are poor, and the negative press and expense of a recall would bankrupt them. Thousands of long-time employees would lose their jobs and pensions.		The cost of settlements with the families of the victims would be much cheaper than the cost of a recall. Not issuing a recall could save the company money and even set Abby up for a promotion.	
d)	Abby thinks carefully, and ultimately decides not to issue the recall. The company saves a great deal of money, but fatal accidents occur as a result.			
	Tragic–Taboo	**Tragic–Control**	**Taboo–Tragic**	**Taboo–Control**
e)	The cost of settlements with the families of the victims would be much cheaper than the cost of a recall. Not issuing a recall could save the company money and even set Abby up for a promotion.	Abby went to the gym next to the office to exercise after work. She had originally planned to run on the treadmill, but they were all occupied so she used the bicycle machine instead.	Abby knows that the finances of the company are poor, and the negative press and expense of a recall would bankrupt them. Thousands of long-time employees would lose their jobs and pensions.	Abby went to the gym next to the office to exercise after work. She had originally planned to run on the treadmill, but they were all occupied so she used the bicycle machine instead.

Scenario 10	
a)	Brock is a clerk working for the Canadian military and can decide to approve or reject draftees that have been referred to him. He is currently considering the case of Aaron, a young man who is eligible to be drafted.
b)	Brock knows that Aaron has experience with engineering and could be put on a bomb defusal squad. This expertise could potentially save the lives of civilians and fellow soldiers.
c)	**Initial Tragic** Brock read that Aaron works with Engineers without Borders. If rejected from the draft, Aaron would continue to build wells in South Africa, giving the poor access to fresh water. **Initial Taboo** Brock was contacted by Aaron's family, who are very influential. They will contact Brock's superiors and get him promoted if he rejects Aaron's file and spares him the draft.
d)	Brock thinks very carefully and ultimately decides to reject Aaron's file. Aaron is not drafted into the army.
e)	**Tragic–Taboo** Brock was contacted by Aaron's family, who are very influential. They will contact Brock's superiors and get him promoted if he rejects Aaron's file and spares him the draft. **Tragic–Control** The following afternoon, Brock attends a meeting along with the other clerks. They discuss a new database program that will help to reduce the amount of paper used in their jobs. **Taboo–Tragic** Brock read that Aaron works with Engineers without Borders. If rejected from the draft, Aaron would continue to build wells in South Africa, giving the poor access to fresh water. **Taboo–Control** The following afternoon, Brock attends a meeting along with the other clerks. They discuss a new database program that will help to reduce the amount of paper used in their jobs.

Scenario 11				
a)	Elizabeth owns and operates an animal shelter that cares for stray dogs. The shelter had signed a contract under a previous owner to supply dogs to a nearby university for research purposes. The contract is up for renewal.			
b)	Elizabeth could refuse to re-sign the contract, in which case the sale of animals to the university would end. The university studies the causes of blindness and tests hazardous chemicals on the animals.			
	Initial Tragic		**Initial Taboo**	
c)	The animal shelter is in poor financial shape. If Elizabeth refuses to re-sign the contract the shelter will likely close. As a result, all of the animals in their care would be turned out into the street.		If Elizabeth re-signs the contract, then the university will increase their payment for the animals to adjust for inflation. The university has also promised to pay Elizabeth a signing bonus of $5000.	
d)	Elizabeth thinks very carefully and ultimately decides to re-sign the contract with the university. Several cats and dogs are taken each month for experimentation.			
	Tragic–Taboo	**Tragic–Control**	**Taboo–Tragic**	**Taboo–Control**
e)	If Elizabeth re-signs the contract, then the university will increase their payment for the animals to adjust for inflation. The university has also promised to pay Elizabeth a signing bonus of $5000.	After work, Elizabeth stops at a friend's house to borrow a movie that her friend had recommended. Elizabeth planned to watch it that night, but got distracted by another program that was on TV.	The animal shelter is in poor financial shape. If Elizabeth refuses to re-sign the contract the shelter will likely close. As a result, all of the animals in their care would be turned out into the street.	After work, Elizabeth stops at a friend's house to borrow a movie that her friend had recommended. Elizabeth planned to watch it that night, but got distracted by another program that was on TV.

<table>
<tr><th>Scenario 12</th><th colspan="4"></th></tr>
<tr><td><b>a)</b></td><td colspan="4">Angela is a 40-year-old mother of two children, aged 12 and 14. She is has been approached by two women (Sandra and Megan) to act as a surrogate mother, and is considering whether to accept either offer.</td></tr>
<tr><td><b>b)</b></td><td colspan="4">By being a surrogate mother for Sandra, Angela would be helping a close friend who has always wanted a child but cannot conceive on her own. Sandra is willing to pay for Angela’s healthcare costs, so that the pregnancy will not cost Angela anything.</td></tr>
<tr><td></td><td colspan="2"><b>Initial Tragic</b></td><td colspan="2"><b>Initial Taboo</b></td></tr>
<tr><td><b>c)</b></td><td colspan="2">Angela wants to send her children to college, but she does not have the money for a college savings fund. Megan has promised to pay Angela generously for carrying her child. Without this money, Angela may not be able to send her children to college.</td><td colspan="2">By being a surrogate mother for Megan, Angela will be generously compensated. Megan is a wealthy acquaintance and in addition to the large payment, has offered to buy Angela box seats at the Metropolitan Opera, of which Angela is an enormous fan.</td></tr>
<tr><td><b>d)</b></td><td colspan="4">Angela thinks carefully and ultimately decides to act as a surrogate mother for Megan. She is generously compensated for her trouble.</td></tr>
<tr><td></td><td><b>Tragic–Taboo</b></td><td><b>Tragic–Control</b></td><td><b>Taboo–Tragic</b></td><td><b>Taboo–Control</b></td></tr>
<tr><td><b>e)</b></td><td>By being a surrogate mother for Megan, Angela will be generously compensated. Megan is a wealthy acquaintance and in addition to the large payment, has offered to buy Angela box seats at the Metropolitan Opera, of which Angela is an enormous fan.</td><td>While Angela’s children are at school she reads her favorite book. She notices that she is nearly finished and drives to a local bookstore to pick out something new. The store was having a sale, and so she picks out two books instead of just the one she planned to buy.</td><td>Angela wants to send her children to college, but she does not have the money for a college savings fund. Megan has promised to pay Angela generously for carrying her child. Without this money, Angela may not be able to send her children to college.</td><td>While Angela’s children are at school she reads her favorite book. She notices that she is nearly finished and drives to a local bookstore to pick out something new. The store was having a sale, and so she picks out two books instead of just the one she planned to buy.</td></tr>
</table>

Scenario 13

a) Gregory is the captain of a fishing vessel that operates off the coast of Cape Cod. He is considering implementing a new fishing method for himself and his crew.

b) The new method involves specialized nets that release larger creatures caught in them. If used, it would decrease the number of dolphins that are accidentally caught and strangled in the netting.

c)

Initial Tragic

Gregory knows that by implementing the new method he would be forced to lay off a third of his crew due to the related expenses. These people would have a very difficult time finding other jobs.

Initial Taboo

Gregory has run a profitable business on the side where he sells dolphin fins to natural medicine distributors. If he implemented the new fishing method, he would need to shut down this business.

d) Gregory thinks carefully and ultimately decides not to implement the new fishing method. The vessel continues to kill several dolphins per month.

e)

Tragic–Taboo

Gregory has run a profitable business on the side where he sells dolphin fins to natural medicine distributors. If he implemented the new fishing method, he would need to shut down this business.

Tragic–Control

On Saturday, Gregory drives to Connecticut to spend the weekend with his parents. The traffic is very light and Gregory arrives at his parent's house two hours earlier than he had expected to.

Taboo–Tragic

Gregory knows that by implementing the new method he would be forced to lay off a third of his crew due to the related expenses. These people would have a very difficult time finding other jobs.

Taboo–Control

On Saturday, Gregory drives to Connecticut to spend the weekend with his parents. The traffic is very light and Gregory arrives at his parent's house two hours earlier than he had expected to.

Scenario 14				
a)	Sergei is the governor of a small state in an Eastern European country. Sergei is considering whether to pass or veto an amendment banning the death penalty and public executions.			
b)	If Sergei passes the amendment, then the death penalty and public executions will be banned immediately. Based on the estimates of Sergei's staff, this would prevent at least ten executions of innocents per year.			
	Initial Tragic		**Initial Taboo**	
c)	Sergei knows that the state uses the income from tickets sold to public executions. Banning the death penalty would eliminate funding for several ongoing investigations into gang violence, leaving citizens in danger.		The public executions are very popular among Sergei's supporters. Sergei will have a much better chance of reelection if he vetoes the proposal and allows both public executions and the death penalty to continue.	
d)	Sergei thinks very carefully and ultimately decides to veto the proposal. The death penalty is maintained.			
	Tragic–Taboo	**Tragic–Control**	**Taboo–Tragic**	**Taboo–Control**
e)	The public executions are very popular among Sergei's supporters. Sergei will have a much better chance of reelection if he vetoes the proposal and allows both public executions and the death penalty to continue.	Sergei attends a briefing on the accounting department of the local government. The head accountant arrives late, but the meeting still ends early as most of the plans had been discussed previously at a related meeting.	Sergei knows that the state uses the income from tickets sold to public executions. Banning the death penalty would eliminate funding for several ongoing investigations into gang violence, leaving citizens in danger.	Sergei attends a briefing on the accounting department of the local government. The head accountant arrives late, but the meeting still ends early as most of the plans had been discussed previously at a related meeting.

Scenario 15	
a)	Brian is a contestant in a game show in which pairs of participants answer questions for money but are humiliated if they fail. Brian’s partner has just failed the final question, and Brian has the option to “save” him.
b)	If Brian decides to save his partner, his partner will not be dropped into a tank of human excrement, and Brian will end the game with half of the $10,000 he has earned.
c)	**Initial Tragic** Earlier, both Brian and his partner stated their reasons for being on the show. Brian’s cousin was diagnosed with lung cancer, and Brian wants help pay for his treatment. **Initial Taboo** By abandoning his partner to be dropped into the tank of human excrement, Brian will keep all of his money and be given a new Mercedes Benz car as a bonus prize.
d)	Brian thinks carefully and ultimately decides to abandon his partner. Brian’s partner is dropped into the human excrement, and Brian collects his earnings.
e)	**Tragic–Taboo** By abandoning his partner to be dropped into the tank of human excrement, Brian will keep all of his money and be given a new Mercedes Benz car as a bonus prize. **Tragic–Control** Brian is browsing Facebook later that night and notices that a former coworker has recently gotten married. Brian tells his wife who is surprised to hear the news. **Taboo–Tragic** Earlier, both Brian and his partner stated their reasons for being on the show. Brian’s cousin was diagnosed with lung cancer, and Brian wants help pay for his treatment. **Taboo–Control** Brian is browsing Facebook later that night and notices that a former coworker has recently gotten married. Brian tells his wife who is surprised to hear the news.

Scenario 16

a) Cassandra is a member of the transportation board in a large American city. The board is considering the addition of a lane to a dangerous section of the freeway at the edge of the city.

b) Cassandra knows that this stretch of highway is notorious for causing vehicles to lose control, and that there have been fatal accidents year round at it. Adding an additional lane would prevent approximately 50 deaths due to accidents per year.

c)

Initial Tragic

Cassandra recently spoke to the mayor, who told her that the money for the lane would need to come from the education budget. By rejecting the repairs to the highway, Cassandra could prevent the city from having to lay off 100 teachers.

Initial Taboo

Cassandra's husband is a personal injury lawyer who makes most of his income from settling accidents. Approving the additional lane would create less business for Cassandra's husband, decreasing their combined income by a great deal.

d) Cassandra thinks carefully and ultimately decides not to approve the construction of the new highway lane.

e)

Tragic–Taboo

Cassandra's husband is a personal injury lawyer who makes most of his income from settling accidents. Approving the additional lane would create less business for Cassandra's husband, decreasing their combined income by a great deal.

Tragic–Control

When Cassandra arrives home she notices that the sports magazine "NFL Monthly" is in her mailbox. She does not subscribe to this magazine, and has received magazines that were intended for a former tenant of her apartment since she moved in.

Taboo–Tragic

Cassandra recently spoke to the mayor, who told her that the money for the lane would need to come from the education budget. By rejecting the repairs to the highway, Cassandra could prevent the city from having to lay off 100 teachers.

Taboo–Control

When Cassandra arrives home she notices that the sports magazine "NFL Monthly" is in her mailbox. She does not subscribe to this magazine, and has received magazines that were intended for a former tenant of her apartment since she moved in.

<table>
<tr><th>Scenario 17</th><th colspan="4"></th></tr>
<tr><td><b>a)</b></td><td colspan="4">Dr. Robertson is the medical administrator at St. Luke's hospital. The hospital has received a $10,000 donation and she must decide how it will be spent.</td></tr>
<tr><td><b>b)</b></td><td colspan="4">Dr. Robertson could use the $10,000 to save the life of Caleb, an 8 year-old boy who has a rare, but curable, degenerative disease. Caleb's family cannot afford the treatment on their own, and if it were given then Caleb would be cured immediately.</td></tr>
<tr><td></td><td colspan="2"><b>Initial Tragic</b></td><td colspan="2"><b>Initial Taboo</b></td></tr>
<tr><td><b>c)</b></td><td colspan="2">Dr. Robertson knows that many patients who come to St. Luke's hospital cannot afford basic immunizations. Setting up a free immunization clinic with the $10,000 could provide an essential service and save hundreds of lives in the long term.</td><td colspan="2">Dr. Robertson knows that the hospital board has promoted past administrators who implemented new programs. Creating a free immunization clinic with the money would set Dr. Robertson up for promotion.</td></tr>
<tr><td><b>d)</b></td><td colspan="4">Dr. Robertson thinks very carefully and ultimately uses the $10,000 to set up the free immunization clinic.</td></tr>
<tr><td></td><td><b>Tragic–Taboo</b></td><td><b>Tragic–Control</b></td><td><b>Taboo–Tragic</b></td><td><b>Taboo–Control</b></td></tr>
<tr><td><b>e)</b></td><td>Dr. Robertson knows that the hospital board has promoted past administrators who implemented new programs. Creating a free immunization clinic with the money would set Dr. Robertson up for promotion.</td><td>Dr. Robertson later attends a seminar on a database the hospital is implementing. The database will help to coordinate organ transplants with other area hospitals. The meeting runs late and Dr. Robertson arrives home after dark.</td><td>Dr. Robertson knows that many patients who come to St. Luke's hospital cannot afford basic immunizations. Setting up a free immunization clinic with the $10,000 could provide an essential service and save hundreds of lives in the long term.</td><td>Dr. Robertson later attends a seminar on a database the hospital is implementing. The database will help to coordinate organ transplants with other area hospitals. The meeting runs late and Dr. Robertson arrives home after dark.</td></tr>
</table>

Scenario 18

a) Peter is a lieutenant in the United States Army during World War I, returning from a reconnaissance mission. He is behind enemy lines with Arthur, a private from his unit who was injured by a landmine.

b) By taking Arthur back with him Peter would be saving his life. Due to his injury, Arthur cannot walk and would be killed by the Germans if Peter were to leave him.

c)

Initial Tragic	Initial Taboo
Peter knows that a German sneak attack is about to launch. Unless Peter leaves Arthur behind he will not reach the trenches in time to warn and save his unit.	By leaving Arthur behind, Peter will avoid having to pay the money he owes Arthur. The soldiers have been playing cards in the trenches and Peter owes Arthur $1,000.

d) Peter thinks very carefully and ultimately decides to leave Arthur behind.

e)

Tragic–Taboo	Tragic–Control	Taboo–Tragic	Taboo–Control
By leaving Arthur behind, Peter will avoid having to pay the money he owes Arthur. The soldiers have been playing cards in the trenches and Peter owes Arthur $1,000.	Peter comes across an abandoned German truck on his way back to the trenches. He checks if it can be hot-wired, but quickly realizes that the truck is inoperable.	Peter knows that a German sneak attack is about to launch. Unless Peter leaves Arthur behind he will not reach the trenches in time to warn and save his unit.	Peter comes across an abandoned German truck on his way back to the trenches. He checks if it can be hot-wired, but quickly realizes that the truck is inoperable.

Scenario 19

a) Angelo is a treasure hunter in the Amazon, who is returning to camp with his partner Galeno after finding a powerful magical artifact: the mystic ankh. On the path to their camp, a trap knocks Galeno and the ankh into quicksand.

b) Angelo could save Galeno by pulling him from the quicksand. The mystic ankh is sinking, and if Angelo saves Galeno it will be lost forever.

c)

Initial Tragic

The villagers that hired Angelo and Galeno to find the ankh need its magical power to grow their crops. Unless it is returned to them they will all starve.

Initial Taboo

By saving the ankh and letting Galeno sink, Angelo could avoid having to split the pay that was promised by the villagers that hired them.

d) Angelo thinks very carefully and ultimately decides to let Galeno sink .He pulls the mystic ankh from the quicksand.

e)

Tragic–Taboo

By saving the ankh and letting Galeno sink, Angelo could avoid having to split the pay that was promised by the villagers that hired them.

Tragic–Control

On his way back to camp, Angelo sees movement in the jungle. He remains still in case it is a predator, and then continues to the camp.

Taboo–Tragic

The villagers that hired Angelo and Galeno to find the ankh need its magical power to grow their crops. Unless it is returned to them they will all starve.

Taboo–Control

On his way back to camp, Angelo sees movement in the jungle. He remains still in case it is a predator, and then continues to the camp.

Scenario 20				
a)	Justice Adams is a judge in London, England in the year 1878. He is considering his verdict on the case of Vincent. Vincent is the son of a rich merchant and killed a man while he was drunk.			
b)	Justice Adams could convict Vincent. The sentence for murder in cold blood is death. Vincent would be kept in a cell overnight and hanged at dawn the next morning.			
c)	**Initial Tragic** Justice Adams knows that Vincent owns a business, employing 100 men. These men and their families would almost certainly starve if Vincent were convicted, as the business would need to close.		**Initial Taboo** A wealthy colleague of Vincent's offered to pull strings to promote Justice Adams to a superior court. This offer is on the condition that Justice Adams pardons Vincent.	
d)	Justice Adams thinks very carefully and ultimately decides to pardon Vincent.			
e)	**Tragic–Taboo** A wealthy colleague of Vincent's offered to pull strings to promote Justice Adams to a superior court. This offer is on the condition that Justice Adams pardons Vincent.	**Tragic–Control** Justice Adams is scheduled to hear another murder trial after this case. The court cannot find the key witness in time, and the trial is postponed until tomorrow.	**Taboo–Tragic** Justice Adams knows that Vincent owns a business, employing 100 men. These men and their families would almost certainly starve if Vincent were convicted, as the business would need to close.	**Taboo–Control** Justice Adams is scheduled to hear another murder trial after this case. The court cannot find the key witness in time, and the trial is postponed until tomorrow.

Scenario 21

a) Dr. Ingris is a professor at McAdams University. Dr. Ingris employs Eric, a graduate student, and is part of an international project investigating the causes of cancer. A fire has broken out in her lab and when she arrived the lab was filled with smoke.

b) Eric is lying unconscious on the floor. Dr. Ingris could drag her graduate student to safety, saving his life. This would leave the lab server, which stores all of the lab's data, to be destroyed in the fire.

c)

Initial Tragic
The lab server contains five years of research into the causes of cancer. Losing this data would set back progress in treatment by several years and lead to hundreds of patients' needless deaths.

Initial Taboo
Dr. Ingris has stored her own research on the server, which is unrelated to the cancer work. If she left the lab server behind then it could take several more years for her to obtain tenure.

d) Dr. Ingris thinks very carefully and ultimately decides to save the lab server.

e)

Tragic–Taboo
Dr. Ingris has stored her own research on the server, which is unrelated to the cancer work. If she left the lab server behind then it could take several more years for her to obtain tenure.

Tragic–Control
Outside the building, Dr. Ingris learns that the fire began in the lab beside hers after unstable chemicals were left too close to a hotplate. The fire had then spread through the walls into Dr. Ingris' lab.

Taboo–Tragic
The lab server contains five years of research into the causes of cancer. Losing this data would set back progress in treatment by several years and lead to hundreds of patients' needless deaths.

Taboo–Control
Outside the building, Dr. Ingris learns that the fire began in the lab beside hers after unstable chemicals were left too close to a hotplate. The fire had then spread through the walls into Dr. Ingris' lab.

Scenario 22

a) Matthew is 32 years old and lives in Argentina, next door to his brother. He is woken in the night by a tornado warning. There is a storm shelter in the basement of Matthew's apartment building.

b) Matthew's brother is disabled and cannot move without help. Unless Matthew helps him, he will be killed when the tornados hit.

	Initial Tragic	**Initial Taboo**
c)	Matthew's neighbors left their two children home alone for the night. If Matthew saves his brother he will not have time to get them to shelter.	Matthew is listed as the beneficiary in his brother's life insurance. If his brother dies then Matthew will receive a large compensation package.

d) Matthew thinks very carefully and ultimately decides not to save his brother.

	Tragic–Taboo	**Tragic–Control**	**Taboo–Tragic**	**Taboo–Control**
e)	Matthew is listed as the beneficiary in his brother's life insurance. If his brother dies then Matthew will receive a large compensation package.	Days later, Matthew learns that there were five other tornados that night. He hears that international aid might help with the rebuilding effort.	Matthew's neighbors left their two children home alone for the night. If Matthew saves his brother he will not have time to get them to shelter.	Days later, Matthew learns that there were five other tornados that night. He hears that international aid might help with the rebuilding effort.

Scenario 23

a) Carver is a contractor for a private company hired by the Federal Emergency Management Agency (FEMA). Carver is in charge of containing a nuclear meltdown in a major American city.

b) Fixing the nuclear reactor will expose Carver's crew to radiation. By waiting on special equipment for defusal, Carver could guarantee the safety of his crew when they are sent inside to contain the reaction.

c)

Initial Tragic
By sending in his crew immediately, Carver could be sure that the radiation is stopped before it could contaminate the city's water supply. If the water supply is contaminated the city will be uninhabitable for years.

Initial Taboo
Carver's contract makes it clear that future work is only guaranteed if the situation can be resolved quickly. By sending his crew in immediately Carver would bring more business to the company and possibly be promoted.

d) Carver thinks very carefully and ultimately decides not to wait for special equipment, sending his crew in immediately.

e)

Tragic–Taboo
Carver's contract makes it clear that future work is only guaranteed if the situation can be resolved quickly. By sending his crew in immediately Carver would bring more business to the company and possibly be promoted.

Tragic–Control
Carver sends a team to collect radiation readings throughout the city. Hundreds of samples must be collected and sent back to the laboratory. This information will determine how the clean-up proceeds next.

Taboo–Tragic
By sending in his crew immediately, Carver could be sure that the radiation is stopped before it could contaminate the city's water supply. If the water supply is contaminated the city will be uninhabitable for years.

Taboo–Control
Carver sends a team to collect radiation readings throughout the city. Hundreds of samples must be collected and sent back to the laboratory. This information will determine how the clean-up proceeds next.

Scenario 24				
a)	Andrei is 40 years old and lives in Latvia. He is driving home from the market when he comes across the scene of a terrible accident and sees an injured man lying in the street.			
b)	Andrei could drive the man to a hospital. The hospital is a 30-minute drive away. The man would almost certainly survive if he was able to quickly get medical attention.			
	Initial Tragic		**Initial Taboo**	
c)	Andrei lives with and cares for his father, who suffers from Alzheimer's. Andrei left him sleeping at home and unless he returns immediately his father may wake and wander into the street.		The injured man is covered in blood and it will get on the seats if Andrei helps him. Andrei had planned to sell his car soon and this would lower the car's value.	
d)	Andrei thinks very carefully and ultimately decides to leave the injured man.			
	Tragic–Taboo	**Tragic–Control**	**Taboo–Tragic**	**Taboo–Control**
e)	The injured man is covered in blood and it will get on the seats if Andrei helps him. Andrei had planned to sell his car soon and this would lower the car's value.	On the drive home Andrei sees that he is running low on gas. He stops at the next gas station to fill his tank, as he will not pass another until he reaches home.	Andrei lives with and cares for his father, who suffers from Alzheimer's. Andrei left him sleeping at home and unless he returns immediately his father may wake and wander into the street.	On the drive home Andrei sees that he is running low on gas. He stops at the next gas station to fill his tank, as he will not pass another until he reaches home.

Appendix B. By-stimuli best linear unbiased predictors (BLUPs).

Scenario	Description	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	DMPFC	VMPFC	PC	RTPJ	LTPJ
Tragic-Taboo											
1	Drug trial	0.1505	-0.4360	1.4758	-1.8834	-0.8203	0.0575	-0.0217	0.0466	-0.0246	0.1776
2	EPA vote	1.3138	-0.0567	-0.7408	-1.6045	0.7643	-0.0062	0.0565	0.0444	0.0259	0.2055
3	Expanding farm	1.1065	-0.2988	-0.0291	-0.9230	0.8282	-0.0317	0.0714	0.0431	0.0477	0.0828
4	Tiger preservation	1.2608	1.3875	-0.2842	-0.3801	0.8234	-0.0402	0.1437	0.0346	-0.0666	-0.0893
5	Education budget	0.4736	-0.0083	-0.5134	1.4894	2.0359	-0.0041	0.0125	0.0448	-0.0377	0.0354
6	Paid adoption	0.7384	0.8854	0.2269	-1.4888	0.8453	-0.0326	-0.0126	0.1009	0.0404	0.0133
7	Euthanize mother	1.2122	1.0979	0.0683	0.0678	0.6727	-0.0267	0.1570	0.0825	0.0563	-0.0064
8	Pregnant medical condition	0.4783	-0.4699	0.7120	-0.3247	-2.2520	-0.0386	0.0494	0.0023	-0.0088	0.0683
9	Car recall	-0.6043	1.6330	0.7173	-0.5896	2.9266	0.0035	0.0991	0.0698	0.0185	0.0473
10	Military draft	1.8932	-0.1472	-1.0550	1.1406	-0.9889	-0.0418	0.1422	0.1668	0.0382	0.0114
11	Animal shelter	0.4669	0.5320	0.1705	-0.1119	1.4362	-0.0314	-0.0314	-0.0159	-0.0539	0.0225
12	Surrogate mother	-0.4396	-2.2091	0.4868	1.4532	1.3789	-0.0408	0.0711	0.0659	-0.0075	0.0157
13	Fishing captain	0.8879	1.0157	-0.5004	0.6157	1.2384	0.0143	0.0767	0.0465	0.0347	-0.0405
14	Banning death penalty	0.5447	0.7758	0.1707	-0.8851	0.2467	-0.0196	0.0645	0.0807	0.0126	0.0245
15	Humiliating game show	0.8010	-0.0688	0.2105	-0.4218	-1.1703	-0.0090	-0.0077	-0.0001	-0.0526	-0.0576
16	Highway lane	1.6641	1.0280	-0.4369	-0.2916	-1.2172	0.0042	0.1303	0.0787	0.0139	0.0970
17	Hospital donation	0.6934	-0.4719	0.4098	-0.6476	1.1408	-0.0431	0.0813	-0.0372	-0.0026	-0.0339
18	WWI soldier	1.9435	0.8957	-0.9612	1.6938	-1.4412	0.0131	0.1075	-0.0074	-0.0303	0.0278
19	Magic Ankh	0.5815	1.3557	-0.3110	-0.0032	0.9704	-0.0297	0.1482	0.0647	0.0037	0.1213
20	London judge	1.2172	1.3089	0.5656	0.6588	-1.2748	-0.0410	0.1452	0.0970	0.0782	0.0193
21	Lab fire	-0.0811	0.5218	0.6902	1.1457	2.4347	-0.0425	0.0445	-0.0026	-0.0622	0.1697
22	Tornado shelter	1.0566	1.2399	0.7360	0.8376	-0.4485	-0.0181	0.0916	-0.0100	-0.0422	0.0631

23	Nuclear meltdown	1.1939	0.2574	0.0427	0.0683	-0.0597	-0.0255	0.0370	0.0172	-0.0810	-0.0268
24	Road accident	1.0657	1.3433	0.4913	-0.4635	-1.7822	-0.0084	0.0850	-0.0080	-0.0171	-0.0186
Taboo-Tragic											
1	Drug trial	-0.0631	-1.2230	1.8706	-1.1677	1.0138	-0.1139	-0.1071	-0.1761	-0.0552	-0.1405
2	EPA vote	1.1647	-1.2069	0.3010	-0.5339	1.5040	-0.0791	-0.0142	-0.0137	-0.0484	-0.0994
3	Expanding farm	1.4583	-0.9820	-0.0448	0.5873	-0.8396	-0.1300	0.0229	-0.1272	-0.0905	-0.0408
4	Tiger preservation	0.8268	0.5238	0.7415	-0.5780	-0.5949	0.1202	0.1894	0.1525	0.0324	0.0523
5	Education budget	0.1923	-1.0728	0.6032	0.8111	1.6277	-0.0592	-0.0619	-0.1597	-0.0574	-0.0515
6	Paid adoption	0.2809	0.1996	0.7858	-1.3206	0.1995	-0.1249	-0.0569	0.1138	0.0712	-0.0095
7	Euthanize mother	1.7293	0.3770	0.6874	-0.9022	-0.7487	-0.0839	0.1832	0.2050	0.0158	0.0265
8	Pregnant medical condition	0.5893	-1.3522	0.8952	-0.6139	0.4928	0.0047	0.0223	-0.1013	-0.0209	-0.0648
9	Car recall	-0.0193	0.7778	1.1429	-0.3560	0.7429	-0.0372	0.1113	0.1423	0.0340	0.0282
10	Military draft	1.7214	-1.4300	-0.9208	0.4151	0.6987	-0.0251	0.1559	0.1202	0.0197	-0.0029
11	Animal shelter	1.0544	-0.2153	0.5983	-0.6826	-0.6921	-0.1143	-0.0981	-0.0894	-0.0431	-0.1231
12	Surrogate mother	-0.4022	-2.4092	2.0058	0.1392	0.5257	-0.0701	0.0710	0.1141	0.0019	-0.0247
13	Fishing captain	1.0864	0.3505	-0.1250	-0.9104	0.0671	0.0850	0.0915	0.1158	0.0377	0.0452
14	Banning death penalty	0.9169	0.0660	0.1776	0.0521	-0.3610	-0.0192	0.0608	0.0781	-0.0180	0.0048
15	Humiliating game show	1.3239	-1.2947	1.2102	0.2900	-0.4405	-0.0524	-0.0286	0.1909	-0.0094	-0.0123
16	Highway lane	1.5365	-0.4764	-0.8424	-0.6419	0.8150	-0.0514	0.1177	0.0813	0.0812	0.0118
17	Hospital donation	0.4696	-1.3582	0.9436	-0.6026	1.0051	0.0242	0.1139	0.0956	0.1088	-0.0075
18	WWI soldier	1.7786	-0.5818	-0.2867	0.9575	0.6662	-0.0987	0.1276	0.1393	0.1290	-0.0439
19	Magic Ankh	1.4679	0.1012	-0.4153	1.2791	-1.5258	0.0180	0.1823	0.2813	0.1465	-0.0426
20	London judge	1.0933	0.0594	0.3537	0.3139	-0.6779	-0.0409	0.1750	0.2572	0.0573	0.0773
21	Lab fire	0.4096	-0.3649	0.9469	1.4535	0.6915	-0.1014	-0.0093	0.1271	-0.0664	-0.0934
22	Tornado shelter	1.4710	-0.2312	0.2910	1.1000	0.0186	-0.1620	0.1043	0.1253	0.0322	0.0108

23	Nuclear meltdown	1.3189	-1.0155	0.2529	1.1491	0.3950	-0.0323	-0.0088	-0.0032	-0.0320	-0.0403
24	Road accident	1.0253	0.4156	0.3088	-0.7112	0.4637	0.0037	0.0895	0.1921	0.0816	0.0576
Tragic-Control											
1	Drug trial	-0.6985	-0.9015	-0.0926	-1.5469	-1.4533	-0.0536	0.0203	-0.0070	-0.0129	0.1015
2	EPA vote	-0.7260	-1.1220	-1.6364	-0.7380	0.7167	-0.106	0.0195	-0.0097	-0.0174	-0.0857
3	Expanding farm	-0.7939	-0.9507	-0.2856	0.5583	-0.5454	-0.1403	0.0091	0.0018	0.0891	-0.0960
4	Tiger preservation	-0.6249	0.0972	-1.3332	-0.5912	-0.0064	-0.0971	-0.0362	-0.0250	-0.0152	-0.0199
5	Education budget	-0.6605	-1.5528	-2.6268	1.0609	1.0821	-0.1000	0.0169	-0.0034	-0.0047	-0.0251
6	Paid adoption	-0.8336	0.1490	0.1414	-1.5735	-0.5885	-0.1400	0.0036	-0.0893	-0.0590	-0.0635
7	Euthanize mother	-0.6596	0.3054	-1.2926	-1.0193	0.0268	-0.1261	-0.0184	-0.0755	0.0132	-0.1051
8	Pregnant medical condition	-0.4861	-1.1289	-0.6679	-0.7612	-2.4576	-0.1191	-0.0033	0.0353	0.0455	0.0097
9	Car recall	-0.6992	1.2571	-1.3774	-0.2444	0.4159	-0.0884	-0.0170	-0.0597	-0.0160	-0.1420
10	Military draft	-0.7533	-1.7831	-1.2257	0.4896	-0.4508	-0.1283	-0.0178	-0.1650	-0.0573	-0.0875
11	Animal shelter	-0.6952	-0.0320	-1.1454	-0.8914	-0.0772	-0.1368	0.0084	0.0422	0.0560	0.142
12	Surrogate mother	-0.6942	-2.4295	-0.6822	-0.2672	-0.6096	-0.1366	-0.0134	-0.0515	0.0150	0.0762
13	Fishing captain	-0.6756	0.2381	-1.5725	-0.8937	0.1691	-0.0534	-0.0248	-0.0302	-0.0095	0.0203
14	Banning death penalty	-0.5001	-0.0726	-1.6578	0.1097	-0.8410	-0.1063	-0.0135	-0.0660	-0.0153	-0.0801
15	Humiliating game show	-0.4024	-0.9692	-2.5167	0.2576	-1.1810	-0.1032	-0.0072	-0.0062	0.0145	-0.0126
16	Highway lane	-0.6596	-0.7933	-1.1230	-0.8074	-0.9267	-0.0907	0.0002	-0.0586	0.0144	-0.1450
17	Hospital donation	-1.0643	-1.7030	0.3133	-0.7277	0.3924	-0.1194	-0.0268	0.0535	0.0408	0.0840
18	WWI soldier	-0.6799	-0.8698	-0.7763	1.3878	-0.7784	-0.0920	-0.0167	0.0133	-0.0059	0.0132
19	Magic Ankh	-0.8677	0.3772	0.0694	1.7133	-1.1921	-0.1082	-0.0171	-0.0676	-0.0834	-0.1679

20	London judge	-0.7670	-0.1495	0.1196	0.3012	-1.8022	-0.1307	-0.0244	-0.0956	0.0165	-0.1808
21	Lab fire	-1.2216	0.0228	1.0551	1.8310	0.6076	-0.1445	0.0082	0.0159	0.0334	-0.0935
22	Tornado shelter	-1.0541	-0.4750	1.3668	1.4458	-1.0586	-0.1341	-0.0142	0.0264	0.0067	-0.1542
23	Nuclear meltdown	-1.1247	-1.0083	2.1975	1.2473	-0.9295	-0.1145	0.0028	0.0019	0.0237	-0.0117
24	Road accident	-0.7794	0.9166	0.9790	-0.6107	-1.9420	-0.0912	-0.0113	0.0125	0.0398	-0.1776
Taboo-Control											
1	Drug trial	-1.3640	-0.3846	1.7239	-2.0113	-0.2000	0.1299	0.0941	-0.0123	-0.0018	-0.0082
2	EPA vote	-0.8701	0.5313	-0.9949	-1.2952	1.0218	-0.0210	0.0279	-0.0473	-0.0117	-0.0424
3	Expanding farm	-0.8441	-0.1110	-0.0577	0.1815	-0.0830	-0.0299	-0.0193	-0.0370	-0.0282	-0.0520
4	Tiger preservation	-0.9362	0.9683	-0.2946	-0.4567	0.2942	-0.2453	-0.2571	-0.0459	0.0390	-0.0462
5	Education budget	-0.8883	0.2145	-1.7888	1.4702	1.1502	-0.0328	0.0446	-0.0209	0.0009	-0.0421
6	Paid adoption	-1.0814	0.6554	0.2642	-1.7908	0.2729	-0.0356	0.1103	-0.0201	0.0255	-0.0496
7	Euthanize mother	-1.0723	0.7580	0.4278	-0.4723	0.3255	-0.0568	-0.1012	-0.0410	0.0046	-0.0580
8	Pregnant medical condition	-0.8937	-0.4822	-0.0291	-0.7022	-0.8119	-0.1503	-0.0232	-0.0414	0.0084	-0.0327
9	Car recall	-1.1016	1.7166	-0.0322	-0.0234	0.7282	-0.0357	-0.0622	-0.0337	0.0191	-0.0609
10	Military draft	-0.6603	-0.0245	-1.8691	0.7256	0.3427	-0.1329	-0.1703	0.0168	0.0080	-0.0544
11	Animal shelter	-0.9864	0.8816	-0.1790	-0.6675	0.1199	-0.0418	0.0444	-0.0278	0.0118	-0.0125
12	Surrogate mother	-1.2639	-2.1233	0.3500	0.3827	0.4387	-0.0952	-0.0345	-0.0358	0.0151	-0.0213
13	Fishing captain	-0.9061	0.9887	-1.2438	-0.3465	0.6290	-0.1120	-0.1462	-0.0456	0.0168	-0.0339
14	Banning death penalty	-0.7499	0.6222	-0.9922	-0.1421	-0.0542	-0.0946	-0.0543	-0.0222	0.0023	-0.0513
15	Humiliating game show	-0.8462	-0.1156	-0.5139	-0.0687	-0.4605	-0.0477	0.0891	-0.0351	0.0220	-0.0454
16	Highway lane	-0.6727	1.4279	-1.0795	-0.9316	-0.1423	-0.0229	-0.0093	-0.0344	0.0379	-0.0580

17	Hospital donation	-1.2777	-0.0518	0.9815	-0.9935	0.5974	-0.1746	-0.0323	-0.0643	0.0524	-0.0231
18	WWI soldier	-0.7531	1.5042	-0.8756	1.8387	-0.1537	0.0319	-0.0021	-0.0470	0.0646	-0.0343
19	Magic Ankh	-0.8792	1.4683	-0.6540	1.6701	-0.5686	-0.1437	0.0231	-0.0475	0.0600	-0.0637
20	London judge	-1.0591	0.7814	0.7341	0.6548	-0.7202	-0.1188	-0.1637	-0.0399	0.0141	-0.0687
21	Lab fire	-1.4976	0.8465	1.6413	2.3171	0.7233	-0.0734	-0.0204	-0.0570	0.0049	-0.0468
22	Tornado shelter	-1.2156	1.6833	1.9677	1.7229	-0.5693	0.0218	0.0191	-0.0647	0.0338	-0.0630
23	Nuclear meltdown	-1.0596	0.6280	1.5080	1.0941	-0.1835	-0.0956	-0.0407	-0.0273	0.0211	-0.0441
24	Road accident	-1.0491	1.4266	0.9526	-0.5648	-0.6008	-0.0912	0.0189	-0.0590	0.0463	-0.0724

Supplemental Materials

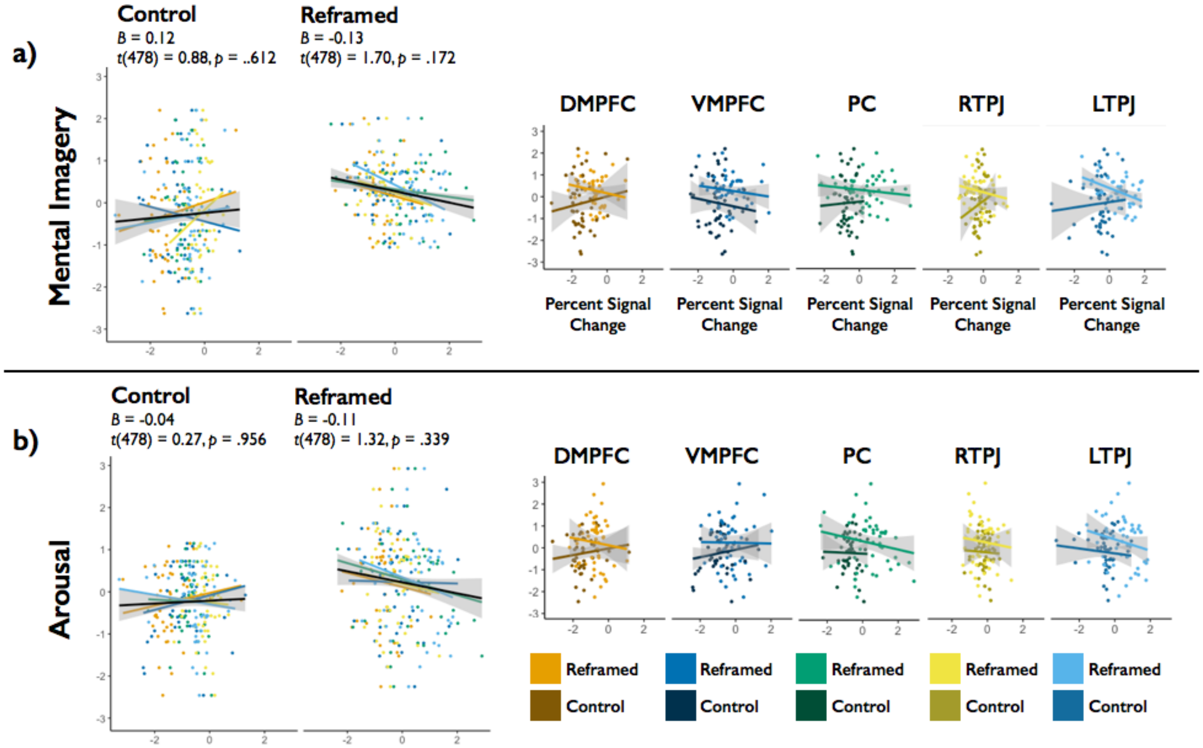


Figure S1. Additional ToMN behavioral component scores relationships. Main effects of ToMN activity (left) and their breakdown by ROI (right). Interactions with ROI were not significant, but are presented for ease of interpretation. Neither mental imagery (a) nor arousal (b) was related to ToMN activity in either reframed (i.e. tragic–taboo/taboo–tragic), or control (i.e. tragic–control/taboo–control) scenarios. Component scores were varimax rotated. Relationships for person-based prediction error, moral/normative prediction error, and descriptive prediction err are reported in the main paper (Figure 3).

Table S1. All Study 1 Principal Components Analyses.

2-Factor PCA

Measure	**Factor 1**	**Factor 2**	**Communality**
Mental Imagery	0.71		0.5
Mental State Inference	0.94		0.89
Valence		0.89	0.79
Arousal	0.71	-0.40	0.66
Descriptive Frequency		0.47	0.23
Prescriptive Norm Violation		-0.94	0.89
Belief Violation	0.96		0.93
Desire Violation	0.93		0.86
Impression Violation	0.95		0.91
Moral Judgment		0.94	0.89
Component loading	4.61	2.94	
Proportional variance	0.46	0.29	
Cumulative variance	0.46	0.76	

3-Factor PCA

Measure	**Factor 1**	**Factor 2**	**Factor 3**	**Communality**
Mental Imagery	0.71			0.51
Mental State Inference	0.94			0.89
Valence		0.93		0.87
Arousal	0.72	-0.39		0.67
Descriptive Frequency		0.20	0.97	0.98
Prescriptive Norm Violation		-0.88	-0.33	0.89
Belief Violation	0.96			0.93
Desire Violation	0.92			0.86
Impression Violation	0.95			0.91
Moral Judgment		0.96		0.93
Component loading	4.61	2.77	1.07	
Proportional variance	0.46	0.28	0.11	
Cumulative variance	0.46	0.74	0.85	

4-Factor PCA

Measure	Factor 1	Factor 2	Factor 3	Factor 4	Communality
Mental Imagery	0.42		0.89		0.97
Mental State Inference	0.84		0.44		0.91
Valence		0.93			0.87
Arousal	0.65	-0.39	0.31		0.68
Descriptive Frequency		0.20		0.97	0.99
Prescriptive Norm Violation		-0.88		-0.33	0.89
Belief Violation	0.96				0.96
Desire Violation	0.97				0.94
Impression Violation	0.96				0.95
Moral Judgment		0.96			0.93
Component loading	4.12	2.76	1.14	1.07	
Proportional variance	0.41	0.28	0.11	0.11	
Cumulative variance	0.41	0.69	0.80	0.91	

5-Factor PCA

Measure	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Communality
Mental Imagery	0.41		0.90			0.99
Mental State Inference	0.82		0.44		0.21	0.91
Valence		0.94				0.89
Arousal	0.50	-0.27			0.80	1.00
Descriptive Frequency				0.97		0.99
Prescriptive Norm Violation		-0.87		-0.34		0.89
Belief Violation	0.95					0.97
Desire Violation	0.97					0.96
Impression Violation	0.96					0.96
Moral Judgment		0.96				0.94
Component loading	2.85	2.68	1.11	1.08	0.78	
Proportional variance	0.38	0.27	0.11	0.11	0.08	
Cumulative variance	0.38	0.65	0.76	0.87	0.95	

6-Factor PCA							
Measure	**Fact or 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**	**Factor 6**	**Communality**
Mental Imagery	0.40		0.90				0.99
Mental State Inference	0.81		0.44		0.22		0.92
Valence		0.92				0.36	0.99
Arousal	0.49	-0.27			0.80		1.00
Descriptive Frequency		0.20		0.98			1.00
Prescriptive Norm Violation		-0.88		-0.30		0.26	0.96
Belief Violation	0.95						0.97
Desire Violation	0.96						0.96
Impression Violation	0.96						0.96
Moral Judgment		0.96					0.95
Component loading	3.84	2.68	1.11	1.06	0.79	0.22	
Proportional variance	0.38	0.27	0.11	0.11	0.08	0.02	
Cumulative variance	0.38	0.65	0.76	0.87	0.95	0.97	

PCA was performed on by-stimuli BLUPs, extracted from Study 1 models for each measure. Factor loadings with an absolute value < .2 are omitted from tables for ease of interpretation. Principal components were varimax rotated.

Table S2. ToMN ROI peak coordinates.

Region	**Subjects**	**x**	**y**	**z**	**k**
		M (SD)	***M (SD)***	***M (SD)***	***M (SD)***
DMPFC	17/20	1.5 (4.1)	56.0 (4.6)	25.9 (7.6)	123 (78.4)
VMPFC	16/20	1.4 (3.6)	55.2 (4.8)	-8.1 (6.3)	133.8 (74.1)
PC	19/20	0.9 (4.3)	-56.0 (5.2)	36.4 (5.2)	247.3 (80.8)
RTPJ	20/20	50.8 (6.1)	-53.4 (4.5)	25.4 (5.3)	263.4 (83.7)
LTPJ	19/20	49.7 (6.8)	57.1 (5.3)	26.9 (4.8)	220.2 (83.5)

Table S3. ToMN ROI models.

DMPFC

Model:

PSC ~ InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch
+ (1+ InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch | Subject)
+ (0 + Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch | Scenario)

REML criterion at convergence: 156.7

Epoch: Dummy coded (first pass = 0; second pass = 1)
InitialCond: Contrast coded (initial tragic = -0.5; initial taboo = +0.5)
ReframeCond: Contrast coded (reframed = -0.5; control = +0.5)

Random effects structure (by-subject)

	Variance	**St.Dev**	**Correlations**					
Intercept	0.009	0.097	-					
InitialCond	0.001	0.031	-0.35	-				
Epoch	0.010	0.101	-0.64	0.65	-			
InitialCond*Epoch	0.004	0.064	-0.69	0.82	0.51	-		
ReframeCond*Epoch	0.009	0.097	0.57	-0.42	-0.35	-0.71	-	
InitialCond*ReframeCond*Epoch	0.016	0.128	0.12	-0.31	0.21	-0.39	-0.26	-

Random effects structure (by-stimuli)

	Variance	**St.Dev**	**Correlations**		
Epoch	0.002	0.042	-		
ReframeCond*Epoch	0.007	0.085	-0.16	-	
InitialCond*ReframeCond*Epoch	0.043	0.208	-0.09	1.00	-

Residual

Variance	**St.Dev**
0.062	0.248

Fixed effects	***B* (SE)**	***t*(df)**	***p***
Intercept	0.10 (0.03)	t(16.3) = 3.59	.002 **
InitialCond	-0.02 (0.03)	t(74.5) = -0.61	.546
Epoch	-0.16 (0.03)	t(18.0) = -5.00	< .001 ***
InitialCond*Epoch	0.02 (0.04)	t(42.4) = 0.56	.575
ReframeCond*Epoch	0.06 (0.04)	t(18.8) = 1.48	.156
InitialCond*ReframeCond*Epoch	-0.07 (0.07)	t(19.4) = -0.97	.345

VMPFC

Model:

PSC ~ InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch
+ (1 + InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch | Subject)
+ (0 + InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch | Scenario)

REML criterion at convergence: 503.1

Epoch: Dummy coded (first pass = 0; second pass = 1)
InitialCond: Contrast coded (initial tragic = -0.5; initial taboo = +0.5)
ReframeCond: Contrast coded (reframed = -0.5; control = +0.5)

Random effects structure (by-subject)

	Variance	**St.Dev**	**Correlations**					
Intercept	0.018	0.134	-					
InitialCond	0.010	0.099	0.87	-				
Epoch	0.035	0.187	-0.89	-0.89	-			
InitialCond*Epoch	0.009	0.093	-0.49	-0.70	0.83	-		
ReframeCond*Epoch	0.003	0.051	0.19	-0.23	-0.22	-0.06	-	
InitialCond*ReframeCond*Epoch	0.011	0.103	-0.81	-0.91	0.66	0.35	0.42	-

Random effects structure (by-stimuli)

	Variance	**St.Dev**	**Correlations**				
InitialCond	0.015	0.122	-				
Epoch	0.001	0.036	0.64	-			
InitialCond*Epoch	0.045	0.212	-0.93	-0.67	-		
ReframeCond*Epoch	0.012	0.110	-0.91	-0.36	0.72	-	
InitialCond*ReframeCond*Epoch	0.046	0.215	0.23	0.03	0.13	-0.53	-

Residual	**Variance**	**St.Dev**
	0.102	0.320

Fixed effects	***B* (SE)**	***t*(df)**	***p***
Intercept	0.09 (0.04)	t(15.0) = 2.34	0.034 *
InitialCond	-0.03 (0.05)	t(23.6) = -0.54	0.594
Epoch	-0.11 (0.05)	t(15.6) = -2.01	0.063 †
InitialCond*Epoch	-0.0006 (0.07)	t(23.2) = -0.01	0.993
ReframeCond*Epoch	0.06 (0.04)	t(23.9) = 1.33	0.195
InitialCond*ReframeCond*Epoch	-0.06 (0.08)	t(24.8) = -0.75	0.461

PC

Model:
PSC ~ InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch
+ (1 + InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch | Subject)
+ (1 + InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch | Scenario)

REML criterion at convergence: 8

Epoch: Dummy coded (first pass = 0; second pass = 1)
InitialCond: Contrast coded (initial tragic = -0.5; initial taboo = +0.5)
ReframeCond: Contrast coded (reframed = -0.5; control = +0.5)

Random effects structure (by-subject)

	Variance	**St.Dev**	**Correlations**					
Intercept	0.002	0.043	-					
InitialCond	0.005	0.068	-0.15	-				
Epoch	0.014	0.117	-0.42	-0.81	-			
InitialCond*Epoch	0.004	0.062	-0.09	-0.74	0.60	-		
ReframeCond*Epoch	0.003	0.053	0.57	-0.79	0.31	0.76	-	
InitialCond*ReframeCond*Epoch	0.015	0.122	0.06	-0.98	0.88	0.65	0.67	-

Random effects structure (by-stimuli)

	Variance	**St.Dev**	**Correlations**					
Intercept	0.002	0.042	-					
InitialCond	0.024	0.155	0.04	-				
Epoch	0.001	0.028	-0.65	-0.70	-			
InitialCond*Epoch	0.045	0.212	0.27	-0.95	0.50	-		
ReframeCond*Epoch	0.013	0.112	0.81	-0.25	-0.17	0.51	-	
InitialCond*ReframeCond*Epoch	0.030	0.175	-0.14	-0.98	0.81	0.91	0.26	-

Residual	**Variance**	**St.Dev**
	0.05	0.22

Fixed effects	***B* (SE)**	***t*(df)**	***p***
Intercept	0.12 (0.02)	t(21.4) = 7.22	<.001
InitialCond	-0.01 (0.04)	t(26.2) = -0.28	0.785
Epoch	-0.11 (0.03)	t(18.5) = -3.46	0.003 **
InitialCond*Epoch	0.02 (0.05)	t(23.2) = 0.43	0.674
ReframeCond*Epoch	0.09 (0.03)	t(22.3) = 2.69	0.013 *
InitialCond*ReframeCond*Epoch	0.05 (0.06)	t(25.2) = 0.76	0.456

RTPJ

Model:

PSC ~ InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch
+ (1 + InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch | Subject)
+ (0 + InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch | Scenario)

REML criterion at convergence: -262.8

Epoch: Dummy coded (first pass = 0; second pass = 1)
InitialCond: Contrast coded (initial tragic = -0.5; initial taboo = +0.5)
ReframeCond: Contrast coded (reframed = -0.5; control = +0.5)

Random effects structure (by-subject)								
	Variance	**St.Dev**	**Correlations**					
Intercept	0.005	0.072	-					
InitialCond	<0.001	0.022	-0.22	-				
Epoch	0.012	0.108	-0.77	-0.42	-			
InitialCond*Epoch	<0.001	0.019	-0.80	-0.41	0.99	-		
ReframeCond*Epoch	0.004	0.067	0.30	0.82	-0.74	-0.76	-	
InitialCond*ReframeCond*Epoch	0.003	0.055	0.44	-0.96	0.17	0.16	-0.72	-

Random effects structure (by-stimuli)							
	Variance	**St.Dev**	**Correlations**				
InitialCond	0.006	0.080	-				
Epoch	0.001	0.033	-0.59	-			
InitialCond*Epoch	0.021	0.145	-0.93	0.43	-		
ReframeCond*Epoch	0.005	0.070	-0.09	0.56	0.25	-	
InitialCond*ReframeCond*Epoch	0.006	0.078	-0.78	0.31	0.56	-0.50	-

Residual	**Variance**	**St.Dev**
	0.038	0.195

Fixed effects	***B* (SE)**	***t*(df)**	***p***
Intercept	0.06 (0.02)	t(18.9) = 2.96	0.008 **
InitialCond	0.01 (0.02)	t(25.8) = 0.45	0.655
Epoch	-0.05 (0.03)	t(20.8) = -1.63	0.118
InitialCond*Epoch	0.01 (0.04)	t(22.8) = 0.18	0.860
ReframeCond*Epoch	-0.01 (0.03)	t(21.5) = -0.22	0.831
InitialCond*ReframeCond*Epoch	0.01 (0.04)	t(36.1) = 0.17	0.867

LTPJ

Model:

PSC ~ InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch
+ (1 + InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch | Subject)
+ (0 + InitialCond + Epoch + InitialCond*Epoch + ReframeCond*Epoch + InitialCond*ReframeCond*Epoch | Scenario)

REML criterion at convergence: -40.2

Epoch: Dummy coded (first pass = 0; second pass = 1)
InitialCond: Contrast coded (initial tragic = -0.5; initial taboo = +0.5)
ReframeCond: Contrast coded (reframed = -0.5; control = +0.5)

Random effects structure (by-subject)								
	Variance	**St.Dev**	**Correlations**					
Intercept	0.008	0.089	-					
InitialCond	0.003	0.05	-0.29	-				
Epoch	0.008	0.092	-0.44	-0.56	-			
InitialCond*Epoch	0.003	0.051	-0.07	-0.89	0.56	-		
ReframeCond*Epoch	0.007	0.082	0.81	-0.33	0.03	-0.13	-	
InitialCond*ReframeCond*Epoch	0.016	0.128	0.07	-0.09	0.56	-0.22	0.64	-

Random effects structure (by-stimuli)							
	Variance	**St.Dev**	**Correlations**				
InitialCond	0.010	0.102	-				
Epoch	0.001	0.037	-0.39	-			
InitialCond*Epoch	0.021	0.144	-0.99	0.40	-		
ReframeCond*Epoch	0.020	0.140	0.10	0.41	-0.19	-	
InitialCond*ReframeCond*Epoch	0.017	0.131	0.27	0.01	-0.38	0.91	-

Residual	**Variance**	**St.Dev**
	0.047	0.217

Fixed effects	***B* (SE)**	***t*(df)**	***p***
Intercept	0.14 (0.02)	t(18.1) = 6.20	<.001 ***
InitialCond	0.003 (0.03)	t(23.3) = 0.08	0.934
Epoch	-0.12 (0.03)	t(19.7) = -4.38	<.001 ***
InitialCond*Epoch	-0.02 (0.04)	t(23.2) = -0.45	0.66
ReframeCond*Epoch	0.08 (0.04)	t(26.3) = 2.08	0.048 *
InitialCond*ReframeCond*Epoch	0.01 (0.06)	t(20.7) = 0.13	0.896

St.Dev = standard deviation. *p* values for fixed effects were calculated using the Satterthwaite approximation of degrees of freedom, implemented in the *lmerTest* package (Kuznetsova et al., 2015). *** *p* < .001; ** *p* < .01; * *p* < .05; † *p* < .10

Table S4. Study 2 omnibus tests for theory of mind network activity–behavioral component score analysis

		Factor 1		**Factor 2**		**Factor 3**		**Factor 4**		**Factor 5**	
		Person-based prediction error		Moral/normative prediction error		Descriptive prediction error		Mental imagery		Arousal	
	df	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*
InitialCond	*F*(1, 456)	2.18	0.140	0.63	0.429	31.27	< .001 ***	0.26	0.614	11.65	0.001**
ReframeCond	*F*(1, 456)	1721.97	< .001 ***	0.44	0.506	45.75	< .001 ***	0.36	0.551	29.35	< .001 ***
PSC	*F*(1, 456)	4.69	0.031 *	5.28	0.022 *	1.24	0.266	1.03	0.311	1.60	0.207
InitialCond* ReframeCond	*F*(1, 456)	19.06	< .001 ***	178.31	< .001 ***	1.38	0.240	0.15	0.699	16.60	< .001 ***
PSC* ReframeCond	*F*(1, 456)	4.45	0.036 *	9.94	0.002 **	2.64	0.105	1.25	0.264	1.85	0.174
PSC* InitialCond	*F*(1, 456)	0.08	0.774	2.21	0.138	0.03	0.852	0.16	0.688	0.98	0.323
PSC*ROI	*F*(4, 456)	0.70	0.590	0.50	0.733	0.06	0.994	0.52	0.719	0.18	0.950
PSC* InitialCond* ReframeCond	*F*(1, 456)	1.25	0.264	0.01	0.914	1.18	0.278	0.12	0.733	0.20	0.657
PSC* ReframeCond*ROI	*F*(4, 456)	1.35	0.250	0.53	0.710	0.80	0.525	0.59	0.668	0.15	0.963
PSC* InitialCond*ROI	*F*(4, 456)	0.20	0.939	0.68	0.609	0.09	0.985	0.28	0.892	0.12	0.975
PSC* InitialCond* ReframeCond*ROI	*F*(4, 456)	0.44	0.782	1.14	0.336	0.65	0.629	0.24	0.916	0.21	0.931

p values were calculated using the Satterthwaite approximation of degrees of freedom, implemented in the *lmerTest* package (Kuznetsova et al., 2015).
*** *p* < .001; ** *p* < .01; * *p* < .05; † *p* < .10

Part 2

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

Jordan Theriault, Adam Waytz, Larisa Heiphetz, Liane Young

Published as:

Theriault, J., Waytz, A., Heiphetz, L., & Young, L. (in press). Examining overlap in behavioral and neural representations of morals, facts, and preferences. *Journal of Experimental Psychology: General*. http://dx.doi.org/10.1037/xge0000350

Abstract

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

Introduction

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

[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).

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, & 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.

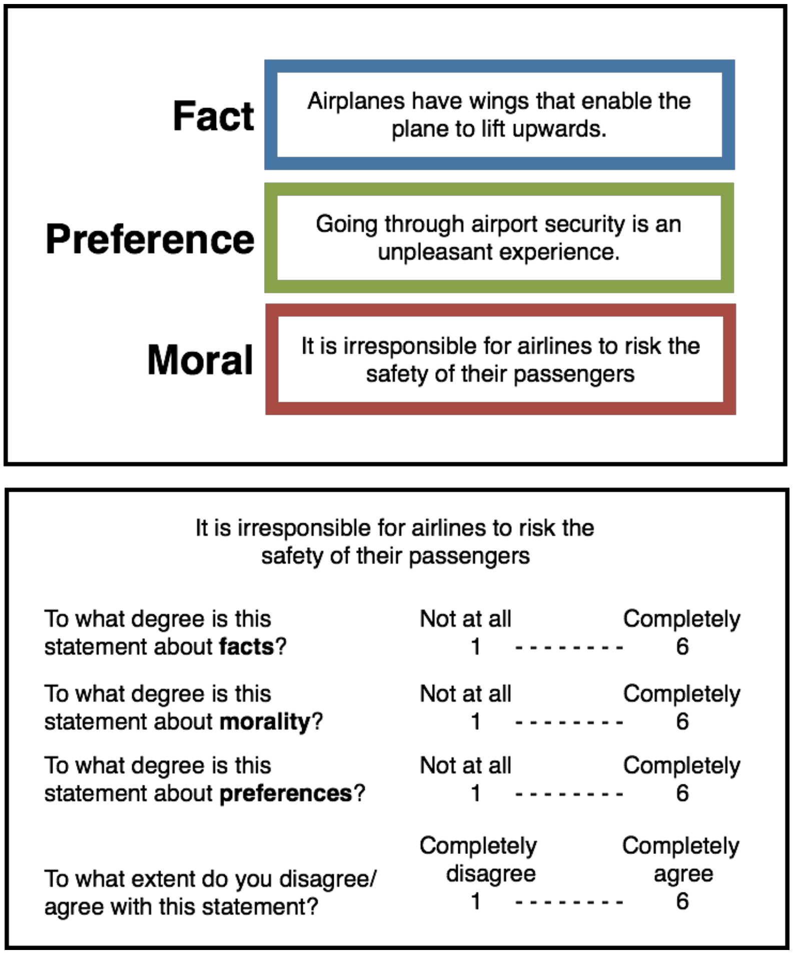
Our central question concerns how people represent moral information (regardless of what makes it moral), and the work that bears the most direct relevance may lead one to predict that people represent morals as objective (i.e. fact-like). The distinction between morals as either fact-like or preference-like 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?



Fact — Airplanes have wings that enable the plane to lift upwards.

Preference — Going through airport security is an unpleasant experience.

Moral — It is irresponsible for airlines to risk the safety of their passengers

It is irresponsible for airlines to risk the safety of their passengers

To what degree is this statement about **facts**?	Not at all 1	Completely 6
To what degree is this statement about **morality**?	Not at all 1	Completely 6
To what degree is this statement about **preferences**?	Not at all 1	Completely 6
To what extent do you disagree/ agree with this statement?	Completely disagree 1	Completely agree 6

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, $M_{\textit{Fact}\text{: Fact-like}} = 5.46$, than moral-like, $M_{\textit{Fact}\text{: Moral-like}} = 1.13$, or preference-like, $M_{\textit{Fact}\text{: Preference-like}} = 1.29$, *t*s > 33, *p*s < .001, *d*s > 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, $M_{\text{Preference: Moral-like}} = 1.21$, *t*s > 36, *p*s < .001, *d*s > 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.

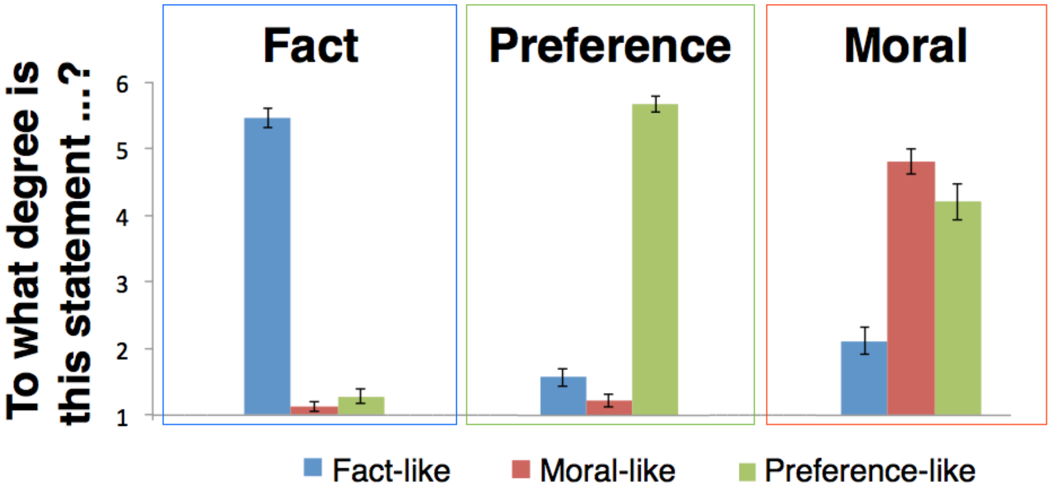


Figure 2. Behavioral Ratings. Claims were rated highest on their content-consistent dimension (e.g., facts were rated as fact-like), but morals were also rated as more preference-like than fact-like. Error bars indicate 95% confidence intervals. For estimates derived from the mixed effects analysis,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

[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?).

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).

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.

[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.

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).

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.

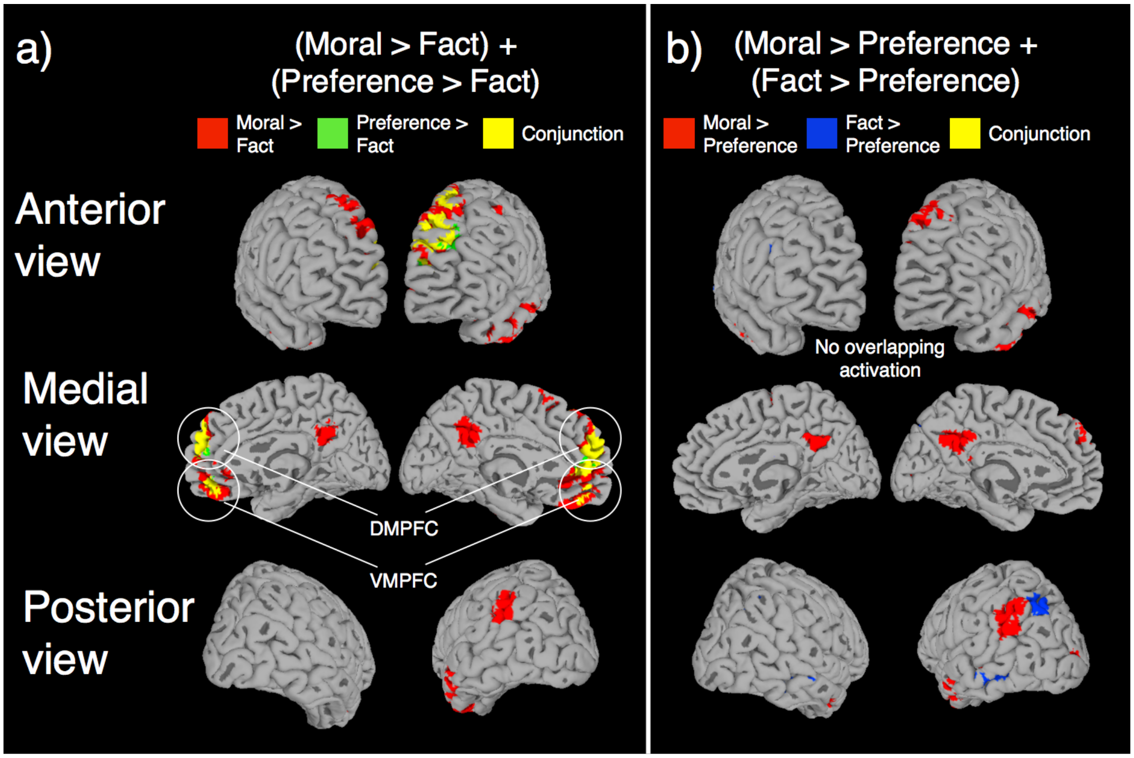


Figure 3. Whole-brain Conjunction Analyses. (a) Morals and preferences, relative to facts, elicited common activity in DMPFC and VMPFC. (b) Morals and facts, relative to preferences, did not elicit any activity in common. Permutation tests (5000 samples) were used to achieve a cluster-corrected familywise 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 ($N_{\text{DMPFC}} = 20/25$; $N_{\text{VMPFC}} = 20/25$; $N_{\text{PC}} = 23/25$; $N_{\text{RTPJ}} = 25/25$; $N_{\text{LTPJ}} = 24/25$). 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 ($p_{\text{corrected}}$ = .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.

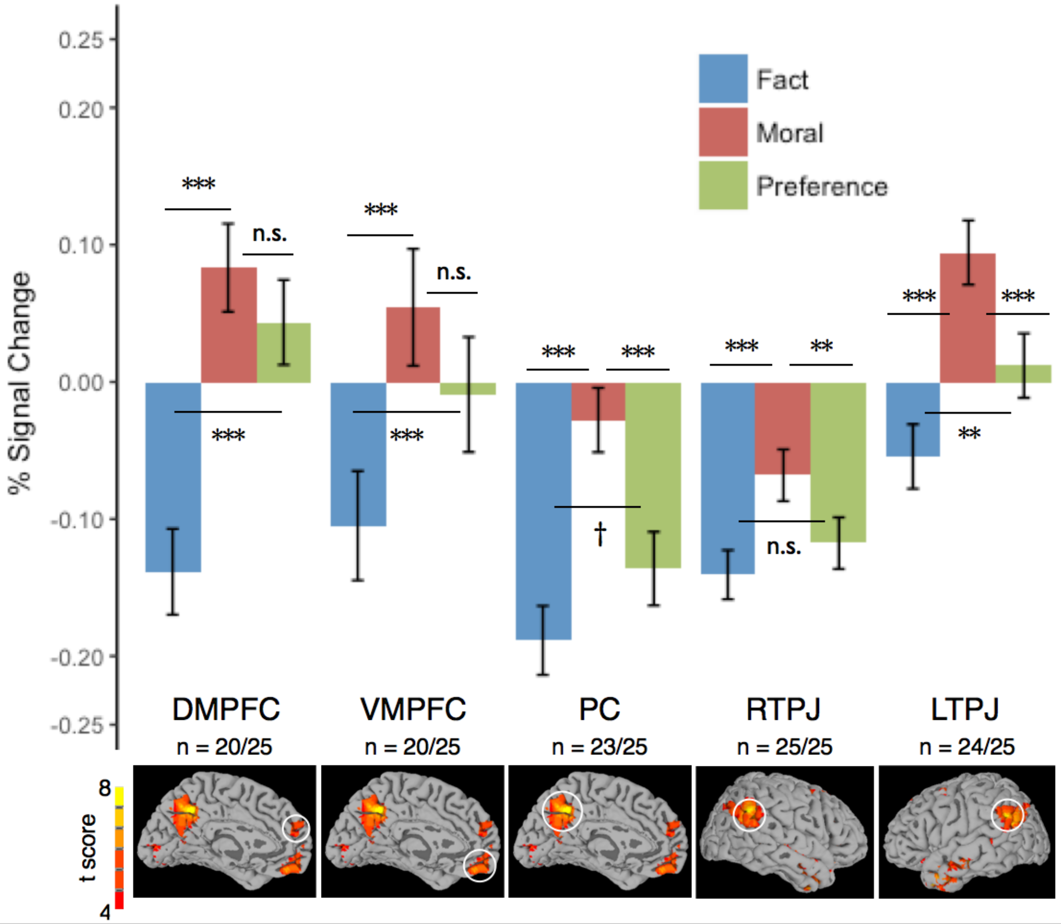


Figure 4. Response Magnitude Across Content (fact/moral/preference) and ROIs. Morals and preferences both elicit greater activity than facts in DMPFC and VMPFC, whereas morals elicit greater activity than both facts and preferences in PC, RTPJ, and LTPJ. ROIs were identified for each individual using 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 α = .05 across five comparisons ($p_{\text{corrected}}$ = .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.

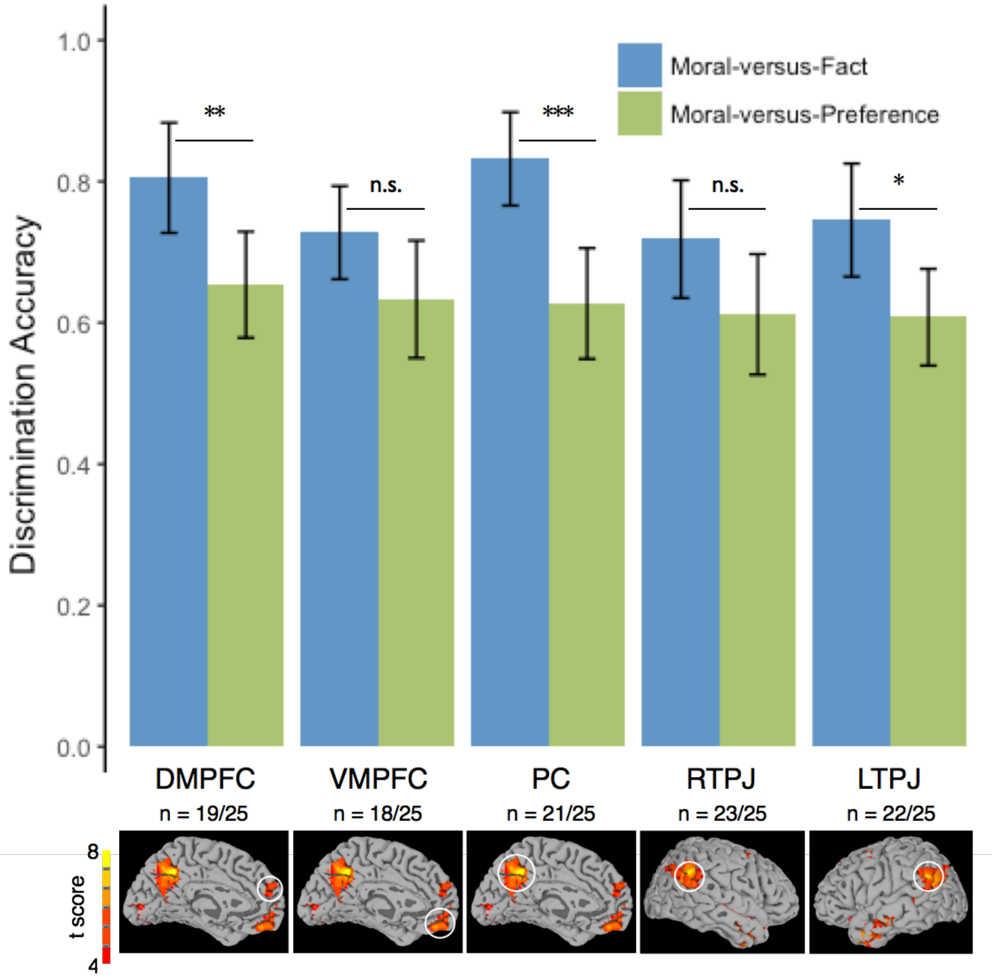


Figure 5. MVPA Discrimination Accuracy for Morals Versus Facts and Preferences. In DMPFC, PC, and LTPJ, morals and facts are more accurately discriminated than morals and preferences, based on the spatial correlation of voxel-wise activity (i.e. independent of mean differences in neural activity, presented 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, $\beta = 0.222$, $t(35.1) = 5.94$, $p < .001$; preferences, $\beta = 0.182$, $t(40.1) = 5.14$, $p < .001$), VMPFC (morals, $\beta = 0.159$, $t(32.8) = 3.91$, $p < .001$; preferences, $\beta = 0.098$, t(33.8) = 2.15, $p = 0.039$), and LTPJ (morals, $\beta = 0.148$, $t(49.5) = 5.00$, $p < .001$; preferences, $\beta = 0.066$, $t(56.3) = 2.40$, $p = .020$). Morals, but not preferences, elicited greater activity than facts in PC (morals, $\beta = 0.158$, $t(58.9) = 4.70$, $p < .001$; preferences, $\beta = 0.051$, $t(58.9) = 1.53$, $p = .132$), and in RTPJ (morals, $\beta = 0.072$, $t(31.9) = 3.55$, $p = .001$; preferences, $\beta = 0.023$, $t(34.6) = 1.35$, $p = .187$).

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:**_ 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}
	Identify potential covariates	lmer(PSC ~ MentalState + (1|Item) + (1|ID))	***_**Mental States:**_ β = 0.078, t(70.0) = 8.74, p = 7.9 x 10^{-13}
		lmer(PSC ~ Arousal + (1|Item) + (1|ID))	***_**Arousal:**_ β = 0.069, t(70.1) = 4.33, p = 4.9 x 10^{-5}
		lmer(PSC ~ NounFamiliarity + (1|Item) + (1|ID))	*_**Noun Familiarity:**_ β = 0.002, t(70.1) = 2.38, p = .020
		lmer(PSC ~ NounConcreteness + (1|Item) + (1|ID))	*_**Noun Concreteness:**_ β = -0.0005, t(69.8) = 2.27, p = .026
		lmer(PSC ~ PersonPresent + (1|Item) + (1|ID))	*_**Person Present:**_ β = 0.077, t(70.0) = 2.19, p = .032
		lmer(PSC ~ NounImageability + (1|Item) + (1|ID))	*_**Noun Imageability:**_ β = -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:*** β = 0.119, t(74.6) = 1.58, p = .118 ***Preference:*** β = 0.098, t(71.2) = 1.54, p = .129 ***Mental States:*** β = 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:*** β = 0.118, t(67.4) = 1.52, p = .132 ***Preference:*** β = 0.097, t(64.8) = 1.43, p = .157 ***Mental States:*** β = 0.037, t(62.2) = 1.30, p = .200 ***Arousal:*** β = 0.004, t(62.5) = 0.19, p = .849 **_**Noun Familiarity:**_ β = 0.002, t(60.8) = 2.70, p = .008 ***Noun Concreteness:***

	β = -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 et al., 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, β = 0.119, *t(*74.6) = 1.58, *p* = .118, and preferences, β = 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, β = 0.091, *t(*70.6) = 0.91, *p* = .368, and preferences, β = 0.043, *t(*70.9) = 0.48, *p* = .629, dropped to non-significance. (c) In LTPJ, after controlling for mental state ratings, both morals, β = 0.051, *t(*74.3) = 0.77, *p* = .443, and preferences, β = -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, $\beta = 0.067$, $t(63.6) = 1.95$, $p = .055$, and PC remained marginal after adding reading ease and reaction time to the model, $\beta = 0.064$, $t(61.0) = 1.64$, $p = .068$. (b) In RTPJ, after controlling for mental state ratings and reaction time, morals dropped to non-significance, $\beta = 0.030$, $t(43.6) = 1.63$, $p = .110$. 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, 2017;

Strohminger & Nichols, 2014; 2015). 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., 2017; 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.

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., Strohminger, N., & Young, L. L. (2017). The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive science*, *41*(3), 744-767. http://dx.doi.org/10.1111/cogs.12354

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. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron, 79*, 836–848. http://dx.doi.org/10.1016/j.neuron.2013.08.020

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

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping, 15*, 1-25. http://dx.doi.org/10.1002/hbm.1058

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. In J. A. Suhr, K. D. Markman, & W. M. P. Klein (eds.) *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 ($p_{\text{corrected}} = .0167$).

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.

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 express oneself.	It is more important to be a team player than to 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. 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.	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. 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. People should be free to decide what group norms or traditions they themselves want to follow.	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.

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	"To what extent do you agree / disagree with this statement?" (1-7; "strongly disagree"-"strongly agree").

	N = 68	
Valence	Scores taken from online norming study. N = 17	Valence was the difference between unipolar positive 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.

Supplemental Materials

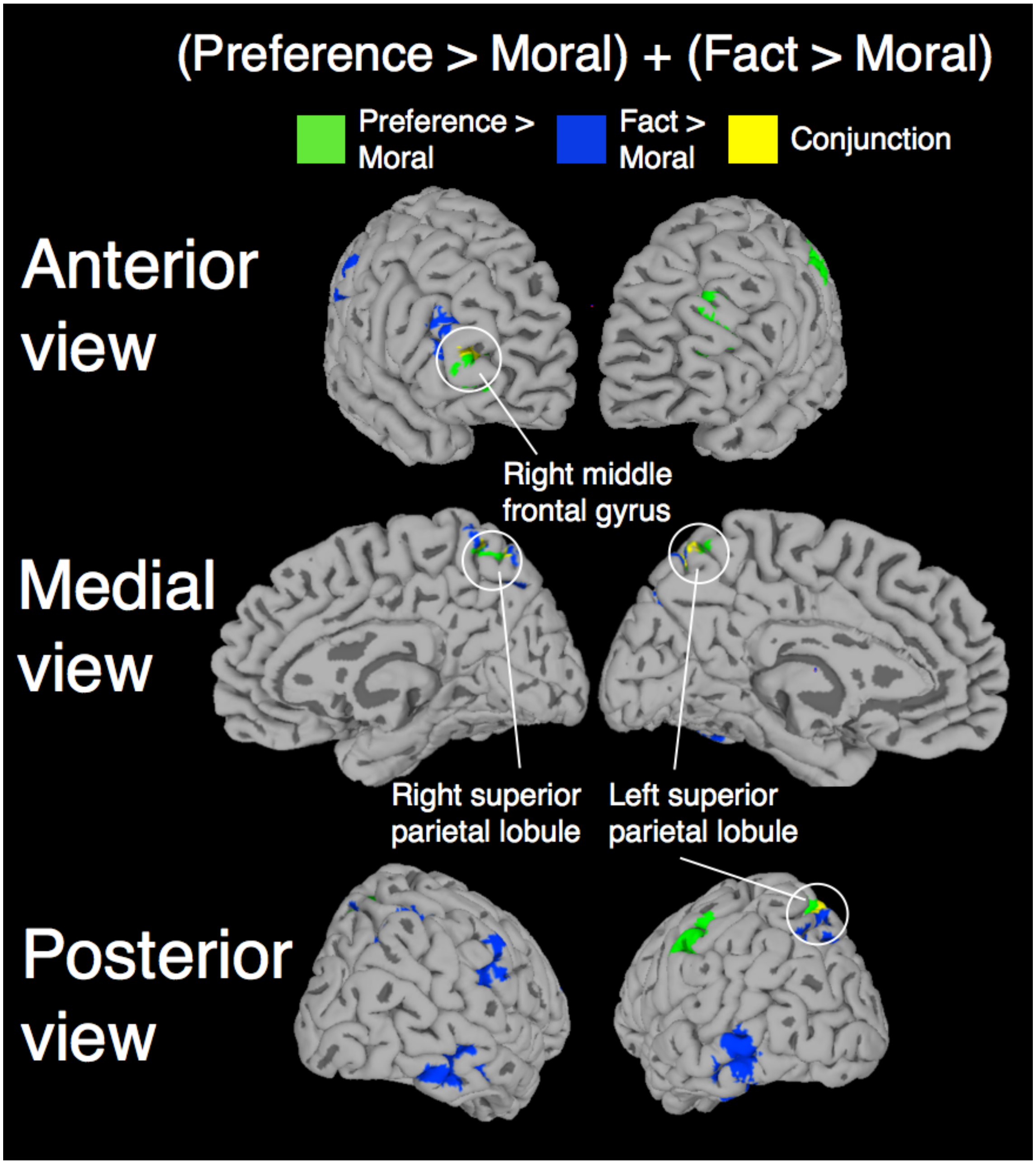

(Preference > Moral) + (Fact > Moral)
Preference > Moral
Fact > Moral
Conjunction
Anterior view
Medial view
Posterior view
Right middle frontal gyrus
Right superior parietal lobule
Left superior parietal lobule

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 α = .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.

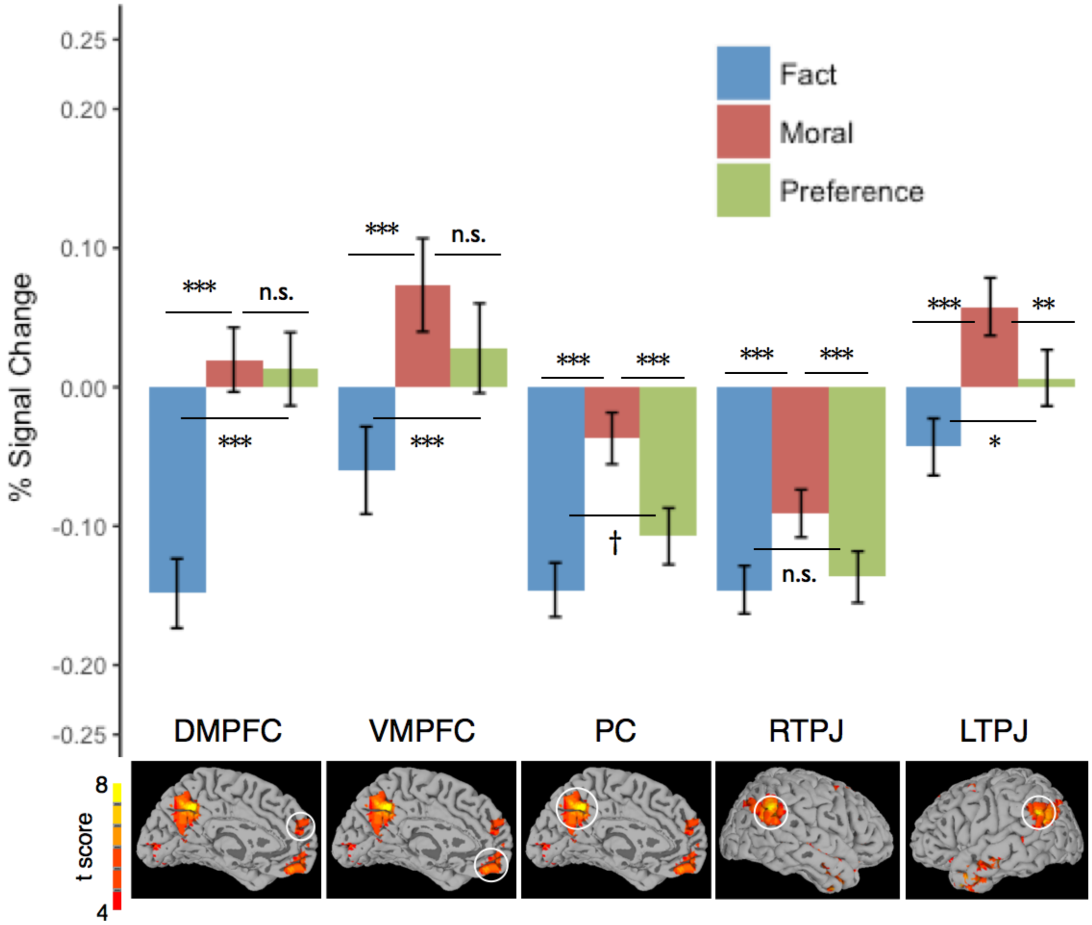


Figure S2. Response Magnitude Across Content (fact/moral/preference) and anatomically defined ToM ROIs. ROIs were identified using the peak coordinates of a whole brain contrast of the localizer contrast (false belief > false photograph) across all participants. Each ROI is defined as a 9mm sphere around 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.

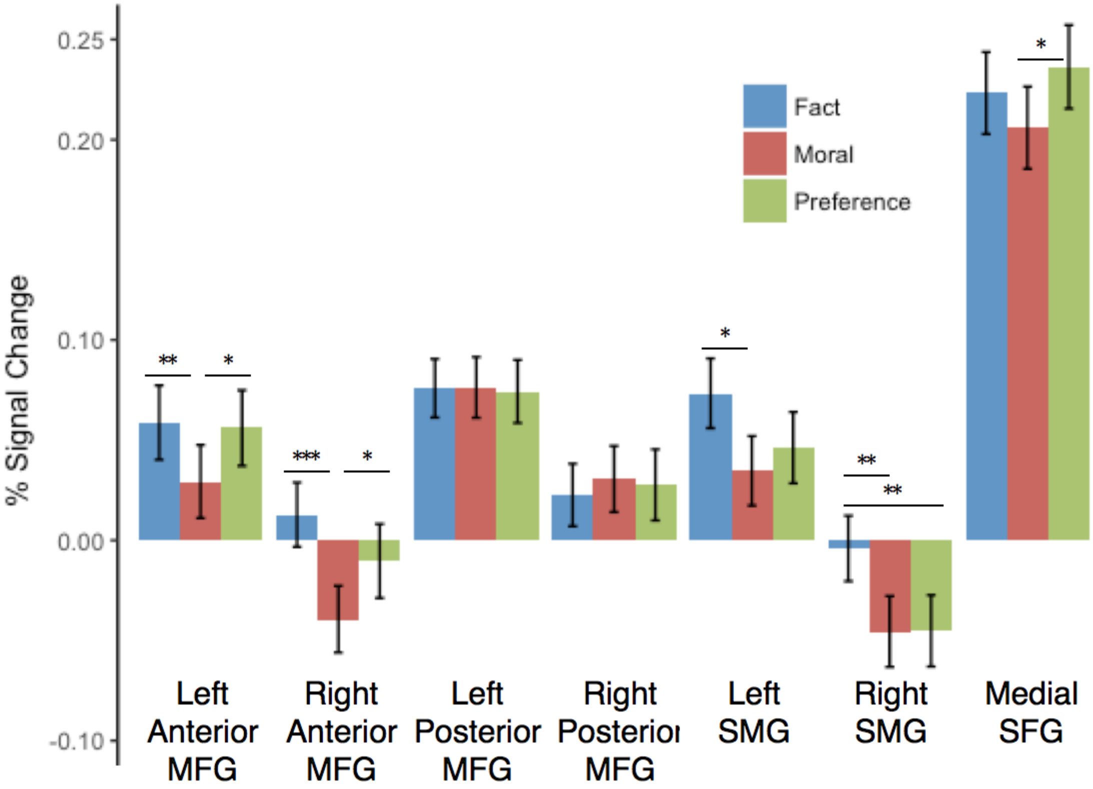


Figure S3. Response Magnitude Across Content (fact/moral/preference) and working memory ROIs. ROIs were identified using the reverse inference map for "working memory" at neurosynth.org (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). MFG = middle frontal gyrus; SMG = supramarginal Gyrus; SFG = 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.

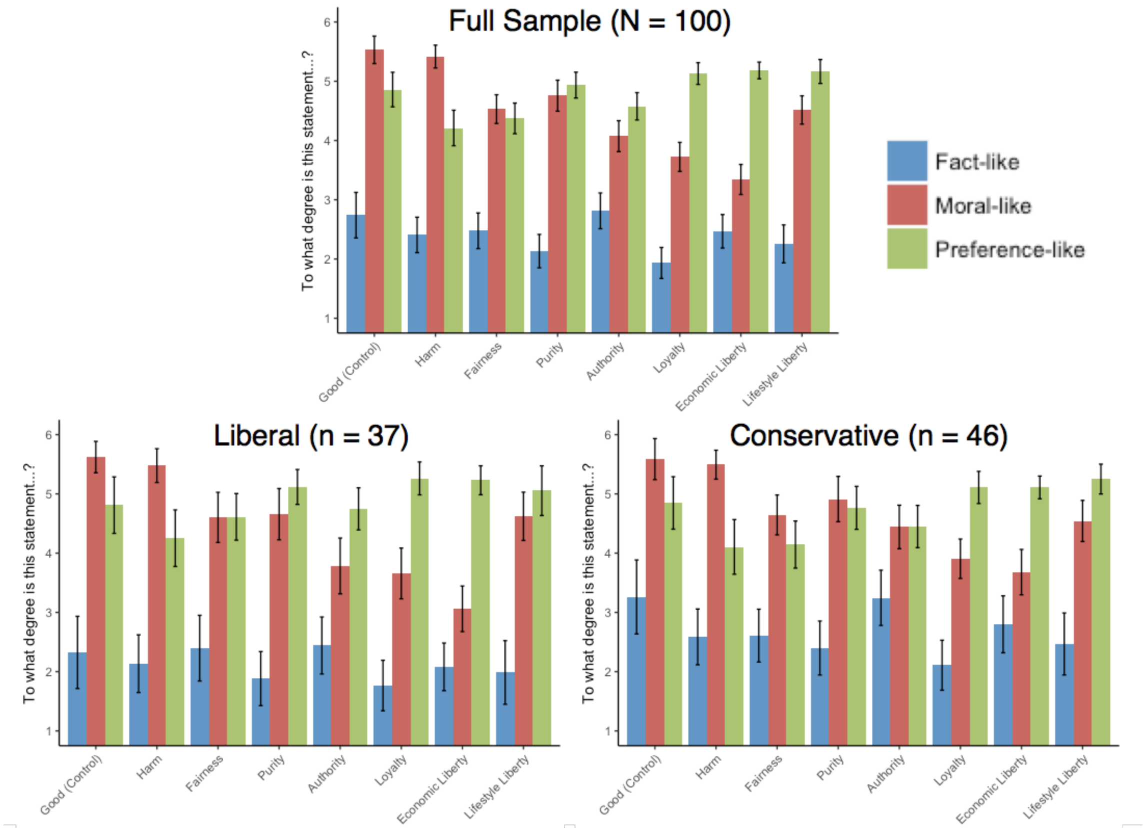


Figure S4. Supplemental Study MFQ Behavioral Ratings. Across all domains, morals were rated as more preference-like than fact-like. This pattern held even when splitting the sample based on political orientation. Participants were grouped as liberal or conservative based on their response to the question: "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 Zone	455	5.45	-10	-74	46
		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:*** $\beta = 0.050$, $t(70.1) = 4.15$, $p = 9.1 \text{ x } 10^{-5}$
		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:*** $\beta = 0.148$, $t(49.5) = 5.00$, $p = 7.5 \text{ x } 10^{-6}$ * ***Preference:*** $\beta = 0.066$, $t(56.3) = 2.40$, $p = .020$
	Identify potential covariates	lmer(PSC ~ MentalState + (1|Item) + (1|ID))	*** ***Mental States:*** $\beta = 0.047$, $t(70.0) = 5.63$, $p = 3.4 \text{ x } 10^{-7}$
		lmer(PSC ~ PersonPresent + (1|Item) + (1|ID))	*** ***Person Present:*** $\beta = 0.113$, $t(69.9) = 4.53$, $p = 2.4 \text{ x } 10^{-5}$
		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:*** $\beta = -0.00005$, $t(60.5) = 0.14$, $p = .887$ *Negation Density: $\beta = 0.001$, $t(64.1) = 2.59$, $p = .012$ ***Flesch Reading Ease*** $\beta = -0.0006$, $t(60.9) = 1.17$, $p = .245$ ***Number of Modifiers:*** $\beta = -0.023$, $t(61.5) = 1.35$, $p = .181$ ***Reaction Time*** $\beta = 0.006$, $t(1578.0) = 0.50$, $p = .495$

PC	***Hypothesis testing***	lmer(PSC ~ Moral + Preference + (1|Item) + (Moral+Preference|ID))
		lmer(PSC ~ Moral + (1|Item) + (Moral|ID))
			***Moral:
			$\beta = 0.158$, $t(58.9) = 4.70$, $p = 1.63 \text{ x } 10^{-5}$
			Preference:
			$\beta = 0.051$, $t(58.9) = 1.53$, $p = .132$
			***Moral:
			$\beta = 0.133$, $t(61.8) = 4.60$, $p = 2.1 \text{ x } 10^{-5}$

	Identify potential covariates	lmer(PSC ~ MentalState + (1|Item) + (1|ID))	***Mental States: $\beta = 0.047$, $t(70.0) = 4.47$, $p = 2.9 \text{ x } 10^{-5}$
		lmer(PSC ~ PersonPresent + (1|Item) + (1|ID))	***Person Present: $\beta = 0.125$, $t(70.0) = 4.14$, $p = 9.7 \text{ x } 10^{-5}$
		lmer(PSC ~ IntentionVerb + (1|Item) + (1|ID))	**Intentional Verb Incidence: $\beta = 0.001$, $t(70.1) = 2.79$, $p = .007$
		lmer(PSC ~ Arousal + (1|Item) + (1|ID))	*Arousal: $\beta = 0.038$, $t(70.0) = 2.44$, $p = .017$
		lmer(PSC ~ ReadingEase + (1|Item) + (1|ID))	*Flesch Reading Ease: $\beta = -0.002$, $t(69.9) = -2.20$, $p = .031$
	Attempt to disprove hypothesis	***Marginal/non-significant model:*** lmer(PSC ~Arousal + IntentionVerb + PersonPresent + MentalState + Moral + (1|Item) + (Moral|ID))	†Moral: $\beta = 0.067$, $t(63.6) = 1.95$, $p = .055$ †Mental States: $\beta = 0.029$, $t(66.2) = 1.87$, $p = .066$ *Person Present: $\beta = 0.069$, $t(66.0) = 2.19$, $p = .032$ **Intention Verb Incidence:** $\beta = 0.0004$, $t(66.2) = 1.18$, $p = .241$

			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	**Morals**	**Preferences**	
	***N*$_{\text{Items}}$ = 24**	***N*$_{\text{Items}}$ = 24**	***N*$_{\text{Items}}$ = 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)	
(Moral-like – Fact-like)						
MFQ Domain	**Diff (SE)**	**z ratio**	**Diff (SE)**	**z ratio**	**Diff (SE)**	**z ratio**
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)						
MFQ Domain	**Diff (SE)**	**z ratio**	**Diff (SE)**	**z ratio**	**Diff (SE)**	**z ratio**
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)						
MFQ Domain	**Diff (SE)**	**z ratio**	**Diff (SE)**	**z ratio**	**Diff (SE)**	**z ratio**
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; $p_{\text{corrected}} = .0167$). $p_{\text{family-wise}} = .05$, $p_{\text{corrected}} = .00208$. *** *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

Part 3

Theory of Mind network activity is associated with metaethical judgment: An item analysis

Jordan Theriault, Adam Waytz, Larisa Heiphetz, Liane Young

Abstract

A single claim can be interpreted at multiple levels: it can be interpreted as correct or incorrect, or it can be interpreted in terms of the information it has expressed (a second-order judgment). For instance, facts provide objective information about some state of the world, whereas preferences provide subjective information about beliefs. For moral claims, second-order judgments (i.e. metaethical judgments) occupy a special case, in that metaethical judgments are variable—some morals are perceived as more objective than others. The present work examined this second-order variability across a set of claims about morals, facts, and preferences—first behaviorally (Study 1), then using fMRI (Study 2). Because subjective claims involve referencing one's own and others' beliefs, we predicted that subjective moral claims would elicit greater activity in the Theory of Mind (ToM) network, a set of regions consistently associated with mental state processing. Using linear mixed effects models, we compared by-stimuli estimates of behavioral ratings (Study 1) with ToM network activity (Study 2). Subjective (i.e. preference-like) moral claims elicited greater activity in the ToM network, and objective (i.e. fact-like) moral claims elicited less activity in this network. Whole brain correlation analyses confirmed that the overlap of these two effects was specific to bilateral TPJ, key nodes in the ToM network. Effects could not be attributed to semantic or syntactic differences across stimuli. Exploratory analyses also identified discrepancies between processing for facts, morals, and preferences; particularly, ToM network activity for facts and preferences was associated with the tendency to consider mental states, but this was not the case for morals. Thus, the present work provides evidence for the representation of metaethical judgment, i.e. second-order judgment in the moral domain. We briefly

speculate on how our findings might apply to accounts of ToM network activity, drawing on recent theories of predictive coding.

Highlights:

- Perceived moral subjectivity is positively associated with ToM network activity.
- Perceived moral objectivity is negatively associated with ToM network activity.
- Relationship between metaethical judgment and BOLD is strongest in bilateral TPJ.
- Multiple by-stimuli, continuous relationships with ToM network were identified.
- Mixed effects analysis compared BOLD activity to independent behavioral estimates.

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

1. Introduction

A single claim can be interpreted at multiple levels. For instance, at one level, people can interpret the claim “Mount Whitney is the tallest mountain in the United States” as correct or incorrect (a first-order judgment; and it is incorrect—the tallest is Mount McKinley). However, on another level (a second-order judgment) people can interpret the claim based on the nature of the information it has expressed: does the claim contain *objective* information about some state of the world, or *subjective* information about what someone believes? Facts and preferences are generally taken to represent opposite extremes of second-order judgment (Goodwin & Darley, 2010; Sayre-McCord, 1986), and the second-order status of other claims can fall between these two extremes—e.g. judgments about morality, aesthetics, or social norms can blur the line between what is objective and what is socially arbitrary. In the present work, we examine metaethical judgement (i.e. second-order judgments about morality), and their relationship with activity in the ToM network.

Whether morality is objective or subjective is an area of unresolved debate in moral philosophy, but, when asked, non-philosophers will not give a unified response: they will report that some moral claims are more objective (e.g. “slavery is wrong”), and that others are more subjective (e.g. “eating meat is wrong”). A number of studies have recently highlighted this variability, (Beebe, 2014; Goodwin & Darley, 2008; 2012; Heiphetz & Young, in press; Sarkissian et al., 2011; Wright et al., 2013), but the cognitive and neural processes underlying it are not well understood. Moral judgment is not a singular process, and almost certainly draws on a number of more basic processes (Dungan & Young, 2012). We assume that the same is true for metaethical judgment, which means that an understanding of the processes underlying metaethical variability would not limit conclusions to the moral domain—it could potentially be extended to other domains where second-order status is variable (e.g. social norms; Ruff et al., 2013; Zaki et al., 2011). Thus, in the present work, we are concerned with the nature of the information conveyed by moral claims: is objectivity/subjectivity related to neural activity in known networks of brain regions? And can this activity help us understand what is being represented by this variance?

Because subjective claims involve referencing one's own and others' beliefs, we examined activity in the theory of mind (ToM) network, using a region of interest (ROI) analysis. The ToM network is comprised of brain regions implicated in the representation of internal mental states (e.g. beliefs, intentions; Ciaramidaro et al., 2007; Dodell-Feder et al., 2011; Fletcher et al., 1995; Gallagher et al., 2000; Gobbini et al., 2007; Ruby & Decety, 2003; Saxe & Kanwisher 2003; Saxe & Powell, 2006; Vogeley et al., 2001; Young et al., 2007; 2010; for review see Schurz et al., 2014; Van Overwalle, 2009). If reading subjective moral claims requires that participants represent the speaker's beliefs, then activity in the ToM network may track with this perceived subjectivity. Furthermore, if some moral claims are perceived as more objective or subjective than others, then the most powerful way to examine metaethical variability would be to use an item analysis (e.g. Dodell-Feder et al., 2011; Bedny et al., 2007) to test the relationship between ToM network by-stimuli ratings of perceived objectivity and subjectivity. In Study 1, we collect independent measurements of these by-stimuli rating in an online sample, then compare them with ROI activity in Study 2.

1.1. Present Work

In prior work, moral claims were rated as more objective when supported by a social consensus (Goodwin & Darley, 2012; Heiphetz & Young, in press), and consistent with this, the present work used claims that were designed to vary on the dimension of consensus. Participants read facts, morals, and preferences, that were designed to fit within consensus sub-categories, eliciting either *positive-consensus*, where most people would agree, *negative* consensus, where most people would disagree, or *no-consensus*, where neither agreement nor disagreement was strong (Figure 1; also see 2.1.2). In Study 1, we validated the stimuli and collected behavioral ratings in an online sample, demonstrating that positive-consensus moral claims were perceived as more fact-like, and that metaethical judgments were much more variable among moral claims, compared to facts and preferences. In Study 2, participants read the same set of claims in the scanner. We broke from our *a priori* consensus sub-categories and performed an item analysis across all stimuli, using ToM network activity from Study 2 to predict the by-stimuli variance in the behavioral ratings collected in Study 1. This analysis required that we compare our two samples, and this was made possible by extracting item estimates from maximal mixed effects models in each sample (best linear unbiased predictors; BLUPs; Baayen et al., 2008; Westfall et al., 2017), predicting behavioral ratings, or BOLD (blood-oxygen-level dependent) activity, for each claim. Together, these studies show that metaethical judgments can be predicted by activity throughout the ToM network.

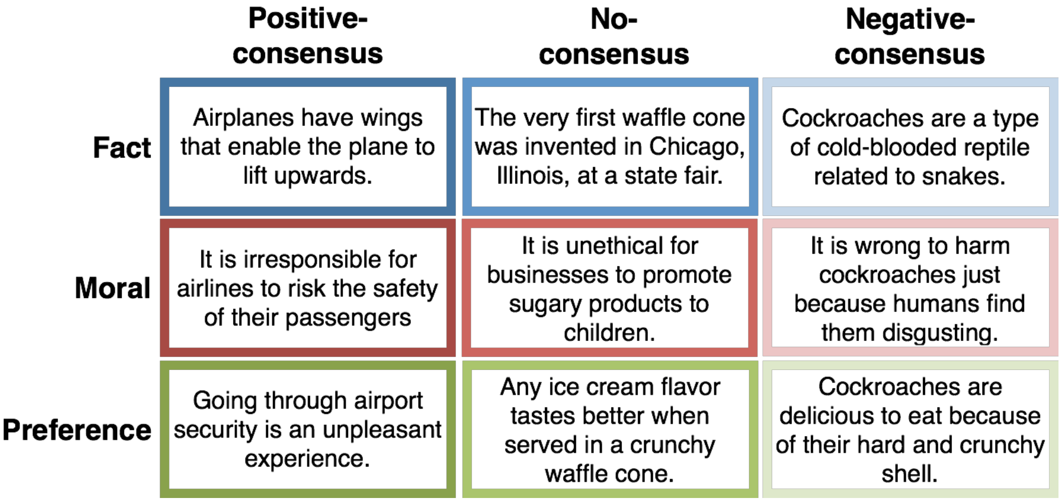
2. Study 1

Study 1 validated our stimuli set in an online sample, and allowed us to extract by-stimuli behavioral rating BLUPs from a sample that was approximately twice the size of our fMRI sample. To measure metaethical judgment, we asked participants to rate each claim on the extent that it was about facts, about morals, and/or about preferences (Theriault et al., in press). This method has several advantages: (a) It validates stimuli conditions—facts should be rated as most strongly fact-like, morals as most moral-like, and preferences as most preference-like; and b) It avoids artificially imposing relationships among ratings—for instance, positive-consensus moral claims may be perceived as more fact-like, less preference-like, or both.

2.1. Method

2.1.1. Participants. Participants were recruited online using Amazon Mechanical Turk (AMT) at an approximate rate of $6/hour, in line with standard AMT compensation rates. The final sample consisted of 49 adults (25 female, 1 unspecified; M_{Age} = 35.5 years, SD_{Age} = 10.7 years), after excluding two participants for failing an attention check that asked them to describe any claim they had read. The Boston College Institutional Review Board approved Studies 1 and 2, and each participant provided consent before beginning.

2.1.2. Procedure. Participants read a series of claims (e.g., "It is irresponsible for airlines to risk the safety of their passengers"; see Appendix A for all claims), and, for each, rated a) their agreement ("To what extent do you disagree/agree; 1–7, "completely disagree"—"completely agree"), and b) the extent that the claim was about facts, about morals, *and* about preferences (*Rating-type:* fact-like/moral-like/preference-like; "To what degree is this statement about … [facts, morality, preferences]"; 1–7, "not at all"—"completely"). The order of rating-types was counterbalanced across participants. Claims were designed to be interpreted as either facts, morals, or preferences and were evenly divided between categories (n_{Fact} = 24, n_{Moral} = 24, $n_{\text{Preference}}$ = 24). Each category included three consensus subcategories: a) *positive-consensus*, where most people would agree with the claim (*n* = 6); b) *negative-consensus*, where most people would disagree (*n* = 6); and c) *no-consensus*, where there would be no strong positive or negative consensus (*n* = 12). No-consensus claims (as opposed to controversial claims) were used because the feature of interest was consensus; in other words, no-consensus claims were intended to elicit a unipolar, non-skewed distribution of agreement. By contrast, controversial claims (e.g. "abortion is wrong") would presumably produce a bimodal distribution of agreement, introducing strong individual differences that could decrease the power of our item analyses. The no-consensus subcategory was also larger relative to other subcategories on account of an uninformative distinction that was irrelevant to the final design: six no-consensus facts were true, and six were false. Critically, claims did not contain any mental state markers (e.g., "She thinks," "He believes") that could have elicited neural activity related to mental state processing (i.e. ToM network activity).



	Positive-consensus	No-consensus	Negative-consensus
Fact	Airplanes have wings that enable the plane to lift upwards.	The very first waffle cone was invented in Chicago, Illinois, at a state fair.	Cockroaches are a type of cold-blooded reptile related to snakes.
Moral	It is irresponsible for airlines to risk the safety of their passengers	It is unethical for businesses to promote sugary products to children.	It is wrong to harm cockroaches just because humans find them disgusting.
Preference	Going through airport security is an unpleasant experience.	Any ice cream flavor tastes better when served in a crunchy waffle cone.	Cockroaches are delicious to eat because of their hard and crunchy shell.

Figure 1. Sample stimuli. Claims varied in content (fact/moral/preference) and agreement (positive-consensus/no-consensus/negative-consensus). See Appendix A for the full text of all stimuli.

2.1.3. Statistical Methods. Studies 1 and 2 used mixed effects analyses to model crossed by-subject and by-item random effects (Baayen et al., 2008; Judd et al., 2012; Westfall et al., 2014). This analysis allowed for generalizations beyond our sample of participants and stimuli, compared to standard analyses (e.g. ANOVA), which limit conclusions to the specific stimuli tested. It also allowed us to extract BLUPs for stimuli, predicting stimuli ratings while controlling for variance that could be attributed to subjects in the Study 1 sample (Westfall et al., 2017). Analysis was conducted in R (R Core Team, 2015, using the *lme4* package (Bates et al., 2015), and *p* values for fixed effects were calculated using the Satterthwaite approximation of degrees of freedom, implemented in the *lmerTest* package (Kuznetsova et al., 2015).

2.2. Results

2.2.1. Agreement Validation. Agreement ratings were fit with a maximal mixed effects model: *Fixed Effects*: intercept, category (fact/moral/preference) x consensus (positive-/no-/negative-consensus); *Random Effects (by-subject)*: random intercepts, category x consensus; *Random Effects (by-stimuli)*: random intercepts (for condition means see Table S1 of the online supplemental materials). In this model, we observed a main effect across category, $F(2, 72.15) = 7.35$, $p = .001$, and consensus, $F(2, 55.11) = 55.1$, $p < .001$, but no interaction, $F(4, 66.5) = 0.35$, $p = .843$. In follow-up contrasts, agreement was greater, relative to preferences, for facts, $z = 3.27$, $p = .003$, and morals, $z = 3.33$, $p = .002$, which did not significantly differ from each other, $z = 0.01$, $p = .999$ (p values corrected for 3 comparisons; $\alpha_{\text{familywise}} = .05$; single-step method; *multcomp* package, Hothorn et al., 2008). Critically, agreement was greater for positive-consensus claims, relative to no-consensus claims, $z = 6.48$, $p < .001$, and negative-consensus claims, $z = 10.26$, $p < .001$, and agreement was greater for no-consensus claims relative to negative-consensus claims, $z = 6.80$, $p < .001$. Thus, agreement ratings were consistent with our design. While preferences elicited less agreement in general, we observed no significant interaction between category and consensus, meaning that differences between consensus sub-categories were comparable across facts, morals, and preferences.

2.2.2. Fact-/moral-/preference-like ratings. First, we attempted to fit a maximal mixed effects model—*Fixed Effects*: intercept, category (fact/moral/preference) x rating-type (fact-/moral-/preference-like) x consensus (positive-/no-/negative-consensus); *Random Effects (by-subject)*: random intercepts, category x rating-type x consensus; *Random Effects (by-stimuli)*: random intercepts, rating-type—but this failed to converge (10,000 iterations). We simplified the model, removing consensus sub-categories from by-subject random effects. In this model, we observed a significant 3-way interaction, $F(8, 63.0) = 4.06$, $p < .001$ (for condition means see Table S1 of the online supplemental materials). We performed contrasts within this model to compare consensus sub-categories within each category x rating-type grouping (*p* values corrected for 27 comparisons; $\alpha_{\text{familywise}} = .05$; single-step method). Among morals, positive-consensus claims were perceived as more fact-like than no-consensus, $z = 5.70$, $p < .001$, and negative-consensus claims, $z = 6.26$, $p < .001$. Among facts, negative-consensus claims were perceived as less fact-like than no-consensus, $z = 5.43$, $p < .001$, and negative-consensus claims, $z = 4.72$, $p < .001$. Among preferences, there were no significant differences between consensus categories (Figure 2b).

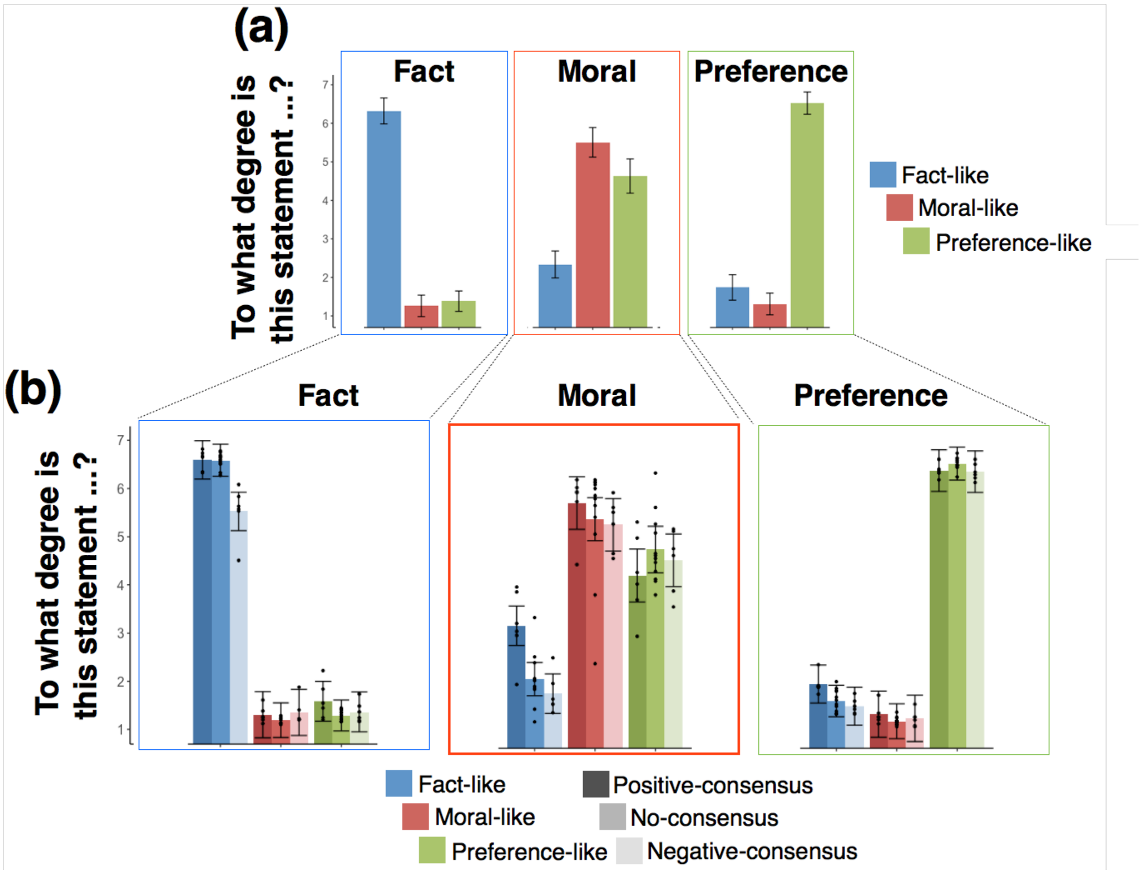


Figure 2. Post-scan behavioral ratings. Participants rated each scenario on the extent that it was fact-like, moral-like, and preference-like (1–7; "not at all" – "completely"). (a) Collapsing across consensus subcategories, ratings were consistent with our *a priori* categories: facts were largely fact-like (left), preferences were largely preference-like (right), and morals were largely moral-like (center). Moral claims were also perceived as largely preference-like (a pattern explored further in Theriault et al., in press). (b) Across consensus subcategories, positive-consensus morals were perceived as more fact-like than no-consensus or negative-consensus morals (center; blue). Note that variance across items was markedly greater for moral claims than for facts or preferences (dots represent item averages for each rating; 72 stimuli x 3 ratings). Error bars represent 95% confidence interval. For condition means, see Table S1 of the online supplemental materials.

Variance across stimuli appeared to be greater among moral claims than among facts and preferences (Figure 2b), so we explored model fit in both the maximal model, and in models limited to each category (Nakagawa & Schielzeth, 2012; implemented in *MuMin* package; Bartoń, 2016). Model fit is represented by $R^2_{marginal}$, which denotes variance accounted for by fixed effects, and $R^2_{conditional}$, which denotes variance accounted for by both fixed and random effects. Model fit was high in the overall model, $R^2_{marginal}$ = 0.683, $R^2_{conditional}$ = 0.838, and marginal fit was high in the fact-only model, $R^2_{marginal}$ = 0.803, $R^2_{conditional}$ = 0.887, and preference-only model, $R^2_{marginal}$ = 0.821, $R^2_{conditional}$ = 0.913, meaning that variance within facts and preferences was largely explained by fixed effects. Marginal fit was much lower in the moral-only model, $R^2_{marginal}$ = 0.331, $R^2_{conditional}$ = 0.666. Likewise, dropping by-stimuli random effects from each model significantly reduced model fit, but the difference was markedly greater for moral claims: facts, χ^2 (6) = 96.81, $p < .001$; preferences, χ^2 (6) = 23.59, $p < .001$; morals, χ^2 (6) = 631.35, $p < .001$. Thus, the rating-type and consensus fixed effects could largely account for the responses within facts and preferences, but among moral claims fixed effects were less able to account for the observed variance, which was particularly strong across stimuli.

2.3. Discussion

Consistent with prior work (Goodwin & Darley, 2012; Heiphetz & Young, in press), people rated positive-consensus moral claims as more fact-like (i.e. more objective) than no-consensus and negative-consensus moral claims. Ratings for moral claims also varied more than for either facts or preferences. Study 2 tested whether this variance among ratings for moral claims could be accounted for by activity in the ToM network, given the hypothesis that evaluating subjective claims would elicit mental state processing[4].

[4] fMRI data used in Study 2 is also used in a separate study (Theriault et al., in press). Analyses are not repeated between the two studies: the present study focuses on by-stimuli relationships between BOLD activity and metaethical judgment. The separate study focuses on contrasts between facts, morals, and preferences, examining why morals and preferences elicit overlapping BOLD activity relative to facts.

3. Study 2

Objective claims refer to some state of the world, whereas subjective claims are mind-dependent: they refer to beliefs (Goodwin & Darley, 2010; Sayre-McCord, 1986). Moral claims can be perceived as both subjective and objective (Goodwin & Darley, 2012; Wright et al, 2013) and Study 1 demonstrated that this variability is not well captured by behavioral measures. In Study 2 we examined the relationship between perceived objective/subjectivity and its relationship to BOLD activity in the ToM network. Discovering a relationship here could provide insight into the neural representation of metaethical judgment, and second-order judgment more generally.

3.1 Method

3.1.1. Participants. Participants were a community sample, recruited through an online posting and paid $65. The final sample consisted of 25 right-handed adults (12 female, 12 male, 1 unspecified; M_{age} = 27.0 years, SD_{age} = 5.2 years). Two more participants were recruited but not analyzed due to excessive movement, identified during spatial preprocessing. Of these 25 participants, two completed only a subset of the scan session runs: one completed five of six runs due to experimenter error, and for the other a movement artifact rendered only the first three runs usable. We were unable to collect post-scan ratings for another of the 25 participants. All participants were native English speakers with no reported history of learning disabilities, previous psychiatric or neurological disorders, or a history of drug or alcohol abuse.

3.1.2. Procedure. Participants completed the study in a single session. Twenty participated at Harvard University's Center for Brain Science Neuroimaging Facility, and five at the Massachusetts Institute of Technology's Martinos Imaging Center. Scanning parameters and equipment were identical between sites (see 3.1.4). In the scanner, participants read claims and rated their agreement with each (ratings were consistent with consensus subcategories; see Table S2). Claims were shown across six runs (12 per run; items were randomized; conditions were counterbalanced to appear equally in each run). Participants read each claim (6 s), rated their agreement (+4 s), and waited during fixation (+12 s). Agreement was provided with a button box (1–4; "Strongly Agree"—"Strongly Disagree"). A thumb press indicated "Don't Know", which was coded as an empty cell. This option was provided to avoid confusion, particularly for no-consensus facts where the answer was generally unknown (across our complete sample, 71.6% of "don't know" responses were for no-consensus facts, followed by 7.3% for no-consensus preferences). Stimuli were presented in white text on a black background using 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. Each experimental run was 4 min 52 s long, totaling 29 min 12 s across six runs. The in-scanner experiment was preceded by a structural scan (6 min 3 s) and a functional localizer (two 4 min 46 s runs; Dodell-Feder et al., 2011; see 3.1.5). The total scan time was 68 min 8 s due to a second study not reported here involving responses to moral dilemmas (29 min 12 s); runs for both studies were interleaved, so that stimuli in the present work were equally likely to appear early or late in the session, across participants. Post-scan, participants gave behavioral ratings for all claims (i.e. fact-/moral-/preference-like ratings) on an Apple Macbook Pro and completed a brief demographics questionnaire.

3.1.3. Stimuli and Measures. Stimuli were identical to those described in Study 1 (see Appendix A). Ratings concerning additional item features were collected in an online sample, in order to explore their relation to ToM network activity. These included several questions used in a prior item analysis of the ToM network (Dodell-Feder et al., 2011), as well as measures of arousal and valence (Kron et al., 2013; see Appendix B). In these online samples, participants were asked one of the following questions: *Mental States* (*n* = 48; "To what extent did this statement make you think about someone's experiences, thoughts, beliefs, and/or desires?"; 1–7; "Not at all"–"Very Much"), *Mental Imagery* (*n* = 46; "To what extent did you picture or imagine what the statement described as you read?"; 1–7–; "Not at all"–"Very much"), *Person Present* (*n* = 48; "Does this statement mention people or a person?"; 0–1; "No"–"Yes"), *Valence* (*n* = 42; the difference between 8-point positive and negative unipolar scales; Kron et al., 2013), *Arousal* (n = 42; the sum of both 8-point positive and negative unipolar scales; Kron et al., 2013).

To ensure that effects were not driven by semantic/syntactic differences across stimuli, several item characteristics were collected using Coh-Metrix 3.0 (Graesser et al., 2004; McNamara et al., 2014). These included features such as word length, reading ease, noun concreteness, familiarity, and imageability, among others (see Appendix B).

3.1.4. 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 the Center for Brain Science Neuroimaging Facility at Harvard University and at the Massachusetts Institute of Technology's Martinos Imaging Center. 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), spatially smoothed using a Gaussian filter (full-width half-maximum = 5 mm kernel), and high-pass filtered (128 Hz).

3.1.5. ToM Localizer Task. An independent functional localizer task identified ToM ROIs (Dodell-Feder et al., 2011). The task consisted of ten stories about mental states (*false-belief*) and ten about physical representations (*false-photograph*). Stories were matched in complexity across conditions; see http://saxelab.mit.edu/superloc.php for the complete set. Each story appeared (10 s) and was followed by a statement about it, rated true or false (+4 s). Typically, to increase power, this contrast is used to select ROIs individually for each participant. However, this approach also means that ROI coordinates cannot be reported in normalized space. Alternatively, we could select ROIs using the peak voxels of a whole brain random effects contrast (belief > photograph) across all participants. Both approaches returned the same results, and so in the interest of providing replicable coordinates we used the latter approach, defining each ROI as a 9mm-radius sphere around the peak voxel (for coordinates see Table S3 of the online supplemental materials).

3.1.6. ROI Analysis. BOLD activity for each functional ROI was estimated using a boxcar regressor, beginning with the appearance of the text, and ending after the agreement rating (10 s total). 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 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), centering each run at mean PSC.

3.1.7. Whole Brain Correlation Analysis. Whole-brain analyses were performed by estimating beta maps for each item, then correlating these with estimates of behavioral ratings (derived from Study 1, see 3.1.8). For each subject, three models correlated beta estimates with fact-like, moral-like, and preference-like ratings. Subject-level beta maps of each correlation were entered into separate second-level analyses across subjects. Each second-level contrast was cluster-corrected by permutation (5000 samples) to achieve a familywise error rate of $\alpha = .05$, thresholding voxels at $p < .001$ (uncorrected; recommended by Woo et al., 2014). Permutation tests were performed using SnPM 13 (http://warwick.ac.uk/snpm; Nichols & Holmes, 2001).

3.1.8. Statistical Methods. We used mixed effects analyses to model behavioral responses and PSC. Model specification and simplification is described below (3.2.2). For ToM ROIs, we identified the model then extracted BLUPs (best linear unbiased predictors) to compare by-stimuli ROI activity with behavioral BLUPs extracted from Study 1 (see Appendix A). BLUPs were desirable for a number of reasons. First, they have the property of shrinkage, i.e. each estimate accounts for the sample distribution and consequently anticipates regression to the mean (Baayen et al., 2008). Second, by-stimuli

BLUPs are estimated separate from by-subject variance, allowing us to use Study 1 estimates that control for by-subject variance. Leveraging these properties allowed us to take advantage of Study 1's larger sample size, providing more accurate estimates of by-stimuli behavioral ratings. Study 1 BLUPs were extracted from the model described in 2.2.2, dropping the fixed effect of consensus categories to minimize the projection of *a priori* categories onto the data (for model details, see Table S4 of the online supplemental materials).

3.2. Results

3.2.1. Behavioral Results. First, we estimated behavioral ratings from our Study 2 sample, to ensure that the patterns in Study 1 were replicated. The maximal mixed effects model used in Study 1 (2.2.2) did not converge, most likely because of Study 2's smaller sample size. There were no parameters which could clearly be dropped (e.g. parameters with zero unique variance, in a model with uncorrelated random effects), so by-stimuli random effects were dropped instead, as we simply wanted to confirm that ratings were similar to those observed in Study 1. We performed contrasts for all comparisons of interest (*p* values corrected for 27 comparisons; $\alpha_{\text{familywise}} = .05$; single-step method), and, consistent with Study 1, people rated positive-consensus morals as more fact-like than no-consensus morals, $z = 6.82$, $p < .001$, and negative-consensus morals, $z = 7.92$, $p < .001$, whereas no significant difference emerged between no-consensus and negative-consensus morals, $z = 2.33$, $p = .378$. The decreased fact-like ratings for negative-consensus facts did not replicate (for condition means see Table S5 of the online supplemental materials).

3.2.2. ToM network model. Initially, a maximal mixed effects model, predicing PSC, was fit across all functional ROIs—*Fixed Efffects*: intercept, ROI (DMPFC/VMPFC/PC/RTPJ/LTPJ) x category (fact/moral/preference); *Random Effects (by-subject)*: random intercepts, ROI x category; *Random Effects (by-stimuli)*: random intercepts, ROI—however, this maximal model failed to converge and had to be simplified. Interactions between random effects were temporarily removed, and all random effects components with zero unique variance were removed (*Random Effects (by-subject)*: moral x (PC/RTPJ), preference x (VMPFC/PC/RTPJ/LTPJ); *Random Effects (by-stimuli)*: PC, RTPJ, LTPJ). The final model included: *Fixed Effects:* ROI x category; *Random Effects (by-subject):* random intercepts, moral, preference, VMPFC, PC, RTPJ, LTPJ, moral x (VMPFC/LTPJ); *Random Effects (by-stimuli):* random intercepts, VMPFC (for model details, see Table S6 of the online supplemental materials).

It is worth noting that simplifying the model provided information about the relations among ROIs. For by-stimuli random effects, we observed high correlations between DMPFC (i.e. intercept), PC, RTPJ, and LTPJ. These correlations precluded calculating unique variances for each ROI, but they also demonstrated that these regions respond in tandem, at least across our set of stimuli. Thus, our mixed effects analysis provided a data driven rationale to treat these regions as a network, while estimating by-item variance for VMPFC separately. In analyses below, we average estimates of DMPFC, PC, RTPJ, and LTPJ, and refer to them collectively as the ToM network; ROI interactions below, therefore, explore differences between the ToM network, and VMPFC.

3.2.3. ToM–behavioral analysis. BLUPs estimating by-stimuli behavioral ratings were extracted from the Study 1 model (see 3.1.8). BLUPs estimating by-stimuli PSC were extracted from the Study 2 model (3.2.2; PSC BLUPs were centered and normalized). Behavioral ratings were fit with a linear model, including the three-way interaction of PSC x category (fact/moral/preference) x rating-type (fact-like/moral-like/preference-like), and the interaction of ROI (ToM/VMPFC) with all other terms. The main effect of ROI was not included in the model, as behavioral ratings were identical across levels (for details, see Table S7 of the online supplemental materials). We observed a 4-way interaction, $F(4, 405) = 5.73$, $p < .001$ between PSC, category, rating-type, and ROI. Follow-up ANOVAs demonstrated that 4-way interactions were significant between morals and facts, $F(2, 270) = 7.42$, $p < .001$, and between morals and preferences, $F(2, 270) = 6.64$, $p = .002$, but non-significant between preferences and facts, $F(2, 270) = 0.01$, $p = .986$.

Within facts, averaging across ToM and VMPFC, PSC did not interact with rating-type, $F(2, 138) = 0.24$, $p = .788$, and its main effect was non-significant, $F(1, 140) = 0.002$, $p = .966$. Likewise, within preferences, averaging across ToM and VMPFC, PSC did not interact with rating type, $F(2, 138) = 0.61$, $p = .543$, and its main effect was also non-significant, $F(1, 140) = 2.67$, $p = .104$.

Within morals, we observed a 3-way interaction between PSC, rating-type, and ROI, $F(2, 135) = 6.03$, $p = .003$. Follow-up ANOVAs demonstrated that the 3-way interactions were significant between preference-like and fact-like ratings, $F(1, 90) = 13.8$, $p < .001$, and between preference-like and moral-like ratings, $F(1, 90) = 5.55$, $p = .022$, but non-significant between fact-like and moral-like ratings, $F(1, 90) = 0.81$, $p =$

.371. The 2-way interaction between PSC and ROI was significant for preference-like ratings, $F(1, 45) = 4.57$, $p = .038$, and for fact-/moral-like ratings, $F(1, 92) = 8.80$, $p = .004$. Thus, among moral claims, the behavioral-BOLD relationship differed between preference-like ratings and fact-/moral-like ratings; however, in both cases, the relationship differed between ToM and VMPFC ROIs. Based on this, our final model included terms for PSC, PSC x preference-like ratings, and their interactions with VMPFC (in addition to a main effect of rating-type; see Table S7 of the online supplemental materials).

Contrasts in the final model demonstrated that, within the ToM network, PSC was negatively related to fact-/moral-like ratings, $B = -1.01$, $t(140) = 4.32$, $p < .001$, and positively related to preference-like ratings, $B = 0.94$, $t(140) = 2.85$, $p = .020$. Within VMPFC, PSC was also negatively related to fact-/moral-like ratings, $B = -0.28$, $t(140) = 3.18$, $p = .007$, and marginally positively related to preference-like ratings, $B = 0.31$, $t(140) = 2.48$, $p = .055$ (p values corrected for 4 comparisons; $\alpha_{\text{familywise}} = .05$; single-step method). Thus, among moral claims, ToM activity was negatively related to fact-/moral-like ratings, and positively related to preference-like ratings, although both relationships were present to a lesser extent in VMPFC (Figure 3).

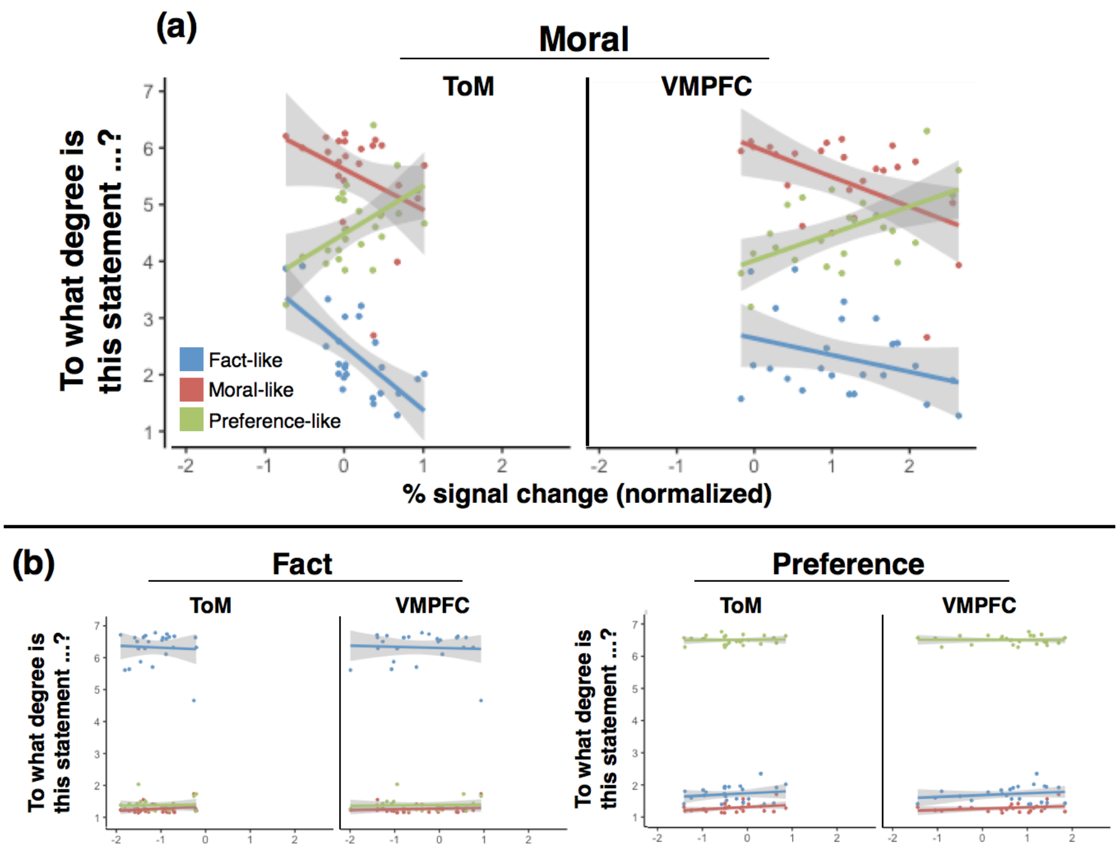


Figure 3. Behavioral–BOLD relationships. BLUPs estimating behavioral ratings were extracted from Study 1 and compared with BLUPs estimating PSC in Study 2. ToM includes averaged estimates for DMPFC, PC, RTPJ, and LTPJ (all by-stimuli random effects were perfectly correlated). (a) Within moral claims, PSC for ToM was positively related to preference-like ratings, and negatively related to fact-/moral-like ratings. These relationships were present to a lesser extent in VMPFC. (b) Within facts and preferences, there was no relationship with PSC. Shaded areas represent 95% confidence intervals.

3.2.4. Model fit comparison. In Study 1, by-stimuli variance in behavioral ratings was high. We tested whether adding the behavioral–BOLD relationships identified in 3.2.3 could partially account for this variance. One linear model predicted behavioral BLUPs (for facts, morals, and preference) using designed contrasts: category x consensus x rating-type. A second model added the four relevant terms identified for moral claims in 3.2.3—a) ToM activity predicting preference-like ratings, b) VMPFC activity predicting preference-like ratings, c) ToM activity predicting fact-/moral-like ratings, and d) VMPFC activity predicting fact-/moral-like ratings. The model comparison was significant, $F(4, 401) = 15.92$, $p = 4.2 \times 10^{-12}$, and remained significant when both models were restricted to only include moral claims, $F(4, 131) = 6.00$, $p = 1.8 \times 10^{-4}$, demonstrating that measures of BOLD activity could account for some of the observed variability.

3.2.5. Whole brain correlation analysis. A whole brain random effects analysis within moral claims clarified whether the observed behavioral–BOLD relationships were specific to the ToM network. We performed three whole brain correlation analyses, testing the relationship between by-stimuli PSC and fact-like, moral-like, and preference-like ratings BLUPs, derived from Study 1 (see 3.1.8). Preference-like ratings were positively correlated with activity in bilateral TPJ (peak coordinates: right [54, -60, 34]; left [-36, -70, 48]), and fact-like ratings were negatively correlated in overlapping regions of bilateral RTPJ (peak coordinates: right [44, -68, 46]; left [-44, -62, -48]). These regions of overlap were slightly dorsal and anterior to the functionally defined ROI positions; however, they did overlap—particularly in RTPJ (Figure 4). Surprisingly, fact-like ratings were negatively correlated with activity in the superior and bilateral middle frontal gyri, and positively correlated with activity in the parietooccipital sulcus. No voxels showed negative associations with preference-like ratings, and no voxels showed either positive or negative associations with moral-like ratings. Thus, fact-like ratings are negatively correlated with activity in several regions, but both fact-like and preference-like ratings were correlated with activity in bilateral TPJ.

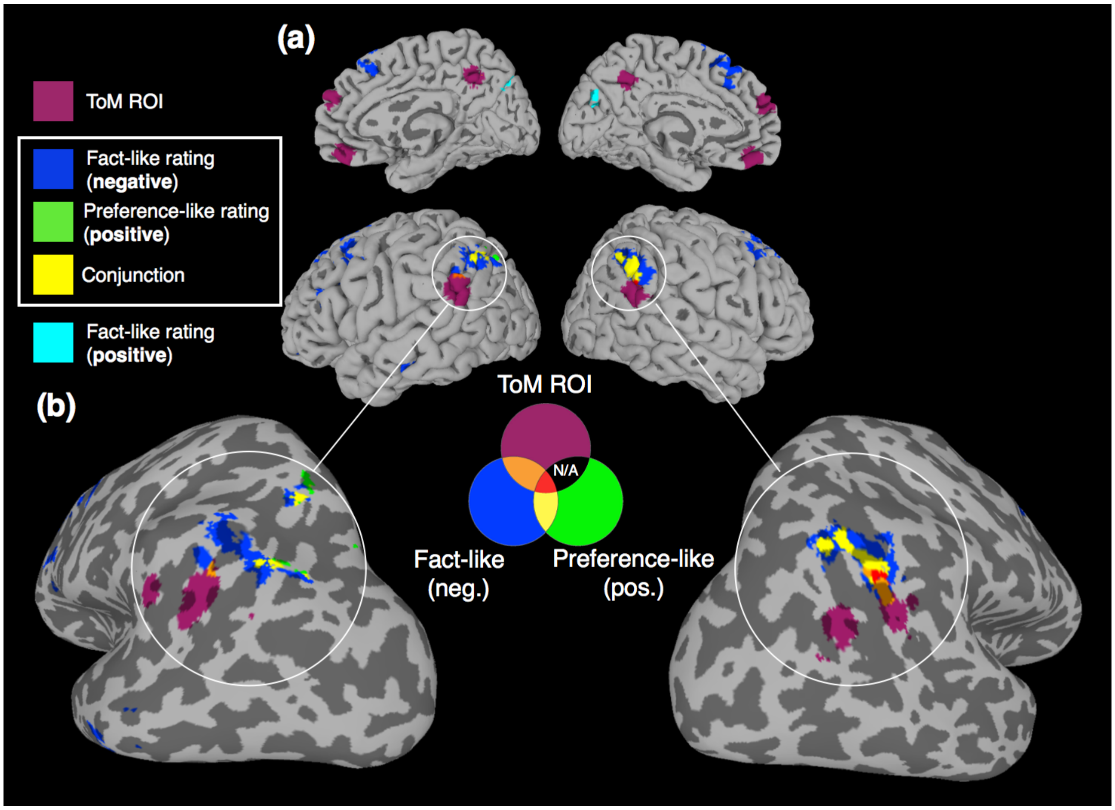

(a)
ToM ROI
Fact-like rating (negative)
Preference-like rating (positive)
Conjunction
Fact-like rating (positive)
(b)
ToM ROI
N/A
Fact-like (neg.)
Preference-like (pos.)

Figure 4. Whole brain behavioral–BOLD correlations, within moral claims. In three separate models, BOLD estimates were correlated with fact-like, moral-like, and preference-like ratings, extracted from Study 1. ToM ROIs are pictured for reference. Fact-like ratings were negatively related to activity in bilateral TPJ, overlapping with regions showing positive correlations with preference-like ratings. These areas of overlap included areas within the defined TPJ ROIs in both hemispheres. Fact-like ratings were also negatively related to BOLD estimates in superior and middle frontal gyri, and positively related in parietooccipital sulcus. For peak coordinates see Table S8 of the online supplemental materials.

3.2.6. ToM–behavioral analysis: controlling for semantic/syntactic features.

Our stimuli could have varied on some other dimension that was not of experimental interest, which could be driving the observed behavioral–BOLD relationship observed in 3.2.3. We collected 13 semantic/syntactic features of our stimuli (e.g. reading ease, noun concreteness; see Appendix B), using Coh-Metrix 3.0 (Graesser et al., 2004; McNamara et al., 2014), as well as in-scanner reaction time. Starting with the ToM network model identified in 3.2.2 (with uncorrelated random effects), we added all features as fixed effects, including their interactions with ROI and category. Features were dropped from the model step-wise. BLUPs were extracted from the resulting model, and the analyses in 3.2.3 were repeated. The results were consistent with the patterns observed in 3.2.3, although we no longer observed an interaction between ToM and VMPFC (for full analysis, see Table S9 of the supplemental online materials)—among moral claims, positive ToM/VMPFC activity was associated with increased preference-like ratings, $B = 1.02$, $t(70) = 2.81$, $p = .013$, and decreased fact-like/moral-like ratings, $B = -0.97$, $t(70) = 3.79$, $p < .001$ (p values corrected for 2 comparisons; $\alpha_{\text{familywise}} = .05$; single-step method).

3.2.7. ToM–behavioral analysis: additional item features. The analyses above demonstrated that, among moral claims, ToM network activity predicts metaethical judgments. Although these findings allow us to examine patterns of ToM network activity *within* the moral domain, they do not allow us to distinguish between processing for morals, facts, and preferences. Behavioral ratings used in the analyses above simply did not vary across facts and preferences, which precluded comparisons between categories.

Fortunately, claims in all categories were designed to elicit a range of agreement, meaning that categories could be compared along this dimension. BLUPs estimating by-stimuli agreement ratings were extracted from Study 1 (from a maximal model, including category x rating-type and all appropriate random effects; consensus sub-categories were not included in the model), and compared with by-stimuli BLUPs estimating ToM network activity (3.2.2). Agreement was fit with a linear model, including the two-way interaction of PSC x category, and the interaction of ROI (ToM/VMPFC) with all other terms. ROI interactions were non-significant and PSC was averaged between ToM and

VMPFC (see Table S10 of the online supplemental materials). We observed an interaction between PSC and category, *F*(2, 66) = 5.92, *p* = .004. Follow-up ANOVAs indicated that relationship between agreement and PSC was significantly different between morals and preferences, *F*(1, 44) = 11.78, was marginal between morals and facts, *F*(1, 44) = 3.71, *p* = .061, and was non-significant between facts and preferences, *F*(1, 44) = 2.32, *p* = .135. Contrasts indicated that ToM (and VMPFC) activity had a negative relationship with agreement among moral claims, *B* = -1.46, *t*(69) = 2.57, *p* = .037, a marginally positive relationship among preferences, *B* = 1.01, *t*(69) = 2.30, *p* = .072, and a non-significant relationship among facts *B* = 0.02, *t*(69) = 0.04, *p* = 1.00 (*p* values corrected for 3 comparisons; $\alpha_{\text{familywise}}$ = .05; single-step method). Thus, for moral claims, agreement was negatively related to PSC across claims and for preferences this relationship was significantly different, and nearly reversed (Figure 5a).

Additional item features, collected in separate online samples (see 3.1.3) contextualized this effect further. These included ratings of *mental states* (*n* = 48), *mental imagery* (*n* = 46), *person present* (*n* = 48), *valence* (*n* = 42), and *arousal* (*n* = 42). BLUPs for these ratings were extracted from maximal mixed effects models, except for person present, which used a generalized linear mixed effects binomial regression, with by-subject random slopes for preferences removed (due to a model conversion failure, stemming from a high correlation with by-subject random intercepts).

For ratings of mental state inference, after collapsing across ROIs (see Table S10 of the online supplemental materials), we observed an interaction between PSC and category, *F*(2, 66) = 5.92, *p* = .004. Follow-up ANOVAs indicated that the relationship was significantly different between morals and preferences, *F*(1, 44) = 12.33, *p* = .001, and between morals and facts, *F*(1, 44) = 5.07, *p* = .029, but non-significant between facts and preferences, *F*(1, 44) = 1.06, *p* = .309. Based on this, we created a model including terms for PSC and PSC x moral-like ratings (in addition to a main effect of category). According to contrasts, for facts and preferences that elicited mental state inferences elicited greater activity throughout the ToM network, *B* = 0.30, *t*(67) = 3.64, *p* = .001, but moral claims did not, *B* = -0.24, *t*(67) = 1.70, *p* = .178 (*p* values corrected for 2 comparisons; $\alpha_{\text{familywise}}$ = .05; single-step method). The model interaction term indicated that the mental-state–ToM relationship was significantly less positive among moral claims, *B* = -0.53, *t*(67) = 3.30, *p* = .001. This was surprising given that the ToM network has been repeatedly shown to be active in response to stories containing information about mental states (Dodell-Feder et al., 2011; Saxe & Kanwisher 2003; Saxe & Powell, 2006; Schurz et al., 2014; Van Overwalle, 2009). However, here we found that ToM activity was related to by-stimuli differences in mental state inference for facts and preferences, but not for morals (Figure 5b). We discuss this finding further in the general discussion.

Of our remaining item features, only the presence of a person was related to ToM network activity (Figure 5c)—we observed a main effect, such that ToM network activity was greater when a person was present, across facts, morals, and preferences, *B* = 1.49, *t*(68) = 2.28, *p* = .026. Other measures were not significantly related to ToM network activity (Figure 5d–e; for analyses, see Table S10 of the online supplemental materials).

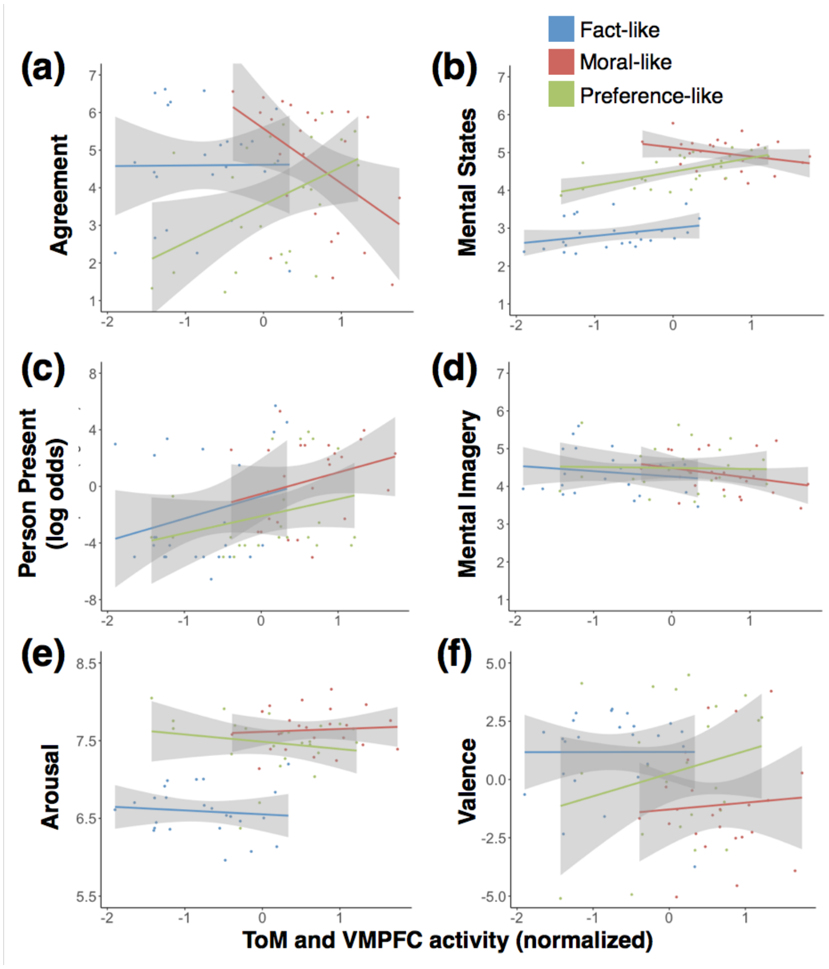


Figure 5. Additional, exploratory behavioral–BOLD relationships. By-stimuli BLUPs predicting agreement were extracted from Study 1, and from independent online studies for the remaining measures. BLUPs estimating ToM and VMPFC activity were extracted from Study 2 (3.2.2) (a) Agreement was negatively associated to ToM/VMPFC activity within moral claims. (b) Mental state inferences were positively associated with ToM/VMPFC activity within facts and preferences, but not within moral claims. (c) The presence of a person in the statement was positively associated with ToM/VMPFC activity for facts, morals, and preferences. (d–f) ToM/VMPFC activity was unrelated to mental imagery, arousal, and valence. Shaded areas represent 95% confidence intervals.

4.1 General Discussion

How is metaethical judgment related to neural activity? We hypothesized that subjective moral claims, which refer to the speaker's beliefs, would elicit greater activity in the ToM network, a set of brain regions active during social cognition and mental state inference (Dodell-Feder et al., 2011; Saxe & Kanwisher 2003; Saxe & Powell, 2006;

Schurz et al., 2014; Van Overwalle, 2009). Consistent with this, preference-like moral claims elicited greater activity throughout the ToM network (Figure 3). Likewise, fact-like moral claims elicited less ToM network activity. Interestingly, some moral claims were rated as more moral-like than others (see Appendix A), and these also elicited less activity throughout the ToM network. In Study 1, by-stimuli variance was high across moral claims (Figure 2b), and the behavioral–neural associations identified in Study 2 explained a significant portion of this variance when added to the model (3.2.4). Whole brain analyses confirmed that these findings were not dependent on our ROI approach: within moral claims, BOLD activity correlated positively with preference-like ratings and negatively with fact-like ratings within overlapping clusters in bilateral TPJ (Figure 4). These findings were also robust to controls for semantic/syntactic features that were not of experimental interest (e.g. reading ease, noun concreteness). Finally, in an exploratory analysis, we observed several distinctions between processing for moral claims, compared to facts and preferences: ToM activity for moral claims, but not facts and preferences, was negatively related to agreement (Figure 5a). Furthermore, for facts and preferences, but not for morals, ToM activity was positively related to mental state inference (Dodell-Feder et al., 2011; *To what extent did this claim make you think about someone's beliefs, thoughts, experiences, or desires?*). This last finding is exploratory, but could be consequential if confirmed in future work, as it suggests that the ToM network may be sensitive to the degree of mental state inference in some domains but not others.

This exploratory finding aside, a great deal of work has established that the ToM network is involved in representing mental states (Schurz et al., 2014; Van Overwalle, 2009), and we take our findings as support for the reverse inference that mental state representation is associated with metaethical judgment. Notably, however, the present work is the first to identify continuous associations between behavioral ratings and by-stimuli ToM network activity: among morals, activity was positively related to preference-like ratings, and negatively related to fact-/moral-like ratings. In prior item analyses bilateral TPJ activity was best accounted for by the belief > photograph contrast used in our functional localizer (PC was related to the number of people per story; DMPFC activity could not be accounted for by any feature; Dodell-Feder et al., 2011). It would be a mistake to conclude from our findings that the function of the ToM network is for metaethical judgment (or that any brain region serves a unique function for moral cognition; Young & Dungan, 2012). However, it has recently been argued that a scientific understanding of ToM could be advanced by "deconstructing and reconstructing" the concept (Schaafsma et al., 2014). ToM network activity tracks with metaethical judgment (a special case of second-order judgment, i.e. judgments about the nature of information expressed), and this relationship may provide clues for future research, taking more direct aim at the underlying cognitive processes represented in the ToM network.

To this end, and in line with the recommendation that ToM be characterized with greater precision (Schaafsma et al., 2014), we briefly speculate about how else to interpret the relationship between ToM network activity and metaethical judgment. Several recent reviews have proposed that activity in social brain regions may be consistent with a prediction error framework (Joiner et al., 2017; Koster-Hale & Saxe, 2013). For instance, social and reinforcement learning may share common mechanisms

(Joiner et al., 2017). Alternatively, prediction error could represent a generalized hierarchical framework, where social information is a high-level abstraction in a system oriented toward predicting incoming sensory information (Clark, 2013; Barrett, 2017; Koster-Hale & Saxe, 2013). In either case, predicted information should elicit less neural activity, while unexpected information should elicit more (in relevant brain regions/networks). Consistent with this, objective, fact-like moral claims may generally be more predictable than subjective, preference-like morals, or preferences. For instance, most people would predict that others hold certain common moral beliefs (e.g. that drinking and driving is bad, that slavery is wrong, etc.). If someone endorses a predictable moral belief, then incoming sensory information matches the prediction, and there is no prediction error to account for (see also Dungan et al., in press).

While promising, this hypothesis must be directly tested in future work. In particular, it is worth exploring the extent that abstract predictions differ between domains (e.g. morals vs. preferences). We observed some evidence of a domain distinction (Figure 5); however, these analyses were exploratory, and outside of metaethical judgment, the central focus of the present work. However, it is worth noting that our design intentionally presented claims in isolation, without an associated speaker or social context. This design feature may be important: moral claims may represent a special case of social prediction, in that moral predictions are generalized across people and contexts in a way that are preferences are not. Regardless of where you are, or whom you are speaking to, you may expect that people will endorse a set of common moral beliefs. Morals are strong social predictions.

4.2 Conclusion

We began by suggesting that subjective moral claims may elicit greater activity in the ToM network, given that they refer to the speaker's beliefs (Goodwin & Darley, 2010; Sayre-McCord, 1986). Our findings are consistent with this hypothesis, but also leave open the possibility that the ToM network implements a more fundamental feature of social processing, such as social prediction error (Joiner et al., 2017; Koster-Hale & Saxe, 2013). Thus, the present work provides evidence for the neural representation of metaethical judgment. We believe that these findings, and our methodological approach, can be harnessed in future work to sharpen accounts of the underlying components of social cognition.

References

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

Bartoń, K (2016). MuMIn: Multi-Model Inference. R package version 1.15.6. Retrieved from https://CRAN.R-project.org/package=MuMIn

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

Bedny, M., Aguirre, G. K., & Thompson-Schill, S. L. (2007). Item analysis in functional magnetic resonance imaging. *Neuroimage, 35(3)*, 1093-1102. http://dx.doi.org/10.1016/j.neuroimage.2007.01.039

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.

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

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

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

Dungan, J., Stepanovic, M., & Young, L. (in press). Theory of mind for processing unexpected events across contexts. *Social Cognitive and Affective Neuroscience*. http://dx.doi.org/10.1093/scan/nsw032

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

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

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

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

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal 50(3)*, 346--363. http://dx.doi.org/10.1002/bimj.200810425

Joiner, J., Piva, M., Turrin, C., & Chang, S. (2017). Social learning through prediction error in the brain. *npj Science of Learning, 2*, 8. http://dx.doi.org/10.1038/s41539-017-0009-2

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

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron, 79*, 836–848. http://dx.doi.org/10.1016/j.neuron.2013.08.020

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).

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

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

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*, 133-142. http://dx.doi.org/10.1111/j.2041-210x.2012.00261.x

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping, 15*, 1-25. http://dx.doi.org/10.1002/hbm.1058

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

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

Ruff, C. C., Ugazio, G., & Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science, 342*, 482-484. http://dx.doi.org/10.1126/science.1241399

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., & 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

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences, 19,* 65–72. http://dx.doi.org/10.1016/j.tics.2014.11.007

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

Theriault, J., Waytz, A., Heiphetz, L., & Young, L. (in press). Examining overlap in behavioral and neural representations of morals, facts, and preferences. *Journal of Experimental Psychology: General*. http://dx.doi.org/10.1037/xge0000350

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

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

Westfall, J., Nichols, T. E., & Yarkoni, T. (2016). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Wellcome open research*, 1. http://dx.doi.org/10.12688/wellcomeopenres.10298.2

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

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

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. (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

Zaki, J., Schirmer, J., & Mitchell, J. P. (2011). Social influence modulates the neural computation of value. *Psychological science, 22(7)*, 894-900. http://dx.doi.org/10.1177/0956797611411057

Appendix A. Experimental Stimuli and BLUPs estimates.

Morals					
Positive-consensus	**ToM BLUPs**	**VMPFC BLUPs**	**Fact-like BLUPs**	**Moral-like BLUPs**	**Preference-like BLUPs**
The goal of sports should be to teach children that respect for others is more important than winning.	0.011	1.1269	3.0221	6.2564	3.8453
Parents should be willing to make sacrifices for the benefit of their baby.	-0.5251	0.5227	3.9144	5.9972	4.0798
It is irresponsible for airlines to risk the safety of their passengers.	0.2148	0.2742	3.2157	5.9815	4.303
Professors should not tolerate students cheating on their exams.	-0.7352	-0.0465	3.8749	6.2122	3.2389
Driving after drinking heavily is a stupid and selfish way to behave.	0.1883	1.5678	3.0313	5.7222	4.8858
The deplorable conditions of Chinese electronics workers should not be ignored.	0.0277	0.9978	2.0062	4.5673	5.3466
No-consensus					
It is unethical for businesses to promote sugary products to children.	0.3939	1.7801	2.5669	6.14	4.6066
It is wrong to cheat when playing games such as Monopoly.	-0.2061	1.1528	3.3337	5.9298	4.1966
Harry Potter should be banned from school libraries for idolizing witchcraft.	-0.229	0.9305	2.497	6.1893	3.9591
People should help their elderly neighbors clear snow from their driveway.	0.4754	0.8605	2.13	6.0441	4.4336
It is wrong to knowingly buy sandals made using sweatshop labor.	0.9306	2.5563	1.9186	5.1095	5.2449
Good Americans buy American cars, such as Hummers.	-0.0027	0.4265	1.947	5.4241	5.0757
Eating fish is acceptable if they were treated humanely when caught or raised.	0.371	2.2236	1.4816	2.6911	6.4033

Music stores should prevent children from buying CDs with violent or sexist lyrics.	0.0083	0.2	2.1262	6.1149	4.5567
It is unjust for businesses to allow apples to rot rather than giving them to the needy.	1.0134	1.664	2.0092	5.6906	4.6681
Destroying the habitats of owls through deforestation is deplorable.	0.673	2.6323	1.2825	3.9891	5.6957
Dog racing is harmful and exploitative to the dogs being raced.	-0.0669	-0.0154	2.1859	6.1189	4.2002
It is wrong to use animals as disposable space shuttle test pilots.	0.0183	2.0761	2.1744	5.8517	4.3915
Negative-consensus					
It is fine for doctors to accidentally kill a small number of patients per year.	0.6888	1.2232	1.6661	5.3421	4.8423
Child labor in coffee bean farming is acceptable because it lowers the market price.	0.4606	1.284	1.6699	4.8386	4.805
Private beaches are immoral, as everyone should be able to share the space.	-0.07	1.4002	2.0179	5.5038	5.1079
Sport fishing to kill and eat fish is barbaric and evil.	-0.0172	0.6177	1.7373	4.6917	5.2054
Universal donors should be obligated to donate their blood.	-0.0665	1.8448	2.5858	5.7527	4.0372
It is wrong to harm cockroaches just because humans find them disgusting.	0.3621	-0.1727	1.5847	6.0397	3.8442
			Preferences		
Positive-consensus					
Afterschool programs involving sports are more fun than most of the alternatives available to children.	-0.9189	0.1251	1.8004	1.1998	6.6632
Babies that are temperamental are aggravating to spend time around.	0.6489	1.6982	1.9433	1.7279	6.4285

Going through airport security is an unpleasant experience.	-0.1557	1.3853	1.9827	1.2996	6.4646
Professors who play videos make their classes more entertaining.	0.3005	1.2072	2.372	1.3584	6.3961
Having a drink every now and then is a good way to relax.	-0.5007	0.7931	1.971	1.4424	6.2982
Using touchscreens is a much more satisfying way to interact with computers.	-0.4045	1.4263	1.9626	1.3536	6.4338
No-consensus					
Any ice cream flavor tastes better when served in a crunchy waffle cone.	-0.1058	0.6164	1.8833	1.3411	6.5164
Many games are better than Monopoly, which is incredibly boring.	0.0287	0.648	1.5868	1.4261	6.5202
The Harry Potter books are engaging and delightful to read, even for adults.	-0.9634	0.5373	1.6953	1.2657	6.5675
In the wintertime, it is fun to catch snowflakes on the tip of your tongue.	-0.5099	0.4378	1.7474	1.1547	6.6263
Because sandals have fewer styles, they are less fun to go shopping for.	0.5904	1.8393	1.4535	1.195	6.5902
Nothing is more awesome than driving in a Hummer.	0.3776	1.0699	1.4302	1.2174	6.665
Sitting in a boat and fishing all day long is boring and a waste of time.	-0.573	1.0348	1.423	1.1657	6.7799
Rock music is pleasing to the ear, and much more agreeable than rap music.	-0.985	0.4179	1.8624	1.2667	6.5586
Green apples are too sour to be an enjoyable lunchtime snack.	-0.5057	0.6827	1.93	1.2962	6.3668
The "hoots" of owls in the woods make camping more enjoyable.	0.8586	1.1436	2.0413	1.3028	6.6514
Dogs are not worth the stress and aggravation it takes to own them.	-0.0913	1.3158	1.4839	1.2129	6.6934
Gazing at planets through a telescope is a satisfying activity.	-1.2491	-1.05	1.6208	1.2306	6.5764
Negative-consensus					

Having a doctor listen attentively to your medical concerns is awful.	-0.4857	-0.5003	1.6712	1.3205	6.6532
Drinking coffee is a miserable experience when you are tired and need energy.	-0.4634	-0.2433	1.5413	1.3137	6.3696
While at a hot beach, it is agonizing to dip your toes in the cool water.	-1.413	-1.4443	1.4348	1.3312	6.5875
Nothing is more appealing than the smell of rotting fish.	-1.3909	-0.9091	1.8264	1.2195	6.301
Having blood drawn is a pleasurable experience.	-0.2406	0.8269	1.5902	1.178	6.478
Cockroaches are delicious to eat because of their hard and crunchy shell.	-0.0912	1.4404	1.4676	1.5881	6.3524
Facts					
Positive-consensus					
In sports-based afterschool programs children participate in sports such as baseball or basketball to name a few.	-1.3451	-1.1744	6.689	1.1949	1.4992
In a full-term human pregnancy, babies spend nine months in a woman's womb.	-1.3908	-1.3893	6.6404	1.5536	1.375
Airplanes have wings that enable the plane to lift upwards.	-0.2091	0.5427	6.3226	1.2109	1.7311
University professors teach classes but also conduct research.	-1.5847	-0.7983	6.6292	1.1615	1.277
A breathalyzer is used to determine whether a driver is intoxicated.	-1.1268	-0.3927	6.7811	1.3927	1.206
Touchscreens are used in a variety of electronics, including smartphones.	-1.5024	-0.9464	6.2851	1.1473	2.0334
No-consensus					
The very first waffle cone was invented in Chicago, Illinois, at a state fair.	-0.8656	-0.2341	6.7416	1.1645	1.2553
Monopoly pieces were made from wood, not metal, during WWI.	-1.9031	-1.3962	6.7146	1.2293	1.3857
The author J.K. Rowling has two younger siblings, one brother and one sister.	-1.2773	-0.0252	6.5033	1.1704	1.221

A town in North Dakota holds the world record for the tallest snowman.	-0.9766	0.02	6.5508	1.2877	1.3402
The oldest sandals in the world were found in Oregon's Paisley Caves.	-0.7217	0.7578	6.3238	1.2247	1.4223
Hummer trucks were first marketed to civilians in 1990.	-1.3589	-1.3786	6.2895	1.2685	1.2493
There are more fish species in the Amazon River than in the Atlantic Ocean.	-0.9593	0.3892	6.6263	1.1715	1.3217
The first CD made for commercial release was the rock CD: "Born in the USA".	-1.5318	-1.2682	6.5087	1.2753	1.4208
Newtown Pippin was the first apple variety exported from the US.	-0.8145	-0.6858	6.6517	1.1539	1.2155
Of all types of birds, owls are the ones that can see the color blue.	-0.226	0.5963	6.6338	1.2753	1.1995
The dog breed, Basenji, is the world's only barkless dog breed.	-0.699	0.4126	6.6678	1.1593	1.2906
Saturn’s moon, Titan, is the only moon known to have clouds.	-0.8793	-0.0489	6.5895	1.2161	1.3635
Negative-consensus					
Medical students at hospitals are able to perform surgeries with little to no training.	-1.7082	-1.072	5.6398	1.2107	1.4144
Coffee beans grow particularly well in freezing cold climates, such as Alaska and Russia.	-1.8077	-1.992	5.6126	1.2347	1.2591
The sand on beaches is usually transported there from nearby deserts.	-1.1876	-0.5137	5.7055	1.4063	1.2456
Fish are able to live outside of water for an extended time.	-0.8853	0.0665	6.11	1.2452	1.2927
In humans, the liver pumps blood throughout the body.	-1.4569	-1.0238	5.8752	1.1975	1.4958
Cockroaches are a type of cold-blooded reptiles related to snakes.	-0.2611	0.9317	4.6601	1.734	1.6692

ToM BLUPs average estimates for DMPFC, PC, RTPJ, and LTPJ, as all by-stimuli random slopes were perfectly correlated. For model details, see Tables S4 and S6 of the supplemental online materials.

Appendix B. List of covariates, with descriptions.

Semantic/syntactic measures (2.2.6)		
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.
Reaction time	In-scanner N = 25	The time from the appearance of the in-scanner agreement rating prompt to the input of a response by the participant.
Online Item Features		
Agreement	Study 1 (N = 49)	“To what extent do you agree / disagree with this statement?” (1-7; “strongly disagree”-“strongly agree”).
Valence	Online sample (N = 42)	Valence was the difference between unipolar positive 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	Online sample (N = 42)	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	Online sample (N = 46)	"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	Online sample (N = 48)	"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	Online sample (N = 48)	"Does this statement mention people or a person?" ("Yes" / "No").

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.

Supplemental Materials

Table S1. Study 1 condition means.

Behavioral ratings

Model:
Rating ~ Category * Rating-type * Consensus +
(1+ Category * Rating-type | ID) +
(1 + Rating-type | Item)

		Consensus		
		Positive-consensus	**No-consensus**	**Negative-consensus**
Category	**Rating-type**	Mean (SE)	Mean (SE)	Mean (SE)
Morals	**About Facts**	3.24 (0.21) ^{A}	2.13 (0.18) ^{B}	1.83 (0.21) ^{B}
	About Morals	5.78 (0.28)	5.46 (0.23)	5.33 (0.28)
	About Preferences	4.28 (0.28)	4.81 (0.25)	4.59 (0.28)
Facts	**About Facts**	6.59 (0.17) ^{A}	6.59 (0.17) ^{B}	5.53 (0.20) ^{B}
	About Morals	1.31 (0.24)	1.19 (0.18)	1.35 (0.24)
	About Preferences	1.58 (0.21)	1.29 (0.16)	1.36 (0.21)
Preferences	**About Facts**	2.03 (0.20	1.68 (0.17)	1.57 (0.20)
	About Morals	1.40 (0.24)	1.26 (0.18)	1.32 (0.24)
	About Preferences	6.46 (0.22)	6.60 (0.17)	6.64 (0.22)

Agreement ratings

Model:
Agreement ~ Category * Consensus +
(1+ Category | ID) +
(1 | Item)

	Consensus		
	Positive-consensus	**No-consensus**	**Negative-consensus**
Category	Mean (SE)	Mean (SE)	Mean (SE)
Fact	6.41 (0.40)	4.76 (0.31)	2.46 (0.42)
Moral	6.05 (0.40)	4.90 (0.29)	2.66 (0.39)
Preference	4.98 (0.40	4.08 (0.28)	1.58 (0.40)

Mean estimates and standard errors are derived from contrasts within the models described in 2.2.1 and 2.2.2. Superscripts denote significant differences within each row (*p* values corrected for 27 comparisons; single-step method; $\alpha_{\text{familywise}}$ = .05; single-step method).

Table S2. Study 2 in-scanner agreement ratings.

	Consensus		
	Positive-consensus	**No-consensus**	**Negative-consensus**
Category	Mean (SD)	Mean (SD)	Mean (SD)
Facts	3.64 (0.06)	2.68 (0.16)	1.47 (0.11)
Morals	3.64 (0.07)	2.81 (0.06)	1.89 (0.11)
Preferences	2.93 (0.07)	2.56 (0.06)	1.46 (0.12)

Mean and standard error (across participants). All comparisons, within content categories, were significant at $p < .001$.

Table S3. Theory of Mind network ROI coordinates.

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). All coordinates are reported in MNI space.

Table S4. Study 1 behavioral rating model details.

ToM network activity

Model:
Rating ~ Category * Rating-type +
(1+ Category * Rating-type | ID) +
(1 + Rating-type | Item)

REML criterion at convergence: 32392.5

Dummy coded control conditions: Facts (category) & Fact-like ratings (rating-type)

Random effects structure (by-subject)

	Variance	St.Dev	Correlations							
			Intercept	Moral	Pref	Moral-like	Pref-like	M* M-like	M* P-like	P* Ml-like
Intercept	0.08	0.89								
Moral	2.52	1.59	-0.86							
Preference	2.37	1.54	-0.89	0.94						
Moral-like	1.94	1.39	-0.95	0.84	0.90					
Preference-like	2.24	1.50	-0.97	0.88	0.94	0.98				
M*Moral-like	7.23	2.69	0.84	-0.89	-0.87	-0.88	-0.87			
M*Pref-like	6.85	2.62	0.73	-0.82	-0.84	-0.77	-0.77	0.74		
P*Moral-like	2.34	1.53	0.91	-0.94	-1.00	-0.91	-0.95	0.89	0.85	
P*Pref-like	7.70	2.78	0.89	-0.90	-0.98	-0.94	-0.94	0.88	0.86	0.98

Random effects structure (by-stimuli)

	Variance	St.Dev	Correlations	
			Intercept	Moral-like
Intercept	.301	.549		
Moral-like	.349	.590	-0.58	
Pref-like	.750	.866	-0.90	0.27

Residual

Variance	St.Dev
1.10	1.05

Fixed Effects			
Name	**B (SE)**	**t(df)**	**p**
Intercept	6.32 (0.17)	*t*(101.1) = 36.85	< .001 ***
Moral	-3.99 (0.28)	*t*(88.2) = 14.25	< .001 ***
Preference	-4.58 (0.27)	*t*(90.0) = 16.70	< .001 ***
Moral-like	-5.06 (0.24)	*t*(79.0) = 21.38	< .001 ***
Preference-like	-4.94 (0.28)	*t*(98.6) = 17.59	< .001 ***
Moral*Moral-like	8.24 (0.42)	*t*(66.1) = 19.40	< .001 ***
Preference*Moral-like	4.63 (0.28)	*t*(91.9) = 16.32	< .001 ***
Moral*Preference-like	7.23 (0.45)	*t*(86.2) = 15.93	< .001 ***
Preference*Preference-like	9.72 (0.47)	*t*(82.8) = 20.57	< .001 ***

St.Dev = standard deviation. *** *p* < .001; ** *p* < .01; * *p* < .05; † *p* < .10

Table S5. Study 2 behavioral rating means.

		Consensus		
		Positive-consensus	**No-consensus**	**Negative-consensus**
Category	**Rating-type**	Mean (SE)	Mean (SE)	Mean (SE)
Morals	**About Facts**	3.14 (0.24) $^{\mathbf{A}}$	2.30 (0.22) $^{\mathbf{B}}$	2.01 (0.24) $^{\mathbf{B}}$
	About Morals	6.35 (0.18) $^{\mathbf{A}}$	5.88 (0.17) $^{\mathbf{B}}$	5.87 (0.18) $^{\mathbf{B}}$
	About Preferences	4.23 (0.31)	4.31 (0.30)	4.34 (0.31)
Facts	**About Facts**	6.59 (0.19)	6.65 (0.18)	6.38 (0.19)
	About Morals	1.55 (0.21)	1.29 (0.20)	1.57 (0.21)
	About Preferences	1.85 (0.22) $^{\mathbf{A}}$	1.37 (0.21) $^{\mathbf{B}}$	1.68 (0.22) $^{\mathbf{AB}}$
Preferences	**About Facts**	2.52 (0.23)	1.95 (0.21)	2.14 (0.23)
	About Morals	2.11 (0.25)	1.74 (0.24)	1.90 (0.25)
	About Preferences	6.51 (0.19) $^{\mathbf{A}}$	6.58 (0.18) $^{\mathbf{B}}$	6.32 (0.19) $^{\mathbf{AB}}$

Mean and standard error are estimated using contrasts within the model defined in 3.2.1. Superscripts denote significant differences within each row (*p* values corrected for 27 comparisons; single-step method; $\alpha_{\text{familywise}}$ = .05; single-step method).

Table S6. Study 2 ToM network mixed effects model.

ToM network activity

Model:
PSC ~ Category * ROI +
(1+ Moral + Preference + VMPFC + PC + RTPJ + LTPJ + Moral*(VMFPC+LTPJ) | ID) +
(1 + VMPFC | Item)

REML criterion at convergence: 2271

Dummy coded control conditions: Facts (category) & DMPFC (ROI)

Random effects structure (by-subject)

	Variance	St.Dev	Correlations							
			Intercept	**VMPFC**	**PC**	**RTPJ**	**LTPJ**	**Moral**	**Pref**	**M* VMPFC**
Intercept	.008	.091								
VMPFC	.010	.101	0.01							
PC	.008	.090	-0.60	-0.13						
RTPJ	.011	.105	-0.57	-0.20	0.71					
LTPJ	.007	.085	-0.46	-0.22	0.54	0.53				
Moral	.003	.053	-0.27	0.24	-0.34	-0.30	-0.41			
Pref	.002	.042	-0.07	0.18	-0.14	0.07	-0.11	0.52		
M*VMPFC	.002	.042	-0.14	-0.08	-0.43	-0.18	-0.41	0.62	0.66	
M*LTPJ	.002	.041	-0.33	0.30	0.59	0.35	0.64	-0.08	0.37	-0.30

Random effects structure (by-stimuli)

	Variance	St.Dev	Correlations
			Intercept
Intercept	.002	.048	
VMPFC	.004	.064	.03

Residual

Variance	St.Dev
.071	.027

Fixed Effects			
Name	**B (SE)**	**t(df)**	**p**
Intercept	-0.147 (0.023)	*t*(46) = 6.27	< .001 ***
Moral	0.166 (0.023)	*t*(128) = 7.05	< .001 ***
Preference	0.160 (0.022)	*t*(169) = 7.11	<. 001 ***
VMPFC	0.084 (0.029)	*t*(51) = 2.93	.005 **
PC	0.005 (0.024)	*t*(47) = 0.20	.846
RTPJ	0.005 (0.026)	*t*(42) = 0.18	.858
LTPJ	0.105 (0.023)	*t*(45) = 4.55	< .001 ***
Moral*VMPFC	-0.035 (0.030)	*t*(97) = 1.18	.241
Moral*PC	-0.059 (0.022)	*t*(844) = 2.68	.007 **
Moral*RTPJ	-0.113 (0.022)	*t*(844) = 5.12	< .001 ***
Moral*LTPJ	-0.068 (0.024)	*t*(162) = 2.88	.004 **
Pref*VMPFC	-0.074 (0.029)	*t*(114) = 2.56	.012 *
Pref*PC	-0.123 (0.022)	*t*(844) = 5.56	< .001 ***
Pref*RTPJ	-0.152 (0.022	*t*(844) = 6.89	< .001 ***
Pref*LTPJ	-0.11 (0.022)	*t*(844) = 5.05	< .001 ***

St.Dev = standard deviation. *** p < .001; ** p < .01; * p < .05; † p < .10

Table S7. Study 2 ToM–behavioral analysis.

DV: Behavioral ratings (Study 1 BLUPs)			
Term		***F statistic***	***p***
PSC x 2 (ROI: ToM/VMPFC) x 3 (rating-type (fact-/moral-/preference-like) x 3 (category: fact/moral/preference)		*F*(2, 405) = 5.73	< .001 ***
PSC x 2 ROI x 3 rating-type x 2 (fact/preference)		*F*(2, 270) = 0.01	.986
PSC x 2 ROI x 3 rating-type x 2 (fact/moral)		*F*(2, 270) = 7.42	< .001 ***
PSC x 2 ROI x 3 rating-type x 2 (moral/preference)		*F*(2, 270) = 6.64	.002 **
Within moral claims			
PSC x 2 ROI x 3 rating-type		*F*(2, 135) = 6.03	.003**
PSC x 2 ROI x 2 (fact-like/preference-like)		*F*(1, 90) = 13.77	< .001 ***
PSC x 2 ROI x 2 (fact-like/moral-like)		*F*(1, 90) = 0.81	.371
PSC x 2 ROI x 2 (moral-like/preference-like)		*F*(1, 90) = 5.47	.022 *
Within preference-like ratings			
PSC x ROI		*F*(1, 45) = 4.57	.038 *
Within fact-/moral-like ratings			
PSC x ROI		*F*(1, 92) = 8.80	.004 **
Model: rating-type + (PSC x ROI) + (PSC x ROI x preference-like)	***B (SE)***	***t statistic***	**p**
Intercept (Fact-like rating)	2.57 (0.11)	*t*(137) = 22.80	< .001 ***
Moral-like	3.17 (0.14)	*t*(137) = 23.00	< .001 ***
Preference-like	1.80 (0.17)	*t*(137) = 10.62	< .001 ***
PSC (within fact-like/moral-like)	-1.01 (0.23)	*t*(137) = 4.32	< .001 ***
PSC x preference-like (within ToM)	1.94 (0.40)	*t*(137) = 4.81	< .001 ***
PSC x ROI (interaction for VMFPC, within fact-/moral-like)	0.72 (0.23)	*t*(137) = 3.11	.002 **
PSC x preference-like x ROI (interaction for VMPFC, for preference-like)	-1.35 (0.40)	*t*(137) = 3.35	.001 **
Contrasts:	***B (SE)***	***t statistic***	**p**
Fact-/moral-like–ToM relationship	-1.01 (0.23)	*t*(140) = 4.32	< .001 ***
Preference-like–ToM relationship	0.94 (0.33)	*t*(140) = 2.85	.020 *
Fact-/moral-like-VMPFC relationship	-0.28 (0.09)	*t*(140) = 3.18	.007 **
Preference-like–VMPFC relationship	0.31 (0.12)	*t*(140) = 2.48	.055 †

Contrast *p* values corrected for 4 comparisons; single-step method; $\alpha_{\text{familywise}}$ = .05; single-step method. *** *p* < .001; ** *p* < .01; * *p* < .05; † *p* < .10

Table S8. Whole brain correlation peak coordinates.

Contrast	Name	Cluster Size	Peak T	x	y	z
Fact-like rating (negative)	M Superior frontal gyrus	1037	7.60	-8	24	58
			6.76	-18	20	62
			5.40	16	44	48
	L Middle frontal gyrus	658	7.22	-28	10	50
			6.24	-42	12	44
			5.33	-48	28	24
	L Angular gyrus	398	6.53	-44	-62	-48
			5.60	-38	-72	48
			5.23	-34	-62	48
	L Middle temporal gyrus	142	6.51	-58	-32	-16
			5.57	-66	-26	-12
	R Angular gyrus	412	6.43	44	-68	46
			5.76	50	-56	32
			5.74	52	-62	40
	R Middle frontal gyrus	239	5.58	40	20	44
			5.14	32	28	42
	L Medial caudate nucleus	132	5.18	-14	14	10
			5.16	-12	6	14
Fact-like rating (positive)	M Parietooccipital sulcus	145	4.85	8	-78	36
			4.83	-6	-80	18
			4.14	6	-84	44
Preference-like rating (positive)	R Angular gyrus	124	5.12	54	-60	34
			4.84	50	-62	42
			3.95	44	-68	46
	L Angular gyrus	102	4.98	-36	-70	48
			4.36	-30	-64	54
			4.14	-30	-76	46

First level models produced a beta map for each item, for each participant. For each participant, 3 models, predicting by-item estimates were created, with fact-like, moral-like, and preference-like ratings as respective predictors. Beta maps for ratings from each model were 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 S9. Study 2 ToM–behavioral analysis: controlling for semantic/syntactic features

Model:
PSC ~ Category*ROI*NounConcreteness + ROI*LeftEmbeddedness + NounFamiliarity +
(1+ Moral + Preference + VMPFC + PC + RTPJ + LTPJ + Moral*(VMFPC+LTPJ) | ID) +
(1 + VMPFC | Item)

REML criterion at convergence: 2378.2

Dummy coded control conditions: Facts (category) & DMPFC (ROI)

DV: Behavioral ratings (Study 1 BLUPs)
PSC corrected for syntactic/semantic features

Term		***F statistic***	***p***
PSC x 2 (ROI: ToM/VMPFC) x 3 (rating-type (fact-/moral-/preference-like) x 3 (category: fact/moral/preference)		*F*(4, 405) = 0.86	.486
PSC averaged across ROI			
PSC x 3 rating-type x category		*F*(4, 198) = 9.54	<.001 ***
PSC x 3 rating-type x category (fact/preference)		*F*(2, 132) = 0.61	.542
PSC x 3 rating-type x category (fact/moral)		*F*(2, 132) = 9.81	<.001 **
PSC x 3 rating-type x category (moral/preference)		*F*(2, 132) = 14.17	<.001 **
Within moral claims			
PSC x 3 rating-type		*F*(2, 66) = 9.89	<.001 ***
PSC x 2 (fact-like/preference-like)		*F*(1, 44) = 18.20	<.001 ***
PSC x 2 (fact-like/moral-like)		*F*(1, 44) = 0.001	.972
PSC x 2 (moral-like/preference-like)		*F*(1, 44) = 13.97	< .001 ***
Model: rating-type + PSC + (PSC x preference-like)	***B (SE)***	***t statistic***	**p**
Intercept (Fact-like rating)	2.18 (0.14)	*t*(67) = 15.31	< .001 ***
Moral-like	3.18 (0.19)	*t*(67) = 16.50	< .001 ***
Preference-like	2.61 (0.21)	*t*(67) = 12.72	< .001 ***
PSC (within fact-like/moral-like)	-0.97 (0.26)	*t*(67) = 3.79	< .001 ***
PSC x preference-like	1.99 (0.44)	*t*(67) = 4.48	< .001 ***
Contrasts:	***B (SE)***	***t statistic***	**p**
Fact-/moral-like–ToM relationship	-0.97 (0.26)	*t*(70) = 3.79	< .001 ***
Preference-like–ToM relationship	1.02 (0.36)	*t*(70) = 2.81	.013 *

Contrast *p* values corrected for 2 comparisons; single-step method; $\alpha_{\text{familywise}}$ = .05; single-step method. *** *p* < .001; ** *p* < .01; * *p* < .05; † *p* < .10

Table S10. Model simplification for item features.

DV: Agreement (Study 1 BLUPs)

Term		***F statistic***	***p***
PSC x 3 (category: fact/moral/preference) x 2 (ROI: ToM/VMPFC)		*F*(2, 135) = 1.27	.284
PSC x 2 (ROI)		*F*(1, 137) = 0.30	.587
PSC averaged across ROI			
PSC x 3 (category)		*F*(2, 66) = 5.92	.004 **
PSC x 2 (category: moral/fact)		*F*(1, 44) = 3.71	.061 †
PSC x 2 (category: moral/preference)		*F*(1, 44) = 11.78	.001 **
PSC x 2 (category: fact/preference)		*F*(1, 44) = 2.32	.135
Model: PSC x 3 (category)	***B (SE)***	***t statistic***	**p**
Contrast: PSC *within Facts*	0.02 (0.50)	*t*(69) = 0.04	1.00
Contrast: PSC *within Morals*	-1.46 (0.59)	*t*(69) = 2.59	.037 *
Contrast: PSC *within Preferences*	1.01 (0.44)	*t*(69) =2.30	.072 †

DV: Mental State (Online Sample BLUPs)

Term		***F statistic***	***p***
PSC x 3 (category: fact/moral/preference) x 2 (ROI: ToM/VMPFC)		*F*(2, 135) = 1.36	.261
PSC x 2 (ROI)		*F*(2, 137) = 0.04	.844
PSC averaged across ROI			
PSC x 3 (category)		*F*(2, 66) = 5.92	.004 **
PSC x 2 (category: moral/fact)		*F*(1, 44) – 5.07	.029 *
PSC x 2 (category: moral/preference)		*F*(1, 44) = 12.33	.001 **
PSC x 2 (category: fact/preference)		*F*(1, 44) = 1.06	.309
Model: DV ~ category + PSC + (moral x PSC)	***B (SE)***	***t statistic***	**p**
PSC	0.29 (0.08)	*t*(67) = 3.64	< .001 ***
PSC x moral	-0.53 (0.16)	*t*(67) = 3.30	.001 **
Contrast: PSC *within facts/preferences*	0.30 (0.08)	*t*(67) = 3.64	.001 **
Contrast: PSC *within morals*	-0.24 (0.14)	*t*(67) = 1.70	.178

DV: Person Present (Online Sample BLUPs)

Term		***F statistic***	***p***
PSC x 3 (category: fact/moral/preference) x 2 (ROI: ToM/VMPFC)		*F*(2, 135) = 0.61	.546
PSC x 2 (ROI)		*F*(1, 137) = 0.22	.637
PSC averaged across ROI			
PSC x 3 (category)		*F*(2, 66) = .04	.962
Model: DV ~ category + PSC	***B (SE)***	***t statistic***	**p**
Main effect: PSC (*within* moral/preference/fact)	1.49 (0.65)	*t*(68) = 2.28	.026 *

DV: Mental Imagery (Online Sample BLUPs)

Term		***F statistic***	***p***
PSC x 3 (category: fact/moral/preference) x 2 (ROI: ToM/VMPFC)		*F*(2, 135) = 0.31	.732
PSC x 2 (ROI)		*F*(1, 137) = 0.07	.797
PSC averaged across ROI			
PSC x 3 (category)		*F*(2, 66) = 0.39	.680
Model: DV ~ category + PSC	***B (SE)***	***t statistic***	**p**
Main effect: PSC (*within* moral/preference/fact)	-0.13 (0.11)	*t*(68) = 1.15	.253

DV: Arousal (Online Sample BLUPs)

Term		***F statistic***	***p***
PSC x 3 (category: fact/moral/preference) x 2 (ROI: ToM/VMPFC)		*F*(2, 135) = 0.11	.892
PSC x 2 (ROI)		*F*(1, 137) = 0.26	.611
PSC averaged across ROI			
PSC x 3 (category)		*F*(2, 66) = 0.35	.703

Model: DV ~ category + PSC	***B (SE)***	***t statistic***	**p**
Main effect: PSC (*within* moral/preference/fact)	-0.05 (0.06)	*t*(68) = 0.77	.443
DV: Valence (Online Sample BLUPs)			
Term	***F statistic***		***p***
PSC x 3 (category: fact/moral/preference) x 2 (ROI: ToM/VMPFC)		*F*(2, 135) = 0.13	.868
PSC x 2 (ROI)		*F*(1, 137) = 0.03	.857
PSC averaged across ROI			
PSC x 3 (category)		*F*(2, 66) = 0.03	.857
Model: DV ~ category + PSC	***B (SE)***	***t statistic***	**p**
Main effect: PSC (*within* moral/preference/fact)	0.51 (0.47)	*t*(68) = 1.09	.281

*** *p* < .001; ** *p* < .01; * *p* < .05; † *p* < .10

General Discussion

This dissertation began by examining the relationship between social prediction and activity in the theory of mind network (ToMN). Prior work in social psychology has established that social behavior can be predicted through multiple sources, such as dispositional and normative information (Cialdini et al., 1990; Gilbert & Malone, 1995), and in **Part 1** we demonstrated that ToMN activity is related to prediction error in both of these domains. **Part 2** demonstrated that moral claims are perceived as more subjective/social than prior work would anticipate; but furthermore, moral claims actually elicited greater activity across the ToMN compared to preferences, which were initially chosen as a benchmark of social content. **Part 3** examined the relationship between ToMN activity and metaethical variability. Prior work has established that some moral claims are perceived as more objective than others (Beebe, 2014; Goodwin & Darley, 2012; Wright et al., 2013), and we observed that subjective moral claims elicited greater ToMN activity, whereas objective moral claims elicited less. We argued that this relationship may be best understood in terms of prediction: Moral claims may represent a special case of social prediction, in that they are generalized across people and are socially relevant in a way that preferences (typically) are not.

Item Analysis as a Middle Path

It is worth briefly commenting on the analytic approach taken in **Part 1** and **Part 3**. Social behavior is complicated, and it can be difficult to rule out every conceivable confound between experimental conditions. This is especially true of social neuroscience research. Subtle differences between conditions (and the stimuli that comprise them), can confound simple BOLD subtraction analyses, and complex social stimuli multiply these subtleties dramatically. At the other end of the spectrum, computational modeling can allow researchers to characterize neural activity at an extremely fine-grained level; however, applying these techniques to social behavior typically requires abstracting the behavior from its real world analogue (e.g. as in economic games). Item analyses can offer a middle path. By generating a complex set of stimuli (various claims and scenarios in the present work), researchers can allow for emergent, naturalistic variability in their stimuli, while at the same time testing multiple fine-grained hypotheses, directed at the underlying processes. Perhaps even more importantly, these stimuli sets can be preserved as a resource, and passed to other researchers. There is a growing consensus that neuroscientific research will require us to move beyond single study designs and toward aggregation (Woo et al., 2017); item analysis could provide a simple way to foster cumulative research in social neuroscience.

Morality as a Scaffold for Social Prediction

Predictive coding approaches argue that the brain is fundamentally organized around predicting incoming sensory information (Barrett, 2017; Clark, 2013; Friston, 2010). This is a radical claim, and future work is necessary to verify it. However, if we accept the theory (tentatively, and for the sake of argument), then it is worth exploring its implications for moral psychology. In particular, it is worth exploring its interaction with our interpretation of **Part 3**, where we suggested that decreased ToMN activity for objective moral claims may stem from these claims being predictable, i.e. objective moral claims elicit less social prediction error.

Research in moral psychology is largely dominated by theoretical debates about “which” actions are moralized. For instance, moral foundations theorists (Graham et al., 2011; Iyer et al., 2012) argue that morality is comprised of several foundations (harm, fairness, purity, loyalty, and authority), whereas harm theorists argue that all of these foundations are reducible to harm and the relationship between victim and perpetrator (Gray et al., 2012; Schein & Gray, 2015). These theories are about the content of morality; about what does, or does not, fall within the semantic purview of the moral domain. By extension, these debates are also about what motivates moral behavior: either a collection of distinct values, or a singular aversion to harm. By contrast, a predictive account of morality raises the possibility of a different motivation: that people may feel compelled to behave in accordance with moral norms (or even non-moral norms more generally) because doing so maintains structure in one's social world, simplifying the process of social prediction. I unpack this idea below.

At any given moment, I have the power to radically change my social environment. Sitting in a coffee shop, I could scream at the top of my lungs. If I did, then I would instantly draw the attention of every person in the room. Some might ask me what's wrong, some might avoid eye contact, some might yell at me, but in every case their behavior is more unpredictable than it was a moment ago, before I screamed.

A predictive coding account suggests that our brains are fundamentally designed to minimize prediction error (Friston, 2010), and our social environment (other independently acting agents) is a large potential source of prediction error. I propose that social norms may depend on a positive feedback loop of mutual social prediction error minimization. If we infer how others predict we will behave and then act in accordance with these predictions, then we can minimize their social prediction error. By minimizing their social prediction error, we reduce the potential that they will alter their behavior, which in turn makes them more predictable. Likewise, it is in their interest to do the same for us. Violating our predictions about their behavior comes at a cost, as it simultaneously makes us less predictable to them—i.e. they could lie, steal, or cheat, but doing so runs the risk of disturbing their social environment, making it less predictable. Importantly, this cost is independent of any actual punishment issued by an agent. Put another way: conformity keeps the social world stable, and every moment that I continue to conform, I am acting on my environment to keep it predictable. I also rely on everyone else to do the same.

Consider the question: "If God does not exist, then why would anyone behave morally?", i.e. if there is no clear standard for what is right and wrong, then why can't I decide for myself what I can and cannot do? The answer, on the proposed account, is: Because everyone else has predictions about how you will behave, and if you violate them, then you forfeit the structure that they provide. Also on this account, morality fits within a broader spectrum of normative expectations, rather than as a distinct domain (as in the moral/conventional distinction, Turiel, 1978; Wainryb et al., 2004). If objective moral claims are highly predictable, as suggested by our findings in **Part 3**, then objective moral claims may be interpreted as social predictions that are most critical for regulating social relationships, within one's social context.

This proposal is also consistent with prior work, which has argued that morality is fundamentally about relationship regulation (Rai & Fiske, 2011; Fiske & Rai, 2014). The present work adds a potential mechanism giving morality its normative force. Philosophers have argued that this normative force may stem from reason (Kant, 1785/2005), or emotion (Hume, 1739–1740/1969), but the present proposal suggests that we are compelled to observe moral (and normative) rules because doing so keeps our social environment predictable. In other words, moral rules may serve as a scaffold for social prediction.

Conclusion

This dissertation has presented three papers, showing that: a) ToMN activity is related to social prediction error; b) the ToMN is particularly active when people read moral claims; and c) this activity is related to metaethical status, where subjective claims elicit more ToMN activity, and objective claims elicit less. We suggested that objective moral claims may represent strong social predictions—the moral beliefs that we expect others to hold by default. Although a unified, predictive coding account is a radical and contentious claim, adopting it may shed light on why we are compelled to act in accordance with moral norms: because by conforming to the social predictions of others, I can act on my social environment to keep it predictable. Morality then, may be a shared projection that makes social coordination possible.

References

Baron-Cohen, S. (2001). Theory of mind in normal development and autism. *Prisme, 34,* 174–183.

Barrett, L. F. (2017). *How emotions are made*. Macmillan.

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.

Blakemore, S. J., Boyer, P., Pachot-Clouard, M., Meltzoff, A., Segebarth, C., & Decety, J. (2003). The detection of contingency and animacy from simple animations in the human brain. *Cerebral Cortex*, *13*, 837-844. http://dx.doi.org/10.1093/cercor/13.8.837

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*, 1015–1026. http://dx.doi.org/10.1037/0022-3514.58.6.1015

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

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

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

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

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*, 21. http://dx.doi.org/10.1037/0033-2909.117.1.21

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. (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

Gopnik, A. (2003). The theory theory as an alternative to the innateness hypothesis. *Chomsky and his Critics*, 238-254. http://dx.doi.org/10.1002/9780470690024.ch10

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

Hume, D. (1739–1740/1969). *A treatise of human nature.* London: Penguin.

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

Joiner, J., Piva, M., Turrin, C., & Chang, S. (2017). Social learning through prediction error in the brain. *npj Science of Learning, 2*, 8. http://dx.doi.org/10.1038/s41539-017-0009-2

Kant, I. (1785/2005). *Groundwork for the metaphysics of morals.* Toronto: Broadview Press.

Kircher, T., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J., van Os, J., & Krach, S. (2009). Online mentalising investigated with functional MRI. *Neuroscience Letters*, *454*, 176-181. http://dx.doi.org/10.1016/j.neulet.2009.03.026

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron, 79*, 836–848. http://dx.doi.org/10.1016/j.neuron.2013.08.020

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*, *110*, 5648-5653. http://dx.doi.org/10.1073/pnas.1207992110

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

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255-258. http://dx.doi.org/10.1126/science.1107621

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*, 515-526. http://dx.doi.org/10.1017/S0140525X00076512

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

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*, 79–87. http://dx.doi.org/10.1038/4580

Saxe, R. (2009). The happiness of the fish: Evidence for a common theory of one's own and others' actions. In J. A. Suhr, K. D. Markman, & W. M. P. Klein (eds.) *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

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences, 19,* 65–72. http://dx.doi.org/10.1016/j.tics.2014.11.007

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

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

Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*. The MIT Press.

Westfall, J., Nichols, T. E., & Yarkoni, T. (2016). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Wellcome Open Research*, 1. http://dx.doi.org/10.12688/wellcomeopenres.10298.2

Woo, C. W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: brain models in translational neuroimaging. *Nature neuroscience*, *20*, 365-377. http://dx.doi.org/10.1038/nn.4478

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

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., 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., & 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. (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