fMRIprep preprocessing details

All fMRI data were preprocessed using fMRIPrep version 1.3.2 (Esteban et al., 2019), a Nipype based tool (Gorgolewski et al., 2011). Each T1w (T1- weighted) volume was corrected for intensity non-uniformity using N4BiasFieldCorrection v2.1.0 (Tustison et al., 2010) and skull-stripped (with the OASIS template) using antsBrainExtraction.sh v2.1.0 (Avants et al., 2008). Brain surfaces were reconstructed using recon-all from FreeSurfer v6.0.1 (Dale, Fischlm Sereno, 2009), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (Klein et al., 2017).

Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009) was performed through nonlinear registration with the antsRegistration tool of ANTs v2.1.0, using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed with FSL on the brainextracted T1w (Zhang, Brady & Smith, 2001). Functional data was slice-time corrected using 3dTshift from AFNI v16.2.07 (Cox, 1996) and motion corrected using FSL's mcflirt (Jenkinson et al., 2002). This was followed by co-registration to the corresponding T1w using boundary-based registration with 9 degrees of freedom (Greve & Fischl, 2009), using bbregister (FreeSurfer v6.0.1). Motion correcting transformations, BOLD-to-T1w transformation and T1w-to-template (MNI) warp were concatenated and applied in a single step using antsApplyTransforms (ANTs v2.1.0) using Lanczos interpolation. Images were smoothed in SPM with a 8mm-FWHM Gaussian kernel.

Physiological noise regressors were extracted applying CompCor (Behzadi, Restom, Liau, Liu, 2007). Principal components were estimated for the anatomical CompCor variant (aCompCor). A mask to exclude signal with cortical origin was obtained by eroding the brain mask, ensuring it only contained subcortical structures. For aCompCor, six components were calculated within the intersection of the subcortical mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run. Framewise displacement was calculated for each functional run using the implementation of Nipype (Power et al., 2014).

Many internal operations of FMRIPREP use Nilearn (Abraham et al., 2014), principally within the BOLD-processing workflow. For more details of the pipeline see https://fmriprep.readthedocs.io/en/latest/workflows.html.

Figure S1. Theory of Mind ROIs identified using a functional localizer. ROIs were defined as all voxels within a 9-mm radius of the peak voxel that passed threshold in the contrast *'false belief > false photo'* ($p < 0.001$, uncorrected; $k > 16$). Viewed at x = 0, y = -58, z = 28.

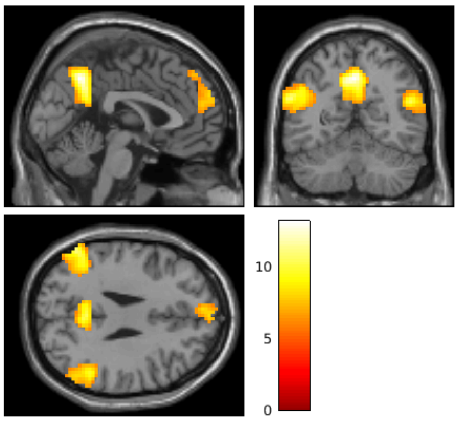

Table S1. Peak coordinates of ToM ROIs identified using a functional localizer task.

Region	x	y	z	t value	# voxels	N
DMPFC	-6	53	28	8.09 (subpeak)	1519	21
RTPJ	51	-55	22	11.28 (subpeak)	2155	27
LTPJ	-57	-58	25	10.87	2098	27
PC	0	-58	43	13.16	932	27

Figure S2. Regions tracking magnitude of behavioral updating (LVLPFC and LIFG). Viewed at x = -48, y = 26, z = -11.

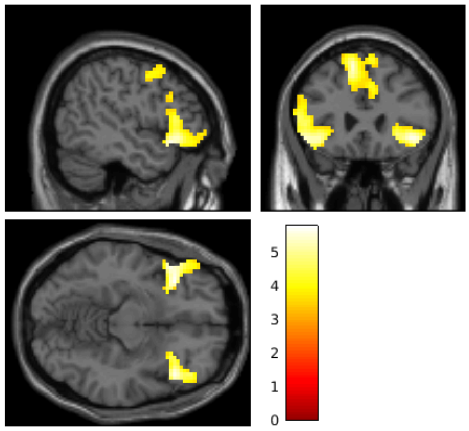

Table S2. Neural activity by position and sequence type. Looking within sequence type, we tested whether ToM ROIs exhibited greater activity to post-switch behaviors than to pre-switch behaviors. †: no by-item random intercepts. ‡: no by-subject random intercepts.

	DMPFC	RTPJ	LTPJ	PC
strong neg→pos	ns $\chi^2(1) = 1.291, p = 0.257$	**post > pre** †, ‡ $F(1,518) = 4.562, p = 0.033$	ns $\chi^2(1) = 1.319, p = 0.251$	ns ‡ $\chi^2(1) = 0.396, p = 0.529$
weak neg→pos	ns ‡ $\chi^2(1) = 2.647, p = 0.104$	ns $\chi^2(1) = 0.310, p = 0.578$	**pre > post** † $\chi^2(1) = 15.628, p < 0.001$	ns †, ‡ $F(1,517) = 0.13, p = 0.719$
strong pos→neg	**post > pre** $\chi^2(1) = 39.867, p < 0.001$	**post > pre** † $\chi^2(1) = 28.792, p < 0.001$	**post > pre** ‡ $\chi^2(1) = 20.557, p < 0.001$	**post > pre** † $\chi^2(1) = 9.738, p = 0.002$
weak pos→neg	**post > pre** † $\chi^2(1) = 23.456, p < 0.001$	ns † $\chi^2(1) = 0.616, p = 0.433$	**post > pre** ‡ $\chi^2(1) = 23.142, p < 0.001$	ns $\chi^2(1) = 1.963, p = 0.161$

Table S3. Neural activity for non-updating sequences. To test whether ToM ROIs are exhibiting a mere time effect — that is, greater activity to behaviors presented later on in the sequence — we compared (1) activity to the first two behaviors and activity to the middle two behaviors, in strong prior sequences; (2) activity to the middle two behaviors and activity to the last two behaviors, in control sequences. RTPJ and PC (marginal) exhibited greater activity to the last two behaviors vs. the middle two behaviors in Positive Control sequences; otherwise, ToM ROIs did not show increases in activity over time for non-updating sequences.

†: no by-item random intercepts. ‡: no by-subject random intercepts.

	DMPFC	**RTPJ**	**LTPJ**	**PC**
strong neg→pos	**pre < post** †, ‡ $F(1,399) = 7.042$, $p = 0.008$	**pre < post** $\chi^2(1) = 4.398$, $p = 0.036$	**pre < post** $\chi^2(1) = 11.068$, $p = 0.001$	ns $\chi^2(1) = 2.501$, $p = 0.114$
strong pos→neg	ns † $\chi^2(1) = 2.384$, $p = 0.123$	ns †, ‡ $F(1, 520) = 0.884$, $p = 0.348$	ns † $\chi^2(1) = 1.835$, $p = 0.176$	ns † $\chi^2(1) = 0.383$, $p = 0.536$
neg control	ns $\chi^2(1) = 0.091$, $p = 0.763$	ns †, ‡ $F(1, 252) = 0.015$, $p = 0.904$	ns $\chi^2(1) = 0.282$, $p = 0.596$	ns $\chi^2(1) = 0.131$, $p = 0.717$
pos control	ns $\chi^2(1) = 0.626$, $p = 0.429$	**post > pre** $\chi^2(1) = 12.214$, $p < 0.001$	ns $\chi^2(1) = 0.299$, $p = 0.584$	**post > pre** † $\chi^2(1) = 3.453$, $p = 0.063$

Table S4. Updating vs. mere time effect. To test for updating-related activity in response to meaningful changes in behavior, above and beyond a mere time effect (i.e., greater activity to behaviors presented later on in the sequence), we compared the changes in activity for expectation-violation sequences with analogous changes in activity for control sequences. For example, we compared the neural updating measure for Weak Positive-to-Negative trials with an analogous measure for Positive Control trials: average of the middle two behaviors minus average of the first two behaviors. If changes in PSC in an ROI reflect updating in response to meaningful changes in behavior, then we would expect to see a greater change in activity on updating trials compared to non-updating trials, when ordinal position is held constant. These analyses suggest that increases in PSC in DMPFC and LTPJ, for pos→neg sequences, reflect meaningful updating-related activity above and beyond the effect of time found in non-updating sequences.

†: no by-item random intercepts. ‡: no by-subject random intercepts.

	DMPFC	**RTPJ**	**LTPJ**	**PC**
strong neg→pos		ns †, ‡ $F(1,385) = 1.085, p = 0.298$		
weak neg→pos				
strong pos→neg	**updating > non-updating** † $\chi^2(1) = 9.705, p = 0.002$	ns †, ‡ $F(1,383) = 0.031, p = 0.86$	**updating > non-updating** $\chi^2(1) = 3.302, p = .069$	ns ‡ $\chi^2(1) = 0.19, p = 0.663$
weak pos→neg	**updating > non-updating** $\chi^2(1) = 3.205, p = 0.073$		**updating > non-updating** † $\chi^2(1) = 6.78, p = .009$	

Neural updating measure for expectation-violation sequences

As an alternative way to examine neural activity associated with impression updating, we also computed a *neural updating measure* for expectation-violation trials (see **Methods**). Neural updating for each ToM ROI was fit using linear mixed effects models, allowing for by-subject and by-item random intercepts. For some of the models, the random effects structure was simplified to address singular fit or non-convergence. The model for LTPJ was simplified by removing by-subject random intercepts. PC was simplified by removing by-item random intercepts. RTPJ was simplified by removing both by-subject and by-item random intercepts, resulting in a regular linear model. No simplification was necessary for DMPFC.

DMPFC: There was a significant main effect of update direction (pos→neg > neg→pos, $\chi^2(1) = 27.883$, $p < 0.001$), and a significant main effect of prior strength (strong > weak, $\chi^2(1) = 4.678$, $p = 0.031$).

RTPJ: There was a marginal main effect of update direction (pos→neg > neg→pos, $F(1) = 2.948$, $p = 0.086$), and a significant main effect of prior strength (strong > weak, $F(1) = 7.733$, $p = 0.006$).

LTPJ: There was a significant main effect of update direction (pos→neg > neg→pos, $\chi^2(1) = 34.415$, $p < 0.001$), and no main effect of prior strength ($\chi^2(1) = 1.614$, $p = 0.204$).

PC: There was a marginal main effect of update direction (pos→neg > neg→pos, $\chi^2(1) = 3.257$, $p = 0.071$), and no main effect of prior strength ($\chi^2(1) = 1.071$, $p = 0.301$).

Brain-behavior analyses

Looking within sequence type, neural updating was associated with behavioral updating for: Strong Positive-to-Negative sequences in RTPJ (beta = −0.08, $\chi^2(1) = 3.861$, $p = 0.049$); Weak Negative-to-Positive sequences in LTPJ (beta = 0.079, $\chi^2(1) = 4.597$, $p = 0.030$); Strong Positive-to-Negative sequences in PC (beta = −0.092, $\chi^2(1) = 6.334$, $p = 0.012$); and Weak Negative-to-Positive sequences in IFG (beta = 0.088, $\chi^2(1) = 7.127$, $p = 0.008$).

Table S5. Additional encoding model analyses (see **Table 2**). We sought to identify regions across the whole brain that respond to: positive vs. negative valence; negative vs. positive valence; and any change in valence from the previous behavior to the current behavior. All regions survived cluster-level correction (FWE, $p < 0.05$).

Region name	x	y	z	t value	# voxels	ToM	VLPFC/IFG
Model C: preferential response to positive valence							
right lingual gyrus	9	-79	-5	9.48	230		
Model C: preferential response to negative valence							
left middle temporal gyrus	-60	-55	16	7.56	1018	LTPJ	
left superior frontal gyrus (medial)	-6	53	28	6.57	1125	DMPFC	
left inferior frontal gyrus (orbital part)	-45	29	-8	6.54	760		LVLPFC, LIFG
right inferior occipital gyrus	39	-94	-2	6.25	118		
left lingual gyrus	-6	-82	-5	6.02	174		
right middle temporal gyrus	54	-34	-2	5.62	174		
right superior parietal gyrus	18	-67	58	4.88	162		
Model E: any valence change occurred from previous behavior to current behavior							
right angular gyrus	51	-55	37	5.68	103		
right insula	39	26	-5	4.95	152		
left inferior frontal gyrus (triangular part)	-54	26	7	4.82	156		LIFG

Whole-brain analyses

In addition to the encoding models, we ran condition-based GLM analyses, which identify regions across the whole brain that are sensitive to different types of expectation violations. We also sought to identify regions sensitive to different types of updating.

Participants' preprocessed images were analyzed in SPM12. For each participant, images from all runs were concatenated to produce a single-session model, to allow for analysis of the control sequences (as only one of each control sequence type was presented on each run). The SPM function "spm_fmri_concatenate" was used to adjust high-pass-filtering and add session regressors to the concatenated model.

Each behavior position in each sequence type was treated as a single condition, resulting in 36 (6 positions * 6 sequence types) such conditions.

Thirty-six regressors of interest were convolved with a canonical hemodynamic response function; six aCompCor components (Muschelli et al. 2014) were also included as nuisance regressors. Parameter estimates were generated in each voxel for all conditions. To correct for multiple comparisons, images from group-level analyses were subjected to a voxel-wise threshold of $p < 0.001$ (uncorrected) and a cluster extent threshold ensuring $p < 0.05$ (FWE-corrected).

We were primarily interested in identifying brain regions displaying: (1) a main effect of updating (i.e., an increase in activity from pre-switch to post-switch behaviors), (2) a main effect of prior strength on updating, and (3) a main effect of update direction on updating.

Table S6. Regions that respond more to post-switch behaviors than pre-switch behaviors, collapsing across sequence type.

Region name	x	y	z	t value	# voxels	ToM	VLPFC/IFG
post>pre							
right superior frontal gyrus	18	23	58	11	7639	DMPFC	
right caudate nucleus	9	11	10	8.68	752		
left angular gyrus	-45	-61	49	7.22	682		
precuneus	6	-67	46	6.8	482	PC	
right middle temporal gyrus	57	-34	-2	7.25	1215	RTPJ	
left middle temporal gyrus	-63	-28	-11	5.4	230		

Table S7. Regions that respond more to post-switch behaviors, by sequence type.

Region name	x	y	z	t value	# voxels	ToM	VLPFC/IFG
Strong Pos→Neg post>pre							
left superior frontal gyrus	0	41	43	13.36	8473	DMPFC	
left middle temporal gyrus	-51	-34	1	7.96	1165	LTPJ	
right middle temporal gyrus	63	-19	-14	7.94	1896	RTPJ	
Strong Neg→Pos post>pre							
right middle frontal gyrus	36	53	16	6.94	2159		
right middle cingulate	9	-25	25	6.59	711		
right inferior parietal gyrus	51	-55	46	5.48	297		
right inferior frontal gyrus (orbital part)	42	20	-14	5.44	212		RVLPFC
Weak Pos→Neg post>pre							
left SMA	-9	17	67	12.75	2404		
left inferior frontal gyrus (orbital part)	-42	23	-14	8.21	1466		LVLPFC, LIFG
right inferior frontal gyrus (orbital part)	48	26	-11	6.52	640		RVLPFC
left caudate nucleus	-12	8	13	8.01	332		
left middle temporal gyrus	-51	-31	1	7.26	874	LTPJ	
right middle temporal gyrus	60	-34	-2	6.99	369		
Weak Neg→Pos post>pre							
precuneus	3	-70	43	5.58	215		
right angular gyrus	45	-58	49	4.64	143		

Table S8. Regions sensitive to different types of updating.

Region name	x	y	z	t value	# voxels	ToM	VLPFC/IFG
Pos→Neg updating > Neg→Pos updating (voxelwise threshold .0001)							
left SMA	-9	17	67	13.4	1734		
left inferior frontal gyrus (orbital part)	-42	23	-14	12.37	1355		LVLPFC, LIFG
right inferior frontal gyrus (orbital part)	48	26	-8	6.74	107		RVLPFC
left middle temporal gyrus	-60	-55	10	9.27	770	LTPJ	
right middle temporal gyrus	66	-49	10	5.52	26		
right caudate	12	2	10	7.65	148		
left calcarine fissure	-6	-85	-5	7.16	158		
right superior frontal gyrus	24	-1	61	6.96	114	DMPFC	
right superior parietal gyrus	18	-67	58	6.8	156		
right inferior occipital gyrus	48	26	-8	6.74	107		
right middle frontal gyrus	48	14	52	5.27	59		
right inferior frontal gyrus (triangular part)	57	26	19	4.93	49		RIFG
right middle temporal gyrus	51	-31	-2	6.05	125		
right inferior occipital gyrus	39	-94	-2	6.53	35		
Neg→Pos updating > Pos→Neg updating							
right lingual gyrus	12	-82	-5	12.56	323		
left calcarine fissure	-12	-61	19	5.65	108		
left superior parietal gyrus	-27	-46	76	4.93	252		
right superior temporal gyrus	57	-22	16	4.1	106		
Strong prior updating > Weak prior updating							
right anterior cingulate	9	26	19	5.65	526		
right insula	36	23	1	4.76	145		
Weak prior updating > Strong prior updating							
right postcentral gyrus	21	-25	58	5.54	209		
right superior temporal gyrus	63	-4	4	4.96	139		