Social prediction in the Theory of Mind Network

Minjae Kim[*1], Jordan Theriault[*2], & Liane Young[3]

* denotes equal contribution

[1]Johns Hopkins University
Department of Psychological & Brain Sciences

[2]Northeastern University
Department of Psychology

[3]Boston College
Department of Psychology & Neuroscience

Corresponding author:
Minjae Kim
mkim19@jhu.edu

Theory of mind refers to the ability to represent internal mental states (Premack & Woodruff, 1978). The Theory of Mind Network (ToMN) is a network of brain regions that has been consistently implicated in tasks involving mental inference (for review see Amodio & Frith, 2006, Schurz et al., 2014; Van Overwalle, 2009), such as reading comics and stories about beliefs (Ciaramidaro et al., 2007; Dodell-Feder et al., 2011; Fletcher et al., 1995; Gallagher et al., 2000; Saxe & Kanwisher 2003; Saxe & Powell, 2006; Young et al., 2007; 2010), watching social animations (Blakemore et al., 2003; Gobbini et al., 2007), perspective taking (Ruby & Decety, 2003; Vogeley et al., 2001), strategic decision making in economic games (Kircher et al., 2009), trait inference (Harris et al., 2005; Ma et al., 2012), impression formation (Baron et al., 2011; Bhanji & Beer, 2013; Cloutier et al., 2011; Ma et al., 2012; Mende-Siedlecki & Todorov, 2016; Mende-Siedlecki et al., 2013; Mitchell et al., 2005; Schiller et al., 2009), distinguishing categories of mental content (e.g., belief vs. preference, Jenkins & Mitchell, 2010; morals vs. preferences and facts; Theriault et al., 2017), and even the shared experience of narrative among listeners (Yeshurun et al., 2017). The ToMN also overlaps (at least partially; Mars et al., 2012, also see Schurz et al., 2017) with the default mode network, a network of regions active during rest that are thought to be critical for generating an internal model of the world (Barrett, 2017a; Buckner, 2012; Hassabis & Maguire, 2010). Although regions in this network are reliably activated by social tasks, less is known about the computations they might implement.

This broad, overlapping activation across a diverse population of tasks (especially overlap near default mode regions, thought to underlie high-level representations of the world; Barrett, 2017a; Buckner, 2012; Hassabis & Maguire, 2010) raises the possibility that ToMN activity may relate to a more general underlying process, rather than mental state representation specifically (and granted, an underlying processes tightly linked with social cognition). Consistent with this, some researchers have argued that theory of mind is often erroneously treated as a monolithic process (Schaafsma et al., 2014; Heyes, 2014), meaning that "theory of mind" may be an overly general functional explanation for ToMN activity. For instance, to represent a mind, one would presumably need to (at the very least) distinguish oneself from others, track goals and intentions, and understand causality (Schaafsma et al., 2014). Others have noted that the temporoparietal junction (TPJ), a critical region in the ToMN, is functionally well-positioned to play a high-level integrative role, sitting at the nexus of regions implicated in memory, attention, social cognition, and language (Carter & Huettel, 2013). Nonetheless, the dilemma remains: ToMN is strongly associated with social tasks and information, but it is unclear how best to characterize the processes underlying ToMN activity.

A promising hypothesis comes from recent theoretical work connecting theory of mind with a predictive coding framework (Koster-Hale & Saxe, 2013; see also Joiner et al., 2017). Predictive coding models have proposed a unifying framework for neural computation: that the brain is a "hierarchical prediction machine" (A. Clark, 2013). That is, the brain builds an internal model of the world to predict incoming sensory information (Friston, 2010). This model is efficient, taking up only sensory information that is informative, i.e. information that was not already anticipated by the model. Input to the brain-system, then, exists in the form of "prediction error"—the difference between the model's prediction (i.e. its prior), and the actual, raw sensory information arriving at the retina, nose, skin, etc. After sensory information enters the brain, it ascends a cortical hierarchy, with each layer receiving only prediction error from an earlier level, and only passing its own prediction error to the next. In this framework, social predictions could be characterized as high-level, abstracted predictions (A. Clark, 2013), about high-level derivatives of sensory information (Koster-Hale & Saxe, 2013). Some preliminary

work is consistent with this hypothesis, where unexpected or unpredicted social information elicits greater ToMN activity (e.g., Dungan et al., 2017; Mende-Siedlecki et al., 2013; Park, Fareri, Delgado, & Young, 2020; Kim, Mende-Siedlecki, Anzellotti, & Young, 2021).

At the same time, social predictions can be made (and violated) in a number of ways, as documented in classic social psychological work. For instance, we can attribute behavior to dispositional or situational sources (Gilbert & Malone, 1995)—e.g., "Dave was short with the waiter because Dave is a jerk" vs. "Dave was short with the waiter because the parking meter was running out". Furthermore, within a given social context, multiple norms can guide predictions: prescriptive norms describe expectations based on moral or social values, whereas descriptive norms refer to expectations based on statistical frequency (Brauer & Chaurand, 2010; Cialdini et al., 1990). For instance, descriptively, most people will drive a short distance rather than walk; however, prescriptively, this is frowned upon (i.e., people often do it, but they shouldn't; Brauer & Chaurand, 2010). Recent work has argued that predictive coding can be reconciled as a high-level explanation of this classic work, and in particular, as an explanation of the centrality of person and situation-based predictions (Bach & Schenke, 2017). In sum, information about people (i.e. dispositional information) and about descriptive/prescriptive norms could be used to generate social priors: predictions about specific people, or people in general within a given context. These sources of social prediction are examined in the present work.

Present Work

The present work tested the relationship between multiple forms of social prediction error and ToMN activity. Impression formation has often been studied by presenting discrete, contradictory pieces of information about individuals (e.g., Kim et al., 2021), an approach which can examine broad characteristics of the information (e.g., its valence), but which does not operationalize contextually dependent factors, such as prescriptive and descriptive norms. In the present work, we used a series of detailed narratives. Each narrative elicited an initial moral judgment, then, subsequently, induced participants to update that judgment (or not, in control conditions). The scenarios described an agent facing a moral dilemma: they described the background and potential outcomes (e.g., a hospital administrator must choose to save one sick child, or create a larger immunization program), then presented the agent's decision (e.g., the administrator creates the immunization program), then reframed the dilemma with additional information (e.g., the hospital board has promoted past administrators who began new programs). Importantly, the additional information reframed the scenario by introducing a previously hidden motive, rather than directly contradicting earlier information (Mann & Ferguson, 2015; Kim et al., 2022); this type of paradigm encourages participants to revise how they view the agent's decision, rather than their estimation of the reliability of the information or other meta-narrative features.

In Study 1, we characterized the narrative stimuli along multiple dimensions by first collecting ratings of each scenario on relevant measures (in an online sample) and then identifying the underlying dimensions (via principal components analysis). Within our stimulus set, dispositional, prescriptive, and descriptive sources of social prediction were clearly distinguished. In Study 2, we examined the relationship between Study 1 component scores and ToMN activity.

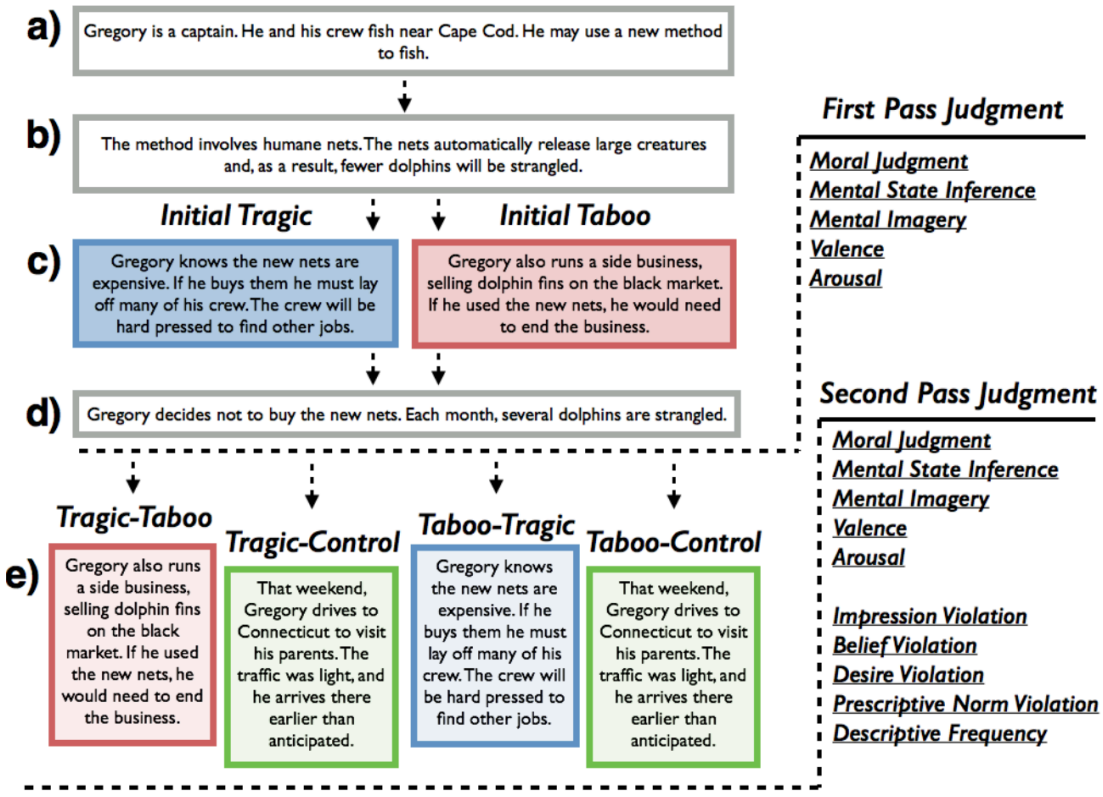
Study 1

In Study 1, we collected ten feature ratings for a set of 24 narrative scenarios, and conducted a PCA to identify the underlying dimensions of the stimulus set. Questions were selected to test: the extent to which the agent's decision violated *prescriptive* and *descriptive norms*, and the extent to which additional information violated *predictions about the agent* (Table 1). Questions also measured more general stimulus features used in prior item analyses of the Theory of Mind Network (*mental state inference* and *mental imagery*, Dodell-Feder et al., 2011), in addition to *valence*, *arousal*, and *moral judgment*.

Method

Participants. Participants were recruited online using Amazon Mechanical Turk (AMT) in two cohorts (each cohort responded to different sets of measures). Cohort one consisted of 239 adults (131 female, 107 male, 1 unspecified; M_{Age} = 36.4 years, SD_{Age} = 12.1 years), after excluding four participants for failing a simple attention check asking them to briefly describe any of the scenarios they had read. Cohort two consisted of 315 adults (140 female, 173 male, 2 unspecified; M_{Age} = 33.6 years, SD_{Age} = 9.4 years), after excluding seven participants for failing the same attention check. The study was approved by the Boston College Institutional Review Board; all participants provided informed consent and were compensated at an approximate rate of $6/hour (in line with standard rates at the time of data collection, 2012-2013).

Figure 1. Example scenario. Scenarios consisted of segments which could be substituted or rearranged to form four conditions: *tragic–taboo*, *tragic–control*, *taboo–tragic*, and *taboo–control* (1e). For each scenario, participants made either: *initial* and *final* judgments, or only a *final* judgment. The text above is abbreviated. Twenty-four scenarios were used in total; for full text see Appendix A.



Stimuli. Stimuli consisted of 24 root scenarios, adapted from or inspired by prior work (Critcher et al., 2012; Lichtenstein et al., 2007; Tetlock et al., 2000; Uhlmann et al., 2013). See Appendix A for full text of scenarios (appendix_a.pdf). Scenarios described *tragic* and *taboo* dilemmas. Tragic dilemmas forced agents to choose between two moral outcomes, whereas taboo dilemmas forced agents to choose between a moral outcome and a selfish outcome. For example, in a tragic dilemma, Gregory, a fishing boat captain, could save the jobs of his crew members, but at the expense of killing more dolphins; in a taboo dilemma, he could save the dolphins, but at a personal cost (Fig. 1a-c).

Scenarios were initially presented as either tragic or taboo; later, additional information was presented that reframed the scenarios. For instance, in the *tragic–taboo* version of the example scenario, participants first learn that Gregory must choose between saving dolphins (by purchasing expensive nets) and saving his crew's jobs. He chooses not to buy the new nets, saving his crew (a relatively moral motive; Fig. 1d). Next, participants learn that this decision actually sustained Gregory's side-business: selling dolphin fins on the black market (Fig. 1e). In the *taboo–tragic* version of this scenario, participants first learn that Gregory must choose between his side-business and saving dolphins. He chooses not to buy the new nets, preserving his side business (a selfish motive; Fig. 1d). Next, participants learn that this decision saved Gregory's crew (Fig. 1e). Thus, content for *tragic–taboo* and *taboo–tragic* versions of scenarios was identical, only presented in different orders. These *reframed* conditions were intermixed with *control* conditions, which appended morally irrelevant information following the agent's decision (Fig. 1e). All 24 root scenarios were used to form all four conditions: *tragic–taboo*, *tragic-control*, *taboo-tragic*, and *taboo-control* (96 total items).

For each scenario, there were two timepoints at which participants provided judgments. *Initial* judgments were made after the agent's decision in the dilemma was presented (Fig. 1d), and *final* judgments were made after reframing or control information was appended (Fig. 1e). In cohort one, all scenario segments were presented sequentially; in cohort two, segments a–d were presented simultaneously, followed by segment e. This change was made to reduce excessive page loading times.

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

Principal Component Analysis (PCA). In order to reduce the dimensionality of our feature ratings, we conducted a PCA on the ten collected measures using the *psych* package in R. Only *final* judgments, which were collected for all measures, were entered into the PCA. We tested 2-factor, 3-factor, 4-factor, 5-factor, and 6-factor varimax rotated solutions. To fit all variables, we sought to ensure that communalities (proportion of variance explained for each variable) were high in all cases.

To examine principal component scores for each scenario condition, we performed mixed effects analyses using the *lme4* package in R (Bates et al., 2015). Separate linear mixed effects models were fit to predict each of the five component scores (extracted for the *final* timepoint

only). Models specified fixed effects as: A + B + AxB, where A = Initial Condition (*tragic*, *taboo*), and B = Reframing Condition (*reframed*, *control*). Random effects parameters were chosen by first fitting all necessary by-scenario random slopes and intercepts (Barr, Levy, Scheepers, & Tily, 2013), then removing random effects components that showed near-zero variance in an uncorrelated model until convergence could be achieved. *P*-values for fixed effects were obtained via Satterthwaite's degrees of freedom method in the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017).

Data Sharing. All data and code to reproduce analyses are available on OSF (https://osf.io/sr9cn/?view_only=845a79ec0b68438e8bfce57be98ba0e9).

Results

Feature Ratings by Condition. Across cohorts, we collected ten feature ratings of our scenarios: moral judgment, mental state inference, mental imagery, valence, arousal, impression violation, belief violation, desire violation, prescriptive norm violation, and descriptive frequency. Table 1 reports means and standard errors for: *initial* judgments of scenarios (when they are first cast as *tragic* or *taboo*), and *final* judgments of *tragic–taboo*, *tragic–control*, *taboo-tragic*, and *taboo-control* scenarios.

Table 1. Feature ratings by scenario condition and timepoint.

		Initial Tragic	Initial Taboo	Tragic–Taboo	Tragic–Control	Taboo–Tragic	Taboo–Control
Measure	**Text**	***M (SE)***	***M (SE)***	***M (SE)***	***M (SE)***	***M (SE)***	***M (SE)***
Moral Judgment (*N* = 315)	"Are <agent>'s actions moral?" 1 (not at all) – 7 (completely)	4.083 (0.137)	2.617 (0.144)	3.303 (0.189)	4.116 (0.191)	4.011 (0.178)	2.786 (0.199)
Mental State Inference (*N* = 80)	"To what extent did this story make you think about someone's experiences, thoughts, beliefs, and/or desires?" 1 (very little) – 7 (very much)	5.381 (0.055)	5.309 (0.062)	5.364 (0.090)	3.807 (0.109)	5.405 (0.093)	3.798 (0.110)
Mental Imagery (*N* = 80)	"To what extent did you picture or imagine the events of the story happening as you read?" 1 (very little) – 7 (very much)	5.471 (0.059)	5.399 (0.050)	5.041 (0.078)	4.529 (0.113)	5.221 (0.080)	4.747 (0.110)
Valence (*N* = 79)	"Please rate your feelings regarding this statement on two scales:" 1 (no unpleasant feelings) – 8 (strong unpleasant feelings) 1 (no pleasant feelings) – 8 (strong pleasant feelings) *Valence indexed as difference of unipolar scales*	-1.861 (0.278)	-3.611 (0.263)	-2.890 (0.413)	-1.635 (0.373)	-1.283 (0.407)	-3.035 (0.331)
Arousal (*N* = 79)	"Please rate your feelings regarding this statement on two scales:" 1 (no unpleasant feelings) – 8 (strong unpleasant feelings) 1 (no pleasant feelings) – 8 (strong pleasant feelings) *Arousal indexed as sum of unipolar scales*	8.137 (0.094)	8.352 (0.067)	8.374 (0.117)	7.821 (0.159)	8.324 (0.126)	8.100 (0.101)
Impression Violation (*N* = 63)	"Is this new information **inconsistent** with your previous impression of <agent>?" 1 (not at all) – 7 (very much so)	na	na	3.790 (0.166)	2.126 (0.069)	4.096 (0.162)	2.129 (0.076)
Belief Violation (*N* = 61)	"Is this new information **inconsistent** with what you previously thought <agent> believed?" 1 (not at all) – 7 (very much so)	na	na	4.082 (0.171)	2.198 (0.076)	4.453 (0.154)	1.974 (0.096)
Desire Violation (*N* = 67)	"Is this new information **inconsistent** with what you previously thought <agent> desired?" 1 (not at all) – 7 (very much so)	na	na	3.823 (0.140)	2.359 (0.101)	3.921 (0.167)	2.345 (0.113)
Prescriptive Norm Violation (*N* = 62)	"**With this new information in mind,** to what extent is <agent>'s decision deviant (i.e., to what extent does it go against the norms of our society)?"	na	na	4.234 (0.182)	3.080 (0.197)	3.225 (0.112)	4.084 (0.191)

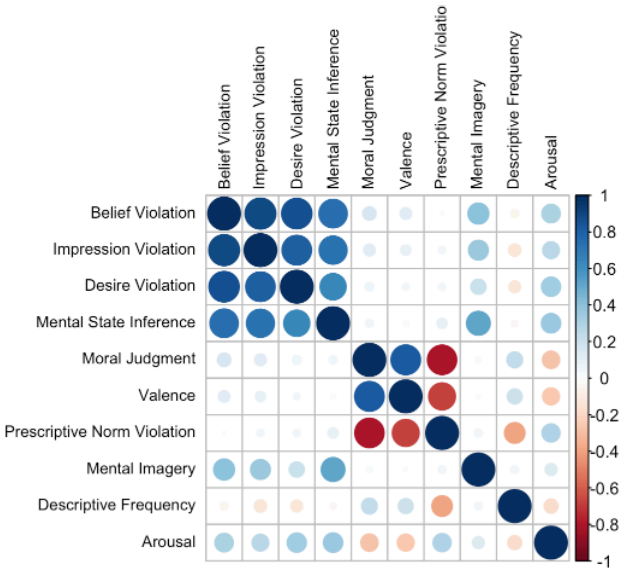
	1 (not at all) – 7 (very much so)						
Descriptive Frequency (*N* = 62)	"**With this new information in mind,** to what extent is <agent>'s decision common (i.e., to what extent is it frequently observed in our society)?" 1 (not at all) – 7 (very much so)	na	na	4.145 (0.116)	4.217 (0.121)	4.169 (0.112)	4.111 (0.134)

Principal Component Analysis. A PCA reduced the dimensionality of our measures. A 2-factor solution was a poor fit for descriptive frequency (*h2* = .21). A 3-factor solution offered an improvement, but remained a poor fit for arousal (*h2* = .38). A 4-factor solution was a reasonable fit for the data, with the worst fit being for arousal (*h2* = .58); however, communalities were high across all variables in the 5- factor solution ($h2_{\min}$ = .80), and a 6-factor solution offered modest improvement ($h2_{\min}$ = .85). Based on this, we decided to use the 5-factor solution (Table 2). Factors were interpreted as: 1) *dispositional prediction error* (belief violation, impression violation, desire violation, and mental state inference); 2) *prescriptive prediction error* (moral judgment, valence, and prescriptive norm violation; factor loadings were reverse scored); 3) *mental imagery*; 4) *descriptive prediction error* (descriptive frequency, reverse scored); and 5) *arousal*. Factor loadings and variance explained are reported in Table 2, and correlations among measures are visualized in Figure 2.

Table 2. Principal component analysis; 5-factor solution. Factor loadings with an absolute value <.2 are omitted for ease of interpretation. Principal components were varimax rotated.

Measure	**Factor 1 ("dispositional prediction error")**	**Factor 2 ("prescriptive prediction error")**	**Factor 3 ("mental imagery")**	**Factor 4 ("descriptive prediction error")**	**Factor 5 ("arousal")**	***h2***
Belief Violation	0.94	0.09	0.18	-0.01	0.11	0.93
Impression Violation	0.93	0.05	0.15	-0.10	0.02	0.91
Desire Violation	0.93	0.01	-0.06	-0.05	0.13	0.90
Mental State Inference	0.75	-0.04	0.44	0.04	0.18	0.80
Moral Judgment	0.09	0.95	0.03	0.05	-0.10	0.92
Valence	0.05	0.91	0.01	-0.01	-0.09	0.84
Prescriptive Norm Violation	0.05	-0.87	0.06	-0.28	0.08	0.84
Mental Imagery	0.24	-0.01	0.95	0.03	0.03	0.96
Descriptive Frequency	-0.07	0.19	0.04	0.97	-0.06	0.98
Arousal	0.24	-0.22	0.05	-0.07	0.94	1.00
Component loading	3.33	2.57	1.16	1.04	0.98	
Proportional variance	0.33	0.26	0.12	0.10	0.10	
Cumulative variance	0.33	0.59	0.71	0.81	0.91	

Figure 2. Correlations among feature ratings. Dimension reduction was achieved using PCA; the 5-factor solution fit all variables well and was clearly interpretable. Factor 1 was interpreted as *dispositional prediction error*; factor 2 as *prescriptive prediction error* (reverse scored); factor 3 as *mental imagery*; factor 4 as *descriptive prediction error* (reverse scored); and factor 5 as *arousal*.



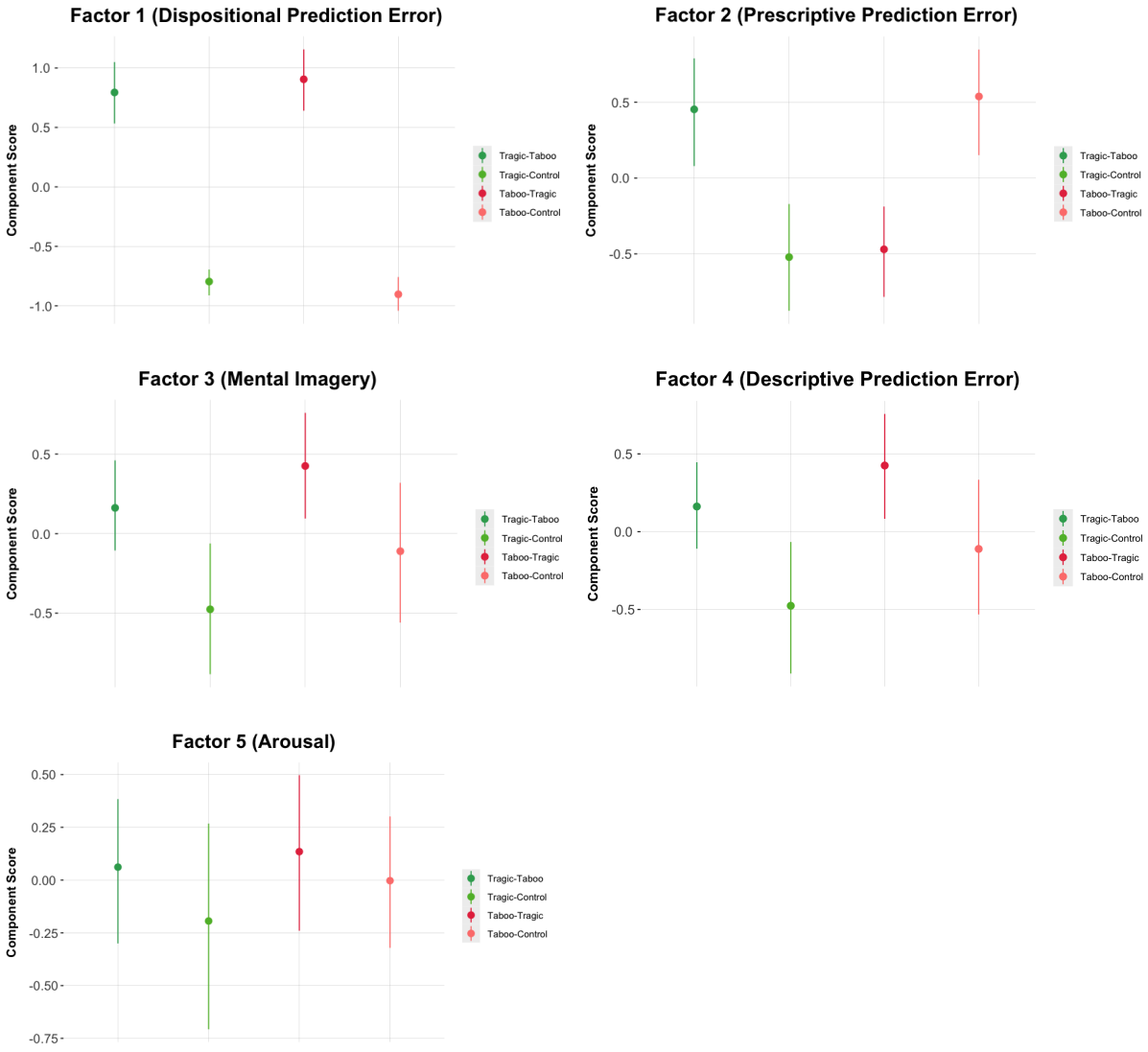
Component Scores by Condition. Linear mixed effects models were fit to predict each of the five components, including as fixed effects: Initial Condition (*tragic*, *taboo*), Reframing Condition (*reframed*, *control*), and their interaction (see Supplementary Table 1 for condition means). *P*-values for main effects and interactions were adjusted using the Holm-Bonferroni method (15 total comparisons).

We found that for Factor 1 (dispositional prediction error) and Factor 3 (mental imagery), there was a main effect of *reframed* > *control* scenarios (dispositional prediction error: Estimate = 1.70, SE = 0.11, t(24.96) = 15.28, $p < 0.0001$; mental imagery: Estimate = 0.59, SE = 0.17, t(69) = 3.52, p = 0.010; Figure 3).

For Factor 2 (prescriptive prediction error), there was a significant two-way interaction between Initial Condition and Reframing Condition (Estimate = 1.98, SE = 0.17, t(69) = 11.64, $p < 0.0001$; Figure 3), and no main effects. Follow-up tests revealed the following significant contrasts (Tukey-adjusted for comparing a family of 4 estimates): Tragic–Taboo > Taboo–Tragic (t(69) = 7.66, $p < 0.0001$); Taboo–Control > Tragic–Control (t(69) = 8.79, $p < 0.0001$); Tragic–Taboo > Tragic–Control (t(69) = 8.09, $p < 0.0001$); and Taboo–Control > Taboo–Tragic (t(69) = 8.37, $p < 0.0001$). Thus, there was relatively more prescriptive prediction error when *taboo* information reframed initially *tragic* information, or when *control* information followed initially *taboo* information. That is, prescriptive prediction error was greater when the most recent available motive for the agent's decision was selfish.

For Factor 4 (descriptive prediction error) and Factor 5 (arousal), no main effects or interactions were significant.

Figure 3. Component scores by scenario condition.



Discussion

Study 1 identified the dimensions underlying our stimulus set. PCA distinguished between dispositional, prescriptive, and descriptive prediction error (Figure 2; Table 2). Mental imagery and arousal comprised additional dimensions. We found that *reframed* scenarios, compared to *control* scenarios, appeared to elicit greater dispositional prediction error and mental imagery. We relate these measures to ToMN activity in Study 2.

Study 2

In Study 2, we examined how the Theory of Mind Network (ToMN) responds to social information that may prompt observers to update their initial impressions and moral judgments (i.e., information that may elicit social prediction error.) Participants read the scenarios analyzed in Study 1 while undergoing fMRI. Four ToM regions of interest were identified in each participant using an independent functional localizer: dorsomedial prefrontal cortex (DMPFC),

right/left temporoparietal junction (RTPJ/LTPJ), and precuneus (PC). We tested whether neural activity in the ToMN differs as a function of: (1) the initial moral valence of an agent's decision (either a relatively moral decision in a tragic dilemma, or a selfish decision in a taboo dilemma); (2) the nature of additional information (either a selfish motive behind a seemingly moral decision, a moral motive behind a seemingly selfish decision, or morally irrelevant information); (3) component scores from Study 1 (interpreted as dispositional, prescriptive, & descriptive prediction error, mental imagery, and arousal); and (4) participants' moral judgments.

Method

Sample size. The sample size ($N \approx 20$) was chosen in advance to be consistent with fMRI studies of social cognition at the time of data collection, 2012–2013 (e.g., Fourie et al., 2014: $N = 22$; Koster-Hale et al., 2013: $N_{S1} = 23$, $N_{S2} = 16$, $N_{S3} = 14$, $N_{S4} = 16$; Mende-Siedlecki et al., 2013: $N = 24$; Ratner et al., 2012: $N = 17$; Young & Saxe, 2011: $N = 17$). We note that participants in the current sample were tested on another task during the same session (involving evaluating claims about facts, morals, and preferences); analyses of these data have been reported in two publications (Theriault, Waytz, Heiphetz, & Young, 2017; Theirault, Waytz, Heiphetz, & Young, 2020). As our sample size is small compared to that of more recent studies in social neuroscience, and the study was not preregistered, the reported results should all be interpreted with caution due to lower statistical power.

Participants. Participants were a community sample recruited through an online posting. The final sample consisted of 21 participants (10 female, 9 male, 2 unspecified; $M_{age} = 27.4$ years, $SD_{age} = 4.7$ years), after excluding one participant from analysis due to excessive movement (see **Neural Data Exclusion**). Participants were right-handed, native English speakers, and had no reported history of learning disabilities, psychiatric or neurological disorders, or drug or alcohol abuse. The study was approved by the Boston College Institutional Review Board; all participants provided informed consent and were compensated $65.

Stimuli and Measures. Stimuli were identical to those used in Study 1. Each participant read 24 scenarios (6 *tragic–taboo*; 6 *tragic–control*; 6 *taboo–tragic*; 6 *taboo–control*), presented in a semi-random order; scenario–condition combinations were counterbalanced across participants. Participants were told that they would read about different characters and provide a moral judgment at the end of each story, and were encouraged to keep their judgment in mind as they read. At the end of each scenario, participants provided a single moral judgment using a button box ("How morally wrong?" 1 ("not at all") – 4 ("very")). We collected one final judgment (as opposed to both initial and final judgments) to avoid interrupting participants as they read.

In-scanner Presentation. Stimuli were presented in white text on a black background, using Matlab 7.7.0 (R2008b) on a Macbook Pro. Four scenarios (one of each condition) were presented in each of six runs. Scenarios appeared cumulatively in five segments (10 s each, 50 s total). Moral judgments were probed on a separate screen (4 s), followed by fixation (12 s). Runs were 4.6 minutes each, and total scan time was 64.6 minutes, due to the inclusion of a second study reported elsewhere (Theriault et al., 2017; Theriault et al., 2020).

MRI Data Acquisition and Processing. The MRI data were collected using a 12-channel head coil in a 3.0 T Siemens Tim Trio scanner at the Center for Brain Science Neuroimaging Facility at Harvard University. Thirty-six slices with 3 mm isotropic voxels, with a 0.54 mm gap between slices to allow for full brain coverage, were collected using gradient-echo planar imaging (TR = 2000 ms, TE = 30 ms, FA = 90°, FOV = 216 x 216 mm;

interleaved acquisition). Anatomical data were collected with T1-weighted MEMPRAGE sequences (1 mm isotropic voxels, 0.5 mm gap between slices, TR = 2530 ms, TE = 1.64 ms, FA = 7°, FOV = 256 x 256 mm). Data processing and analysis were performed using fMRIPrep (Esteban et al., 2019), SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/), and R (R Core Team). 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 = 8 mm kernel), and high-pass filtered (128 s).

Neural Data Exclusion. Individual functional runs were removed from further analysis if > 20% of timepoints in the run exhibited > 1 mm framewise displacement, or if any one timepoint exhibited > 3 mm movement. Participants were excluded if more than ⅓ of functional runs were dropped. These criteria resulted in the exclusion of one participant.

ToMN Localizer. An independent localizer task (Dodell-Feder et al., 2011) was used to functionally define four ToM regions of interest (ROIs) in each participant: DMPFC ($N = 15$), RTPJ ($N = 21$), LTPJ ($N = 21$), and PC ($N = 20$). ROIs were defined as all voxels within a 9-mm radius of the peak voxel that passed threshold in the contrast "*false belief stories > false photo stories*" ($p < 0.001$, uncorrected; $k > 16$, computed via 1000 iterations of a Monte-Carlo simulation, Slotnick et al., 2003). We used the same ROI selection parameters as previous neuroimaging research examining the ToMN (Tsoi et al., 2018; Dungan & Young, 2019).

PSC Calculation. Within each ROI, the percent signal change (PSC) relative to baseline was calculated for each timepoint for each condition, spatially averaging across all voxels in the ROI. Baseline response, calculated separately for each run, was the average response to fixation. PSC for each timepoint for each condition was calculated as: 100[(average response for condition at time *t* - baseline)/baseline]. Timepoints that exhibited >1 mm framewise displacement were removed prior to further analysis.

PSC values were temporally averaged within each of four timewindows (offset by 4 s from presentation time to adjust for hemodynamic lag; Dodell-Feder et al., 2011): *background segment* (4–24 s, before scenarios diverged into initially tragic or taboo); *initial segment* (24–44 s, after scenarios diverged into initially tragic or taboo); *final segment* (44–54 s, after scenarios were reframed and diverged into four conditions); and *response segment* (54–58 s, participant's moral judgment). Analyses focused on two time windows: the *initial segment*, where we examined responses to initially tragic vs. taboo scenarios, and the *final segment*, where we examined responses to reframed vs. control, tragic vs. taboo scenarios.

Condition-based Analyses. Mixed effects analyses were performed using the *lme4* package in R (Bates et al., 2015; see Supplementary Materials for model specifications in R). A linear mixed effects model was fit to predict PSC in the ToMN during the *initial segment*, including as fixed effects: Initial Condition (*tragic*, *taboo*), ROI (*DMPFC*, *RTPJ*, *LTPJ, PC*), and their interaction. A separate model was fit to predict PSC during the *final segment*, including as fixed effects: Initial Condition, Reframing Condition (*reframed*, *control*), ROI, PSC during the initial segment, and all interactions between Initial Condition, Reframing Condition, and ROI.

Random effects parameters were chosen by first including: (1) by-subject and by-scenario random intercepts; (2) by-subject and by-scenario random slopes for all main effects; and (3) by-subject and by-scenario random slopes for significant interaction effects. Then, we removed random effects components that showed near-zero variance in an uncorrelated model until convergence could be achieved. We took this approach because fitting maximal models was prohibitively computationally intensive. *P*-values for fixed effects were obtained via Satterthwaite's degrees of freedom method in the *lmerTest* package (Kuznetsova, Brockhoff, &

Christensen, 2017). To obtain *p*-values for interactions with ROI, we conducted likelihood ratio tests of the full model against a model omitting ROI.

Correlation Analyses. We examined the relationship between Study 1 component scores and ToMN activity during the final segment (when either reframing or control information was appended to scenarios). Linear mixed effects models were fit to predict PSC during the *final segment*, including as fixed effects: one of the five components (dispositional prediction error, prescriptive prediction error, descriptive prediction error, mental imagery, and arousal), ROI, their interaction, and PSC during the initial segment.

We also examined the relationship between ToMN activity during the final segment and in-scanner moral judgments. A linear mixed effects model was fit to predict moral judgments, including as fixed effects: PSC during the final segment, ROI, their interaction, and PSC during the initial segment. Random effects structures for correlation models were reduced as described above for condition-based analyses.

Data Sharing. All PSC data and code to reproduce analyses are available on OSF (https://osf.io/sr9cn/?view_only=845a79ec0b68438e8bfce57be98ba0e9).

Results

PSC for initial segment. Condition means for each ROI are visualized in Figure 4. There was no main effect of Initial Condition (*tragic*, *taboo*) in the ToMN (Estimate = 0.0003, SE= 0.013, t(1575) = 0.024, p = 0.981), and no interaction with ROI (likelihood ratio test, Chisq(3) = 0.671, p = 0.880). Thus, initially tragic vs. taboo information did not elicit differences in ToMN activity.

PSC for final segment. We observed a main effect of Reframing Condition, such that there was greater ToMN activity for *reframed* vs. *control* scenarios (Estimate = 0.066, SE = 0.015, t(1650) = 4.447, p < 0.0001); this effect was qualified by an interaction with ROI (Chisq(3) = 19.035, p = 0.0003).

There was a marginal main effect of Initial Condition (*tragic* > *taboo*; Estimate = 0.027, SE = 0.015, t(1644) = 1.843, p = 0.066), and a marginal 2-way interaction between Initial Condition and Reframing Condition (Estimate = 0.049, SE = 0.029, t(1645) = 1.655, p = 0.098). Follow-up comparisons revealed marginally greater ToMN activity for *tragic–taboo* vs. *taboo–tragic* scenarios, but no difference in ToMN activity for *tragic–control* vs. *taboo–control* scenarios (for full statistics see Supplementary Materials).

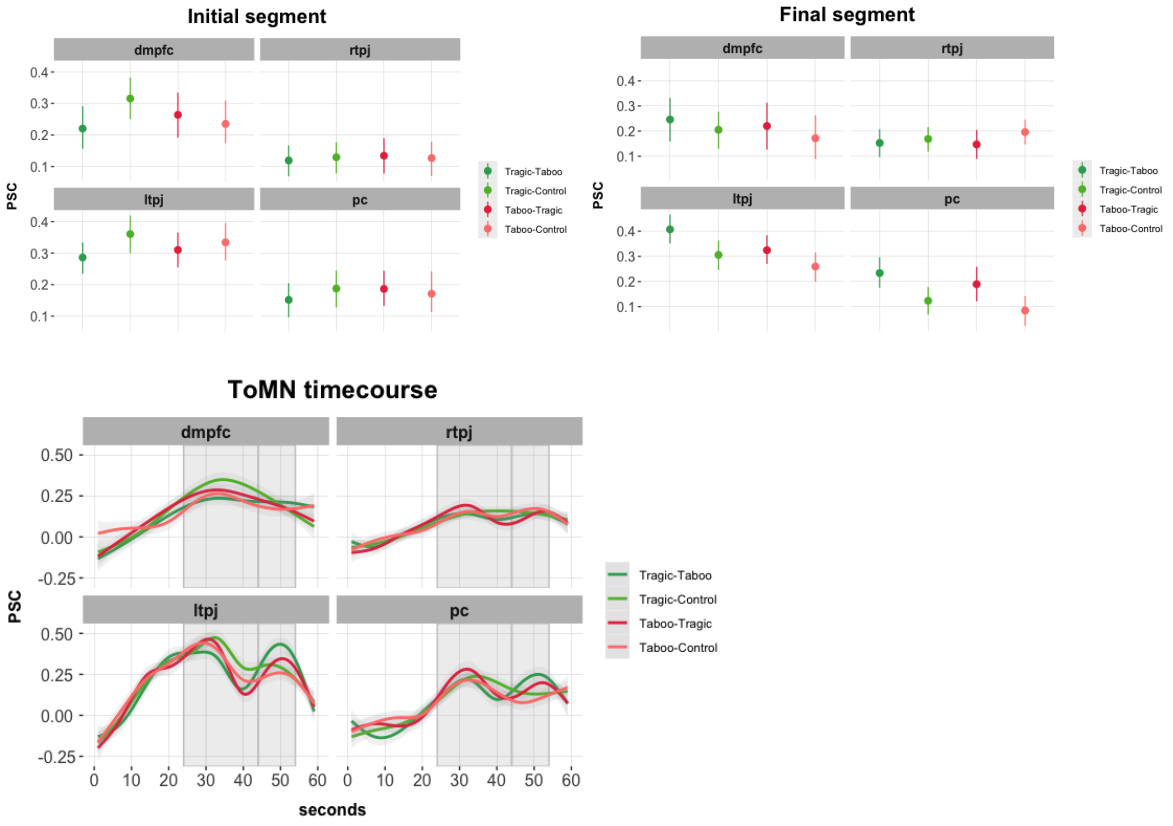
We next fit models to predict PSC in each individual ROI, including as fixed effects: Initial Condition, Reframing Condition, their interaction, and PSC for the initial segment. *P*-values for main effects and interactions were adjusted using the Holm-Bonferroni method (12 total comparisons). In LTPJ and PC, there was a main effect of Reframing Condition, such that there was greater activity for *reframed* vs. *control* scenarios (see Table 3 for statistics). Additionally in LTPJ, there was a marginal main effect of Initial Condition, such that there was greater activity for the final segments of initially tragic (vs. taboo) scenarios (Table 3). There were no other main effects or 2-way interactions in individual ROIs.

In sum, the ToMN exhibited greater activity when additional information revealed an unexpected motive behind an agent's initial decision, driven by activity in LTPJ and PC. In contrast, ToMN activity during both the initial and final segments was largely unmodulated by whether the scenario started out tragic vs. taboo.

Table 3. Effects of Initial Condition, Reframing Condition, and their interaction on neural activity during the final segment. P-values for within-ROI effects were adjusted using the Holm-Bonferroni method. For condition-wise means see Supplementary Table 2.

	DMPFC	**RTPJ**	**LTPJ**	**PC**	**ToMN**
Initial Condition (*tragic > taboo*)	Estimate = 0.020, SE = 0.037, t(304.11) = 0.550, p = 1.000	Estimate = -0.007, SE = 0.023, t(395.71) = -0.308, p= 1.000	Estimate = 0.066, SE = 0.025, t(411.29) = 2.597, p = 0.097	Estimate = 0.045, SE = 0.028, t(372.53) = 1.623, p = 0.928	Estimate = 0.027, SE = 0.015, t(1644) = 1.843, p = 0.066
Reframing Condition (*reframed > control*)	Estimate = 0.061, SE = 0.037, t(305.88) = 1.635, p = 0.928	Estimate = -0.029, SE = 0.031, t(19.09) = -0.937, p = 1.000	Estimate = 0.102, SE = 0.026, t(413.14) = 3.972, p = 0.001	Estimate = 0.114, SE = 0.032, t(15.90) = 3.574, p = 0.028	Estimate = 0.066, SE = 0.015, t(1650) = 4.447, p < 0.0001
Initial Condition X Reframing Condition	Estimate = 0.054, SE = 0.074, t(305.21) = 0.726, p = 1.000	Estimate = 0.035, SE = 0.046, t(391.54) = 0.763, p = 1.000	Estimate = 0.050, SE = 0.051, t(417.11) = 0.982, p = 1.000	Estimate = 0.016, SE = 0.056, t(373.76) = 0.280, p = 1.000	Estimate = 0.049, SE = 0.029, t(1645) = 1.655, p = 0.098

Figure 4. Top: Percent signal change in the Theory of Mind Network, by scenario condition and timepoint. Bottom: Timecourse of percent signal change. Shaded rectangles indicate the two timewindows that were analyzed: *initial segment* (24–44 s) and *final segment* (44–54 s).



Behavioral Component Score–ToM Activity Analysis. We tested the relationship between component scores, derived in Study 1, and ToMN activity during the final segment (collapsing across scenario conditions). Mixed effects models were fit to PSC for the final segment and included as fixed effects: one of the five components, ROI, their interaction, and $PSC_{initial}$. For each component, we also tested whether it predicted activity within each ROI (*p*-values were adjusted for 4 comparisons using the Holm-Bonferroni method).

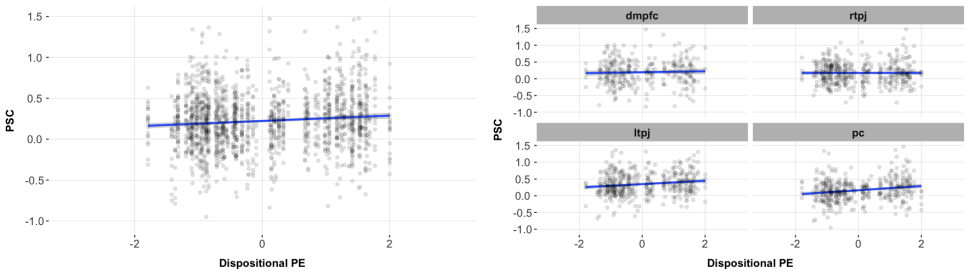
We observed a significant association between PSC_{final} and dispositional prediction error (Estimate = 0.038, SE = 0.007, t(1648) = 5.175, p < 0.0001), such that greater ToMN activity predicted greater dispositional prediction error. There was also a significant interaction with ROI (Chisq(3) = 16.091, p = 0.001).

Looking within each ROI, we found that in LTPJ and PC, activity during the final segment was significantly associated with greater dispositional prediction error (LTPJ: Estimate = 0.056, SE = 0.013, t(396.64) = 4.44, p < 0.0001; PC: Estimate = 0.057, SE = 0.015, t(22.13) = 3.83, p = 0.004; Figure 5). There was a marginal association in DMPFC (Estimate = 0.036, SE = 0.018, t(303.83) = 2.03, p = 0.087) and no such association in RTPJ (Estimate = -0.007, SE = 0.014, t(20.22) = -0.533, p = 0.600). Thus, new social information that was inconsistent with specific impressions of the agent elicited greater activity in the ToMN, principally within LTPJ and PC.

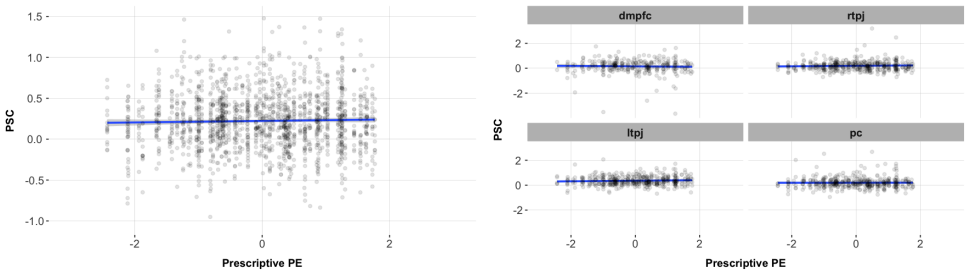
In contrast, activity in ToMN and in each ROI was not significantly associated with prescriptive prediction error, descriptive prediction error, mental imagery, or arousal (Figure 5; Supplementary Table 3).

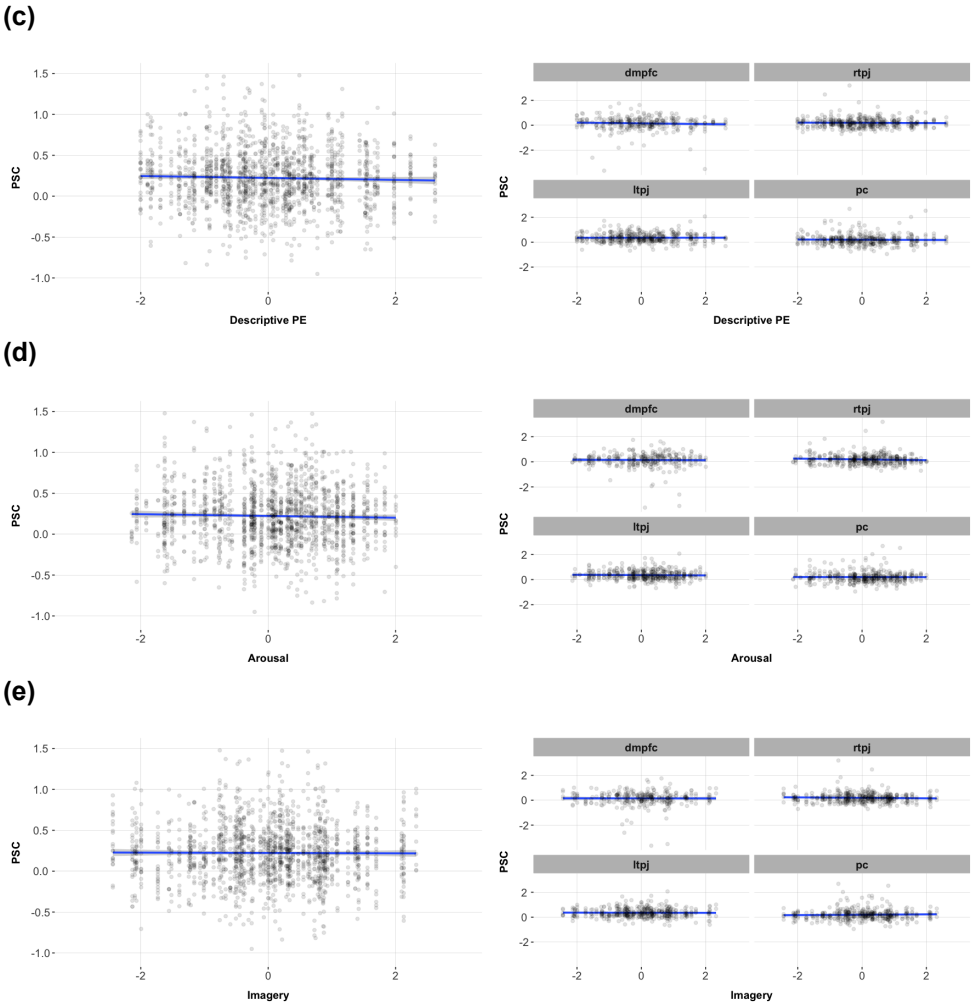
Figure 5. Relationships between behavioral component scores (derived from a separate set of participants) and ToMN activity. Main effects of ToMN activity (**left**) and their breakdown by ROI (**right**). **(a)** Dispositional prediction error was related to activity in the ToMN overall, and in LTPJ and PC. **(b–e)** Prescriptive prediction error, descriptive prediction error, mental imagery, and arousal showed no relationship with ToMN activity.

(a)



(b)





Mediation Analysis. We performed a mediation analysis to examine whether increased ToMN activity for *tragic–taboo* and *taboo–tragic* scenarios was driven by greater dispositional prediction error. Bayesian multilevel models (brms package; Bürkner, 2017) tested whether the magnitude of dispositional prediction error mediates the relationship between Reframing Condition and ToM activity. Default, uninformative priors were used, and all R_{hat} values were < = 1.01, suggesting that the models had converged.

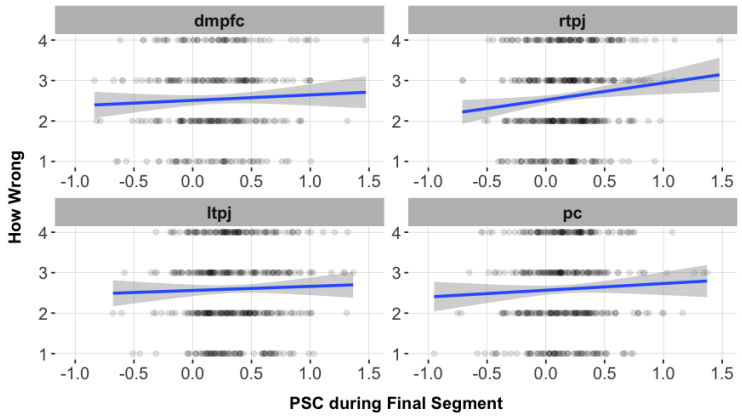
We found credible evidence for an indirect effect of Reframing Condition on ToMN activity via dispositional prediction error. The mean estimated indirect effect was b = 0.208, Bayesian Credible Interval = [0.034, 0.375]; the total effect was b = 0.214 [0.033, 0.386], and the direct effect was b = 0.008 [-0.226, 0.258].

Moral Judgment–ToMN Activity Analysis. Mixed effects models were fit to participants' moral judgments (Figure 6) and included PSC_{final}, ROI, their interaction, and $PSC_{initial}$ as fixed effects. We observed a significant relationship between PSC_{final} and moral judgments, such that greater ToMN activity predicted greater moral wrongness judgments (Estimate = 0.183, SE = 0.086, t(1611) = 2.131, p = 0.033).

Looking within each ROI, activity in only RTPJ predicted moral wrongness (Estimate = 0.399, SE = 0.163, t(422.83) = 2.448, p = 0.059). This effect is marginally significant after correcting for multiple comparisons and should be interpreted with caution. However, we report

it here because it is consistent with an extensive literature showing a role for RTPJ in incorporating inferences of negative intent into moral judgments (for review see Anzellotti & Young, 2019).

Figure 6. Relationships between neural activity in each ROI and in-scanner moral judgments. Increased RTPJ activity is (tentatively) associated with increased wrongness judgments.



General Discussion

In the present work, we tested the relationship between ToMN (Theory of Mind Network) activity and several forms of social prediction error. We found that ToMN activity was associated with dispositional prediction error – the degree to which new information contradicted specific impressions of the agent – but not with prescriptive or descriptive prediction error (Figure 5). Measures of prediction error were aggregated component scores (extracted from a principal component analysis in Study 1). These findings are broadly consistent with the hypothesis that the ToMN is associated with social prediction error (Koster-Hale & Saxe, 2013; Joiner et al., 2017).

Task-specific predictive coding

Our results suggest that ToMN activity is related to a specific type of social prediction error, at least in the context of moral dilemmas. The component score for dispositional prediction error was derived from four features: whether the new information evoked mental state inference, and whether the new information was inconsistent with (1) the agent's desires, (2) the agent's beliefs, and (3) previous impressions of the agent. Thus, dispositional prediction error in the present study was largely dependent on inferences formed in the context of the experiment (i.e., the agent's mental states and traits given the first half of each story). On the other hand, prescriptive prediction error and descriptive prediction error in the present study were largely dependent on expectations formed outside of the experimental setting – what people *ought* to do and what people *typically* do (Brauer & Chaurand, 2010). Our results suggest that these three forms of social prediction error may be distinguished by their relationship to theory of mind (at least in the current context). Descriptive prediction error and prescriptive prediction error can conceivably arise in the absence of mentalizing: descriptive prediction error may arise from discrepancies between the current observation and expectations based on statistical knowledge of people's behaviors, and prescriptive prediction error may arise from discrepancies between the current observation and expectations based on internalized rules about morality. Thus, one possibility is that participants in the current study were more likely to engage cognitive processes

related to theory of mind when thinking about what the *agent* will do, and less likely to do so when thinking about what *people in general* will do. We expect that in contexts where preexisting (non-experimental) priors are more important for the task at hand, observers will engage in more mentalizing about *people in general*. Indeed, in a recent fMRI study examining moral statements, factual statements, and preference statements, we found a significant association between the subjectivity of moral statements and ToMN activity, such that low-consensus moral statements (e.g., *It is unethical for businesses to promote sugary products to children*) elicited more ToMN activity than high-consensus moral statements (e.g., *It is irresponsible for airlines to risk the safety of their passengers*). Here, social consensus was a key determinant of predictability and thus ToMN activity. Thus, across two studies, we found that the ToMN can track both dispositional prediction error and consensus information; overall, our findings are broadly consistent with a predictive account of neural activity, and suggest that the type of prediction error being tracked in the social brain will be task-dependent.

Recent work supports a flexible account of ToMN function. McManus and colleagues (2023) examined whether prediction error signals in the ToMN depend on (1) the type of social information used to make predictions (mental states vs. behaviors), and (2) the type of social information that confirms or violates those predictions (mental states vs. behaviors). There was an expectancy effect in DMPFC, LTPJ, and RTPJ (greater activity to unexpected vs. expected information), but this effect did not vary based on prior information type or outcome information type. That is, ToM regions appear to flexibly support the formation and monitoring of multimodal social predictions, at least when those predictions pertain to specific individuals. We speculate that in most task contexts, predictions about specific people will dominate; predictions about people in general may come into play in the absence of identifiable targets, or if predictions about a target rely specifically on social norm information.

Scope of mentalizing

In McManus and colleagues' paradigm, both expectation-confirming and expectation-violating information were directly *relevant* to the initial information that formed the basis of the expectation (e.g., Mr. Johnson is expected to go to the beach, but he goes to the movies instead). In the current study, reframing information provided an alternative, counter-valenced motive behind the agent's initial action, but control information was *irrelevant* to the initial action. The latter was not necessarily expectation-confirming; rather, it was not expectation-violating. Mediation analyses suggested that heightened ToMN responses to *reframed* (vs. *control*) scenarios can be largely explained by differences in dispositional prediction error. This gives confidence that our experimental manipulation worked as intended, and that our feature ratings are an appropriate index of the implicit experience of prediction error. The mediation result also suggests that mentalizing in the present paradigm may have had a circumscribed scope: the agent's decision in the moral dilemma. When additional information that was irrelevant to the decision was presented (e.g., Gregory drove to his parents' house), participants did not view it as inconsistent with their prior impressions, likely because they had not made any relevant mental state or trait inferences. An open question is what would happen if the additional information violates impressions without reframing moral judgment – for instance, if Gregory the fisherman was afraid of swimming – we expect that we will still see increased ToMN activity in these cases, as initial impressions can likely accommodate identity- or profession-related expectations.

Valence asymmetries in reframing and updating paradigms

We found that neural activity did not differ as a function of the direction of moral updating (from moral to immoral vs. from immoral to moral). This can be contrasted with past neuroimaging work on moral impression updating, which typically finds a negativity bias (greater neural activity when immoral information follows moral information; Mende-Siedlecki et al., 2013; Kim et al., 2021). The present paradigm, wherein observers learned about agents' decisions in a moral dilemma, then came across a previously hidden motive behind the agent's original decision, is qualitatively distinct from typical impression updating paradigms. In typical updating paradigms, targets are often described as performing a series of distinct behaviors that is internally consistent or inconsistent. In contrast, in our *reframing* paradigm, targets are described as engaging in one central behavior (a seemingly selfless decision or a seemingly selfish decision in a moral dilemma); impression updating occurs because a new motivation behind the central behavior gets revealed (a hidden economic incentive or a hidden prosocial outcome). These paradigm differences are important for understanding differences between the present findings and findings from past work. In a recent behavioral study employing the same paradigm as the current study (Kim, Theriault, Hirschfeld-Kroen, & Young, 2022), we found an unexpected positivity bias, where there was greater moral impression updating when scenarios were reframed from immoral to moral (vs. from moral to immoral). This positivity bias was partly explained by the extent to which reframing information elicits external (vs. internal) causal attributions. We speculated that the inherent difficulty of reframed tragic dilemmas moved participants toward a state of greater uncertainty, which destabilized their initial evaluation, and prompted a consideration of the broader context. In the current work, there was no valence effect on ToMN activity. While we did not observe a positivity bias, it is still notable that the negativity bias was not present; we hypothesize that reframed tragic dilemmas and reframed taboo dilemmas equally prompted participants to consider the broader context, and in future work we hope to examine causal attribution ratings and relate them to both neural activity and impression updating (which cannot be computed for the present paradigm, as only final moral ratings were collected).

Mental state representations and social prediction

The ToMN is broadly involved in social cognition, and, based on the present study alone, it is impossible to draw conclusions about the precise relationship between mental state representation and social prediction; for instance, does all mental state representation boil down to prediction? Or is prediction only one part of the broader construct? Despite this open question, it is worth pointing out that we included a measure of mental state representation (used in prior work; Dodell-Feder et al., 2011), and it loaded heavily onto our factor tracking dispositional prediction error (Figure 2; Table 2). Intuitively, this makes sense, as people should be inclined to consider an agent's mental states when that agent does not behave as predicted. In other words, when an entity does not behave according to a transparent system of input and output (i.e. when it acts according to an "intentional stance"; Dennett, 1987; Theriault & Young, 2014; Waytz, Morewedge, Epley, Monteleone, Gao, & Cacioppo, 2010), attributing mental states may help to generate predictions about how the entity will behave. Thus, the predictive utility of both mental state attributions and dispositional inferences may be a common feature underlying the observed ToMN activity.

Conclusion

ToMN activity is associated with social prediction error, consistent with a predictive coding account of social cognition (Koster-Hale & Saxe, 2013). Although future research is required to better characterize the relationship between ToMN activity and social cognition, the

present work is consistent with an account of the brain as a "hierarchical prediction machine" (A. Clark, 2013), which uses information about people and contexts to anticipate its social environment. These social predictions, and the ToMN activity they evoke, appear to be especially sensitive to dispositional information in the context of moral dilemmas.

References

Amodio DM, Frith CD. 2006. Meeting of minds: The medial frontal cortex and social cognition. Nat Rev Neurosci. 7:268–277.

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

Baayen RH, Davidson DJ, Bates DM. 2008. Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang. 59:390–412.

Bach P, Schenke KC. 2017. Predictive social perception: Towards a unifying framework from action observation to person knowledge. Soc Personal Psychol Compass. 11:e12312.

Baron SG, Gobbini MI, Engell AD, Todorov A. 2010. Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions. Soc Cogn Affect Neurosci. 6:572-581.

Barr DJ, Levy R, Scheepers C, Tily HJ. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. J Mem Lang. 68:255–278.

Barrett LF. 2017a. The theory of constructed emotion: an active inference account of interoception and categorization. Soc Cogn Affect Neurosci. 12:1-23.

Barrett LF. 2017b. How emotions are made: The secret life of the brain. Boston (MA): Houghton Mifflin Harcourt.

Bates D, Maechler M, Bolker B, Walker S. 2015. Fitting linear mixed effects models using lme4. J Stat Softw, 67:1–48.

Bhanji JP, Beer JS. 2013. Dissociable neural modulation underlying lasting first impressions, changing your mind for the better, and changing it for the worse. J Neurosci. 33:9337- 9344.

Blakemore SJ, Boyer P, Pachot-Clouard M, Meltzoff A, Segebarth C, Decety J. 2003. The detection of contingency and animacy from simple animations in the human brain. Cereb Cortex. 13:837-844.

Brauer M, Chaurand N. 2010. Descriptive norms, prescriptive norms, and social control: An intercultural comparison of people's reactions to uncivil behaviors. Eur J Soc Psychol. 40:490-499.

Buckner RL. 2012. The serendipitous discovery of the brain's default network. Neuroimage. 62:1137-1145.

Carter RM, Huettel SA. 2013. A nexus model of the temporal–parietal junction. Trends Cogn Sci. 17:328-336.

Cialdini RB, Reno RR, Kallgren CA. 1990. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. J Pers Soc Psychol. 58:1015–1026.

Ciaramidaro A, Adenzato M, Enrici I, Erk S, Pia L, Bara BG, Walter H. 2007. The intentional network: How the brain reads varieties of intentions. Neuropsychologia. 45:3105–3113.

Clark A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci. 36:181–204.

Clark H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. J Verbal Learning Verbal Behav. 12:335–359.

Cloutier J, Gabrieli J. D, O'Young D, Ambady N. 2011. An fMRI study of violations of social expectations: when people are not who we expect them to be. NeuroImage. 57, 583-588.

Critcher CR, Inbar Y, Pizarro DA. 2013. How quick decisions illuminate moral character. Soc Psychol Personal Sci. 4:308-315.

Dennett DC. 1989. The intentional stance. Cambridge (MA): MIT press.

Dodell-Feder D, Koster-Hale J, Bedny M, Saxe R. 2011. fMRI item analysis in a theory of mind task. NeuroImage. 55:705–712.

Dungan J, Stepanovic M, Young L. 2017. Theory of mind for processing unexpected events across contexts. Soc Cogn Affect Neurosci. *11*:1183–1192.

Eklund A, Nichols TE, Knutsson H. 2016. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci U S A. 113:7900–7905.

Fletcher PC, Happé F, Frith U, Baker SC, Dolan RJ, Frackowiak RS, Frith, CD. 1995. Other minds in the brain: A functional imaging study of “theory of mind” in story comprehension. Cognition. 57:109–128.

Fodor JA. 1983. The modularity of mind: An essay on faculty psychology. Cambridge (MA): MIT press.

Fourie MM, Thomas KG, Amodio DM, Warton CM, Meintjes EM. 2014. Neural correlates of experienced moral emotion: an fMRI investigation of emotion in response to prejudice feedback. Soc Neurosci. 9:203-218.

Friston K. 2010. The free-energy principle: a unified brain theory? Nat Rev Neurosci. 11:127-138.

Gallagher HL, Happé F, Brunswick N, Fletcher PC, Frith U, Frith CD. 2000. Reading the mind in cartoons and stories: An fMRI study of ‘theory of mind’ in verbal and nonverbal tasks. Neuropsychologia. 38:11–21.

Gelman A, Hill J, Yajima M. 2012. Why we (usually) don't have to worry about multiple

comparisons. J Res Educ Eff. 5:189-211.

Gilbert DT, Malone PS. 1995. The correspondence bias. Psychol Bull. 117:21–38.

Gobbini MI, Koralek AC, Bryan RE, Montgomery KJ, Haxby JV. 2007. Two takes on the social brain: A comparison of theory of mind tasks. J Cogn Neurosci. 19:1803–1814.

Harris LT, Todorov A, Fiske ST. 2005. Attributions on the brain: Neuro-imaging dispositional inferences, beyond theory of mind. NeuroImage. 28:763–769.

Hassabis D, Maguire EA. 2009. The construction system of the brain. Philos Trans R Soc Lond

Hothorn T, Bretz F, Westfall P. 2008. Simultaneous Inference in General Parametric Models. Biom J. 50:346–363.

Jenkins AC, Mitchell JP. 2010. Mentalizing under uncertainty: Dissociated neural responses to ambiguous and unambiguous mental state inferences. Cereb Cortex. 20:404–410.

Joiner J, Piva M, Turrin C, Chang S. 2017. Social learning through prediction error in the brain. NPJ Sci Learn. 2:8.

Judd CM, Westfall J, Kenny DA. 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. 103:54–69.

Kim M, Mende-Siedlecki P, Anzellotti S, Young L. 2021. Theory of mind following the violation of strong and weak prior beliefs. Cerebral Cortex, 31(2), 884-898.

Kim M, Theriault J, Hirschfeld-Kroen J, Young L. 2022. Reframing of moral dilemmas reveals an unexpected “positivity bias” in updating and attributions. Journal of Experimental Social Psychology, 101, 104310.

Kircher T, Blümel I, Marjoram D, Lataster T, Krabbendam L, Weber J, van Os J, Krach S. 2009.

Koster-Hale J, Saxe R. 2013. Theory of Mind: A Neural Prediction Problem. Neuron. 79:836–848.

Koster-Hale J, Saxe R, Dungan J, Young LL. 2013. Decoding moral judgments from neural

Kron A, Goldstein A, Lee DH-J, Gardhouse K. 2013. How are you feeling? Revisiting the quantification of emotional qualia. Psychol Sci. 24:1503–1511.

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

Lichtenstein S, Gregory R, Irwin J. 2007. What's bad is easy: Taboo values, affect, and cognition. Judgm Decis Mak. 2:169–188.

Ma N, Vandekerckhove M, Van Hoeck N, Van Overwalle F. 2012. Distinct recruitment of temporo-parietal junction and medial prefrontal cortex in behavior understanding and trait identification. Soc Neurosci. 7:591–605.

Mars RB, Neubert FX, Noonan MP, Sallet J, Toni I, Rushworth MF. 2012. On the relationship between the “default mode network” and the “social brain”. Front Hum Neurosci. 6:189.

McManus R, Dungan J, Jiang K, Young L. 2023. How unexpected events are processed in theory of mind regions: A conceptual replication. Social Neuroscience, 18(3), 155-170.

Mende-Siedlecki P, Todorov A. 2016. Neural dissociations between meaningful and mere inconsistency in impression updating. Social Soc Cogn Affect Neurosci. 11:1489-1500.

Mende-Siedlecki P, Baron SG, Todorov A. 2013. Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. J Neurosci. 33:19406- 19415.

Mitchell JP, Banaji MR, Macrae CN. 2005. General and specific contributions of the medial prefrontal cortex to knowledge about mental states. NeuroImage. 28:757–762.

Premack D, Woodruff G. 1978. Does the chimpanzee have a theory of mind? Behav Brain Sci. 1:515-526.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ratner KG, Kaul C, Van Bavel JJ. 2012. Is race erased? Decoding race from patterns of neural activity when skin color is not diagnostic of group boundaries. Soc Cogn Affect Neurosci. 8:750-755.

Ruby P, Decety J. 2003. What you believe versus what you think they believe: A neuroimaging study of conceptual perspective-taking. Eur J Neurosci. 11:2475–2480.

Saxe R, Kanwisher N. 2003. People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind”. NeuroImage. 19:1835–1842.

Saxe R, Powell LJ. 2006. It’s the thought that counts: Specific brain regions for one component of theory of mind. Psychol Sci. 17:692–699.

Schaafsma SM, Pfaff DW, Spunt RP, Adolphs R. 2015. Deconstructing and reconstructing theory of mind. Trends Cogn Sci. 19:65–72.

Schiller D, Freeman JB, Mitchell JP, Uleman JS, Phelps EA. 2009. A neural mechanism of first impressions. Nat Neurosci. 12:508-514.

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.

Schurz M, Tholen MG, Perner J, Mars RB, 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. 38:4788–4805.

Tetlock PE, Kristel OV, Elson SB, Green MC, Lerner JS. 2000. The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. J Pers Soc Psychol. 78:853–870.

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

Theriault J, Waytz A, Heiphetz L, Young L. 2017. Examining overlap in behavioral and neural representations of morals, facts, and preferences. J Exp Psychol Gen. *146*:1586–1605.

Theriault J, Waytz A, Heiphetz L, Young L. 2020. Theory of Mind network activity is associated with metaethical judgment: An item analysis. Neuropsychologia, 143, 107475.

Uhlmann EL, Zhu LL, Tannenbaum D. 2013. When it takes a bad person to do the right thing. Cognition. 126:326-334.

Van Overwalle F. 2009. Social cognition and the brain: A meta-analysis. Hum Brain Mapp. 30:829–858.

Vogeley K, Bussfeld P, Newen A, Herrmann S, Happé F, Falkai P. ... Zilles K. 2001. Mind reading: Neural mechanisms of theory of mind and self-perspective. NeuroImage. 14:170–181.

Waytz A, Morewedge CK, Epley N, Monteleone G, Gao JH, Cacioppo JT. 2010. Making sense by making sentient: effectance motivation increases anthropomorphism. J Pers Soc.

Westfall J, Nichols TE, Yarkoni T. 2016. Fixing the stimulus-as-fixed-effect fallacy in task fMRI. Wellcome Open Res. 1:23.

Young L, Saxe R. 2009. An fMRI investigation of spontaneous mental state inference for moral judgment. J Cogn Neurosci. 21:1396–1405.

Yeshurun Y, Swanson S, Simony E, Chen J, Lazaridi C, Honey CJ, Hasson U. 2017. Same story, different story: the neural representation of interpretive frameworks. Psychol Sci. 28:307- 319.

Young 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 U S A. 104:8235–8240.

Young L, Camprodon J, Hauser M, Pascual-Leone A, Saxe R. 2010. Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. Proc Natl Acad Sci U S A. 107:6753–6758.