Routledge
Taylor & Francis Group

Social Neuroscience

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/psns20

How unexpected events are processed in theory of mind regions: A conceptual replication

Ryan M. McManus, James A Dungan, Kevin Jiang & Liane Young

To cite this article: Ryan M. McManus, James A Dungan, Kevin Jiang & Liane Young (2023) How unexpected events are processed in theory of mind regions: A conceptual replication, Social Neuroscience, 18:3, 155-170, DOI: 10.1080/17470919.2023.2218620

To link to this article: https://doi.org/10.1080/17470919.2023.2218620

SOCIAL NEUROSCIENCE
2023, VOL. 18, NO. 3, 155–170
https://doi.org/10.1080/17470919.2023.2218620

Routledge
Taylor & Francis Group

RESEARCH PAPER

How unexpected events are processed in theory of mind regions: A conceptual replication

Ryan M. McManus[a], James A Dungan[b], Kevin Jiang[a] and Liane Young[a]

[a]Department of Psychology and Neuroscience, Boston College, Boston, MA, USA; [b]Booth School of Business, University of Chicago, Chicago, IL, USA

ABSTRACT

Recent research in social neuroscience has postulated that Theory of Mind (ToM) regions play a role in processing social prediction error (PE: the difference between what was expected and what was observed). Here, we tested whether PE signal depends on the type of prior information people use to make predictions – an agent's prior mental states (e.g., beliefs, desires, preferences) or an agent's prior behavior – as well as the type of information that confirms or violates such predictions. That is, does prior information about mental states (versus behavior) afford stronger predictions about an agent's subsequent mental states or behaviors? Additionally, when information about an agent's prior mental states or behavior is available, is PE signal strongest when information about an agent's subsequent mental state (vs behavior) is revealed? In line with prior research, results suggest that DMPFC, LTPJ, and RTPJ are recruited more for unexpected than expected outcomes. However, PE signal does not seem to discriminate on the basis of prior or outcome information type. These findings suggest that ToM regions may flexibly incorporate any available information to make predictions about, monitor, and perhaps explain, inconsistencies in social agents.

ARTICLE HISTORY
Received 8 December 2022
Revised 16 April 2023
Published online 31 May 2023

KEYWORDS
Theory of mind; prediction error; social neuroscience

Introduction

Consider making the following prediction: "My father is going to pick me up from the airport when I fly home to visit tomorrow." Given this prediction, you may be extremely surprised if he forgets to pick you up. Conversely, if your prediction was that he would forget to pick you up (perhaps because you just phoned him and he did not mention your travel plans), you may be especially surprised when he shows up at the airport right on time. In either case, you would experience social prediction error (i.e., a difference between what was expected and what was observed, within a social context).

There is a variety of information on which people base their predictions of others. This information can include but is not limited to: an agent's prior behavioral history (Dungan et al., 2016; Heil et al., 2019), their prior mental states (Dungan et al., 2016), and descriptive or prescriptive norms (see J. E. Theriault et al., 2021). These sources can be used to predict not only an agent's future behavior, but also their future mental states. For example, you might predict that your father will pick you up (or simply wants to pick you up) from the airport based on: his calling you to talk about your travel plans, explicit knowledge of his desire to pick you up (perhaps because your mother told you this), or the idea that parents tend to pick up their children from the airport. The current paper zeroes in on prior behavior and prior mental states as sources of social prediction, investigating whether the type of information on which people base their predictions, as well as the type of information that is being predicted, affects how surprising an outcome is. Specifically, we examine the brain regions implicated in processing agents' unexpected behavior and mental states as a function of their prior behavior and mental states.

Research on social prediction error has been steadily increasing, due to a seminal review article linking predictive coding models to theory of mind (ToM) tasks in neuroscience (Koster-Hale & Saxe, 2013). Predictive coding, put simply, is the idea that neuronal activity contains not only information about sensory input, but also information about the difference between expected and actual sensory input (e.g., Fiorilla et al., 2003; Wacongne et al., 2012). The crucial idea behind the marriage of predictive coding and ToM is that "most experiments on ToM depend on predictions based on prior expectations and an internal model of human behavior" (Koster-Hale & Saxe, 2013). As in the opening

CONTACT Ryan M. McManus mcmanurd@bc.edu Department of Psychology and Neuroscience, Boston College, McGuinn 300, 140 Commonwealth Ave, Chestnut Hill, Boston, MA 02467, USA

Supplemental data for this article can be accessed online at https://doi.org/10.1080/17470919.2023.2218620

example, you may make predictions about your father's mental states (or behavior) based on his prior mental states (or behavior). In line with Koster-Hale and Saxe's (2013) theorizing, if your predictions are not borne out, your experience of prediction error may be due to ToM regions (i.e., DMPFC, PC, LTPJ, and RTPJ) preferentially responding to information that is inconsistent with your predictions.

Extant research supports this notion, showing that predictions about a person's mental states or behaviors are influenced by social knowledge. For example, Saxe and Wexler (2005) presented participants with stories about social agents whose mental states were consistent or inconsistent with expectations that follow from social norms within the target agent's culture (e.g., your friend from high school, who has a happy marriage, confides in you that he [really hates the idea that/would find it fun if] his wife might ever have an affair). For participants, and for their imagined target friend, the cultural norm was that happily married people do not find it exciting if their partner wants a relationship with another person. RTPJ activity was stronger when the target's cultural norms were violated, and this result held even when participants' own cultural norms differed from the target's. Similarly, bilateral TPJ and MPFC were recruited more when participants learned about politicians whose political desires were incongruent (versus congruent) with their party identity (e.g., a democrat who wants a [smaller/larger] government), suggesting that these regions preferentially respond to expectation-violating information (Cloutier et al., 2011). Even knowledge of an agent's ability (e.g., a novice versus an experienced bowler) is enough to produce prediction error in mentalizing regions when performance is inconsistent with that ability (Heil et al., 2019).

In total, it seems that ToM regions flexibly encode different information types to make predictions and process expectation violations. This flexibility account finds additional support in recent work on metaethical judgments. J. Theriault et al. (2020) demonstrated that moral statements judged as more preference-like elicit greater ToM activity, whereas moral statements judged as more fact-like elicit less ToM activity. In each case, ToM activity was related not to an agent, but to social consensus regarding metaethics. However, even though past research is generally consistent with an account in which ToM regions flexibly encodes and uses the available types of social information to make predictions and process their outcomes, individual studies have tended to focus on only a single type of information when engendering predictions or revealing outcomes.

More targeted research has attempted to isolate expectation-based ToM effects that result from nonsocial and social knowledge sources. Dungan et al. (2016) presented participants with stories that varied the source of an expectation (i.e., prior information about a nonsocial object, and prior information about a social agent [prior mental state vs prior behavior]), and whether subsequent behavior was consistent with that expectation. For nonsocial objects, results suggested that ToM regions were recruited less overall when compared to social agents. Moreover, ToM regions did not show effects at the univariate level for nonsocial objects' violation of expectations, nor did these regions discriminate between nonsocial objects' unexpected versus expected behavior in multi-voxel pattern analyses (MVPA). For social agents, however, results showed that DMPFC and bilateral TPJ were preferentially recruited for unexpected over expected behaviors at the univariate level, and that these regions discriminated between unexpected and expected behavior in MVPA. Overall, these results suggest that ToM regions are selectively recruited for prediction error that is social in nature. However, simple effects analyses revealed that the social prediction error effects in bilateral TPJ were evident only when prior information was behavior.

Current research

Because Dungan et al. (2016) varied only the prior information type (i.e., behavior versus mental state) and not the outcome information (i.e., all outcomes were behaviors), it was impossible to know whether ToM regions would also preferentially respond to unexpected mental state outcomes. Similarly, participants might have experienced especially strong prediction error if they had instead learned about a mental state that was inconsistent with a prior mental state. Additionally, because Dungan et al's expectedness effect in bilateral TPJ occurred only when prior information was about behavior, it was impossible to know whether the lack of an effect in the other prior information condition (i.e., mental state) was simply due to a mismatch between prior and outcome types. For example, if your father did not show up at the airport to pick you up, and your prior information was that he simply *wanted* to pick you up, you may have considered the fact that even though people sometimes want to do something, they often do not (or cannot) do it. Therefore, you may not find your father's behavior particularly surprising. For these reasons, the current paper uses a novel paradigm to examine whether social prediction error occurs across brain regions typically implicated in thinking about others' minds, as well as whether these regions' sensitivity to social prediction error differs as a function of the type of information that is used to make predictions and the type of information that confirms or violates them.

fMRI experiment

Method

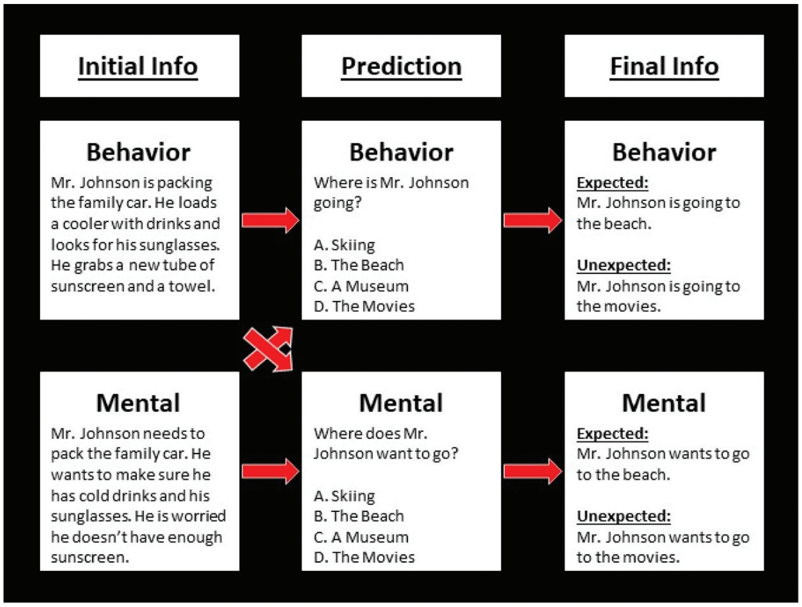
Participants

Participants were 25 right-handed adults were recruited from the Greater Boston Area. One participant, unable to remain still, was removed from the scanner partway through the study, resulting in a final $N = 24$ (age: $M = 24.08$, $SD = 4.11$; 50% female). All participants were native English speakers, had normal or corrected-to-normal vision, and gave written informed consent in accordance with the Boston College Internal Review Board. Additionally, participants reported no psychiatric disorders or history of learning disabilities. Sample size was determined by available resources at the time data were collected; we note that although this sample size is small, it is typical of the time these data were collected in 2015.

Procedure and materials

Participants were scanned while reading and responding to 64 vignettes, learning and making predictions about 64 different agents (see Supplemental Online Materials [SOM]). Each story was presented in three sequential segments: Initial Info, Prediction, and Final Info (see Figure 1). During the Initial Info segment, participants read background information to establish an expectation about how an agent would likely think or behave in the future. During the Prediction segment, participants were presented with a multiple-choice question asking them to make a prediction about the agent's future thoughts or behavior. Four options were provided: one that was expected based on the Initial Info segment, and three others that would be relatively unexpected. Participants responded to this question by using a button-box. During the Final Info segment, the vignette's outcome was presented, which corresponded to one option from the Prediction segment's multiple-choice question.

Critically, we varied the type of information presented in each segment. The Initial Info segment consisted of either the agent's 1) prior behavior, or 2) prior mental states. Similarly, during the Prediction segment, participants made a prediction about either the agent's 1) subsequent behavior, or 2) subsequent mental state. Last, the Final Info segment presented participants with either the agent's 1) actual subsequent behavior, or 2) actual subsequent mental state, which was the same type of information queried in the Prediction segment. Also, the Final Info segment presented was either expected or unexpected based on the information provided in the Initial Info segment. For the Final Info segment, an unexpected ending was shown on half of all



Mr. Johnson is packing the family car. He loads a cooler with drinks and looks for his sunglasses. He grabs a new tube of sunscreen and a towel.

Where is Mr. Johnson going?

A. Skiing
B. The Beach
C. A Museum
D. The Movies

Expected:
Mr. Johnson is going to the beach.

Unexpected:
Mr. Johnson is going to the movies.

Mr. Johnson needs to pack the family car. He wants to make sure he has cold drinks and his sunglasses. He is worried he doesn't have enough sunscreen.

Where does Mr. Johnson want to go?

A. Skiing
B. The Beach
C. A Museum
D. The Movies

Expected:
Mr. Johnson wants to go to the beach.

Unexpected:
Mr. Johnson wants to go to the movies.

Figure 1. Example experimental stimulus and its variants. Participants first read the initial information segment of the stimulus (Behavior or Mental), then made a prediction (Behavior or Mental), and last learned the final information. Importantly, final information type (Behavior or Mental) always matched prediction type. Participants never saw the same stimulus across conditions; different participants saw different experimental variations of the same stimulus.

trials, whereas an expected ending was shown on the remaining trials. Crossing these dimensions (Initial Info, Final Info, and Expectedness) yielded 8 conditions in a 2 (Initial Info: behavior, mental state) x 2 (Final Info: behavior, mental state) x 2 (Expectedness: expected, unexpected) design. These three experimental factors were also manipulated within stimuli, constituting a fully within-subject/within-stimulus design. The order of conditions and pairing of conditions and vignettes were randomized across participants. An online behavioral sample, that completed the same task as the fMRI participants did, verified that we successfully manipulated expectedness (see SOM Table 1).

The vignettes were presented in a pseudo-randomized order in white font on a black background via an Apple MacBook Pro running MATLAB 8.5 (2015) with Psychophysics Toolbox. The Initial Info segment was presented on-screen for 10 seconds, the Prediction segment for 8 seconds, and the Final Info segment for 4 seconds. To analyze Initial Info and Final Info segments separately, 0, 2, or 4 seconds of jittered fixation were included between each story segment. Stimulus presentation was divided into 8 equal runs (8 stimuli per run, 1 per condition) lasting 4 minutes and 4 seconds each.

Functional localizer

Participants also completed a theory of mind (ToM) functional localizer task (Dodell-Feder et al., 2011) consisting of 10 stories about mental states (e.g., false-belief condition) and 10 stories about physical representations (e.g., false-photograph condition). The task was presented in two 4.5-minute runs, interleaved between experiment runs.

fMRI Data Acquisition and Preprocessing

The fMRI data were collected using a 16-channel head coil in a 3T Siemens scanner at the Athinuoula A. Martinos Imaging Center, Massachusetts Institute of Technology. Data were acquired in 36 near-axial slices (3 mm isotropic voxels, 0.54 mm gap). Standard gradient echo planar imaging (EPI) procedures were used (TR = 2000 ms; TE = 30 ms; flip angle = 90°; FOV = 216 × 216; interleaved acquisition). Anatomical data were collected with T1-weighted multi-echo magnetization prepared rapid acquisition gradient echo image sequences (MEMPRAGE) using the following parameters: TR = 2530 ms; TE = 1.64 ms; FA = 7°; 1 mm isotropic voxels; 0.5 mm gap between slices; FOV = 256 × 256. Data processing and analysis were performed using fMRIPrep (Esteban et al. (2019); see Supplementary Materials p. 1 for details), SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/), and custom software. The functional data were realigned, coregistered to the anatomical image, normalized onto a common brain space (Montreal Neurological Institute, MNI, template), spatially smoothed using a Gaussian filter (fullwidth half-maximum = 8 mm kernel), and high-pass filtered (128 Hz). Neural responses were modeled in an event-related design using a general linear model (GLM), with conditions modeled as boxcar functions convolved with a canonical hemodynamic response function (HRF). The GLM included the six components of the anatomical CompCor variant (aCompCor) as nuisance regressors (Behzadi et al., 2007).

Analytic approach

Whole-brain and regions of interest (ROI) analyses were conducted. For whole-brain analyses, we first conducted a whole-brain random effects analysis (voxel-wise threshold: $p < .001$, uncorrected; $k > 16$; cluster-wise threshold: $p < .05$, FWE-corrected) of behavior over mental conditions during the Initial Info segment. Second, we conducted a whole-brain random effects analysis of mental over behavior conditions during the Initial Info segment. Third, we conducted a whole-brain random effects analysis of expected over unexpected conditions during the Final Info segment. Fourth, we conducted a whole-brain random effects analysis of unexpected over expected conditions during the Final Info segment. For all whole-brain analyses, assignments of coordinates to brain regions were aided by use of the *label4MRI* package in R (Chuang, 2020), which performs automatic anatomic labeling (AAL) based on the most recently updated atlas, AAL3 (Rolls et al., 2020; see our OSF page for an RMarkdown file). Next, we describe the ROI analyses.

A whole-brain contrast of false-belief versus false-photograph stories in the ToM localizer task (Dodell-Feder et al., 2011) was used to identify ROIs implicated in ToM: DMPFC ($N = 20$), PC ($N = 23$), LTPJ ($N = 22$), and RTPJ ($N = 23$). ROIs were selected for each participant

Table 1. Average Peak MNI coordinates for ToM in Functional Localizer Task.

		MNI Coordinates				
ROI	N	X	Y	Z	Voxels	t-value
RTPJ	23	53	−55	23	93	8.07
LTPJ	22	−49	−55	24	75	6.58
PC	23	−2	−58	34	85	6.40
DMPFC	20	3	51	23	61	5.47

individually and defined as contiguous voxels within a 9-mm radius of the peak voxel that passed contrast threshold (see Table 1 for by-ROI coordinate information). Within each ROI, the average percent signal change (PSC) relative to runwise baseline (PSC = 100×raw BOLD magnitude for (condition−fixation)/raw BOLD magnitude for fixation) was calculated for each condition at each time point (averaging across all voxels in the ROI and all blocks of the same condition). Initial Info and Final Info segments were modeled separately. Timepoints were shifted by 6 seconds to account for hemodynamic lag. The Prediction segment was not analyzed.

Importantly, for all ROI analyses, we analyzed only trials on which participants selected the correct outcome prediction (i.e., the option that we determined, a priori, was the most likely to occur given the information that participants received in the Initial Info segment). This was done to ensure that our results were uncontaminated by the possibility of participants' lack of attention, random responding, or their own expectations being different from the paradigm's intended expectations. The frequency of incorrect predictions was similar across Initial Info x Final Info conditions (i.e., BB, BM, MB, MM): 16%, 12%, 15%, and 13% of each condition's total trials were incorrect, respectively. Further, we removed images according to the following criteria: individual scans, along with their two temporally adjacent scans, were excluded if framewise displacement (FD) (Power et al., 2012) exceeded 0.5 mm; individual runs were excluded if either (1) FD for more than two-thirds of the scans in that run exceeded 0.5 mm or (2) FD for any scan in that run exceeded 3 mm (see results' table notes for final analyzable Ns). With the above data transformations/exclusions in mind, we next explain how we arrived at analyzable data. Specifically, for each participant's stimuli, multiple segments were presented (e.g., participant 1's Initial Info segment for vignette 1). Within each participant's stimulus, we averaged across time course activity during the segment of interest (accounting for hemodynamic lag). Finally, we used those averages in linear mixed effects models, meaning that each participant, for each stimulus, had a single PSC value for a specific segment of the experimental design.

For all ROI analyses, linear mixed effects models were constructed in R (R Core Team, 2021) to simultaneously account for variability across participants and stimuli (Judd et al., 2012). Within each ROI, we attempted to fit a maximal model that allowed all main effects and interactions to vary over participants and stimuli. If the maximal model failed to converge or yielded a singular fit, we followed guidelines to avoid false positives (see Barr et al., 2013; Barr, 2013; Singmann & Kellen, 2019). First, we simplified the random effects structure by removing all correlations between random effects. Next, if the zero-correlation model failed to converge or converged with a singular fit, we further simplified the random effects structure by removing variance components that were estimated as zero. If this further reduced model resulted in non-convergence, or convergence with a singular fit, we repeated the second step. If there were no remaining variances estimated as zero, we removed the smallest variance components, one at a time, until the model converged with a non-singular fit. Last, when appropriate, we attempted to add random effects' correlations back into the model (see Bates et al., 2018). If this extended model converged with a non-singular fit, we retained it as our final model. However, if this extended model did not converge, or converged with a singular fit, we retained the non-extended model as our final model. We report only our final models here. The entire model selection process (i.e., specifications and simplifications) can be found on our OSF page: https://osf.io/tf852/.

Whole-brain results

As described above, we first conducted a whole-brain random effects analysis of behavior over mental conditions during the Initial Info segment. This contrast revealed only one peak cluster in the right angular gyrus [48, −64, 31]. Second, we conducted a whole-brain random effects analysis of mental over behavior conditions during the Initial Info segment. However, no peak clusters emerged for this contrast. Third, we conducted a whole-brain random effects analysis of expected over unexpected conditions during the Final Info segment. Fourth, we conducted a whole-brain random effects analysis of unexpected over expected conditions during the Final Info segment. See Table 2 for peak clusters revealed for these final two contrasts. We note that none of the ToM regions passed threshold in whole-brain contrasts for unexpected over expected outcomes.

ROI results

Initial info

Here, we investigated the effect of the Initial Info manipulation on neural activity during the Initial Info segment. We note that model reduction within

Table 2. Regions passing threshold in whole-brain analysis during Final Info segment.

		MNI Coordinates				
Contrast	Region	X	Y	Z	*t*-Value	Cluster Size
Expected > Unexpected						
	L middle occipital gyrus	−48	−79	1	9.22	704
	L superior temporal gyrus	−63	−31	13	7.83	446
	R precuneus	12	−76	55	7.36	1864
	R superior frontal gyrus (dorsolateral)	21	−1	73	5.43	84
	R inferior temporal gyrus	51	−67	−5	5.33	210
	R lobule VI cerebellar hemisphere	21	−64	−14	5.00	169
	L lingual gyrus	−12	−88	−11	4.93	116
Unexpected > Expected						
	L supplementary motor area	−6	17	64	11.34	3567
	R inferior frontal gyrus pars orbitalis	48	26	−11	9.14	1019
	R thalamus	12	−7	7	7.67	129
	L lobule VI/V cerebellar hemisphere	−18	−37	−29	6.94	166
	R middle temporal gyrus	48	−22	−11	5.90	157
	L middle temporal gyrus	−57	−34	−8	5.82	224
	R lobule VI cerebellar hemisphere	33	−61	−29	5.31	192
	L angular gyrus	−45	−55	31	4.90	147

Table 3. DMPFC activity during Initial Info segment.

Final Model:
PSC ~ Initial + (1 | Item) + (1 | Subject)
Coding:
Initial Info: Contrast coded (Mental = −0.5; Behavior = +0.5)

Random Effects	Var.	SD			
Item					
Intercept	.003	.052			
Subject					
Intercept	.015	.124			
Residual					
	.116	.341			
Fixed Effects			***b*** ***(SE)***	***t*** **(df)**	***p-value***
Intercept			−.03 (.03)	−1.01 (21)	.325
Initial			.01 (.02)	0.53 (997)	.593

Note. Analysis included 1063/1280 observations from 20 subjects and 64 items. Degrees of freedom were Satterthwaite-approximated and rounded to the nearest integer for all analyses.

Table 4. PC activity during Initial Info segment.

Final Model:
PSC ~ Initial + (1 + Initial || Item) + (1 | Subject)
Coding:
Initial Info: Contrast coded (Mental = −0.5; Behavior = +0.5)

Random Effects	Var.	SD				
Item						
Intercept	.002	.048				
Initial	.001	.031				
Subject						
Intercept	.006	.079				
Residual						
	.103	.322				
Fixed Effects			***b*** ***(SE)***	***t*** **(df)**	***p-value***	
Intercept			−.04 (.02)	−2.07 (24)	.049	*
Initial			−.00 (.02)	−0.01 (60)	.992	

Note. Analysis included 1234/1472 observations from 23 subjects and 64 items.

each ROI sometimes led to different random effects structures across ROIs. We chose this strategy of conservatism (rather than an anti-conservative strategy which held random effects structures constant across ROIs) to avoid false positives in some ROIs. Within each ToM ROI, there was no effect of

Table 5. LTPJ activity during Initial Info segment.

Final Model:
PSC ~ Initial + (1 + Initial | Item) + (1 | Subject)
Coding:
Initial Info: Contrast coded (Mental = −0.5; Behavior = +0.5)

Random Effects	Var.	SD	Correlations			
Item						
Intercept	.003	.051	-			
Initial	.005	.067	.23	-		
Subject						
Intercept	.049	.222	-			
Residual						
	.087	.294				
Fixed Effects			***b* (*SE*)**	***t* (df)**	***p-value***	
Intercept			.12 (.05)	2.54 (22)	.019	*
Initial			−.01 (.02)	−0.67 (65)	.508	

Note. Analysis included 1189/1408 observations from 22 subjects and 64 items.

Table 6. RTPJ activity during Initial Info segment.

Final Model:
PSC ~ Initial + (1 | Item) + (1 | Subject)
Coding:
Initial Info: Contrast coded (Mental = −0.5; Behavior = +0.5)

Random Effects	Var.	SD			
Item					
Intercept	.001	.036			
Subject					
Intercept	.012	.108			
Residual					
	.071	.267			
Fixed Effects			***b* (*SE*)**	***t* (df)**	***p-value***
Intercept			−.02 (.02)	−0.68 (23)	.504
Initial			.01 (.02)	0.61 (1161)	.539

Note. Analysis included 1234/1472 observations from 23 subjects and 64 items.

Table 7. ToM Network activity (averaged across ROIs) during Initial Info segment.

Final Model:
PSC ~ Initial + (1 | Item) + (1 | Subject)
Coding:
Initial Info: Contrast coded (Mental = −0.5; Behavior = +0.5)

Random Effects	Var.	SD			
Item					
Intercept	.002	.039			
Subject					
Intercept	.009	.095			
Residual					
	.047	.216			
Fixed Effects			***b* (*SE*)**	***t* (df)**	***p-value***
Intercept			.01 (.02)	0.30 (24)	.768
Initial			.00 (.01)	0.13 (1158)	.897

Note. Analysis included 1234 observations from 23 subjects and 64 items.

the Initial Info manipulation on neural activity during the Initial Info segment. See Tables 3–6 for detailed information about all final models (i.e., coding scheme, random effects structure, random effects estimates, and fixed effects estimates). These patterns held when analyzing neural activity averaged across the entire ToM network (see Table 7).

Final info

Here, we investigated the effect of the Initial Info, Final Info, and Expectedness manipulations on neural activity during the Final Info segment. See Figure 2 for results plotted by ROI and Tables 8–11 for detailed information about all final models. We note, here too, that model reduction within each ROI sometimes led to different random effects

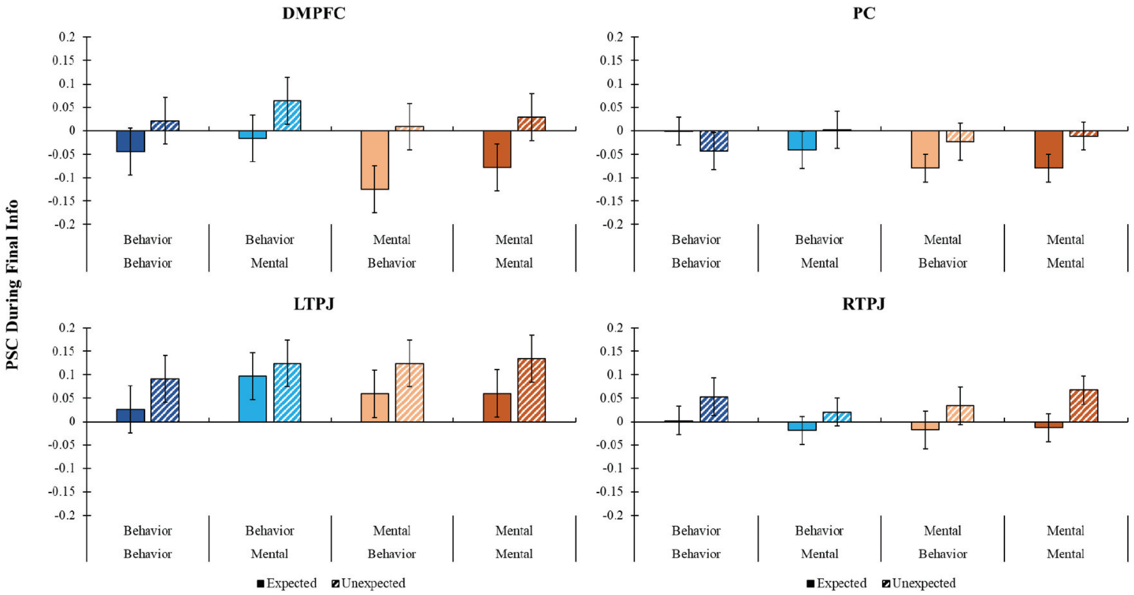


Figure 2. PSC during Final Info segment within each ROI. On the x-axis, the top factor (Behavior vs Mental) refers to the Initial Info manipulation, whereas the bottom factor (Behavior vs Mental) refers to the Final Info manipulation. Solid bars show expected outcome estimates, whereas patterned bars show unexpected outcome estimates. Estimates are predicted means from each ROI's linear mixed effects model. Error bars represent ± 1 SE for the predicted mean.

Table 8. DMPFC activity during Final Info segment.

Final Model:
PSC ~ Initial*Final*Expected +
(1 + Final + Initial:Final + Initial:Expected + Final:Expected || Item) +
(1 + Expected + Initial:Final:Expected || Subject)
Coding:
Initial Info: Contrast coded (Mental = −0.5; Behavior = +0.5)
Final Info: Contrast coded (Mental = −0.5; Behavior = +0.5)
Expectedness: Contrast coded (Expected = −0.5; Unexpected = +0.5)

Random Effects	Var.	SD
Item		
Intercept	.002	.045
Final	.000	.094
Initial:Final	.014	.119
Initial:Expected	.008	.091
Final:Expected	.013	.113
Subject		
Intercept	.027	.163
Expected	.007	.081
Initial:Final:Expected	.004	.064
Residual		
	.123	.350

Fixed Effects	*b (SE)*	*t* **(df)**	*p-value*	
Intercept	−.02 (.04)	−0.45 (20)	.659	
Initial	.05 (.02)	2.13 (869)	.033	*
Final	−.03 (.03)	−1.36 (60)	.178	
Expected	.10 (.03)	3.35 (17)	.004	**
Initial:Final	−.00 (.05)	−0.04 (63)	.965	
Initial:Expected	−.05 (.05)	−1.06 (60)	.294	
Final:Expected	.01 (.05)	0.12 (62)	.903	
Initial:Final:Expected	−.04 (.09)	−0.46 (14)	.653	

Note. Analysis included 1021/1280 observations from 20 subjects and 64 items.

Table 9. PC activity during Final Info segment.

Final Model:
PSC ~ Initial*Final*Expected +
(1 + Expected + Initial:Final + Initial:Expected + Initial:Final:Expected || Item) +
(1 + Initial:Expected || Subject)
Coding:
Initial Info: Contrast coded (Mental = −0.5; Behavior = +0.5)
Final Info: Contrast coded (Mental = −0.5; Behavior = +0.5)
Expectedness: Contrast coded (Expected = −0.5; Unexpected = +0.5)

Random Effects	Var.	SD			
Item					
Intercept	.002	.044			
Expected	.000	.001			
Initial:Final	.016	.128			
Initial:Expected	.005	.069			
Initial:Final:Expected	.043	.209			
Subject					
Intercept	.010	.100			
Initial:Expected	.002	.046			
Residual					
	.104	.323			
Fixed Effects			*b (SE)*	*t* **(df)**	*p-value*
Intercept			−.03 (.02)	−1.47 (24)	.156
Initial			.03 (.02)	1.51 (991)	.140
Final			−.00 (.02)	−0.21 (989)	.829
Expected			.03 (.02)	1.68 (63)	.100
Initial:Final			.00 (.04)	0.10 (61)	.924
Initial:Expected			−.06 (.04)	−1.55 (17)	.139
Final:Expected			−.05 (.04)	−1.30 (979)	.194
Initial:Final:Expected			−.07 (.08)	−0.91 (59)	.368

Note. Analysis included 1179/1472 observations from 23 subjects and 64 items.

Table 10. LTPJ activity during Final Info segment.

Final Model:
PSC ~ Initial*Final*Expected +
(1 + Initial + Expected + Initial:Final + Initial:Expected + Final:Expected +
Initial:Final:Expected || Item) +
(1 + Expected || Subject)
Coding:
Initial Info: Contrast coded (Mental = −0.5; Behavior = +0.5)
Final Info: Contrast coded (Mental = −0.5; Behavior = +0.5)
Expectedness: Contrast coded (Expected = −0.5; Unexpected = +0.5)

Random Effects	Var.	SD				
Item						
Intercept	.002	.040				
Initial	.002	.047				
Expected	.005	.069				
Initial:Final	.029	.169				
Initial:Expected	.013	.113				
Final:Expected	.022	.148				
Initial:Final:Expected	.052	.228				
Subject						
Intercept	.037	.192				
Expected	.001	.027				
Residual						
	.093	.306				
Fixed Effects			*b (SE)*	*t* **(df)**	*p-value*	
Intercept			.09 (.04)	2.10 (21)	.048	*
Initial			−.01 (.02)	−0.48 (58)	.631	
Final			−.03 (.02)	−1.54 (894)	.125	
Expected			.06 (.02)	2.72 (17)	.015	*
Initial:Final			−.05 (.04)	−1.08 (61)	.284	
Initial:Expected			−.02 (.04)	−0.62 (61)	.550	
Final:Expected			.01 (.04)	0.35 (59)	.726	
Initial:Final:Expected			.05 (.08)	0.60 (60)	.553	

Note. Analysis included 1134/1408 observations from 22 subjects and 64 items.

Table 11. RTPJ activity during Final Info segment.

Final Model:
PSC ~ Initial*Final*Expected +
(1 + Initial:Final + Initial:Final:Expected || Item) +
(1 + Expected + Initial:Expected + Final:Expected + Initial:Final:Expected || Subject)
Coding:
Initial Info: Contrast coded (Mental = –0.5; Behavior = +0.5)
Final Info: Contrast coded (Mental = –0.5; Behavior = +0.5)
Expectedness: Contrast coded (Expected = –0.5; Unexpected = +0.5)

Random Effects	Var.	SD				
Item						
Intercept	.001	.024				
Initial:Final	.008	.087				
Initial:Final:Expected	.002	.069				
Subject						
Intercept	.014	.117				
Expected	.000	.015				
Initial:Expected	.005	.068				
Final:Expected	.001	.030				
Initial:Final:Expected	.026	.162				
Residual						
	.083	.229				
Fixed Effects			***b (SE)***	***t* (df)**	***p-value***	
Intercept			.02 (.03)	0.60 (22)	.552	
Initial			–.00 (.02)	–0.21 (986)	.828	
Final			.00 (.02)	0.21 (988)	.836	
Expected			.06 (.02)	3.24 (21)	.004	**
Initial:Final			.05 (.04)	1.31 (60)	.195	
Initial:Expected			–.02 (.04)	–0.55 (18)	.590	
Final:Expected			–.01 (.03)	–0.24 (22)	.813	
Initial:Final:Expected			.04 (.08)	0.52 (17)	.611	

Note. Analysis included 1179/1472 observations from 23 subjects and 64 items.

Table 12. ToM Network activity (averaged across ROIs) during Final Info segment.

Final Model:
PSC ~ Initial*Final*Expected +
(1 + Final + Expected + Initial:Final + Final:Expected +
Initial:Final:Expected || Item) +
(1 | Subject)
Coding:
Initial Info: Contrast coded (Mental = –0.5; Behavior = +0.5)
Final Info: Contrast coded (Mental = –0.5; Behavior = +0.5)
Expectedness: Contrast coded (Expected = –0.5; Unexpected = +0.5)

Random Effects	Var.	SD				
Item						
Intercept	.001	.035				
Final	.001	.032				
Expected	.000	.016				
Initial:Final	.007	.083				
Final:Expected	.007	.082				
Initial:Final:Expected	.021	.146				
Subject						
Intercept	.012	.109				
Residual						
	.052	.227				
Fixed Effects			***b (SE)***	***t* (df)**	***p-value***	
Intercept			.01 (.02)	0.40 (23)	.696	
Initial			.01 (.01)	1.07 (953)	.284	
Final			–.01 (.01)	–1.06 (61)	.294	
Expected			.06 (.01)	4.36 (61)	<.001	***
Initial:Final			.00 (.03)	0.08 (61)	.938	
Initial:Expected			–.03 (.03)	–1.25 (949)	.212	
Final:Expected			–.01 (.03)	–0.41 (61)	.684	
Initial:Final:Expected			–.00 (.06)	–0.02 (64)	.985	

Note. Analysis included 1179 observations from 23 subjects and 64 items.

structures across ROIs. We chose this strategy of conservatism (rather than an anti-conservative strategy which held random effects structures constant across ROIs) to avoid false positives in some ROIs.

In DMPFC, LTPJ, and RTPJ, there were main effects of Expectedness, such that neural activity was higher when the final information was unexpected compared to expected (DMPFC: $b = .10$, $SE = .03$, $p = .004$; LTPJ: $b = .06$, $SE = .02$, $p = .015$; RTPJ: $b = .06$, $SE = .02$, $p = .004$). This main effect of Expectedness held when analyzing neural activity averaged across the entire ToM network (see Table 12). To investigate the robustness of these unexpectedness effects, we investigated how many participants showed them, finding that most participants showed these effects within each ROI (DMPFC = 16/20; LTPJ = 13/22; RTPJ = 17/23). Additionally, in DMPFC only, there was a main effect of Initial Info, such that neural activity during the Final Info segment was higher when Initial Info was behavior (compared to mental), $b = .05$ ($SE = .02$), $p = .033$. No other main effects or interactions were observed.

General discussion

The current work investigated whether social prediction error (i.e., the difference between what was expected and what was observed) occurs across brain regions typically implicated in thinking about others' minds, as well as whether these regions' sensitivity to social prediction error differs as a function of the type of information that is used to make predictions and confirm or violate them. When people learned about an agent's unexpected mental states or behavior, DMPFC, LTPJ, and RTPJ activity was higher than when people learned about an agent's expected mental states or behaviors. But no region showed differential effects of expectedness based on the type of information that was used to make predictions or confirm/violate them (i.e., mental states versus behaviors), suggesting that these regions may flexibly incorporate any available social information to make predictions about and monitor or explain social inconsistencies. These findings add to a growing literature (e.g., Cloutier et al., 2011; Dungan et al., 2016; Heil et al., 2019; J. Theriault et al., 2020; Saxe & Wexler, 2005) supporting the link between predictive coding and ToM activity (Koster-Hale & Saxe, 2013). However, some of our results are consistent with prior research, whereas others are not.

In the current experiment, effects of expectedness in DMPFC, LTPJ and RTPJ are consistent with prior work. Specifically, such effects have been found in investigations of unexpected behaviors (Dungan et al., 2016; Heil et al., 2019) and unexpected mental states (Cloutier et al., 2011). Additionally, that we did not find an effect of expectedness in PC is also consistent with prior research (Dungan et al., 2016). However, we caution readers that this lack of detection (as well as other null effects) could be interpreted in the following ways: (1) no effect occurs in PC, (2) a theoretically meaningful but small effect occurs in PC but could not be detected, or (3) any effect that occurs in PC is too small to be theoretically meaningful. Because there will be disagreement about what is theoretically meaningful to all researchers, we do not take a firm position on this issue. We do note the following: in the current experiment, standardized effect sizes for expectedness in DMPFC ($d = 0.21$), LTPJ ($d = 0.11$), and RTPJ ($d = 0.15$) ranged from more than 1.5× − 3× the standardized effect size for expectedness in PC ($d = 0.07$) (All d's were calculated using variance estimates, as described in Brysbaert & Stevens, 2018; Westfall et al., 2014). These estimates, combined with Dungan et al.'s reported null effect, suggests that PC may diverge from other ToM regions in monitoring expectation violations.

Other findings in the current experiment are somewhat surprising considering prior research. In particular, expectedness effects in ToM regions were not moderated by the type of information that engendered or violated predictions. On one popular account of cultural learning, people attend to "credibility-enhancing" displays to determine their degree of confidence in someone else's beliefs (Henrich, 2009), being more confident in someone else's beliefs when a costly behavior reflects the purported belief. Applying this logic to the current data, one possibility is that prior behaviors could have served as stronger predictors of subsequent behaviors/mental states because behaviors are interpreted as better signals of one's current beliefs or intentions. For example, community organizers who had themselves installed solar panels were more effective in recruiting new residents to install solar panels than organizers who had not done so, as the former engaged in costly behavior which was inferred as an honest signal of their belief in the technology's benefit (Kraft-Todd et al., 2018).

This possibility is also consistent with research by Dungan et al. (2016) in which an expectedness effect in RTPJ was driven by one condition (see Figure 2 of Dungan et al.). In the outcome segment, Dungan et al. reported that RTPJ showed an expectedness effect only when prior information was behavior (but not mental), suggesting that RTPJ may be recruited specifically when an agent's subsequent behavior is inconsistent with their prior behavior, perhaps in order to generate mental states that explain the behavioral inconsistency (see Decety & Lamm, 2007). In the current experiment, however, when outcome information was behavior (as was always true in Dungan et al., 2016), we did not replicate this finding. More specifically, both expectedness simple effects were identical in magnitude in our data (initial behavior d = 0.14; initial mental d = 0.14). Using all of our data, we also failed to detect a three-way interaction among prior information, outcome information, and expectedness, suggesting that Dungan et al. (2016) findings in RTPJ may not be due to mismatches between prior information and outcome information.

On the other hand, on the logic that people can never be certain about the inferred mental states of others, explicit mental states could have served as stronger predictors of subsequent behaviors/mental states. That participants might make the most reliable predictions about future mental states based on prior mental states is also consistent with recent theoretical work arguing that people can track the transitional probabilities between mental states (Tamir & Thornton, 2018). Even though people do not come with thought bubbles above their heads in the real world, participants in the current experiment were given explicit access to others' mental states in a thought-bubble-like way. Therefore, the current data's lack of prior-by-outcome information moderation on prediction error signals is surprising. However, it has also been argued that social outcomes are inherently much less predictable than nonsocial outcomes (FeldmanHall & Shenhav, 2019), which may explain why social information type does not moderate the effects here. Further experimentation, and much larger (and therefore higher-powered) fMRI paradigms are needed to better understand if, when, and how social prediction error interacts with the type of social information that predictions and violations were based on, as the observed null interactions could simply be false negatives.

There are multiple methodological features that may explain inconsistencies between the current work and Dungan et al.'s findings specifically. First, a strength of the current work is that stimuli were constructed for both behavioral and mental conditions of the prior information segment (see Figure 1 for an example), whereas stimuli in Dungan et al. were nested in a particular prior information condition (i.e., completely different stimuli constituted behavior versus mental conditions). Therefore, Dungan et al. "s effects of initial behavior versus initial mental states may have been driven by stimulus differences rather than a true distinction between information types. Second, data in the current experiment were analyzed with linear mixed effects models as opposed to traditional repeated-measures ANOVAs. For multilevel data (e.g., responses nested within participants/stimuli), linear mixed effects models better control Type I error rates by retaining the true variability in the data and adjusting standard errors of test statistics to account for the possibility that some participants/stimuli will respond (to an experimental manipulation) differently than other participants/stimuli. Therefore, some of Dungan et al."s effects may have been due to specific participants or stimuli behaving in ways that led to a group-level effect which was not representative of most participants or stimuli. These methodological changes can also explain another discrepancy between the present data and Dungan et al.'s. In the prior information segment, Dungan et al. detected an effect in which RTPJ activity was higher when initial information was behavior (versus mental), suggesting that RTPJ may play a special role in mental state *inference* based on witnessed behavior rather than processing mental states directly. However, in the current experiment, we found no evidence of this effect (d = 0.03). Sample size issues notwithstanding, we believe that the methods of the current experiment offer the best tests of the ideas under investigation thus far. Therefore, that we successfully replicated prior research's effects of expectedness lends especially strong evidence to the idea that DMPFC, LTPJ, and RTPJ are ToM regions coding for discrepancies between social predictions and their outcomes.

Limitations and future directions

Although the current experiment improved on prior work, it has important limitations. First, our design focused on people's predictions about unknown others based on a single prior behavior or mental state. However, people more typically interact with and make predictions about agents they know. Social prediction error may occur more strongly or weakly depending on one's relationship to the agent who is the object of prediction. For example, people believe that there are stronger obligations to help family members compared to non-family members (Marshall et al., 2020, 2022; McManus et al., 2020, 2021), leading to neglect of family members being judged as more unexpected (see SOM of McManus et al., 2020). Additionally, people seem to experience stronger prediction error when they imagine witnessing close (versus distant) others commit crimes (Berg et al., 2021), which may be due to their having stronger positive priors about close others (see Hughes, Ambady, et al., 2017; Hughes, Zaki, et al., 2017; Kim, Park, et al., 2020). Future research can shed more light on the neural mechanisms involved in using relationship information to make and monitor predictions about an agent's mental states and behavior.

Second, the current experiment did not systematically vary the social context in which expectations were confirmed or violated, which may alter if and how social prediction error occurs. For example, imagine that an agent thinks to herself, "I want to speak up the next time I hear a sexist joke about women." However, the next time she hears a sexist joke is at her workplace where all of her colleagues are men. When she fails to speak up, you may be unsurprised, perhaps because you understand that she might experience additional negativity at her workplace in the future. Conversely, if at least half of her colleagues were women, you may be extremely surprised when she fails to speak up, perhaps because you believe she would have wanted to alleviate the possible discomfort experienced by her same-gendered colleagues. Such an example suggests that there are many potential features of the social context that can affect predictions and therefore what is considered unexpected, such as an agent's reputational concerns, the demographic composition of surrounding others, and more.

Additionally, this example demonstrates that the time point at which a prediction is made, and whether this prediction is updated, is a crucial factor. If you were to predict the woman's behavior at the exact time a sexist joke occurred, you might make starkly different predictions based on her social context. If, however, you predicted the woman's behavior at an earlier time (e.g., the time at which you first learned she imagined speaking up), you may or may not update your prediction based on the context in which the sexist joke occurs. While recent work has documented that people update their (moral) impressions of others through learning more about the agents' past behaviors (see Brambilla et al., 2019; Ferguson & Mann, 2017; Kim, Park, et al., 2021), less is known about how people update their predictions of an agent's single future behavior over time. Future research could investigate the conditions under which predictions are updated and how this relates to prediction error (see Bach & Schenke, 2017, for a detailed discussion).

Third, although we failed to detect interaction effects based on the sources of information used to make and violate predictions, it is unlikely that such effects would never occur. For example, imagine learning that someone thought to himself, "I really hate my coworkers and my role in this company. I want to quit." When he continues to go into work, you may not be very surprised because you realize that people often think and want to do things that they do not do. However, if he instead sent out a company-wide e-mail stating, "I really hate my coworkers and my role in this company. I'm going to quit," you might be extremely surprised when he continues to go into work. The critical difference between these cases is that the agent's prior behavior seems to leave no doubt about his near-future intentions. We may not have seen such effects in the current data because most of the agents' prior behaviors did not yield near-certain predictions. Interestingly, follow-up behavioral and behavior-brain analyses of our data suggest that vignette-level prediction confidence indeed varied substantially and was positively related to the difference in vignette-level neural activity during expectation confirmation/violation (see SOM Table 3). Unfortunately, we did not have enough condition-specific data to address whether prior information moderated these relations.

It is possible that our experimental paradigm was responsible for the observed null interactions. That is, after participants saw the first few stimuli, they would have become aware that they would have to continually make predictions about future behavior or future mental states based on both prior behavior and prior mental states. Once this awareness set in, it is possible (even likely) that participants were generating future behavior and future mental state predictions for each stimulus. This could have led to non-interactions between prior and future information on unexpectedness-related neural activity. Therefore, although our experimental paradigm was designed to address potential alternative explanations for prior research (i.e., Dungan et al., 2016), it may have fundamentally altered the psychological experience that we intended to study. Moreover, perhaps our paradigm exaggerated the effect of expectation violations. Specifically, in the real world, people may not engage in explicit predictions in the way that our participants did. Therefore, perhaps prompting explicit predictions led to artificial but equally strong predictions across information types, leading to the observed null interactions.

We opted for the explicit prediction methodology for two reasons: first to ensure that participants indeed engaged in prediction, and second to constrain the kinds of predictions made so that different participants would not be making different predictions. However, even though we attempted to control the experimental environment and constrain the types of predictions made as much as possible (i.e., behavior vs mental state), participants were likely to spontaneously generate mental state inferences regardless of the information presented. The reason for this, we propose, is that our experimental task (i.e., social prediction) may fundamentally rest on mental state inferences, as trying to predict a target's future mental states or behavior requires inferring the beliefs of the target. As an additional example of an experimental task that supports this point, prior research shows that people engage in spontaneous mental state inference when making moral judgments (Young & Saxe, 2009), as making moral judgments demands inferring the beliefs and intentions of a transgressor. This possibility results in two falsifiable predictions. Specifically, if participants spontaneously generate mental state inferences, then neural activity should have been similar for behavior and mental state stimuli immediately following the Initial Info segment. Additionally, participants should be equally good at choosing "correct" mental state outcomes during the Prediction segment, regardless of the Initial Info type. Both predictions were supported in our data. First, in the Initial Info segment, ToM recruitment was similar in magnitude for behavioral versus mental state information. Second, in the Prediction segment, participants were similarly likely to make the "correct" mental state prediction following Initial Info that was behavioral versus mental state (i.e., being correct on 88% vs 87% of trials, respectively). Therefore, our data suggest one of two theoretically plausible possibilities: ToM regions flexibly incorporate any available social information in order to make predictions and monitor their violation/confirmation, or, regardless of what type of social information is perceived, ToM regions spontaneously generate mental state inferences to use for prediction and then monitor their violation/conformation. Consequently, future, more targeted, research is needed to better understand if and when prior behavior (or mental state) information leads to stronger predictions and its consequences on prediction error.

Last, there have been continued calls to communicate "constraints on generalizability" in psychology and neuroscience research (Simons et al., 2017; Yarkoni, 2020). In addition to the above limitations, it is unclear if our recruited fMRI participants are representative of most people. Recent work suggests that fMRI research suffers from generalizability issues. Specifically, fMRI samples tend to be lower in trait anxiety than behavioral samples (Charpentier et al., 2021), suggesting a self-selection bias. Since past research has linked anxiety to ToM abilities (Washburn et al., 2016) and difficulty in understanding/completing ToM tasks (Lenton-Brym et al., 2018), future research on social prediction error would benefit from considering the role of anxiety and other individual differences.

Conclusion

The current research found that brain regions implicated in theory of mind (ToM: DMPFC, LTPJ, and RTPJ) are especially responsive to an agent's unexpected behavior or mental states based on knowledge of their prior behavior or mental states. These findings also suggest that ToM regions may not discriminate in their sensitivity to expectedness based on information type, though additional research is needed to conclusively provide evidence for or against this possibility. Overall, these findings replicate recent research consistent with a predictive coding account of the neural computations underlying ToM (Koster-Hale & Saxe, 2013), and lay the foundation for future research investigating when, how, and for whom certain kinds of prior social knowledge give rise to robust predictions and therefore shape prediction error signals.

Acknowledgements

We thank current and former members of the Morality Lab at Boston College for feedback on an original version of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the John Templeton Foundation and the National Science Foundation

ORCID

Ryan M. McManus http://orcid.org/0000-0002-1793-1595

References

Bach, P., & Schenke, K. (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

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328. https://doi.org/10.3389/fpsyg.2013.00328

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Reinhold, K., Vasishth, S., & Baayen, R. H. (2018). Parsimonious mixed models. 1506.04967v2.pdf (arxiv.org)

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Berg, M. K., Kitayama, S., & Kross, E. (2021). How relationships bias moral reasoning: Neural and self-report evidence. *Journal of Experimental Social Psychology*, *95*, 104156. https://doi.org/10.1016/j.jesp.2021.104156

Brambilla, M., Carraro, L., Castelli, L., & Sacchi, S. (2019). Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology*, *82*, 64–73. https://doi.org/10.1016/j.jesp.2019.01.003

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), Article 9. https://doi.org/10.5334/joc.10

Charpentier, C. J., Faulkner, P., Pool, E. R., Ly, V., Tollenaar, M. S., Kluen, L. M., Fransen, A., Yamamori, Y., Lally, N., Mkrtchian, A., Valton, V., Huys, Q. J. M., Sarigiannidis, I., Morrow, K. A., Krenz, V., Kalbe, F., Cremer, A., Zerbes, G., Kausche, F. M. … O'Doherty, J. P. (2021). How representative are neuroimaging samples? Large-scale evidence for trait anxiety differences between fMRI and behaviour-only research participants. *Social Cognitive and Affective Neuroscience*, *16* (10), 1057–1070. https://doi.org/10.1093/scan/nsab057

Chuang, Y. (2020). *label4MRI: MRI-labeling*.

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

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *The Neuroscientist*, *13*(6), 580–593. https://doi.org/10.1177/1073858407304654

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.

Dungan, J. A., Stepanovic, M., & Young, L. (2016). Theory of mind for processing unexpected events across contexts. *Social Cognitive and Affective Neuroscience*, *11*(8), 1183–1192. https://doi.org/10.1093/scan/nsw032

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMriprep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behavior*, *3*(5), 426–435. https://doi.org/10.1038/s41562-019-0590-x

Ferguson, M. J., & Mann, T. C. (2017). Reversing implicit first impressions through reinterpretation after a two-day delay. *Journal of Experimental Social Psychology*, *68*, 122–127. https://doi.org/10.1016/j.jesp.2016.06.004

Fiorilla, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898–1902. https://doi.org/10.1126/science.1077349

Heil, L., Colizoli, O., Harstra, E., Kwisthout, J., van Pelt, S., van Rooij, I., & Bekkering, H. (2019). Processing of predictions errors in mentalizing areas. *Journal of Cognitive Neuroscience*, *31*(6), 900–912. https://doi.org/10.1162/jocn_a_01381

Henrich, J. (2009). The evolution of costly displays, cooperation and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior*, *30*(4), 244–260. https://doi.org/10.1016/j.evolhumbehav.2009.03.005

Hughes, B. L., Ambady, N., & Zaki, J. (2017). Trusting outgroup, but not ingroup members, requires control: Neural and behavioral evidence. *Social Cognitive Affective Neuroscience*, *12*(3), 372–381. https://doi.org/10.1093/scan/nsw139

Hughes, B. L., Zaki, J., & Ambady, N. (2017). Motivation alters impression formation and related neural systems. *Social Cognitive Affective Neuroscience*, *12*(1), 49–60. https://doi.org/10.1093/scan/nsw147

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality & Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

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. https://doi.org/10.1093/cercor/bhaa263

Kim, M., Park, B., & Young, L. (2020). The psychology of motivated versus rational impression updating. *Trends in Cognitive Sciences*, *24*(2), 101–111. https://doi.org/10.1016/j.tics.2019.12.001

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

Kraft-Todd, G. T., Bollinger, B., Gillingham, K., Lamp, S., & Rand, D. G. (2018). Credibility-enhancing displays promote the provision of non-normative public goods. *Nature*, *563* (7730), 245–248. https://doi.org/10.1038/s41586-018-0647-4

Lenton-Brym, A. P., Moscovitch, D. A., Vidovic, V., Nilsen, E., & Friedman, O. (2018). Theory of mind ability in high socially anxious individuals. *Anxiety, Stress & Coping*, *31*(5), 487–499. https://doi.org/10.1080/10615806.2018.1483021

Marshall, J., Gollwitzer, A., Mermin-Bunnell, N., Shinomiya, M., Retelsdorf, J., & Bloom, P. (2022). How development and culture shape intuitions about prosocial obligations. *Journal of Experimental Psychology: General*, 151(8), 1866–1882.

Marshall, J., Wynn, K., & Bloom, P. (2020). Do children and adults take social relationship into account when evaluating other peoples' actions?. *Child Development*, *91*, 1395–1835.

McManus, R. M., Kleiman-Weiner, M., & Young, L. (2020). What we owe to family: The impact of special obligations on moral judgment. *Psychological Science*, *31*(3), 227–242. https://doi.org/10.1177/0956797619900321

McManus, R. M., Mason, J. E., & Young, L. (2021). Re-examining the role of family relationships in structuring perceived helping obligations, and their impact on moral evaluation. *Journal of Experimental Social Psychology*, *96*, 104182. https://doi.org/10.1016/j.jesp.2021.104182

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage*, *59*(3), 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018

Rolls, E. T., Huang, C., Lin, C., Feng, J., & Joliot, M. (2020). Automated anatomical labelling atlas 3. *NeuroImage*, *206*, 116189. https://doi.org/10.1016/j.neuroimage.2019.116189

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

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. *New Methods in Cognitive Psychology*, 4–31. https://doi.org/10.4324/9780429318405-2

Tamir, D. I., & Thornton, M. (2018). Modeling the predictive social mind. *Trends in Cognitive Science*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

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. https://doi.org/10.1016/j.neuropsychologia.2020.107475

Theriault, J. E., Young, L., & Barrett, L. F. (2021). The sense of should: A biologically-based framework for modeling social pressure. *Physics of Life Reviews*, *36*, 100–136. https://doi.org/10.1016/j.plrev.2020.01.004

Wacongne, C., Changeux, J. P., & Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience*, *32*(11), 3665–3678. https://doi.org/10.1523/JNEUROSCI.5003-11.2012

Washburn, D., Wilson, G., Roes, M., Rnic, K., & Harkness, K. L. (2016). Theory of mind in social anxiety disorder, depression, and comorbid conditions. *Journal of Anxiety Disorders*, *37*, 71–77. https://doi.org/10.1016/j.janxdis.2015.11.004

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. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, *45*. https://doi.org/10.1017/S0140525X20001685

Young, L., & Saxe, R. (2009). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Nueroscience*, *21*(7), 1396–1405. https://doi.org/10.1162/jocn.2009.21137