Neuropsychologia 143 (2020) 107475

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: http://www.elsevier.com/locate/neuropsychologia

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

Jordan Theriault [a,*], Adam Waytz [b], Larisa Heiphetz [c], Liane Young [d]

[a] Northeastern University, Department of Psychology, Boston, MA, 02115, USA
[b] Northwestern University, Kellogg School of Management, Evanston, IL, 60208, USA
[c] Columbia University, Department of Psychology, New York, NY, 10027, USA
[d] Boston College, Department of Psychology, Chestnut Hill, MA, 02467, USA

ARTICLE INFO

Keywords:
Theory of mind
Predictive coding
Mixed effects
Item analysis
Morality
Metaethics

ABSTRACT

The theory of mind network (ToMN) is a set of brain regions activated by a variety of social tasks. Recent work has proposed that these associations with ToMN activity may relate to a common underlying computation: processing prediction error in social contexts. The present work presents evidence consistent with this hypothesis, using a fine-grained item analysis to examine the relationship between ToMN activity and variance in stimulus features. We used an existing dataset (consisting of statements about morals, facts, and preferences) to examine the variability in ToMN activity elicited by moral statements, using metaethical judgments (i.e. judgments of how objective/subjective morals are) as a proxy for their predictability/support by social consensus. Study 1 validated expected patterns of behavioral judgments in our stimuli set, and Study 2 associated by-stimulus estimates of metaethical judgment with ToMN activity, showing that ToMN activity was negatively associated with objective morals and positively associated with subjective morals. Whole brain analyses indicated that these associations were strongest in bilateral temporoparietal junction (TPJ). We also observed additional by-stimulus associations with ToMN, including positive associations with the presence of a person (across morals, facts, and preferences), a negative association with agreement (among morals only), and a positive association with mental state inference (in preferences only, across 3 independent measures and behavioral samples). We discuss these findings in the context of recent predictive processing models, and highlight how predictive models may facilitate new perspectives on both metaethics and the nature of distinctions between social domains (e.g. morals vs. preferences).

1. Introduction

Theory of mind refers to the ability to represent internal mental states (Premack and Woodruff, 1978). The Theory of Mind network (ToMN) is a set of brain regions that are active during mental inference and social cognition, with core regions in this network—medial prefrontal cortex (MFPC), precuneus (PC), and bilateral temporoparietal junction (RTPJ/LTPJ)—showing activation during a variety of social tasks (for review, see Amodio and Frith, 2006; Schurz et al., 2014, 2017; Van Overwalle, 2009). These tasks include reading comics and stories about beliefs (Ciaramidaro et al., 2007; Dodell-Feder et al., 2011; Fletcher, 1995; Gallagher et al., 2000; Saxe and Kanwisher, 2003; Saxe and Powell, 2006; Young et al., 2010a, 2010b), reading stories about moral violations (Young et al., 2010a, 2010b; Young et al., 2007; Young and Saxe, 2009), watching social animations (Blakemore, 2003; Gobbini et al., 2007), taking others' perspectives (Ruby and Decety, 2003; Vogeley et al., 2001), making strategic decisions in economic games (Kircher et al., 2009), inferring personal traits (Harris et al., 2005; Ma et al., 2012a, 2012b), forming impressions (Baron et al., 2011; Bhanji and Beer, 2013; Cloutier et al., 2011; Ma et al., 2012a, 2012b; Mende-Siedlecki et al., 2013; Mende-Siedlecki and Todorov, 2016; Mitchell et al., 2005; Park and Young, 2020; Schiller et al., 2009), and even listening to narratives, in which case the similarity of representations in the ToMN reflects shared experiences across listeners (Yeshurun et al., 2017). The ToMN is also differentially activated by domains of social information—e.g. in bilateral TPJ, moral statements elicit more activity than statements about facts and preferences (Theriault et al., 2017; also, see Jenkins and Mitchell, 2010). However, knowing that social tasks and

* Corresponding author.
E-mail address: jordan_theriault@northeastern.edu (J. Theriault).

https://doi.org/10.1016/j.neuropsychologia.2020.107475
Received 13 January 2020; Received in revised form 2 April 2020; Accepted 20 April 2020
Available online 29 April 2020

stimuli activate the ToMN can only advance a scientific understanding of the ToMN so far: it tells us that particular tasks and stimuli with particular features activate the network, but it leaves us to triangulate between this noisy pattern of associations to understand why it was produced. Ideally, psychologists and neuroscientists would like to understand a more fundamental problem: what underlying computational process is responsible for the activity observed in these socially-sensitive regions?

Answering this question completely is beyond the scope of any one paper, but our aim here is to provide empirical support for one recent and promising hypothesis, which has proposed that ToMN activity (and in particular, activity in TPJ) reflects the updating of predictions in social contexts (Koster-Hale and Saxe, 2013; also, see Kim et al., 2020). That is, assuming that a brain uses prior experience to form and issue predictions about incoming sensory signals (e.g. sights, sounds; Rao and Ballard, 1999), it is hypothesized that the brain performs this same fundamental process in social settings, only at a higher level of abstraction. That is, the brain has been hypothesized to issue predictions about both the observable behaviors enacted by others (as, to an observer, these behaviors are combinations of more basic sensory information) and the latent mental states that motivate behavior (which are not observable, but could be inferred from sensory signals and prior knowledge). Critically, this hypothesis implies that the brain can filter sensory signals efficiently, as predictable signals can be ignored as uninformative, and bottom-up information processing can be limited to encoding *prediction error* (i.e. the difference between the sensory signal as predicted, and the sensory signal as actually received).

This predictive hypothesis for the ToMN is derived from more general predictive coding frameworks (e.g. Barrett and Simmons, 2015; Chanes and Barrett, 2016; A. Clark, 2013, 2015; Denève and Jardri, 2016; Friston et al., 2016, 2017; Hohwy, 2013; Hutchinson and Barrett, 2019; Joiner et al., 2017; Koster-Hale and Saxe, 2013; Rao and Ballard, 1999; Shadmehr et al., 2010; Spratling, 2017; Van de Cruys et al., 2014), which suggest that this computational process—issuing predictions, encoding prediction error, and using prediction error (i.e. information) to form new predictions—is the general computation performed at all hierarchical levels of the brain. That is, each cortical area is thought to issue predictions, filter predictable signals, and pass prediction error up the cortical hierarchy. Signals, here, refer to either incoming sensory signals (e.g. light through the retina, pressure on skin) or their multimodal compressions (e.g. the unique combination of sights, sounds, smells, etc. that constitute a human social interaction). These predictive coding frameworks are grounded in principles of data compression (i.e. information theory), where signals are only informative to the extent that they reduce uncertainty, allowing prediction error to serve as a common currency of encoded information (Shannon and Weaver, 1964/1964).

Returning to the context of mental inference and social cognition: if your brain's predictions about someone's words, behaviors, etc. (i.e. predictions about compressed representations of raw sensory signals) are perfectly accurate, then your brain's predictions about that person's latent mental state need not be updated. It has been proposed that encoding prediction error or updating predictions related to latent mental states may elicit BOLD activity in the ToMN (Koster-Hale and Saxe, 2013), meaning that if there is no information to encode, then ToMN activity is expected to be low, relative to cases where unpredictable mental state information must be encoded. Thus, ToMN activity is expected to be relatively lower when signals (observable behavior, words, etc.) are predictable, and ToMN activity is expected to be relatively higher when these signals are unpredictable (Koster-Hale and Saxe, 2013).

Preliminary work has been consistent with this hypothesis (e.g. Schuwerk et al., 2017; for review, see Koster-Hale and Saxe, 2013); however, it has also been constrained by methods that prevent a simultaneous examination of pre-existing social predictions (i.e. predictions formed outside of the experimental setting), and fine-grained within-subject statistical analyses. Both factors are important if our results are to generalize beyond our experimental context and into more naturalistic settings. If social predictions are not pre-existing (i.e. they are formed in the context of the experiment), then the effect of violating these predictions may be artificially emphasized. Further, by analyzing within-subject variability, we can base inferences on naturally varying features of our stimuli-set (e.g. Westfall et al., 2017; see Section 4.3 for further discussion) as opposed to grounding inference in *a priori* categories, which may homogenize real variance among the examples they contain (e.g. S. A. Gelman and Rhodes, 2012).

These factors introduce some ambiguity into interpretations of prior work. For example, several studies have introduced characters to participants in short vignettes, leading them to form an initial impression and set of expectations. Following these introductions, ToMN activity increased when characters were described as holding contradictory beliefs (Saxe and Wexler, 2005), as engaging in contradictory behaviors (Dungan et al., 2016), or as possessing contradictory traits (Ma et al., 2012a, 2012b; Mende-Siedlecki et al., 2013), relative to initial expectations. However, later work raised the possibility that ToMN activity in these designs may be contingent on the in-lab formation of initial expectations: when characters were not initially familiarized to participants, unpredictable beliefs (e.g. someone believing plants will burst into flame if watered) did not elicit increased ToMN activity (relative to predictable beliefs; Young et al., 2010a, 2010b). Likewise, in studies using videos of behavior, ToMN activity is contingent on the motivation to engage in mental inference: when actors in a video perform unconventional actions (e.g. switching a light on with their knee when their hands are free, as opposed to switching it on while carrying a heavy load), ToMN activity increases (Brass et al., 2007; de Lange et al., 2008), but this effect fails to replicate when participants are not explicitly instructed to attend to the actor's intentions (Ampe et al., 2014). Other studies have leveraged pre-existing social predictions by using characters that were already familiar to participants, either as actual friends (Park and Young, 2020) or as known political figures (Cloutier et al., 2011), but within these studies fine-grained analyses of within-subject variability could not be performed (e.g. Park and Young (2020) examined associations with by-subject averages of RTPJ activity, but their design precluded trial-level analyses).

The present work addresses both of these issues. We leverage pre-existing mental inferences, by using moral statements that were pre-tested to confirm or contradict (to varying degrees) people's preexisting moral beliefs. Further, to leverage analyses of within-subject variability we used item analyses (Bedny et al., 2007; Dodell-Feder et al., 2011; Donnet et al., 2006; Westfall et al., 2017), a powerful (but often overlooked) method for examining by-stimulus variability and its relationship to stimuli features. In particular, unlike by-subject analyses, item analyses allow covariates of interest to be collected in independent samples of participants, allowing us to use larger samples to estimate by-stimulus averages of features of interest. Item analyses can also make use of modern multilevel modeling statistical methods, using information about the structure of the dataset to improve estimates (Baayen et al., 2008; Barr et al., 2013; Judd et al., 2012; Westfall et al., 2014, 2017). To address both issues simultaneously, would require stimuli that leverage pre-existing predictions about beliefs and, at the same time, can be characterized by stimuli features that approximate each stimuli's average predictability across participants. Fortunately, moral statements satisfied both of these constraints, with metaethical judgments serving as a proxy for moral predictability.

1.1. Metaethical variability within the moral domain

In the present work, we performed a secondary analysis of an existing dataset (Theriault et al., 2017), to address the social prediction hypothesis (Koster-Hale and Saxe, 2013). The existing dataset consisted of fMRI data that was previously collected and analyzed to test differences in ToMN activity elicited by factual, preferential, and moral

statements (Theriault et al., 2017). Moral beliefs are methodologically suited to our aims,[1] as (1) they are important to impression formation (i.e. individuals are motivated to predict what moral beliefs others hold), meaning *predictions are formed outside the lab*, and (2) there is some consensus across individuals (within a culture) that some moral beliefs should be endorsed and others opposed, meaning that there is *structured and generalizable (across people) variability in predictions about moral beliefs*. Regarding point (1): as moral beliefs are formed outside the lab, held strongly, and occupy a central position in individuals' identities (Heiphetz et al., 2018; Strohminger and Nichols, 2014, 2015), we expected that participants would be motivated to predict what moral beliefs generic people (i.e. speakers for whom participants have no prior knowledge; as opposed to known friends or politicians, as in prior work; Cloutier et al., 2011; Park and Young, 2020) were likely to hold. By contrast, compared to morals, factual statements contained less information about others' beliefs/mental states, and so were not expected to elicit high ToMN activity. Similarly, preferences are by definition more person-specific (Heiphetz et al., 2014), meaning that high-precision predictions about preferences cannot be made the context of a generic person.[2] We return to the topic of predictive precision in the general discussion (section 4.2), but for present purposes it is sufficient to say that high-precision predictions carry more potential to generate prediction error, and low-precision predictions carry less, as a precise prediction has a tightened probabilistic distribution and makes deviations from the mean more informative (Feldman and Friston, 2010; Kim et al., 2020; Van de Cruys et al., 2014). Initial analyses were consistent with this hypothesis at the categorical level, with moral statements eliciting greater ToMN activity than either facts or preferences (particularly within PC and bilateral TPJ; Theriault et al., 2017). Preferences also elicited greater ToMN activity than facts in this prior work.

Point (2), that some moral consensus exists (within a culture) and varies across moral beliefs, allowed us to examine variability *within* the moral domain. Among moral statements, some are more predictably endorsed than others, and in a moral context, metaethical judgments act as a proxy for this predictability. Metaethical judgments are judgments about what kind of information a moral statement conveys—e.g. whether it is objective (more fact-like, less preference-like), or subjective (more preference-like, less fact-like). Metaethical judgments vary across stimuli (i.e. some views are considered more or less objective/subjective than others; Beebe, 2014; Goodwin and Darley, 2008, 2012; Heiphetz and Young, 2017; Sarkissian et al., 2011; Theriault et al., 2017; Wright et al., 2013) and are highly associated with social consensus (Ayars and Nichols, in press; Beebe, 2014; Goodwin and Darley, 2012; Heiphetz and Young, 2017). That is, moral statements that elicit widespread agreement (e.g. slavery is wrong) are considered to be more objective (and less subjective) than others (e.g. eating meat is wrong). Study 1 verifies that this metaethical variability exists among moral stimuli in our dataset.

Putting this into the context of prediction: when someone makes a moral assertion that is judged by others (on average) as objective, the assertion should be predictable on the basis of social consensus (i.e. most people would endorse it). Further, in the context of information theory (Shannon and Weaver, 1964/1964, discussed above)—where signals are only informative to the extent that they reduce uncertainty—objective moral statements may actually communicate *less information*, as they are statements that most people would already be predicted to believe. By contrast, when someone makes a moral assertion that is judged as subjective, the assertion should be less predictable on the basis of social consensus (i.e. most people would not endorse it). Again, in the context of information theory, subjective moral statements communicate *more information*, as they are statements that most people would not be predicted to believe. In the present work, we exploited this theoretically-grounded equivalence between objectivity/subjectivity and information in our existing dataset, examining the relationship between ToMN activity and judgments of moral objectivity in a diverse sample of stimuli (Theriault et al., 2017).

1.2. Present work

The present work used an existing dataset (Theriault et al., 2017) to examine the relationship between by-stimulus ToMN activity and metaethical judgments (e.g. how objective, or 'fact-like', moral claims are; or alternatively, how subjective, or 'preference-like', moral claims are). As multiple behavioral measures could not be collected in the scanner (agreement was collected on a 4-point scale in the scanner, but not used in any analysis), we examined the relationship between ToMN activity and by-stimulus ratings collected in independent online samples. Study 1 consisted of an online sample (N = 49) where participants read 72 statements (each designed to be read as a statement about facts, morals, or preferences), and rated each on the extent that they agreed/disagreed with it, and the extent that it was "about facts", "about morality", and "about preferences". This online sample accomplished three goals. First, it validated our stimuli design (confirming that our facts were more fact-like than they were moral-like or preference-like). Second, consistent with prior work (Ayars and Nichols, in press; Beebe, 2014; Goodwin and Darley, 2012; Heiphetz and Young, 2017), comparisons between subgroups confirmed that moral statements supported by a social consensus (positive-consensus moral statements) were perceived as more objective than moral statements that were opposed to, or ambiguous with respect to, a social consensus (negative-consensus/no-consensus). Finally, Study 1 provided estimates of by-stimulus metaethical judgments to be used in Study 2: a maximal mixed effects models was fitted to the subjects and stimuli of the online sample (Barr et al., 2013), and by-stimulus estimates of metaethical judgments were extracted (best linear unbiased predictors, i.e. BLUPs; Baayen et al., 2008).

Study 2 examined the association between by-stimulus metaethical judgments and ToMN activity in DMPFC, VMPFC, PC, and bilateral TPJ regions of interest (ROIs). We used maximal mixed effects models to estimate by-stimulus ToMN activity from a sample of 25 participants and 72 stimuli (Baayen et al., 2008; Barr et al., 2013; Westfall et al., 2017), then compared these estimates with by-stimulus estimates of item features, extracted from Study 1. We also explored additional by-stimulus relationships with item feature estimates extracted from seven other online samples, measuring judgments of valence and arousal (N = 42), mental imagery (N = 46), whether a person was present in the scenario (N = 48), and whether a scenario evoked mental state information generally (N = 48), about others' mental states (N = 44), or about one's own mental states (N = 46). For our core analysis, comparing ToMN activity and by-stimulus estimates of metaethical

[1] Moral beliefs are also somewhat unique in the empirical literature examining social predictions, as prior work has generally focused on predictions in the context of impression formation (e.g. Cloutier et al., 2011; Mende-Siedlecki et al., 2013; Park and Young, 2020).

[2] Of course, the boundaries between morals, facts, and preferences are porous and, despite our experimental operationalization of these categories, they should not be taken to represent natural kinds. For example, people may *prefer* pleasure to pain, but if participants are asked whether "pleasure is better than pain" is a moral belief, a fact, or a preference, they may report a preference for pleasure over pain as more fact-like. For this reason, we verified (in Study 1) how our stimuli were interpreted (as fact-like, moral-like, and preference-like). In the same vein, readers may be able to imagine a preference that should be shared by many or most people—however, that such counterexamples can be generated only underscores that morals, facts, and preferences are socially constructed categories. As the present study is concerned with variability among moral claims, it is enough for our purposes to claim that people are generally more motivated and able to deploy predictions about a generic other's moral beliefs compared to their preferences. We are *not* claiming that people are completely unmotivated to predict (or are incapable of predicting) the preferences of generic others.

judgment, we controlled for a variety of syntactic and semantic features that may act as confounds. Overall, our findings demonstrate that the ToMN is responsive to variance among moral statements: more objective moral claims elicit less ToMN activity, and more subjective moral claims elicit greater ToMN activity.

As the present work was a secondary analysis of an existing dataset, we have tried to remain appropriately conservative in our statistical analysis, testing specific relationships (corrected for multiple comparisons) only when justified by omnibus testing (see Tables S7, S9, & S10, in the online supplemental materials). These steps occasionally lead to non-traditional forms of reporting analyses (e.g. grouping ToMN ROIs and analyzing VMPFC separately in Study 2), but these steps were taken to ensure that generalizations were conservative and justified.

2. Study 1

Study 1 validated our stimulus set in an online sample, confirming that moral statements supported by a social consensus were perceived as more objective (Ayars and Nichols, in press; Beebe, 2014; Goodwin and Darley, 2012; Heiphetz and Young, 2017). The stimulus set consisted of 72 statements about facts, morals, and preferences, and were designed to fit within consensus subcategories, 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; Fig. 1; also see 2.1.2). Using an independent online sample allowed us to extract by-stimulus estimates of metaethical judgments in a sample approximately twice the size of the fMRI sample. To measure metaethical judgment, we asked participants to rate each statement on the extent that it was about facts, about morals, and/or about preferences (Theriault et al., 2017). This method has several advantages: (a) it validates stimulus conditions, as facts should be rated as most fact-like, morals as most moral-like, and preferences as most preference-like; and b) it avoids artificially imposing relationships among ratings (unlike bipolar methods, where the design requires that for a statement to be more fact-like, it must necessarily be less preference-like). Our method allows any correlations among ratings to emerge independently, without imposing them by design.

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, 23 male, 1 unspecified; M_{Age} = 33.5 years, SD_{Age} = 10.7 years, $Range_{Age}$ = 19–59 years), after excluding two participants for failing an attention check that asked them to describe any statement they had read. The majority of our sample had either entered or completed college/university (maximum educational attainment of high school = 16.3%; some college/university = 36.7%; completed college/university = 38.8%; completed graduate degree = 8.2%). The majority of the sample was Caucasian (White/Caucasian = 83.7%; Black/African American = 6.1%; Asian = 10.2%; Pacific Islander = 2.0%; Other = 2.0%), and Non-Hispanic (Hispanic = 8.2%; Non-Hispanic = 91.8%). 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 statements (e.g. "It is irresponsible for airlines to risk the safety of their passengers"; see Appendix A for all statements), 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 statement 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. Participants were instructed that they would complete a "statements task", where they would "read short statements and decide whether you agree or disagree with them". Statements were designed to be interpreted as either facts, morals, or preferences, and were evenly divided between categories (n_{Fact} = 24, n_{Moral} = 24, $n_{Preference}$ = 24), but participants were not explicitly alerted to this element of design. Each category included three consensus subcategories: a) *positive-consensus*, where most people would agree with the statement (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 statements (as opposed to controversial statements) were used because the feature of interest was social consensus; in other words, no-consensus statements were intended to elicit a unipolar, non-skewed distribution of agreement. By contrast, controversial statements (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

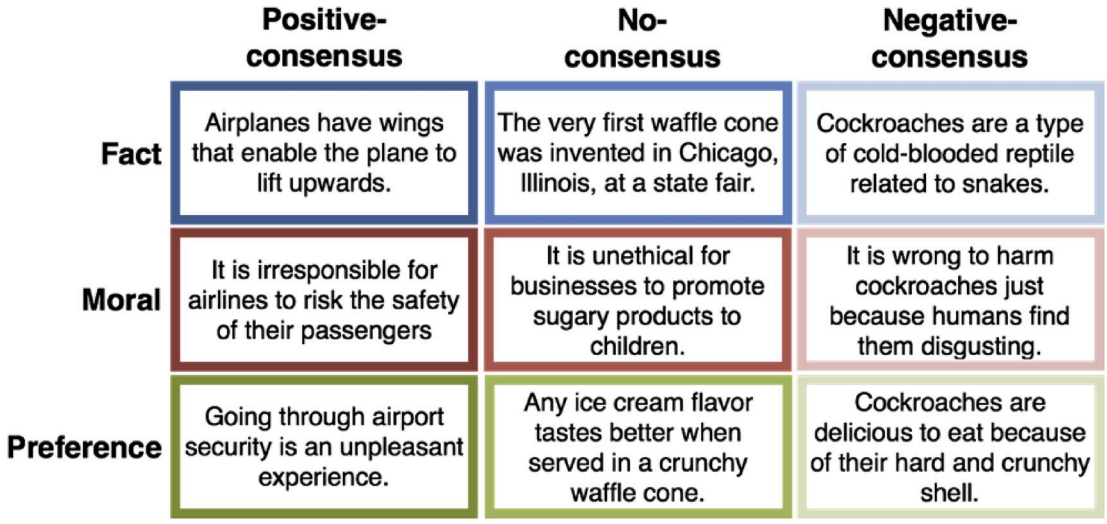

Fig. 1. Sample stimuli. Statements varied in content (fact/moral/preference) and agreement (positive-consensus/no-consensus/negative-consensus). See Appendix A for the full text of all stimuli.

six were false. Statements did not contain any mental state markers (e.g. "She thinks," "He believes").

2.1.3. Statistical methods

Studies 1 and 2 used mixed effects analyses to model crossed by-subject and by-stimulus random effects (Baayen et al., 2008; Judd et al., 2012; Westfall et al., 2014). 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. H. Clark, 1973; Judd et al., 2012; Westfall et al., 2017). Mixed effects analysis also allow for the estimation of BLUPs (best linear unbiased predictors; Baayen, 2008; Baayen et al., 2008). BLUPs are by-stimulus estimates of metaethical judgments for each stimuli, and are preferable to simple by-stimulus averages for two reasons: a) by-stimulus BLUPs are independent from by-subject variance, meaning that estimates were specific to the scenarios (and could be compared with by-stimulus estimates of ToMN activity in Study 2); and b) by-stimulus BLUPs incorporate the sample distribution into the estimate (i.e. they are semi-pooled estimates; A. Gelman et al., 2012), meaning that they anticipate regression to the mean and mitigate against outliers. Analysis was conducted in R (R Core Team, 2016), 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., 2017).

2.1.4. Data and software sharing

De-identified raw behavioral data and code to reproduce all analyses and figures are available at https://osf.io/cx4dp/.

2.2. Results

2.2.1. Validating agreement sub-categories

Agreement ratings were fit with a maximal mixed effects model. As fixed effects, the model included main effects of category (fact/moral/preference), consensus (positive-/no-/negative-consensus), and their interaction. As random effects across subjects, the model included random intercepts, and random slopes for category, consensus, and their interaction. As random effects across stimuli, random intercepts were included. Condition means are presented in Table S1 of the online supplemental materials.

lmer(agreement ~ 1 + category*consensus

+ (1 + category*consensus | subject)

+ (1 | stimuli))

We observed some differences in agreement across categories, but critically, differences among consensus sub-categories were consistent with our design. Main effects were significant for both category, $F(2, 72.15) = 7.35$, $p = .001$, and consensus, $F(2, 55.11) = 55.1$, $p < .001$, but not their interaction, $F(4, 66.5) = 0.35$, $p = .843$. In follow-up contrasts, agreement was greater for facts, $z = 3.27$, $p = .003$, and morals, $z = 3.33$, $p = .003$, relative to preferences, whereas agreement did not differ between facts and morals, $z = 0.01$, $p = .999$ (p values were corrected for 3 comparisons; αfamilywise = 0.05; single-step method; multcomp package Hothorn et al., 2008). Differences among consensus sub-categories were consistent with our design: agreement was greater for positive-consensus statements, relative to no-consensus statements, $z = 6.48$, $p < .001$, and negative-consensus statements, $z = 10.26$, $p < .001$, and agreement was greater for no-consensus statements relative to negative-consensus statements, $z = 6.80$, $p < .001$. Although preferences elicited less agreement in general, there was no significant interaction between category and consensus, meaning that differences between consensus sub-categories were comparable across facts, morals, and preferences.

2.2.2. Examining fact-/moral-/preference-like ratings across consensus sub-categories

An initial, maximal mixed effects model failed to converge using restricted maximum likelihood (REML) estimation (after 10,000 iterations):

lmer(rating ~ 1 + rating_type*category*consensus

+ (1 + rating_type*category*consensus | subject)

+ (1 + rating_type | stimuli))

Given this, we simplified the model, removing consensus sub-categories from by-subject random effects. As fixed effects, the model included main effects of rating type (fact-/moral-/preference-like), category (fact/moral/preference), consensus (positive-/no-/negative-consensus), and all interactions. As random effects across subjects, the model included random intercepts, and random slopes for rating type, category, and their interaction. As random effects across stimuli, the model included random intercepts and random slopes for rating type.

lmer(rating ~ 1 + rating_type*category*consensus

+ (1 + rating_type*category | subject)

+ (1 + rating_type | stimuli))

Consistent with prior work (Ayars and Nichols, in press; Beebe, 2014; Goodwin and Darley, 2012; Heiphetz and Young, 2017), positive-consensus moral statements were perceived as more objective (fact-like) than other moral statements not supported by a social consensus. The model produced a significant 3-way interaction, $F(8, 63.0) = 4.06$, $p < .001$ (for condition means see Table S1 of the online supplemental materials). Contrasts compared consensus sub-categories within each category x rating-type grouping (p values corrected for 27 comparisons; $\alpha_{\text{familywise}} = 0.05$; single-step method). Among morals, positive-consensus statements were perceived as more fact-like than no-consensus, $z = 5.70$, $p < .001$, and negative-consensus statements, $z = 6.26$, $p < .001$. Also, among facts, positive-consensus and no-consensus statements were perceived as more fact-like than negative-consensus statements: $z = 4.72$, $p < .001$, and $z = 5.43$, $p < .001$, respectively (although these differences within facts did not replicate in Study 2; see section 3.2.1). Among preferences, there were no significant differences between consensus categories (Fig. 2b).

2.3. Discussion

Consistent with prior work (Ayars and Nichols, in press; Beebe, 2014; Goodwin and Darley, 2012; Heiphetz and Young, 2017), participants rated positive-consensus moral statements (i.e. moral statements supported by a social consensus) as more fact-like (i.e. more objective) than no-consensus and negative-consensus moral statements. Study 2 examined whether variability in moral objectivity was related to activity in the ToMN, given the hypothesis that regions in this network are associated with processing prediction error in socially relevant contexts (Koster-Hale and Saxe, 2013). According to this hypothesis, moral statements that are predictable on the basis of social consensus (i.e. statements rated as objective) should elicit less ToMN activity, and moral statements that run counter to the social consensus (i.e. statements rated as less objective) should elicit greater ToMN activity.

3. Study 2

Study 2 examined the relationship between the perceived objectivity of moral statements, and activity evoked in ToMN ROI, including right/left temporoparietal junction (R/LTPJ), dorsal-/ventro-medial prefrontal cortex (DMPFC/VMPFC), and precuneus (PC). We identified ToMN

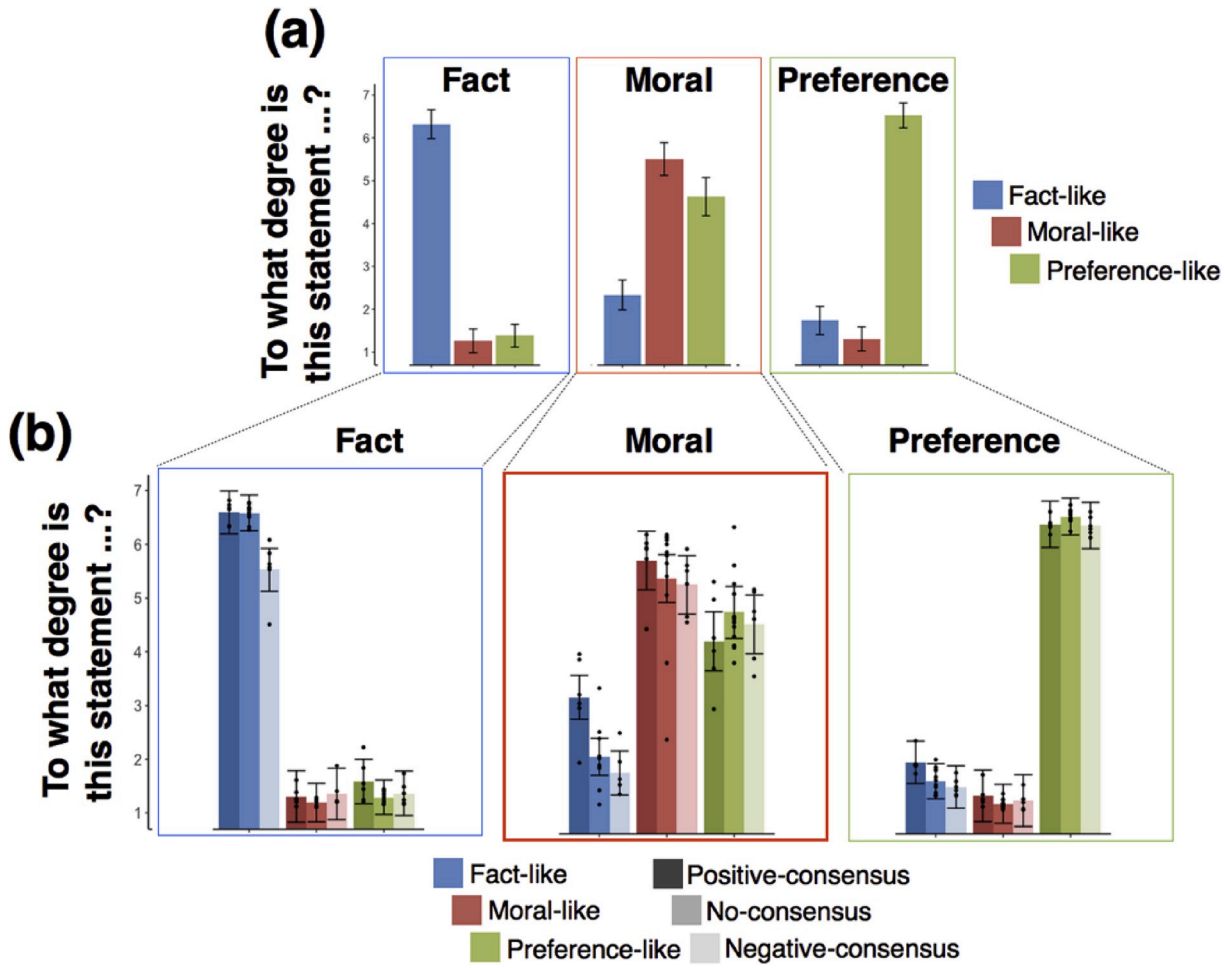

Fig. 2. Study 1 metaethical judgments. Participants rated each scenario on the extent that it was fact-like, moral-like, and preference-like (1–7; "not at all" – "completely"). (a) Collapsing across consensus subcategories, ratings were consistent with our *a priori* categories: facts were largely fact-like (left), preferences were largely preference-like (right), and morals were largely moral-like (center). Moral statements were also perceived as largely preference-like (a pattern explored further in Theriault et al., 2017). (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 statements 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

ROIs using a functional localizer that has been frequently used in prior work (contrasting false beliefs with false photographs), allowing us to compare our present findings with prior work suggesting these same regions encode mental state information (Dodell-Feder et al., 2011). By-stimulus independent ratings of moral objectivity were collected and modeled in Study 1, and compared to by-stimulus estimates of ToMN activity that were collected and modeled in Study 2. In addition, other independent ratings of item features (e.g. the extent stimuli evoked mental state inferences) were collected, modeled, and explored as well.

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; $Range_{Age}$ = 18–35 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 useable. We were unable to collect post-scan ratings for one of the 25 participants. The majority of our sample had either entered or completed college/university (maximum educational attainment of high school = 4%; some college/university = 28%; completed college/university = 52%; completed graduate degree = 12%; no response = 4%). The majority of the sample was Caucasian (White/Caucasian = 56%; Black/African American = 20%; Asian = 20%; no response = 4%), and Non-Hispanic (Hispanic = 4%; Non-Hispanic = 92%; no response = 4%). 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 statements and rated their agreement with each (ratings were consistent with consensus subcategories; see Table S2). Statements were shown across six runs (12 per run; items were randomized; conditions were counterbalanced to appear equally in each run). Participants read each statement (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, 68.5% of "don't know" responses were for no-consensus facts, followed by 7.7% for no-consensus preferences). Stimuli were presented in white text on a black background using a projector, viewable through a mirror mounted on the head coil. 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 rated all statements (i.e. on the extent to which each was fact-/moral-/preference-like) 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 of additional item features were collected in separate online samples. These included questions used in a prior item analysis of the ToMN (Dodell-Feder et al., 2011) as well as measures of arousal and valence (Kron et al., 2013; see Appendix B). In these separate online samples, participants were asked one of the following questions: *Mental Imagery* (n = 46; "To what extent did you picture or imagine what the statement described as you read?"; 1–7–; "Very Little"–"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); and *Mental States* (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"). Given that prior work has established the sensitivity of the ToMN to mental state information (Saxe and Kanwisher, 2003), and given that this *Mental States* measure was ambiguous with respect to whose mental states should be considered (either your own, or the mental states of others), we asked an additional sample of participants two more specific questions: *Mental States (of Others)*, (n = 44; "To what extent did this statement make you think about the experiences, thoughts, beliefs, and/or desires OF OTHER PEOPLE?; 1–7; "Very Little"–"Very Much") and *Mental States (of Self),* (n = 46; "To what extent did this statement make you think about YOUR OWN experiences, thoughts, beliefs, and/or desires?"; 1–7; "Very Little"–"Very Much").

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 S 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 × 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 × 256 mm). Data processing and analysis were performed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) and in-house Matlab modeling scripts (https://github.com/lypsychlab/younglab_scripts). 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. ToMN localizer task

An independent functional localizer task identified ToMN ROIs (Dodell-Feder et al., 2011). The task consisted of ten stories about mental states (*false-belief*) and ten about physical representations (*false-photograph*), presented across two runs. 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 pattern of results (with respect to the significance of relationships between ToMN activity and metaethical judgments), 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). The localizer contrast was thresholded at $p < .001$ (Woo et al., 2014), and k = 10 (to remain consistent with thresholds used in prior work; Dodell-Feder et al., 2011; see Table S3 for k values, and additional clusters identified by the contrast).

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 s 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 first estimating beta maps for each item and then correlating beta maps with estimates of metaethical judgments (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 $\alpha = 0.05$, thresholding voxels at $p < .001$ (as recommended by Woo et al., 2014). Permutation tests were performed using SnPM 13 (http://warwick.ac.uk/snpm; Nichols and Holmes, 2002).

3.1.8. Statistical methods

As in Study 1, mixed effects analyses were used to model behavioral responses and PSC. Model specification and simplification is described below (3.2.2). For ToMN ROIs, BLUPs (best linear unbiased predictors) were extracted from the fitted models and compared to behavioral BLUPs extracted from the online sample in Study 1 (see Table S4 of the online supplemental materials) and additional online samples (see section 3.1.3, and Appendix A). The advantages of mixed effects models are described in detail in section 2.1.3. We began all modeling by including all random effects (Barr et al., 2013), but if this model failed to converge using REML estimation, then correlations between random effects were temporarily removed, the model was inspected, and variance components with zero unique variance were removed (Bates et al., 2015). Any

exceptions to this procedure are noted in the results section, and full models are reported in the text in formula syntax.

3.1.9. Data and software sharing

De-identified ROI timecourse data and code to reproduce all analyses and figures are available at https://osf.io/cx4dp/, as well as raw fMRI data and by-stimulus item estimates in standardized BIDS format (Gorgolewski et al., 2016). Whole-brain unthresholded T maps, corresponding to analyses in Section 3.2.5 and Fig. 4, are available at https://neurovault.org/collections/ZZMRPKDV/.

3.2. Results

3.2.1. Replicating study 1 behavioral results

First, we modeled metaethical judgments in our Study 2 sample to ensure that the patterns were consistent with those observed in Study 1. The maximal mixed effects model used to model behavioral data in Study 1 (2.2.2) did not provide a convergent solution using REML estimation in the Study 2 behavioral data, most likely because of Study 2's smaller sample of participants. There were no parameters which could clearly be dropped (i.e. parameters with zero unique variance), so by-stimulus random effects were dropped instead, as we simply wanted to confirm that ratings were similar to those observed in Study 1. The resulting model included, as fixed effects, main effects of rating type (fact-/moral-/preference-like), category (fact/moral/preference), consensus (positive-/no-/negative-consensus), and all interactions. As random effects across subjects, the model included random intercepts, and random slopes for rating type, category, and their interaction.

lmer(rating ~ 1 + rating_type*category*consensus

+ (1 + rating_type*category | subject)

We performed contrasts for all comparisons of interest (p values corrected for 27 comparisons; $\alpha_{familywise} = 0.05$; single-step method). As in Study 1, participants rated positive-consensus morals as more fact-like than both no-consensus morals, $z = 6.82$, $p < .001$, and negative-consensus morals, $z = 7.92$, $p < .001$, and no significant difference emerged between fact-like ratings for no-consensus and negative-consensus morals, $z = 2.33$, $p = .378$. Unlike in Study 1, there was no significant difference in fact-like ratings within facts across consensus sub-categories (for condition means see Table S5 of the online supplemental materials).

3.2.2. Model fitting and by-stimulus estimates of ToMN activity

Initially, a maximal mixed effects model, predicting PSC, was fit across all functional ROIs:

lmer(PSC ~ 1 + ROI*category

+ (1 + ROI*category | subject)

+ (1 + ROI | stimuli)

However, this maximal model failed to converge using REML and had to be simplified (see section 3.1.8). The final simplified model included, as fixed effects, main effects of ROI (DMPFC/VMPFC/PC/RTPJ/LTPJ), category (fact/moral/preference), and all interactions. As random effects across subjects, the model included random intercepts, and random slopes for ROI, category, and the interactions of VMPFC and LTPJ with the moral category. As random effects across stimuli, the model included random intercepts, and random slopes for VMPFC.

lmer(PSC ~ 1 + ROI*category

+ (1 + ROI + category + moral*VMPFC + moral*LTPJ | subject)

+ (1 + VMPFC | stimuli)

These simplifications from the maximal model are informative with respect to relations among ROIs (for model details, see Table S6 of the online supplemental materials). For by-stimulus random effects, we observed high correlations between DMPFC (i.e. intercept), PC, RTPJ, and LTPJ. These correlations preclude calculating unique variance terms for each ROI, but, at the same time, they suggest that the response of these regions cannot be distinguished within our set of stimuli. Thus, the mixed effects analysis provided a data driven rationale for a more conservative analysis, treating these regions as a network and avoiding overfitting the data to every ROI. At the same time, as by-stimulus variance for VMPFC can be estimated separately, the data licensed separate analyses for this region. As all analyses below concern the by-stimulus estimates of ToMN activity, by-stimulus estimates of DMPFC, PC, RTPJ, and LTPJ are averaged and collectively referred to as the ToMN. Given this, ROI interactions reported below examine differences between the ToMN, and VMPFC.

3.2.3. Among moral statements, ToMN activity is positively associated with preference-like ratings, and negatively associated with fact-like/moral-like ratings

Our core question was whether ToMN activity was related to by-stimulus ratings of moral objectivity. By-stimulus estimates of metaethical ratings (fact-like, moral-like, and preference-like ratings) were extracted from Study 1 (section 2.2.2), and, separately, by-stimulus estimates of ToMN activity (PSC; centered and normalized) were extracted from the Study 2 fMRI data (section 3.2.2). A linear model predicted by-stimulus metaethical judgments on the basis of ToMN activity (for model details, see Table S7 of the online supplemental materials), including main effects and interactions of PSC (estimated from the model in 3.2.2), category (fact/moral/preference), and rating type (fact-/moral-/preference-like), as well as interactions between these terms and ROI (ToMN/VMPFC).

lm(rating ~ 1 + PSC*category*rating_type +

ROI:PSC + ROI:PSC:category + ROI:PSC:rating_type + ROI:PSC:rating_type:category)

The association between metaethical judgments and ToMN activity differed between categories (facts, morals, and preferences): morals showed a distinct pattern of association. We observed a 4-way interaction,[3] $F(4, 405) = 5.73$, $p < .001$ between PSC, category, rating-type, and ROI. Follow-up ANOVAs identified significant 4-way interactions between morals and facts, $F(2, 270) = 7.42$, $p < .001$, and between morals and preferences, $F(2, 270) = 6.64$, $p = .002$, but not between preferences and facts, $F(2, 270) = 0.01$, $p = .986$.

Within morals, ToMN activity was related to metaethical judgments: increased ToMN activity was associated with increases in preference-like ratings, and decreases in both fact-like and moral-like ratings. Following up a 3-way interaction (within morals) between PSC, ROI, and rating type $F(2, 135) = 6.03$, $p = .003$, further 3-way interactions distinguishing between rating types demonstrated distinctions between preference-like and fact-like ratings, $F(1, 90) = 13.8$, $p < .001$, between preference-like and moral-like ratings, $F(1, 90) = 5.55$, $p = .022$, but not between fact-like and moral-like ratings, $F(1, 90) = 0.81$, $p = .371$. Thus, among moral statements, BOLD activity in ToMN and VMPFC shows one relationship with preference-like ratings, and another for moral-like and fact-like ratings. These relationships also differed slightly between ROIs,

[3] A sensitivity analysis (estimated by simulation using the *simr* package; Green and MacLeod, 2016) indicated that this 4-way interaction between PSC, category, rating-type, and ROI could detect a minimum effect size 35% below the effect size observed in the present work, while retaining ~80% power (Arend and Schäfer, 2019; Bloom, 1995). All fixed effects in the model were multiplied by .65, and Monte Carlo simulation was used to compare the model above to an alternative omitting the 4-way interaction, *power* = 83.40%, 95% CI = [80.95%, 85.66%], 1000 simulations.

as 2-way interactions between PSC and ROI were significant among preference-like ratings, $F(1, 45) = 4.57$, $p = .038$, and fact-/moral-like ratings, $F(1, 92) = 8.80$, $p = .004$.

Contrasts tested the strength of each relationship, among morals, between ToMN/VMPFC activity and metaethical judgment. ToMN activity was negatively related to fact-/moral-like ratings, $B = -1.01$, $\beta = -0.46$, $t(140) = 4.32$, $p < .001$, and positively related to preference-like ratings, $B = 0.94$, $\beta = 0.43$, $t(140) = 2.85$, $p = .020$. VMPFC activity was also negatively related to fact-/moral-like ratings, $B = -0.28$, $\beta = -0.13$, $t(140) = 3.18$, $p = .007$, and showed a marginal positive association with preference-like ratings, $B = 0.31$, $\beta = 0.14$, $t(140) = 2.48$, $p = .055$ (p values corrected for 4 comparisons; $\alpha_{familywise} = 0.05$; single-step method). Thus, among moral statements, ToMN activity was negatively related to fact-/moral-like ratings, and positively related to preference-like ratings, with both relationships present, but weaker in VMPFC (Fig. 3a).[4]

Among both facts and preferences, ToMN activity was not related to metaethical judgments. Within facts, averaging across ToMN and VMPFC, PSC did not interact with rating-type, $F(2, 135) = 0.23$, $p = .792$, and PSC showed no significant main effect, $F(1, 135) = 0.01$, $p = .991$. Likewise, within preferences, averaging across ToMN and VMPFC, PSC did not interact with rating type, $F(2, 135) = 0.61$, $p = .550$, and PSC showed no significant main effect, $F(1, 135) = 2.60$, $p = .109$ (Fig. 3b).

3.2.4. The ToMN–metaethical judgment association remains significant after controlling for reaction time and semantic/syntactic features of stimuli

The identified relationship between metaethical judgments and BOLD activity could be driven by variance on some dimension outside of experimental interest, e.g. reaction time, or semantic/syntactic features of the stimuli. Along with reaction time, 13 semantic/syntactic features were collected for each stimulus (e.g. reading ease, noun concreteness; Graesser et al., 2004; McNamara et al., 2014; see Appendix B), and added as fixed effects to the model of ToMN activity identified in 3.2.2 (along with their interactions with ROI and category). Beginning with a maximal model (omitting correlations among random effects), non-significant confounds were dropped from the model step-wise. The final model controlled for stimulus differences in concreteness, words before the main verb, and familiarity (but notably, not reaction time, which was not related to ToMN activity; for model details and full analysis, see Table S9 of the supplemental online materials). By-stimulus estimates of ROI activity were extracted from the final model, and the analyses in 3.2.3 were repeated. The results were consistent with the initial findings, although the interaction between ToMN and VMPFC was no longer significant. Among moral statements, positive ToMN activity was associated with increased preference-like ratings, $B = 1.02$, $\beta = 0.25$, $t(70) = 2.81$, $p = .013$, and decreased fact-like/moral-like ratings, $B = -0.97$, $\beta = -0.23$, $t(70) = 3.79$, $p < .001$ (p values corrected for 2 comparisons; $\alpha_{familywise} = 0.05$; single-step method). These relationships did not change if reaction time, and its interactions with category and ROI, were reintroduced to the behavioral model.

[4] As in footnote 3, a sensitivity analysis (Arend and Schäfer, 2019; Bloom, 1995) was conducted using *simr* (Green and MacLeod, 2016), to identify the minimum effect size we could detect at ~80% power. The negative relationship between ToMN activity and fact-/moral-like ratings could be reduced by 35% from our observed effect, while attaining 80.80% power, 95%CI = [78.22%, 83.20%], 1000 stimulations. The positive relationship between ToMN activity and preference-like ratings could be reduced by 40% from our observed effect, while attaining 82.80% power, 95%CI = [80.32%, 85.09%], 1000 stimulations. Associations with VMPFC activity were less well-powered, and should be considered with less confidence. The association between VMPFC activity and fact-/moral-like ratings could only be reduced 10% while attaining 79.50% power, 95%CI = [76.86%, 81.96%], 1000 stimulations, and the association between VMPFC activity and preference-like ratings could only be reduced 15% while attaining 81.00% power 95%CI = [78.43%, 83.39%], 1000 stimulations.

3.2.5. Whole brain analysis: bilateral TPJ activity is negatively associated with fact-like (objectivity) ratings

A whole brain random effects analysis of moral statements provided context for our analysis of ToMN ROIs, showing that by-stimulus BOLD activity was negatively related to fact-like ratings, and positively related to preference-like ratings in overlapping regions of bilateral TPJ (Fig. 4). We performed three whole brain correlation analyses, testing the relationship between average by-stimulus PSC (modeled in Study 2) and by-stimulus behavioral estimates of fact-like, moral-like, and preference-like ratings (modeled in Study 1). Preference-like ratings were positively correlated with activity in bilateral TPJ (peak MNI coordinates: right [54, −60, 34]; left [-36, −70, 48]), and fact-like ratings were negatively correlated in overlapping regions of bilateral RTPJ (peak MNI coordinates: right [44, −68, 46]; left [-44, −62, 48]). Peak activation for associations with fact-like ratings were slightly dorsal to the functionally defined ROI positions [peak $RTPJ_{localizer}$; [52, −60, 24]; peak $LTPJ_{localizer}$ [-56, −56, 28]; see Table S3 in the online supplemental materials); however, their overlap was particularly noticeable in RTPJ (Fig. 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 correlated (positively or negatively) with moral-like ratings. Thus, fact-like ratings are negatively correlated with activity in several regions, but correlations between BOLD activity and both fact-like and preference-like ratings were present in bilateral TPJ.

3.2.6. Among moral statements, ToMN activity is not associated with the extent of mental state information

Given that we identified an association, across stimuli, between ToMN activity and judgments of objectivity, we also examined whether a by-stimulus relationship exists between ToMN activity and the extent that a statement is judged to contain information about mental states. Unlike metaethical judgments—where low variability among facts and preferences precluded comparisons across facts, morals, and preferences (Fig. 2b)—comparisons across content categories were possible for ratings of mental state inference and other item features. In total, seven additional item features were analyzed in an exploratory analysis: (a) agreement (estimated using Study 1 data); (b) ratings of the presence of *mental states* (either generally, with reference to one's own mental states, or with reference to the mental states of others); (c) ratings of evoked mental imagery; (d) binary judgments of whether a person was present in the statement; (e) ratings of valence; (f) and ratings of arousal. Each rating was fit with maximal mixed effects model, and behavioral by-stimulus estimates were extracted and compared with by-stimulus estimates of ToMN activity (see Table S10 of the online supplemental materials for model details and ANOVA details; see also Fig. S2 for a correlation matrix, displaying linear relationships among all by-stimulus estimates).

Agreement was negatively related to ToMN activity among moral statements, but showed a marginally positively relationship among preferences (Fig. 5a). Correcting for three comparisons ($\alpha_{familywise} = 0.05$; single-step method): among facts, the agreement–ToMN relationship was non-significant, $B = 0.02$, $\beta = 0.01$, $t(69) = 0.04$, $p = 1.00$; among preferences, the agreement–ToMN relationship was positive and marginally significant, $B = 1.01$, $\beta = 0.53$, $t(69) = 2.30$, $p = .072$; and among morals, the agreement–ToMN relationship was negative and significant, $B = -1.46$, $\beta = -0.77$, $t(69) = 2.57$, $p = .037$. That high agreement moral statements elicited less ToMN activity is consistent with our behavioral findings, where fact-like moral statements also elicited less ToMN activity (Fig. 3), and positive-consensus (high agreement) moral statements were perceived as more fact-like (Fig. 2b).

Mental state information (either general, self-oriented, or other-oriented; see 3.1.3), was only positively associated with ToMN activity among preferences, and no significant association was observed among facts or morals (Fig. 5b–d). Among preferences, ToMN activity was

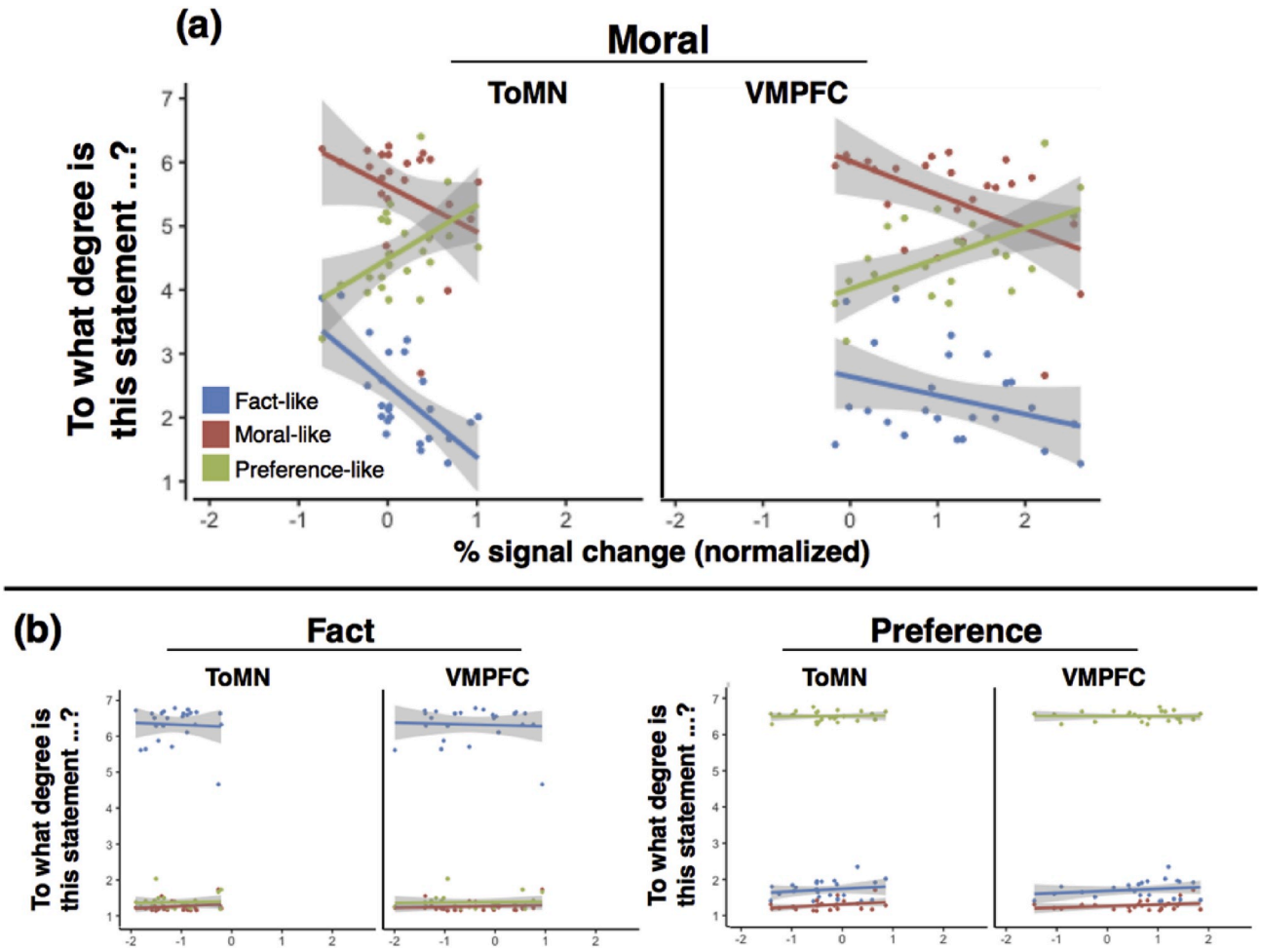

Fig. 3. Behavioral–BOLD relationships. By-stimulus estimates of metaethical judgments were extracted from Study 1 and compared with by-stimulus estimates of PSC (percent signal change), extracted from a model of all ROIs in Study 2. ToMN includes averaged estimates for DMPFC, PC, RTPJ, and LTPJ (all by-stimulus random effects were perfectly correlated). (a) Within moral statements, PSC for ToMN was positively related to preference-like ratings, and negatively related to fact-/moral-like ratings. These relationships were present, but weakened in VMPFC. (b) Within facts and preferences, there was no relationship with PSC. Shaded areas represent 95% confidence intervals. We also performed a supplemental analysis using agreement and metaethical judgments collected within participants, which showed a pattern of results across ROIs consistent with these results (Fig. S1).

positively associated with general mental state information, $B = 0.37$, $\beta = 0.31$, $t(66) = 3.39$, $p = .004$, self-oriented mental state information, $B = 0.51$, $\beta = 0.52$, $t(66) = 2.48$, $p = .047$, and other-oriented mental state information, $B = 0.47$, $\beta = 0.47$, $t(66) = 2.77$, $p = .022$. No significant association was observed among facts or preferences on any measure (Fig. 5b–d; Table S10). Thus, we find some limited support that ToMN activity is sensitive to mental state information among preferences, but we also find that the relationship does not generalize to morals.

Of the remaining item features, only the presence of a person was positively associated with ToMN activity (Fig. 5e): across all categories, if, on average, our online sample agreed that a statement contained a person, then increased ToMN activity was observed, $B = 1.40$, $\beta = 0.35$, $t(68) = 2.28$, $p = .026$. Other measures were not significantly related to ToMN activity (Fig. 5f–h; for analyses, see Table S10 of the online supplemental materials).

4. General discussion

In the present work, we examined the relationship between by-stimulus judgments of moral objectivity/subjectivity (i.e. metaethical judgments) and activity in the ToMN network. We observed that moral statements that were judged as more objective (i.e. more fact-like and less preference-like) elicited less ToMN activity, and moral statements judged as more subjective (i.e. less fact-like and more preference-like) elicited greater ToMN activity. This finding was confirmed in both ROI analyses (Fig. 3; using an established functional localizer; Dodell-Feder et al., 2011) and a whole brain analysis (Fig. 4; for unthresholded T-maps, see https://neurovault.org/collections/ZZMRPKDV/). ROI analyses were robust to controls for reaction time and semantic/syntactic features of the stimuli (section 3.2.4). The whole brain analysis implicated bilateral TPJ as locations in which the observed associations were relatively strong compared to other ToMN ROIs, which is consistent with the emphasis of prior work on TPJ as particularly critical region for processing social or high-level prediction error (e.g. Geng and Vossel, 2013; Schuwerk et al., 2017; but, for a brief discussion of MPFC in a predictive context, see Koster-Hale and Saxe, 2013). In addition, we explored relationships between ToMN activity and additional item features as rated by online samples (Fig. 5), finding that ToMN activity was negatively associated with agreement among morals, positively associated with mental state inferences among preferences, and positively associated with the presence of a person among morals, facts, and preferences. To our knowledge, ours is the first study to identify a relationship between by-stimulus variance in ToMN activity and

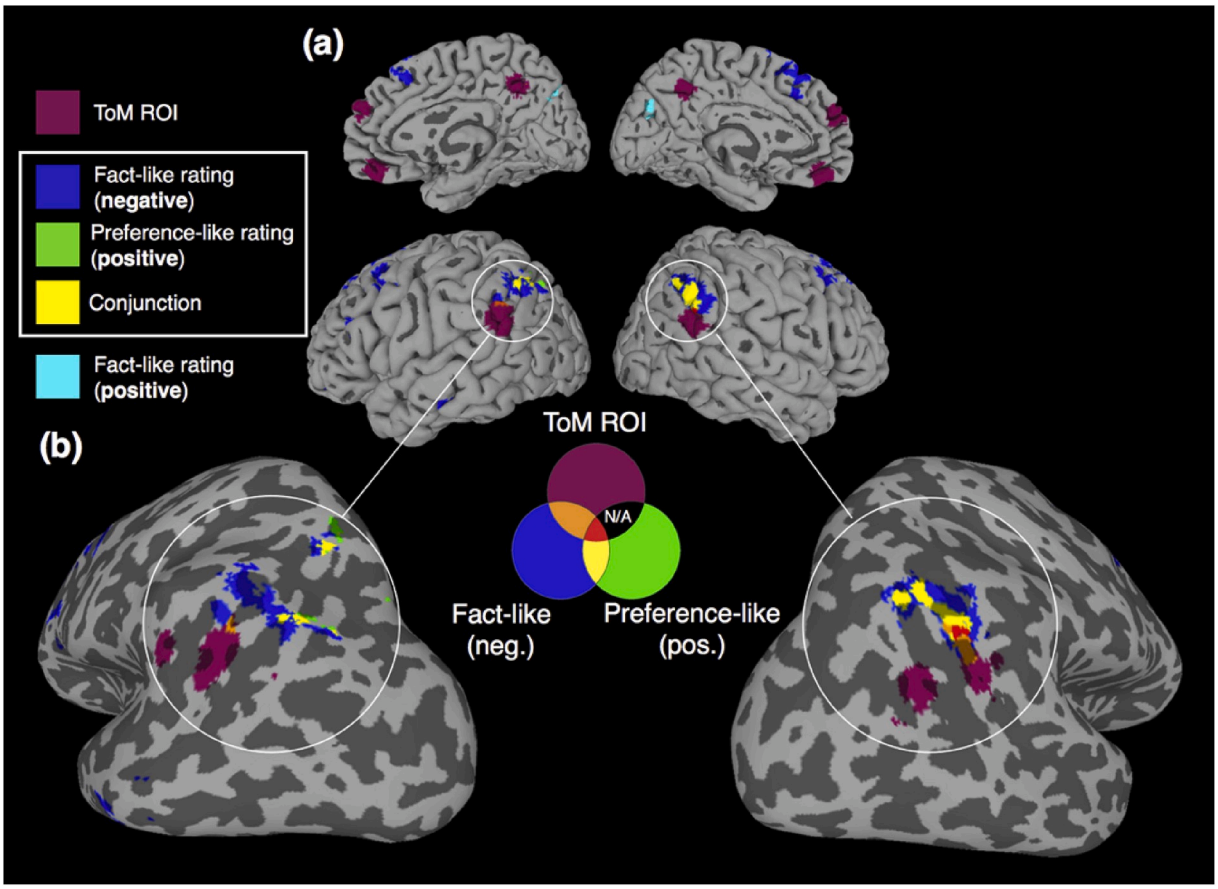

Fig. 4. Whole brain behavioral–BOLD correlations, within moral statements. In three separate models, BOLD estimates were correlated with fact-like, moral-like, and preference-like ratings, extracted from Study 1. ToMN ROIs are pictured for reference. Fact-like ratings were negatively related to activity 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. For unthresholded T maps, see https://neurovault.org/collections/ZZMRPKDV/.

features of individual stimuli (prior work has tried and failed to identify significant by-stimulus relationships within the ToMN; Dodell-Feder et al., 2011).[5] Our study is also one of only a few to leverage pre-existing social expectations in the experimental context (e.g. Cloutier et al., 2011; Park and Young, 2020), and, to our knowledge, the only one to do this while examining by-stimulus variability. Below, we discuss how these results might be integrated into a predictive framework (e.g. Koster-Hale and Saxe, 2013), how they may facilitate new approaches to understanding the social relevance of moral beliefs, and how our methodology for item analysis may be of use in future work.

4.1. Integrating predictive processing accounts and functional associations with ToMN activity

Our core aim was to help address what underlying process is supported by ToMN activity, given the heterogeneity of the tasks and stimuli that activate it. As outlined in our introduction, a variety of social tasks activate the ToMN— e.g. reading stories, watching social animations, taking perspectives, making strategic decisions, inferring traits, forming impressions, reading moral statements, and experiencing shared narratives (for review, see Amodio and Frith, 2006; Schurz et al., 2014, 2017; Van Overwalle, 2009). However, the functional scope of these cortical regions broadens even further when cortex in close proximity to the ToMN is considered. For instance, ToMN regions partially overlap with the default mode network (Buckner, 2012; Mars et al., 2012; Schurz et al., 2017), a network thought to subserve the formation of a predictive model of the external environment (Barrett, 2017; Hassabis and Maguire, 2009). Further, although ToMN regions can be spatially distinguished by peak activation from nearby cortical regions that are responsive to non-social tasks (e.g. attentional reorienting; Scholz et al., 2009; Young et al., 2010a, 2010b), it remains the case that ToMN regions are in close spatial proximity to regions activated by a variety of other abstract non-social tasks (e.g. the TPJ in particular, by episodic memory encoding/retrieval and language; for review, see Cabeza et al., 2012; Carter and Huettel, 2013), and by sensory events occurring over long temporal scales (recruiting ToMN regions more generally; Baldassano et al., 2017; Richardson and Saxe, 2019). To adequately describe the computational processes underlying ToMN activity, a computational account should accommodate the functional heterogeneity in these areas.

In one sense, then, the contribution of the present work might seem small. ToMN activity has been previously characterized by a list of associations, and we added one more: an association between ToMN activity and metaethical judgment, where subjective morals elicited more ToMN activity, and objective morals elicited less. However, in another sense, the present work contributes to more fundamental debates about

[5] Dodell-Feder et al. (2011) did identify an association between activity in right posterior temporal parietal sulcus (a region not typically considered part of the ToMN) and mental state information (the general measure included in our study; Fig. 5b). The effect size was comparable ($\beta = 0.680$), although slightly higher than the effects observed in the present work (e.g. mental state information in general, $\beta = 0.31$, information about one's own mental states, $\beta = 0.47$, and information about others' mental states, $\beta = 0.51$).

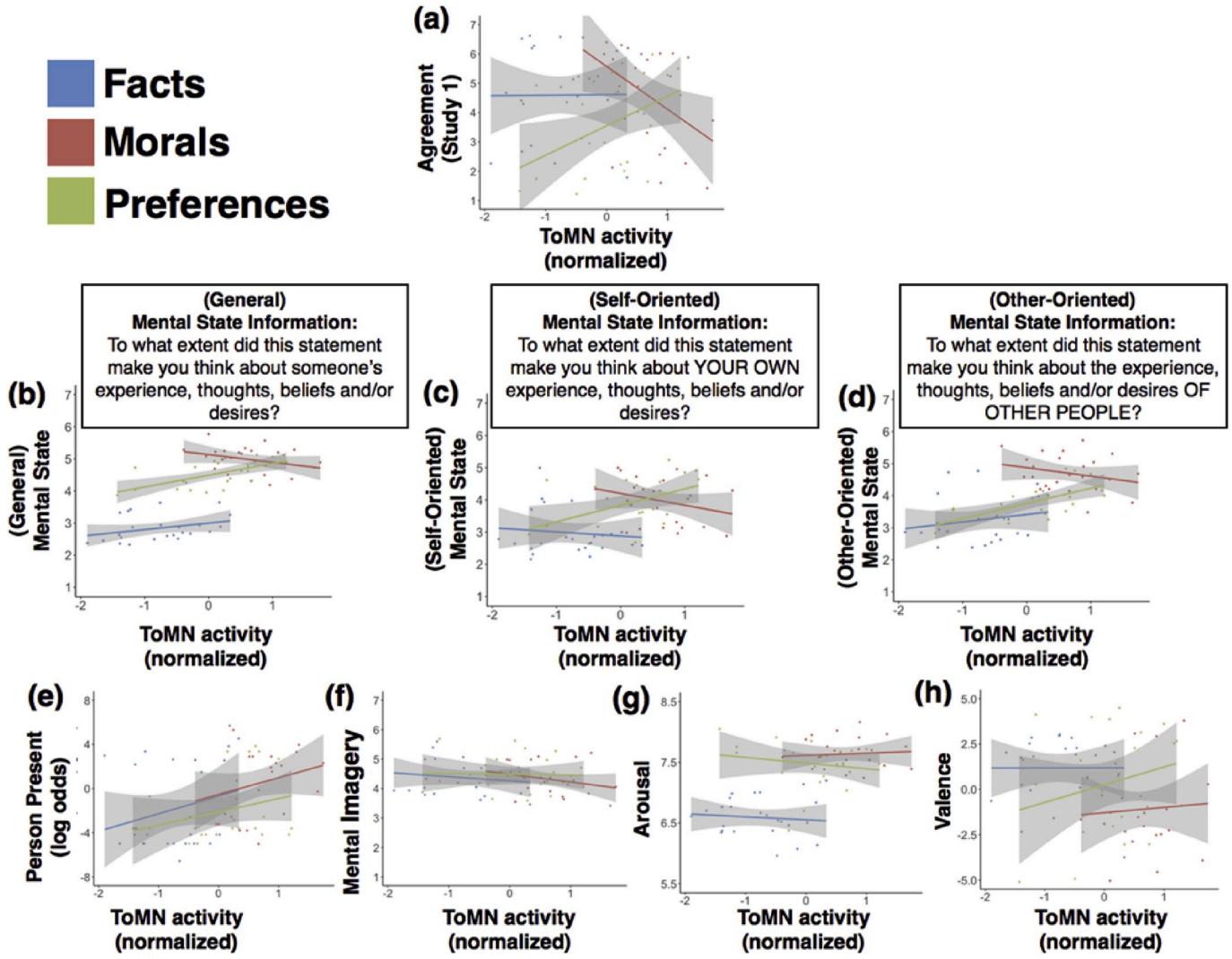

Fig. 5. Exploratory behavioral–BOLD relationships. By-stimulus behavioral estimates of agreement were extracted from Study 1, and from independent online studies for remaining measures. By-stimulus estimates of ToMN activity were extracted from Study 2 (3.2.2), and ToMN and VMPFC estimates were averaged for the figures above, given that no interactions with ROI were observed. (a) Agreement was negatively associated with ToMN activity among moral statements. Mental state inferences were positively associated with ToMN activity among preferences, but not among facts or morals, for each of the three ways the questions were asked: (b) general mental state information, (c) self-oriented mental state information, and (d) other-oriented mental state information. (e) The presence of a person in the statement was positively associated with ToMN activity among facts, morals, and preferences. (f–h) ToMN activity was not significantly associated with mental imagery, arousal, or valence. Shaded areas represent 95% confidence intervals. The full correlation matrix comparing all measures is available in Fig. S2, in the online supplemental materials.

what computational process the ToMN implements. That is, the present work identified an association that is more plausibly incorporated into unifying predictive accounts of ToMN activity (e.g. Koster-Hale and Saxe, 2013), as opposed to an account based around functional modules (Baron-Cohen, 1995; Cosmides and Tooby, 1992; Scholl and Leslie, 2001)—a strong version of which might use the association we have identified to suggest that the brain contains a spatially constrained module for metaethical judgment, as prior work has similarly argued for the existence of functionally localized modules for language, number, and memory retrieval, among others (e.g. Nelson et al., 2012). Perhaps an argument for such a metaethical module could be advanced; but in our opinion, given that metaethical judgment represents judgments about an abstract feature of a moral judgment (which itself, is not clearly localized; Young and Dungan, 2012), interpreting our findings in this way might push against the limits of plausibility.

Likewise, because we did identify an association between ToMN activity and agreement within morals (Fig. 5a), and because agreement was strongly correlated with fact-like, moral-like, and preference-like ratings among morals (Fig. S2), one could argue that our study is better understood as identifying a relationship between agreement with morals and ToMN activity. Within the present design we cannot rule out this confound, but we want to reemphasize that this interpretation would move us toward a theoretical framework that embraces functional specificity and brings the heterogeneity of observed effects into conflict. That is, it seems more probable that the distinct and occasionally opposite sets of associations we observed within the ToMN (e.g. a positive association with agreement among preferences, and a negative association with agreement among morals) all fall under the umbrella of some more general function that is consistently associated with ToMN activity but varies in its implementation across social contexts. It is for this reason that we favor a predictive processing account to explain our results. Future work might use a design or stimuli set that unconfounds agreement and predictability, for example, by presenting a conservative living in a liberal city with a widely shared liberal moral belief (e.g. pro-choice beliefs) and a conservative moral belief held by a minority (e.g. pro-life beliefs). In this case, we would predict relatively

less ToMN activity in a pro-life participant reading the widely shared (but personally disagreeable) pro-choice belief, compared to the minority (but personally agreeable) pro-life belief.

To this end, rather than concluding that subregions within the ToMN (or surrounding the TPJ) each implement a distinct functional computation (i.e. a "fractionated" modular perspective; see Schuwerk et al., 2017), some have put forward more generalized accounts of ToMN activity, hypothesizing that regions within it subserve more general and fundamental aspects of information processing (Cabeza et al., 2012; Carter and Huettel, 2013; Decety and Lamm, 2007; Geng and Vossel, 2013; Lee and McCarthy, 2016; Schuwerk et al., 2017), such as updating context-dependent predictions when those predictions are violated (Geng and Vossel, 2013).[6] These accounts dovetail nicely with predictive processing accounts reviewed in the introduction (e.g. Barrett and Simmons, 2015; Chanes and Barrett, 2016; A. Clark, 2013, 2015; Denève and Jardri, 2016; Friston et al., 2016, 2017; Hohwy, 2013; Hutchinson and Barrett, 2019; Joiner et al., 2017; Koster-Hale and Saxe, 2013, 2013; Rao and Ballard, 1999; Shadmehr et al., 2010; Spratling, 2017; Van de Cruys et al., 2014), where the brain is thought to act as a "hierarchical prediction machine" (A. Clark, 2013), predicting and filtering incoming sensory signals and their multimodal compressions throughout the cortical hierarchy.

Note that this explanation does not dispute the finding that social information reliably elicits activity in spatially constrained cortical regions (Dodell-Feder et al., 2011; Saxe and Kanwisher, 2003; Saxe and Powell, 2006; Saxe and Wexler, 2005; Scholz et al., 2009; Young et al., 2010a, 2010b). Indeed, we observed just this, finding positive associations between ToMN activity and the presence of a person within facts, morals, and preferences, and a positive association between ToMN activity and multiple measures of mental state information within preferences (Fig. 5b–d). We also observed a marginal positive association between ToMN activity and agreement within preferences (the opposite of the pattern observed among morals; Fig. 5a), but this most likely stems from the confounding of agreement and measures of mental state information within our sample of preferences (Fig. S2).

A predictive account, however, unlike a modular perspective, suggests that socially-sensitive brain regions are the particular point of collision between bottom-up multimodal compressions of sensory signals (e.g. the sights, sounds, etc., that collectively form the percept of a person) and top-down predictions about latent causes (i.e. the latent mental states that are predicted to produce observed patterns of behavior, movement, etc.). On this predictive account, then, ToMN activity is thought to reflect the updating of predictions about high-level compressions of sensory information (Bach and Schenke, 2017; Hohwy, 2013; Kilner and Frith, 2008; Koster-Hale and Saxe, 2013; Theriault et al., 2019; also, see Ondobaka et al., 2017).

If this general predictive account were correct, then one would expect to see multiple functional associations within ToMN regions, given that several stimulus features may simultaneously facilitate updating predictions about latent mental states. We observed exactly this: in addition to the association, within morals, between ToMN activity and metaethical judgments, we also observed a positive association between the presence of a person and ToMN activity across facts, morals, and preferences, a negative association between agreement and ToMN activity among morals, and a positive association between mental state inferences and ToMN activity among preferences. On a general predictive interpretation, across facts, morals, and preferences the presence of a person may facilitate predictions about mental states (as their presence provides a referent for mental state predictions), whereas metaethical features may only facilitate predictions about mental states among moral statements (as social consensus is generally more likely to guide predictions in the moral domain than among facts and preferences; see section 4.2). Although our present findings are only *consistent* with this account—and not definitive evidence for it—a general account of ToMN activity may nonetheless be a more attractive way to integrate the heterogeneity of previously observed associations in ToMN regions and the cortex surrounding them, especially when considering the association between ToMN activity and metaethical judgment observed in the present work.

4.2. Morality and metaethics in an informational context

Principles of information theory and predictive processing—e.g. that perfectly predictable signals carry no information—are of obvious use to neuroscience, as they make concrete predictions about how information is represented within the brain. Although the utility of these perspectives may appear less obvious in the context of morality, they may actually help to organize a variety of insights into the nature and importance of moral beliefs under a coherent theoretical framework. Critically, an understanding of social information grounded in prediction, prediction error, and the precision of predictions, may provide new perspectives on metaethical judgment, the importance of morality to social identity, and the demarcation of social domains (i.e. morals vs. preferences). Below, we review each of these potential contributions in turn.

Objective moral beliefs are moral beliefs that are supported by a social consensus (Ayars and Nichols, in press; Beebe, 2014; Goodwin and Darley, 2012; Heiphetz and Young, 2017), making them predictable. By extension, if objective moral beliefs are predictable (i.e. on the basis of social consensus), then affirming them carries less information. For example, when someone tells you they believe that "drinking and driving is wrong", then their beliefs roughly match what you would have already predicted, knowing nothing about them. Conversely, disavowing an objective moral belief provides more information about the person who does the disavowing, e.g. hearing someone say "drinking and driving is good" is informative about them. In this way, metaethics may be productively conceptualized (for some purposes) in informational terms: objective and subjective moral claims may roughly correspond to moral claims that are uninformative and informative, respectively.

The centrality of moral beliefs to people's perceptions of individual identity has been demonstrated in experimental contexts involving hypothetical (Heiphetz et al., 2018; Strohminger and Nichols, 2014) and real (Strohminger and Nichols, 2015) moral changes. For example, changes in a person's moral beliefs are judged as changing their core identity more than changes in memories or physical abilities. In informational terms, if moral beliefs are generally predicted to remain more consistent over time than memories, then moral changes may be understood as representing an informationally important shift in an individual's personality. That is, we expect memories and physical abilities to change; but we do not expect moral changes to the same degree, making such changes more informative. Of course, this leaves unanswered *why* moral beliefs are predicted to remain more consistent over time, a line of questioning that could be pursued in future work.

Finally, behavioral and neural distinctions between social domains (e.g. morals vs. preferences) might be explained by differences in how predictive processes are applied, for example, differences in the precision of predictions that are typically made about the domain. In the context of information theory and predictive processing models, prediction error is determined by both the prediction (e.g. a distribution, with a mean) and its precision (e.g. the width of that distribution; Feldman and Friston, 2010; Kim et al., 2020; Van de Cruys et al., 2014). For example, a slight deviation from the mean of a high-precision prediction (i.e. a tight distribution) would create more prediction error than a slight deviation from the mean of a low-precision prediction. Predictions in some social domains may be more precise than in others. For

[6] Geng and Vossell (2013, p. 2616) are also clear that updating in the context of mental inference may be subserved by distinct sub-regions within the TPJ (as in Scholz et al., 2009) under the umbrella of this more general process of contextual updating.

example, in the context of morality, a culture often accepts only one, or a few, beliefs as acceptable (e.g. on the acceptability of slavery), whereas on questions of preferences, many beliefs are acceptable (e.g. on the best flavor of ice cream).[7] On an information theoretic perspective, this would mean that predictions in the moral domain may be (on average) more precise than predictions in the preference domain. This interpretation that may explain our previous finding that moral statements elicit greater activity than preferences throughout the ToMN (Theriault et al., 2017), as low-precision predictions about the preferences of generic people afford less opportunity for predictions to be violated (and prediction error to be generated). Further, an explanatory framework grounded in predictive processes may provide an avenue for understanding other discrepancies we observed, e.g. that ToMN activity was associated with mental state information among preferences but not among morals (Fig. 5b–d). Rather than assuming that mental inference draws on different brain regions for morals and preferences, we propose that common predictive processes support mental inference across diverse social contexts. Of course, future work might more closely examine factors that vary across these contexts (e.g. predictive precision); but nonetheless, we believe that understanding social cognition in information-based terms may be a productive method to integrate several empirical findings in social psychology, both with each other and with emerging and unifying predictive theories of cortical function.

4.3. Probing BOLD activity through item analysis

Finally, it is worth emphasizing the potential benefits of the item analysis approach used in the present work. Social information is complex, and although one analytic approach would be to fully embrace this complexity, this would also complicate investigations of specific features and their influence on BOLD activity. 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. as in economic games). Item analysis offers a middle path, where researchers can relax constraints on their stimulus set to maximize variance and generalizability (Westfall et al., 2017), but, at the same time, conduct fine-grained analyses across the feature space represented by the population of stimuli. Furthermore, this approach is advantageous in that stimuli can be normed and reused—e.g. researchers can utilize our stimuli (and our by-stimulus estimates) to test their own hypotheses about the dimensions underlying BOLD activity (see Appendix A for the complete stimulus set and paired by-stimulus estimates, and see https://osf.io/cx4dp/for raw functional, anatomical, and behavioral by-stimulus estimates in standardized BIDS format; Gorgolewski et al., 2016). By-stimulus variability is often left unexamined in existing datasets (e.g. in an emotional expression task, amygdala activity varies across emotional faces; Westfall et al., 2017), and future work in neuroimaging may benefit from adopting this underutilized method of independently norming stimuli to examine by-stimulus variance.

[7] One might object that this is an unfair comparison, and that there are fewer possible answers to moral questions (e.g. slavery either is, or is not acceptable, whereas any ice cream flavor might be best); however, we think this view is mistaken. A society might organize itself around any number of arbitrary rules for when slavery is, or is not acceptable (e.g. only for people of a certain race, or for people with a certain accent, or not for children, or only when someone has committed a crime, etc.). That questions of morality intuitively appear as binary, whereas questions of preferences do not, may itself be a product of the precision of predictions in the moral domain. That is, if moral beliefs are predicted precisely, then a sharp contrast may be drawn between the few beliefs that fit the prediction and the many beliefs that do not. By contrast, less precise predictions about preferences may make it more intuitive that a variety of beliefs could answer the question of "which ice cream flavor is best", each fitting to a different degree.

4.4. Conclusion

The theory of mind network is broadly involved in social cognition (for review, see Amodio and Frith, 2006; Schurz et al., 2014, 2017; Van Overwalle, 2009), and it has recently been suggested that a more general computational process may account for this activity: processing prediction error in social contexts (Koster-Hale and Saxe, 2013). Prior work has been consistent with this hypothesis, but this work has generally examined social expectations formed inside the lab (e.g. Dungan et al., 2016; Saxe and Wexler, 2005) or under conditions of explicit instruction (Brass et al., 2007; de Lange et al., 2008), and has not conducted more fine-grained analyses of by-stimulus variance. In the present work, we leveraged an existing dataset (Theriault et al., 2017) to examine by-stimulus variability in ToMN evoked by moral statements, focusing on preexisting expectations as operationalized by metaethical judgments. We observed that ToMN activity was negatively associated with moral objectivity (i.e. how fact-like a moral statement was judged to be, by an independent online sample) and positively associated with moral subjectivity (i.e. how preference-like a moral statement was judged to be, by an independent online sample). This finding is consistent with hypotheses derived from predictive processing models, and, although it is not definitive, it underscores the need for an overarching explanation of ToMN activity that can accommodate the heterogeneity of associations previously observed in these regions and in the cortex surrounding them (e.g. Carter and Huettel, 2013; Geng and Vossel, 2013; Koster-Hale and Saxe, 2013).

CRediT authorship contribution statement

Jordan Theriault: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. **Adam Waytz:** Conceptualization, Writing - review & editing. **Larisa Heiphetz:** Conceptualization, Writing - review & editing. **Liane Young:** Conceptualization, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

We thank Fiery Cushman, Drew Linsley, Sean MacEvoy, Jonathan Phillips, James Russell, Ajay Satpute, Rebecca Saxe, and members of the Morality Lab for feedback. This work was supported by the National Science Foundation 5103831 (L.Y.), the John Templeton Foundation 5107321 (L.Y.), the National Science Foundation SMA-1408989 (L.H.), and Natural Sciences and Engineering Research Council of Canada PGSD3-420445 (J.T.). This work was also supported (in part) by an award from the Russell Sage Foundation (L.H.). Any opinions expressed are those of the authors alone and should not be construed as representing the opinions of the Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2020.107475.

References

Amodio, D.M., Frith, C.D., 2006. Meeting of minds: the medial frontal cortex and social cognition. Nat. Rev. Neurosci. 7 (4), 268–277. https://doi.org/10.1038/nrn1884.

Ampe, L., Ma, N., Van Hoeck, N., Vandekerckhove, M., Van Overwalle, F., 2014. Unusual actions do not always trigger the mentalizing network. Neurocase 20 (2), 144–149. https://doi.org/10.1080/13554794.2012.741251.

Arend, M.G., Schäfer, T., 2019. Statistical power in two-level models: a tutorial based on Monte Carlo simulation. Psychol. Methods 24 (1), 1–19. https://doi.org/10.1037/met0000195.

Ayars, A., Nichols, S., 2020. Rational learners and metaethics: universalism, relativism, and evidence from consensus. Mind Lang. https://doi.org/10.1111/mila.12232 (in press).

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. J. Mem. Lang. 59 (4), 390–412. https://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 (7), e12312. https://doi.org/10.1111/spc3.12312.
Baldassano, C., Chen, J., Zadbood, A., Pillow, J.W., Hasson, U., Norman, K.A., 2017. Discovering event structure in continuous narrative perception and memory. Neuron 95 (3), 709–721. https://doi.org/10.1016/j.neuron.2017.06.041 e5.
Baron, S.G., Gobbini, M.I., Engell, A.D., Todorov, A., 2011. Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions. Soc. Cognit. Affect Neurosci. 6 (5), 572–581. https://doi.org/10.1093/scan/nsq086.
Baron-Cohen, S., 1995. Mindblindness: an Essay on Autism and Theory of Mind. MIT Press.
Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. J. Mem. Lang. 68 (3), 255–278. https://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. Soc. Cognit. Affect Neurosci. 12 (1), 1–23. https://doi.org/10.1093/scan/nsw154.
Barrett, L.F., Simmons, W.K., 2015. Interoceptive predictions in the brain. Nat. Rev. Neurosci. 16 (7), 419–429. https://doi.org/10.1038/nrn3950.
Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Software 67 (1), 1–48. https://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. https://doi.org/10.1016/j.neuroimage.2007.01.039.
Beebe, J.R., 2014. How different kinds of disagreement impact folk metaethical judgments. In: Wright, J.C., Sarkissian, H. (Eds.), Advances in Experimental Moral Psychology. Bloomsbury Academic, pp. 167–187. https://doi.org/10.5040/9781472594150.
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. J. Neurosci. 33 (22), 9337–9344. https://doi.org/10.1523/JNEUROSCI.5634-12.2013.
Blakemore, S.-J., 2003. The detection of contingency and animacy from simple animations in the human brain. Cerebr. Cortex 13 (8), 837–844. https://doi.org/10.1093/cercor/13.8.837.
Bloom, H.S., 1995. Minimum detectable effects: a simple way to report the statistical power of experimental designs. Eval. Rev. 19 (5), 547–556. https://doi.org/10.1177/0193841X9501900504.
Brass, M., Schmitt, R.M., Spengler, S., Gergely, G., 2007. Investigating action understanding: inferential processes versus action simulation. Curr. Biol. 17 (24), 2117–2121. https://doi.org/10.1016/j.cub.2007.11.057.
Buckner, R.L., 2012. The serendipitous discovery of the brain's default network. Neuroimage 62 (2), 1137–1145. https://doi.org/10.1016/j.neuroimage.2011.10.035.
Cabeza, R., Ciaramelli, E., Moscovitch, M., 2012. Cognitive contributions of the ventral parietal cortex: an integrative theoretical account. Trends Cognit. Sci. 16 (6), 338–352. https://doi.org/10.1016/j.tics.2012.04.008.
Carter, R.M., Huettel, S.A., 2013. A nexus model of the temporal–parietal junction. Trends Cognit. Sci. 17 (7), 328–336. https://doi.org/10.1016/j.tics.2013.05.007.
Chanes, L., Barrett, L.F., 2016. Redefining the role of limbic areas in cortical processing. Trends Cognit. Sci. 20 (2), 96–106. https://doi.org/10.1016/j.tics.2015.11.005.
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 (13), 3105–3113. https://doi.org/10.1016/j.neuropsychologia.2007.05.011.
Clark, A., 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. 36 (3), 1–24. https://doi.org/10.1017/S0140525X12000477.
Clark, A., 2015. Surfing Uncertainty: Prediction, Action, and the Embodied Mind. Oxford University Press.
Clark, H.H., 1973. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. J. Verb. Learn. Verb. Behav. 12 (4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3.
Cloutier, J., Gabrieli, J.D.E., 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 (2), 583–588. https://doi.org/10.1016/j.neuroimage.2011.04.051.
Cosmides, L., Tooby, J., 1992. Cognitive adaptations for social exchange. In: Barkow, J., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, pp. 163–228.
de Lange, F.P., Spronk, M., Willems, R.M., Toni, I., Bekkering, H., 2008. Complementary systems for understanding action intentions. Curr. Biol.: Cailiao Baohu 18 (6), 454–457. https://doi.org/10.1016/j.cub.2008.02.057.
Decety, J., Lamm, C., 2007. The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. Neuroscientist 13 (6), 580–593. https://doi.org/10.1177/1073858407304654.
Denève, S., Jardri, R., 2016. Circular inference: mistaken belief, misplaced trust. Current Opinion in Behavioral Sciences 11, 40–48. https://doi.org/10.1016/j.cobeha.2016.04.001.
Dodell-Feder, D., Koster-Hale, J., Bedny, M., Saxe, R., 2011. FMRI item analysis in a theory of mind task. Neuroimage 55 (2), 705–712. https://doi.org/10.1016/j.neuroimage.2010.12.040.
Donnet, S., Lavielle, M., Poline, J.-B., 2006. Are fMRI event-related response constant in time? A model selection answer. Neuroimage 31 (3), 1169–1176. https://doi.org/10.1016/j.neuroimage.2005.08.068.
Dungan, J.A., Stepanovic, M., Young, L.L., 2016. Theory of mind for processing unexpected events across contexts. Soc. Cognit. Affect Neurosci. 11 (8), 1183–1192. https://doi.org/10.1093/scan/nsw032.
Feldman, H., Friston, K.J., 2010. Attention, uncertainty, and free-energy. Front. Hum. Neurosci. 4 https://doi.org/10.3389/fnhum.2010.00215.
Fletcher, P., 1995. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition 57 (2), 109–128. https://doi.org/10.1016/0010-0277(95)00692-R.
Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., Pezzulo, G., 2016. Active inference and learning. Neurosci. Biobehav. Rev. 68, 862–879. https://doi.org/10.1016/j.neubiorev.2016.06.022.
Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., 2017. Active inference: a process theory. Neural Comput. 29 (1), 1–49. https://doi.org/10.1162/NECO_a_00912.
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 (1), 11–21. https://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 (2), 189–211. https://doi.org/10.1080/19345747.2011.618213.
Gelman, S.A., Rhodes, M., 2012. "Two-Thousand years of stasis": how psychological essentialism impedes evolutionary understanding. In: Rosengren, K.S., Brem, S., Evans, E.M., Sinatra, G.M. (Eds.), Evolution Challenges: Integrating Research and Practice in Teaching and Learning about Evolution. Oxford University Press. https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199730421.001.0001/acprof-9780199730421-chapter-1.
Geng, J.J., Vossel, S., 2013. Re-evaluating the role of TPJ in attentional control: contextual updating? Neurosci. Biobehav. Rev. 37 (10, Part 2), 2608–2620. https://doi.org/10.1016/j.neubiorev.2013.08.010.
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. J. Cognit. Neurosci. 19 (11), 1803–1814. https://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. https://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? J. Exp. Soc. Psychol. 48 (1), 250–256. https://doi.org/10.1016/j.jesp.2011.08.006.
Gorgolewski, K.J., Auer, T., Calhoun, V.D., Craddock, R.C., Das, S., Duff, E.P., Flandin, G., Ghosh, S.S., Glatard, T., Halchenko, Y.O., Handwerker, D.A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B.N., Nichols, T.E., Pellman, J., Poldrack, R.A., 2016. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific Data 3 (1), 1–9. https://doi.org/10.1038/sdata.2016.44.
Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z., 2004. Coh-Metrix: analysis of text on cohesion and language. Behav. Res. Methods Instrum. Comput. 36 (2), 193–202. https://doi.org/10.3758/BF03195564.
Green, P., MacLeod, C.J., 2016. SIMR: an R package for power analysis of generalized linear mixed models by simulation. Methods in Ecology and Evolution 7 (4), 493–498. https://doi.org/10.1111/2041-210X.12504.
Harris, L.T., Todorov, A., Fiske, S.T., 2005. Attributions on the brain: neuro-imaging dispositional inferences, beyond theory of mind. Neuroimage 28 (4), 763–769. https://doi.org/10.1016/j.neuroimage.2005.05.021.
Hassabis, D., Maguire, E., 2009. The construction system of the brain. Phil. Trans. Biol. Sci. 364 (1521), 1263–1271. https://doi.org/10.1098/rstb.2008.0296.
Heiphetz, L., Spelke, E.S., Harris, P.L., Banaji, M.R., 2014. What do different beliefs tell us? An examination of factual, opinion-based, and religious beliefs. Cognit. Dev. 30, 15–29. https://doi.org/10.1016/j.cogdev.2013.12.002.
Heiphetz, L., Strohminger, N., Gelman, S.A., Young, L.L., 2018. Who am I? The role of moral beliefs in children's and adults' understanding of identity. J. Exp. Soc. Psychol. 78, 210–219. https://doi.org/10.1016/j.jesp.2018.03.007.
Heiphetz, L., Young, L.L., 2017. Can only one person be right? The development of objectivism and social preferences regarding widely shared and controversial moral beliefs. Cognition 167, 78–90. https://doi.org/10.1016/j.cognition.2016.05.014.
Hohwy, J., 2013. The Predictive Mind. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199682737.001.0001.
Hothorn, T., Bretz, F., Westfall, P., 2008. Simultaneous inference in general parametric models. Biometrical Journal. Biometrische Zeitschrift 50 (3), 346–363. https://doi.org/10.1002/bimj.200810425.
Hutchinson, B., Barrett, L.F., 2019. The power of predictions: an emerging paradigm for psychological research. Curr. Dir. Psychol. Sci. 28 (3), 280–291. https://doi.org/10.1177/0963721419831992.
Jenkins, A.C., Mitchell, J.P., 2010. Mentalizing under uncertainty: dissociated neural responses to ambiguous and unambiguous mental state inferences. Cerebr. Cortex 20 (2), 404–410. https://doi.org/10.1093/cercor/bhp109.
Joiner, J., Piva, M., Turrin, C., Chang, S.W.C., 2017. Social learning through prediction error in the brain. Npj Science of Learning 2 (1). https://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. J. Pers. Soc. Psychol. 103 (1), 54–69. https://doi.org/10.1037/a0028347.
Kilner, J.M., Frith, C.D., 2008. Action observation: inferring intentions without mirror neurons. Curr. Biol. 18 (1), R32–R33. https://doi.org/10.1016/j.cub.2007.11.008.

Kim, M., Park, B., Young, L., 2020. The psychology of motivated versus rational impression updating. Trends Cognit. Sci. 24 (2), 101–111. https://doi.org/10.1016/j.tics.2019.12.001.
Kircher, T., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J., van Os, J., Krach, S., 2009. Online mentalising investigated with functional MRI. Neurosci. Lett. 454 (3), 176–181. https://doi.org/10.1016/j.neulet.2009.03.026.
Koster-Hale, J., Saxe, R., 2013. Theory of mind: a neural prediction problem. Neuron 79 (5), 836–848. https://doi.org/10.1016/j.neuron.2013.08.020.
Kron, A., Goldstein, A., Lee, D.H.-J., Gardhouse, K., Anderson, A.K., 2013. How are you feeling? Revisiting the quantification of emotional qualia. Psychol. Sci. 24 (8), 1503–1511. https://doi.org/10.1177/0956797613475456.
Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest package: tests in linear mixed effects models. J. Stat. Software 82 (1), 1–26. https://doi.org/10.18637/jss.v082.i13.
Lee, S.M., McCarthy, G., 2016. Functional heterogeneity and convergence in the right temporoparietal junction. Cerebr. Cortex 26 (3), 1108–1116. https://doi.org/10.1093/cercor/bhu292.
Ma, N., Vandekerckhove, M., Baetens, K., Van Overwalle, F., Seurinck, R., Fias, W., 2012a. Inconsistencies in spontaneous and intentional trait inferences. Soc. Cognit. Affect Neurosci. 7 (8), 937–950. https://doi.org/10.1093/scan/nsr064.
Ma, N., Vandekerckhove, M., Hoeck, N.V., Overwalle, F.V., 2012b. Distinct recruitment of temporo-parietal junction and medial prefrontal cortex in behavior understanding and trait identification. Soc. Neurosci. 7 (6), 591–605. https://doi.org/10.1080/17470919.2012.686925.
Mars, R.B., Neubert, F.-X., Noonan, M.P., Sallet, J., Toni, I., Rushworth, M.F.S., 2012. On the relationship between the "default mode network" and the "social brain. Front. Hum. Neurosci. 6 https://doi.org/10.3389/fnhum.2012.00189.
McNamara, D.S., Louwerse, M.M., Cai, Z., Graesser, A., 2014. Coh-Metrix Version 3.0. http://cohmetrix.com.
Mende-Siedlecki, P., Cai, Y., Todorov, A., 2013. The neural dynamics of updating person impressions. Soc. Cognit. Affect Neurosci. 8 (6), 623–631. https://doi.org/10.1093/scan/nss040.
Mende-Siedlecki, P., Todorov, A., 2016. Neural dissociations between meaningful and mere inconsistency in impression updating. Soc. Cognit. Affect Neurosci. 11 (9), 1489–1500. https://doi.org/10.1093/scan/nsw058.
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 (4), 757–762. https://doi.org/10.1016/j.neuroimage.2005.03.011.
Nelson, S.M., McDermott, K.B., Petersen, S.E., 2012. In favor of a 'fractionation' view of ventral parietal cortex: comment on Cabeza et al. Trends Cognit. Sci. 16 (8), 399–400. https://doi.org/10.1016/j.tics.2012.06.014.
Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15 (1), 1–25. https://doi.org/10.1002/hbm.1058.
Ondobaka, S., Kilner, J., Friston, K., 2017. The role of interoceptive inference in theory of mind. Brain Cognit. 112, 64–68. https://doi.org/10.1016/j.bandc.2015.08.002.
Park, B., Young, L., 2020. An association between biased impression updating and relationship facilitation: a behavioral and fMRI investigation. J. Exp. Soc. Psychol. 87, 103916. https://doi.org/10.1016/j.jesp.2019.103916.
Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind? Behav. Brain Sci. 1 (4), 515–526. https://doi.org/10.1017/S0140525X00076512.
R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org.
Rao, R.P.N., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2 (1), 79–87. https://doi.org/10.1038/4580.
Richardson, H., Saxe, R., 2019. Development of predictive responses in theory of mind brain regions. Dev. Sci., e12863 https://doi.org/10.1111/desc.12863.
Ruby, P., Decety, J., 2003. What you believe versus what you think they believe: a neuroimaging study of conceptual perspective-taking: a PET study of conceptual perspective-taking. Eur. J. Neurosci. 17 (11), 2475–2480. https://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 Lang. 26 (4), 482–505. https://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 (4), 1835–1842. https://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. Psychol. Sci. 17 (8), 692–699. https://doi.org/10.1111/j.1467-9280.2006.01768.x.
Saxe, R., Wexler, A., 2005. Making sense of another mind: the role of the right temporo-parietal junction. Neuropsychologia 43 (10), 1391–1399. https://doi.org/10.1016/j.neuropsychologia.2005.02.013.
Schiller, D., Freeman, J.B., Mitchell, J.P., Uleman, J.S., Phelps, E.A., 2009. A neural mechanism of first impressions. Nat. Neurosci. 12 (4), 508–514. https://doi.org/10.1038/nn.2278.
Scholl, B.J., Leslie, A.M., 2001. Minds, modules, and meta-analysis. Child Dev. 72 (3), 696–701. https://doi.org/10.1111/1467-8624.00308.
Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E.N., Saxe, R., 2009. Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. PloS One 4 (3), e4869. https://doi.org/10.1371/journal.pone.0004869.
Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J., 2014. Fractionating theory of mind: a meta-analysis of functional brain imaging studies. Neurosci. Biobehav. Rev. 42, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009.
Schurz, M., Tholen, M.G., Perner, J., Mars, R.B., Sallet, J., 2017. Specifying the brain anatomy underlying temporo-parietal junction activations for theory of mind: a review using probabilistic atlases from different imaging modalities. Hum. Brain Mapp. 38 (9), 4788–4805. https://doi.org/10.1002/hbm.23675.
Schuwerk, T., Schurz, M., Müller, F., Rupprecht, R., Sommer, M., 2017. The rTPJ's overarching cognitive function in networks for attention and theory of mind. Soc. Cognit. Affect Neurosci. 12 (1), 157–168. https://doi.org/10.1093/scan/nsw163.
Shadmehr, R., Smith, M.A., Krakauer, J.W., 2010. Error correction, sensory prediction, and adaptation in motor control. Annu. Rev. Neurosci. 33 (1), 89–108. https://doi.org/10.1146/annurev-neuro-060909-153135.
Shannon, C., Weaver, W., 1964. The Mathematical Theory of Communication, tenth ed. The University of Illinois Press (Original work published 1949).
Spratling, M.W., 2017. A review of predictive coding algorithms. Brain Cognit. 112, 92–97. https://doi.org/10.1016/j.bandc.2015.11.003.
Strohminger, N., Nichols, S., 2014. The essential moral self. Cognition 131 (1), 159–171. https://doi.org/10.1016/j.cognition.2013.12.005.
Strohminger, N., Nichols, S., 2015. Neurodegeneration and identity. Psychol. Sci. 26 (9), 1469–1479. https://doi.org/10.1177/0956797615592381.
Theriault, J.E., Waytz, A., Heiphetz, L., Young, L.L., 2017. Examining overlap in behavioral and neural representations of morals, facts, and preferences. J. Exp. Psychol. Gen. 146 (11), 1586–1605. https://doi.org/10.1037/xge0000350.
Theriault, J.E., Young, L.L., Barrett, L.F., 2019. The sense of should: a biologically-based model of social pressure. PsyArXiv. https://doi.org/10.31234/osf.io/x5rbs.
Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., Wagemans, J., 2014. Precise minds in uncertain worlds: predictive coding in autism. Psychol. Rev. 121 (4), 649–675. https://doi.org/10.1037/a0037665.
Van Overwalle, F., 2009. Social cognition and the brain: a meta-analysis. Hum. Brain Mapp. 30 (3), 829–858. https://doi.org/10.1002/hbm.20547.
Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., Maier, W., Shah, N.J., Fink, G.R., Zilles, K., 2001. Mind reading: neural mechanisms of theory of mind and self-perspective. Neuroimage 14 (1), 170–181. https://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 participants respond to samples of stimuli. J. Exp. Psychol. Gen. 143 (5), 2020–2045. https://doi.org/10.1037/xge0000014.
Westfall, J., Nichols, T.E., Yarkoni, T., 2017. Fixing the stimulus-as-fixed-effect fallacy in task fMRI. Wellcome Open Research 1. https://doi.org/10.12688/wellcomeopenres.10298.2.
Woo, C.-W., Krishnan, A., Wager, T.D., 2014. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. Neuroimage 91, 412–419. https://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. Phil. Psychol. 26 (3), 336–361. https://doi.org/10.1080/09515089.2011.633751.
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. Psychol. Sci. 28 (3), 307–319. https://doi.org/10.1177/0956797616682029.
Young, L.L., Camprodon, J.A., Hauser, M., Pascual-Leone, A., Saxe, R., 2010a. Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. Proc. Natl. Acad. Sci. Unit. States Am. 107 (15), 6753–6758. https://doi.org/10.1073/pnas.0914826107.
Young, L.L., Cushman, F., Hauser, M., Saxe, R., 2007. The neural basis of the interaction between theory of mind and moral judgment. Proc. Natl. Acad. Sci. Unit. States Am. 104 (20), 8235–8240. https://doi.org/10.1073/pnas.0701408104.
Young, L.L., Dodell-Feder, D., Saxe, R., 2010b. What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. Neuropsychologia 48 (9), 2658–2664. https://doi.org/10.1016/j.neuropsychologia.2010.05.012.
Young, L.L., Dungan, J., 2012. Where in the brain is morality? Everywhere and maybe nowhere. Soc. Neurosci. 7 (1), 1–10. https://doi.org/10.1080/17470919.2011.569146.
Young, L.L., Saxe, R., 2009. An fMRI investigation of spontaneous mental state inference for moral judgment. J. Cognit. Neurosci. 21 (7), 1396–1405. https://doi.org/10.1162/jocn.2009.21137.