OXFORD

Social Cognitive and Affective Neuroscience, 2016, 1183–1192

doi: 10.1093/scan/nsw032
Advance Access Publication Date: 12 March 2016
Original article

Theory of mind for processing unexpected events across contexts

James A. Dungan, Michael Stepanovic, and Liane Young

Department of Psychology, Boston College, McGuinn 300, 140 Commonwealth Ave., Chestnut Hill, MA 02467, USA

Correspondence should be addressed to James A. Dungan, Department of Psychology, Boston College, McGuinn 300, 140 Commonwealth Ave., Chestnut Hill, MA 02467, USA. E-mail: james.dungan@bc.edu.

Abstract

Theory of mind, or mental state reasoning, may be particularly useful for making sense of unexpected events. Here, we investigated unexpected behavior across both social and non-social contexts in order to characterize the precise role of theory of mind in processing unexpected events. We used functional magnetic resonance imaging to examine how people respond to unexpected outcomes when initial expectations were based on (i) an object's prior behavior, (ii) an agent's prior behavior and (iii) an agent's mental states. Consistent with prior work, brain regions for theory of mind were preferentially recruited when people first formed expectations about social agents *vs* non-social objects. Critically, unexpected *vs* expected outcomes elicited greater activity in dorsomedial prefrontal cortex, which also discriminated in its spatial pattern of activity between unexpected and expected outcomes for social events. In contrast, social *vs* non-social events elicited greater activity in precuneus across both expected and unexpected outcomes. Finally, given prior information about an agent's behavior, unexpected *vs* expected outcomes elicited an especially robust response in right temporoparietal junction, and the magnitude of this difference across participants correlated negatively with autistic-like traits. Together, these findings illuminate the distinct contributions of brain regions for theory of mind for processing unexpected events across contexts.

Key words: theory of mind; social cognition; prediction error; expectations; autism

Introduction

People are remarkable at making sense of the unpredictable world around them. Often, this requires reasoning about the hidden, internal causes behind observable behavior, such as a person's beliefs, goals and intentions. Notably, extensive prior work has shown that reasoning about these mental states, known as theory of mind (ToM), reliably recruits a particular network of brain regions including right and left temporoparietal junction (RTPJ, LTPJ), precuneus (PC) and dorsomedial prefrontal cortex (DMPFC; Fletcher *et al.*, 1995; Saxe and Kanwisher, 2003; Gobbini *et al.*, 2007).

One context where ToM may be particularly useful is when events are unexpected. Whether it is a car that breaks down or a friend who makes an uncharacteristic movie choice, unexpected events trigger learning by signaling when people must update their representations of expected behavior (Niv and Schoenbaum, 2008). Investigating the role of ToM in forming expectations and responding to unexpected outcomes is thus crucial for understanding how people reason and learn about their surroundings.

Prior work has investigated neural responses to unexpected events primarily in non-social contexts (Schultz et al., 1997; Balleine *et al.*, 2007; Herry *et al.*, 2007); however, people make predictions about social agents as well. As is the case for objects, people often base their expectations about social agents on how they behaved in the past. For example, one might expect John to choose comedies over dramas after observing his past movie choices. However, a key difference between agents and objects is that people can also form expectations about agents based on explicit or inferred information about agents'

Received: 4 September 2015; **Revised:** 18 January 2016; **Accepted:** 8 March 2016

mental states. One might form the same expectation that John will choose comedies over dramas without ever having seen a movie with John; instead, one's expectation might be based on the prior knowledge that John prefers light-hearted humor to intense situations. Here, our aim was therefore to investigate ToM across different contexts, both social and non-social, to characterize the precise role of ToM in processing unexpected events.

Existing literature provides mixed evidence as to whether and how unexpected events elicit activity in brain regions for ToM. Studies investigating action understanding have typically found that unusual or unexpected actions recruit ToM regions (for a meta-analysis, see Van Overwalle and Baetens, 2009), such as when a man turns on a light switch using his knee instead of his hand (Brass *et al.*, 2007). However, other studies point to important boundary conditions on this effect. For example, an agent's unexpected behavior may elicit greater activity in ToM regions only when participants receive explicit instructions to infer the intentions underlying an agent's unexpected behavior (de Lange *et al.*, 2008; Ampe *et al.*, 2014) or when a rich social context makes the unusualness of the unexpected behavior more salient (Brass *et al.*, 2007; Van Overwalle and Baetens, 2009; Ampe *et al.*, 2014).

Work directly targeting unexpected beliefs, as opposed to behaviors, also provides mixed evidence. In one study, participants read stories about an object's physical states or an agent's mental states that were either expected (e.g. plants will flower if watered; Maya thinks plants will flower if watered) or unexpected (e.g. plants will burst into flames if watered; Maya thinks plants will burst into flames if watered; Young *et al.*, 2010). Within the ToM network, RTPJ, LTPJ and PC were preferentially recruited for mental states *vs* physical states, but these regions did not discriminate between expected and unexpected outcomes in either the social (mental) or non-social (physical) context (see also, Jenkins and Mitchell, 2010). However, outcomes in this study were unexpected based on participants' own existing knowledge of the world (e.g. plants do not combust when watered), not on information about specific agents' past actions or mental states. In contrast, a number of other studies reveal recruitment of the ToM network when unexpected events reflect internal inconsistencies within the mind of a single agent (Saxe and Wexler, 2005; Cloutier *et al.*, 2011; Ma *et al.*, 2012; see also, Koster-Hale and Saxe, 2013). For example, politicians whose beliefs are incongruent or unexpected based on their political party (Democrat or Republican) elicit enhanced activity within the ToM network, compared with politicians with congruent beliefs (Cloutier *et al.*, 2011). Thus, ToM regions may in fact be recruited for information that is unexpected given prior information about a specific agent's mind.

These discrepant results leave open the question of what role the ToM network plays in processing unexpected events. Although studies have indicated certain boundary conditions on ToM recruitment for unexpected events (Van Overwalle and Baetens, 2009; Ampe *et al.*, 2014), the variability of effects across studies highlights the importance of investigating different contexts within the same paradigm in order to characterize when unexpected events elicit activity in brain regions for ToM. Furthermore, as noted above, expectations about agents can be based on information about agents' past behavior, or on information about agents' mental states, and prior work has neglected this difference. Extending the work discussed above, we directly compare ToM across different contexts, allowing us to test several key hypotheses about the precise role of ToM in processing unexpected events.

One possibility is that behaviors that are unexpected based on an agent's mental states elicit the highest response in the ToM network. Of ToM regions, RTPJ in particular shows the most selective responding to mental state information compared with other socially relevant information about an agent's physical appearance (Saxe and Kanwisher, 2003) or social background (Saxe and Powell, 2006; Kobayashi *et al.*, 2007). Furthermore, in contrast to earlier work showing overlapping activation in RTPJ for attentional reorienting and ToM (Mitchell, 2008), more recent findings demonstrate that these activations are separable (Decety and Lamm, 2007; Scholz *et al.*, 2009) and support the selectivity of RTPJ for mental state reasoning (Jenkins and Mitchell, 2010). These findings suggest that RTPJ, and the ToM network more generally, may respond uniquely to mental state information compared with other information about an agent, including perhaps the agent's behavioral history.

An alternative possibility is that behaviors that are unexpected based on an agent's behavioral history elicit the highest response in ToM regions. People are rarely given explicit information about mental states in typical social interactions. Instead, ToM may be best suited for generating explanations for the internal, unobservable causes of observable behavior (Brass *et al.*, 2007; Van Overwalle and Baetens, 2009). In other words, although the ToM network may not process action information *per se*, it may be recruited when inferring or reasoning about the 'mental states' that motivate observable actions. In line with this possibility, imagining 'why' actions are performed, as opposed to imagining 'how' actions are performed, that is, reflecting on an agent's motive for performing an action, activates the ToM network, including RTPJ, PC and DMPFC (Spunt *et al.*, 2010). Descriptions of behavior and especially unexpected behavior may thus elicit spontaneous ToM (Young and Saxe, 2009).

To investigate the specific role of ToM regions in processing unexpected events across contexts, we presented participants with stories designed to elicit expectations based on three different types of background information: (i) non-social stories describing the behaviors of objects (object-behavior), (ii) social stories describing the behaviors of agents (agent-behavior) and (iii) social stories describing the mental states of agents (agent-mental). Based on background information, participants made predictions about story outcomes by answering multiple-choice questions. Critically, after participants formed expectations, we presented them with expected and unexpected outcomes across different trials, allowing us to test how unexpected events are processed in the ToM network across contexts: social and non-social, and based on information about either behavioral history or mental states.

Notably, although past research established participants' expectations using pre-existing stereotypes (e.g. of Republicans and Democrats; Cloutier *et al.*, 2011), or norms (e.g. moral norms; Ma *et al.*, 2012), the current design used stories describing novel objects and agents. This paradigm has the advantage of decreasing variability across participants in the extent to which different outcomes were unexpected, as expectations were determined primarily by information provided within the stimuli—not participants' own experiences or pre-existing knowledge or beliefs. As participants learn about the novel objects and agents featured in the stimuli, they must reconcile new outcome information with the background information they just read. We could therefore test the role of ToM when people flexibly process new information to build internally coherent representations of objects and agents.

Finally, recent theoretical work suggests that a core deficit in forming predictions and processing unexpected events explains many of the behavioral traits that characterize autism spectrum disorders (ASD; Lawson *et al.*, 2014; Sinha *et al.*, 2014; Van de Cruys *et al.*, 2014). Although the presence of traits associated with ASD predicts social deficits within the neurotypical population (Best *et al.*, 2008; Haffey *et al.*, 2013), little is known about how autistic-like traits affect responses to unexpected events across contexts. One possibility is that autistic-like traits affect how unexpected events are processed primarily in social contexts. Yet, although 'social' deficits may disrupt the ability to form expectations in 'social' contexts, individuals with ASD nevertheless show sensitivity to unexpected events in non-social contexts, demonstrating strong domain-general preferences for predictability and order (Baron-Cohen *et al.*, 2009). Here, we included a measure of traits associated with ASD to explore the possible relationship between autistic-like traits within a neurotypical sample and neural responses to unexpected events across both social and non-social contexts.

Materials and methods

Participants and procedures

Participants were 24 right-handed adults (age: $M = 25.7$, s.d. = 4.2; 12 female) recruited from the Greater Boston Area. 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. Participants were scanned on a 3T Siemens Tim Trio functional magnetic resonance imaging (fMRI) scanner (at the Harvard Center for Brain Science, Cambridge, MA) using thirty-six 3 × 3 × 3 mm near-axial slices (0.54 mm gap) covering the whole brain. Standard gradient echo planar imaging (EPI) procedures were used (TR = 2 s, 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 (TR = 2530 ms; TE = 1.64 ms; FA = 7°; 1 mm isotropic voxels; 0.5 mm gap between slices; FOV = 256 × 256).

Stimuli consisted of 60 stories (see supplementary material), divided into three story types that differed in terms of the kind of information provided in the Background segment. Twenty 'object-behavior' stories described the 'non-social' behaviors or features of objects or places (e.g. trains, the rainforest). 'Social' stories, which described human agents, were divided into 20 'agent-behavior' stories and 20 'agent-mental' stories. Agent-behavior stories described a human agent's physical behavior, whereas agent-mental stories described a human agent's beliefs and desires. Independent ratings from a group of online participants confirmed that stimuli differed on the dimensions associated with each condition (see supplementary material).

In the scanner, each story was presented in three sequential segments: Background, Question and Outcome. During the Background segment, participants read background information about an agent or other entity, intended to create an expectation about the future behavior of the story's subject. During the Question segment, participants were presented with a multiple-choice question asking for a prediction about the story's outcome. Four options were provided: one that followed from the background (expected), and three others that were unlikely to occur given the background (unexpected). Finally, during the Outcome segment, participants were presented with the outcome of the story. On half the trials, the outcome was unexpected (Figure 1). Critically, story outcomes always concerned the future behavior of objects or agents. In other words, in contrast to the different kinds of information presented in each condition during the Background Segment (e.g. past behavior *vs* mental states), outcomes across the three conditions contained similar language, and importantly, none contained any explicit mental state information. Crossing the dimension of story type (object-behavior, agent-behavior, agent-mental) with the dimension of expectedness (expected, unexpected) thus yielded six conditions of interest for the Outcome segment.

All 60 stories were presented in a pseudo-randomized order in white font on a black background via an Apple Macbook Pro running Matlab 2012b with Psychophysics Toolbox. Participants were instructed to read each story and then to answer the multiple-choice questions using a button-box. The background was presented on-screen for 12 s, the question for 8 s, and the outcome for 6 s. In order to analyze Background and Outcome segments separately, 2–6 s of jittered fixation was included between each story segment. Stimulus presentation was divided into five equal runs (12 stimuli per run, 4 per story type) lasting ~7 min and 37 s each.

Participants also completed a 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; see http://saxelab.mit.edu/superloc.php for the task files). The task was presented in two 4.5 min runs, interleaved with the main experiment runs.

Following the scan session, participants rated the expectedness of each story's outcome on a 4-point scale (1 = totally unexpected, 4 = totally expected). Participants rated all 60 stories in the same order as presented in the scanner. All story segments (Background, Question, Outcome) were presented on the same screen. Participants also completed the Autism Quotient (AQ; Baron-Cohen *et al.*, 2001): a 50-item questionnaire measuring traits characteristic of ASD [$M = 14.58$, s.d. = 4.83, range: (8–28)]. Responses on the AQ showed moderate reliability ($\alpha = 0.64$), consistent with previous studies using the AQ (Ingersoll *et al.*, 2011; Ruzich *et al.*, 2015).

Data analyses

MRI data preprocessing and analyses were performed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) and custom software. Each participant's data were corrected for slice timing, realigned to the first EPI, normalized to Montreal Neurological Institute brain space, spatially smoothed using a Gaussian filter (full-width half-maximum = 8 mm kernel) and high-pass filtered (128 Hz). The experimental task was modeled using a boxcar regressor convolved with a canonical hemodynamic response function. The general linear model included movement parameters as nuisance regressors.

Whole-brain and regions of interest (ROIs) analyses were conducted. A whole-brain contrast of false-belief *vs* false-photograph stories in the ToM localizer (Dodell-Feder *et al.*, 2011) revealed ROIs that respond preferentially to mental states ($P < 0.001$, uncorrected, $k > 16$, value computed via 1000 iterations of a Monte Carlo simulation; Slotnick *et al.*, 2003). ROIs were selected for each participant individually and defined as contiguous voxels within a 9 mm radius of the peak voxel that passed contrast threshold. Within each ROI, the average percent signal change (PSC) relative to baseline [PSC = 100 × raw BOLD magnitude for (condition − fixation)/raw BOLD magnitude for

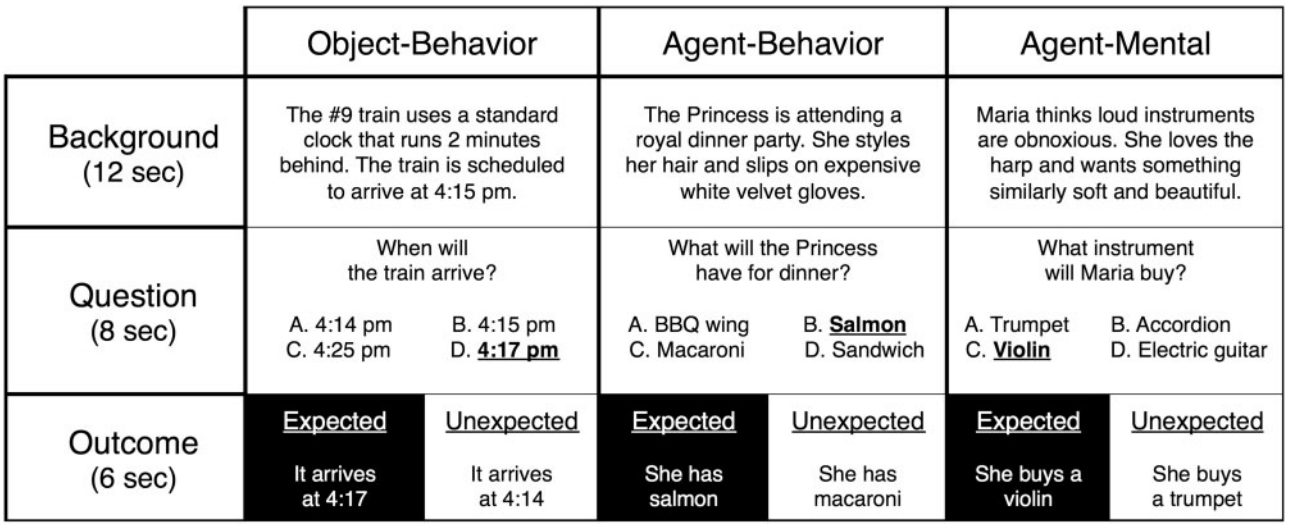

Fig. 1. Abbreviated examples of experimental stimuli. Expected answers to multiple-choice questions are in bold and underlined.

fixation] was calculated for each condition at each time point (averaging across all voxels in the ROI and all blocks of the same condition). Background and Outcome segments were modeled separately (the Question segment was not analyzed).

We also used ROI-based multi-voxel pattern analysis (MVPA) to examine the spatial pattern of activity across voxels within ToM ROIs. Pairs of conditions were contrasted (e.g. unexpected *vs* expected) by calculating a vector of mean centered beta values for each voxel within an ROI for each condition. We split data from 4 of the 5 runs into two equal sets of 2 runs and compared correlations 'within' *vs* 'between' conditions (Haxby *et al.*, 2001; Norman *et al.*, 2006). 'Within' correlations were the correlations of beta value vectors across sets of runs for 'one condition', whereas 'between' correlations were the correlations of beta value vectors across sets of runs for 'different conditions'. These correlations were calculated for all possible iterations of run combinations (2×2 out of 5 runs; 32 total iterations). Classification accuracy was calculated as the percentage of iterations where 'within' correlations exceeded 'between' correlations. Classification was deemed significant if accuracy exceeded chance (50%) in a one-tailed, one sample *t*-test.

Results

Behavioral results

When answering the multiple-choice questions in the scanner, participants chose the expected outcome 69.86% of the time [s.d. = 15.16; significantly above chance accuracy of 25%, $t(23) = 14.494$, $P < 0.001$; $d = 2.96$]. Missed responses accounted for 30% of the errors, consistent with past studies using similar designs (e.g. Ma *et al.*, 2012). Notably, a key feature of our paradigm was to present unexpected outcomes on 50% of trials, which presumably increased the error rate by introducing uncertainty (e.g. even participants choosing expected outcomes on 100% of trials would receive feedback that the story had a different outcome on 50% trials; see supplementary material for a breakdown of error rates across conditions).

Critically, despite any errors made during the Question segment, post-scan ratings indicated that participants' expectations conformed to our intended conditions in the Outcome segment (the primary focus of our analyses). A 3 (story type: object-behavior, agent-behavior, agent-mental) $\times$ 2 (expectedness: expected, unexpected) analysis of variance (ANOVA) revealed the predicted main effect of expectedness [$F(1,23) = 563.275$, $P < 0.001$, $\eta_p^2 = 0.961$] such that unexpected outcomes ($M = 1.57$, s.d. = 0.32) were rated as more unexpected than expected outcomes ($M = 3.44$, s.d. = 0.49) for all three story types (all P's < 0.001).[1] This difference held even when looking only at trials where participants did not choose the expected answer to the multiple-choice question (see supplementary material). Given this, and to maintain the same number of events in analyses across conditions, all trials were included in the subsequent analyses.

FMRI results: functional localizer

A whole-brain analysis of scenarios describing mental states contrasted with scenarios describing physical representations replicated previous findings (Saxe and Kanwisher, 2003), revealing an increased response in four brain regions within the ToM network: RTPJ, LTPJ, PC and DMPFC (Table 1). We localized these regions in the majority of participants: RTPJ (24/24 subjects), LTPJ (23/24), PC (21/24), DMPFC (19/24).

FMRI results: background segment

During the Background segment, we expected increased activity in ToM regions for the two social conditions (agent-behavior, agent-mental) compared with the non-social condition (object-behavior). As expected, a whole-brain random-effects analysis (voxel-wise threshold: $P < 0.001$, uncorrected; $k > 16$; cluster-wise threshold: $P < 0.05$, FWE-corrected) of social over non-social conditions revealed peak clusters in all four regions identified using the

1 Although we found no main effect of story type ($P > 0.80$), we did find an unpredicted story type $\times$ expectedness interaction [$F(2,46) = 18.609$, $P < 0.001$, $\eta_p^2 = 0.447$] driven by ratings of the non-social stories. Ratings were equal across social stories, but the difference between expected and unexpected outcomes was smaller for non-social stories in comparison (see supplementary material). Importantly, though, post-scan ratings were uncorrelated with average response magnitude and classification accuracy, suggesting that neural differences observed across story types (presented below) cannot be attributed to differences in expectedness across story types.

ToM localizer: RTPJ [54, −52, 28], LTPJ [−54, −61, 25], PC [6, −55, 37] and DMPFC [0, 53, 25] (Table 2). No clusters in ToM regions passed threshold in contrasts of the two social conditions (agent-behavior > agent-mental or agent-mental > agent-behavior). Contrasting agent-behavior stories with object-behavior stories revealed clusters near RTPJ [54, −61, 19] and PC [0, −52, 34]. Contrasting agent-mental stories with object-behavior stories revealed clusters near LTPJ [−42, −61, 19], PC [0, −55, 34] and DMPFC [3, 56, 25].

Using an ROI-based approach, we examined the average magnitude of response in ToM ROIs across the three story types. A 4 (ROI: RTPJ, LTPJ, PC, DMPFC) × 3 (story type: object-behavior, agent-behavior, agent-mental) repeated measures ANOVA revealed a main effect of story type [$F(2,34) = 46.681$, $P < 0.001$, η_p^2=0.733]. Social stories about an agent's behavior or mental states recruited the ToM network more than non-social stories about an object's behavior (P's < 0.001; no difference between agent-behavior and agent-mental, $P > 0.10$).

We also observed a main effect of ROI [$F(3,51) = 9.662$, $P < 0.001$, $\eta_p^2 = 0.362$] and an ROI × story type interaction [$F(6,102) = 4.362$, $P = 0.001$, $\eta_p^2 = 0.204$]. The interaction appears to be driven by increased activity for agent-behavior stories *vs* agent-mental stories in RTPJ [$t(23) = 2.794$, $P = 0.010$; $d = 0.59$], but no other ROI (all P's > 0.20). However, consistent with prior research (Saxe and Kanwisher, 2003; Blakemore *et al.*, 2004; Carrington and Bailey, 2009), separate one-way ANOVAs indicated that activity was greater for social (both agent-behavior and agent-mental) *vs* non-social stories in each ROI (all P's < 0.005). Together, these results demonstrate largely the same response pattern across ROIs during the Background segment.

Table 1. Peak MNI coordinates for ToM ROIs identified in the functional localizer

		MNI coordinates				
ROI	N (out of 24)	x	y	z	No. Voxels	t-Value
RTPJ	24	52	−53	24	77	8.21
LTPJ	23	−49	−58	26	81	7.68
PC	21	1	−58	37	80	7.92
DMPFC	19	2	53	33	53	5.47

FMRI results: outcome segment

Our primary analyses focus on the Outcome segment when participants read outcomes that were expected or unexpected based on what they read in the Background segment. A whole-brain random-effects analysis of unexpected over expected outcomes (voxel-wise threshold: $P < 0.001$, uncorrected; $k > 16$; cluster-wise threshold: $P < 0.05$, FWE-corrected) revealed clusters with peak activations in brain regions generally involved in error detection and evaluation: caudate nucleus [12, 5, 4] (O'Doherty *et al.*, 2004; Harris and Fiske, 2010), thalamus [−9, −10, 7] and dorsal anterior cingulate cortex [−6, 26, 40] (Somerville *et al.*, 2006; Table 3). No regions in the ToM network passed threshold in this contrast or the reverse contrast (expected > unexpected).

Next, using an ROI-based approach, we explored how regions within the ToM network process expectedness differently across social and non-social contexts. A 4 (ROI: RTPJ, LTPJ, PC, DMPFC) × 2 (expectedness: unexpected, expected) × 3 (story type: object-behavior, agent-behavior, agent-mental) repeated measures ANOVA revealed a main effect of story type [$F(2,34) = 3.800$, $P = 0.032$, $\eta_p^2 = 0.183$]: outcomes of social stories (agent-behavior: $M = 0.159$, s.d. = 0.096; agent-mental: $M = 0.088$, s.d. = 0.070) elicited greater activity in ToM regions than outcomes of non-social stories ($M = 0.076$, s.d. = 0.081; both P's < 0.005). Notably, we also observed a main effect of expectedness [$F(1,17) = 12.833$, $P = 0.002$, $\eta_p^2 = 0.430$]: unexpected outcomes elicited greater activity in ToM regions ($M = 0.131$, s.d. = 0.072) than expected outcomes ($M = 0.077$, s.d. = 0.059). There was no main

Table 2. Regions passing threshold in a whole-brain random-effects analysis (voxel-wise threshold: $P < 0.001$, uncorrected; $k > 16$; cluster-wise threshold: $P < 0.05$, FWE-corrected) of the Background segment

		MNI coordinates					
Region name	Hemisphere	x	y	z	Z-value	t-Value	Cluster size
Social > non-social							
PC	Right	6	−55	37	5.29	7.51	479
Dorsomedial prefrontal cortex		0	53	25	5.08	7.02	67
Temporoparietal junction	Left	−54	−61	25	4.93	6.67	207
Temporoparietal junction	Right	54	−52	28	4.32	5.44	151
Agent-behavior > agent-mental							
N/A							
Agent-mental > agent-behavior							
Inferior temporal gyrus	Left	−57	−4	−32	5.37	7.71	219
Occipital gyri	Left	−12	−103	13	4.53	5.84	232
Supramarginal gyrus	Left	−54	−58	43	3.99	4.86	35
Agent-behavior > object-behavior							
PC		0	−52	34	4.61	6.37	196
Temporoparietal junction	Right	54	−61	19	3.80	4.74	107
Agent-mental > object-behavior							
Dorsomedial prefrontal cortex	Right	3	56	25	5.26	8.09	61
PC		0	−55	34	4.99	7.33	462
Inferior temporal gyrus	Left	−51	−1	−38	4.79	6.81	84
Middle temporal gyrus	Left	−54	−43	1	4.42	5.96	95
Temporoparietal junction	Left	−42	−61	19	3.76	4.66	222

Table 3. Regions passing threshold in a whole-brain random-effects analysis (voxel-wise threshold: $P < 0.001$, uncorrected; $k > 16$; cluster-wise threshold: $P < 0.05$, FWE-corrected) of the Outcome segment

Region name	Hemisphere	MNI coordinates			Z-value	t-Value	Cluster size
		x	y	z			
Unexpected > expected							
Middle temporal gyrus	Left	−51	−37	−2	5.53	8.15	165
Thalamus	Left	−9	−10	7	5.20	7.30	78
Inferior frontal gyrus, orbital part	Left	−42	23	−14	5.15	7.19	775
Inferior frontal gyrus, triangular part	Right	33	23	−11	4.94	6.69	165
Medial caudate nucleus	Right	12	5	4	4.57	5.92	43
Dorsal anterior cingulate cortex	Left	−6	26	40	4.55	5.88	192
Middle frontal gyrus	Right	54	29	25	4.38	5.55	63
Supramarginal gyrus	Left	−45	−52	46	4.31	5.43	60
Agent-behavior: unexpected > expected							
Supramarginal gyrus	Left	−54	−55	49	4.79	6.81	125
Middle temporal gyrus	Left	−51	−40	−2	4.74	6.69	116
Inferior frontal gyrus, orbital part	Left	−39	23	−11	4.70	6.59	277
Superior frontal gyrus, medial part	Right	6	32	58	4.62	6.40	111
Middle frontal gyrus	Left	−30	14	58	4.55	6.24	102
Putamen	Right	18	5	7	4.44	6.00	56
Inferior frontal gyrus, orbital part	Right	45	32	−17	4.21	5.51	85
Medial caudate nucleus	Left	−9	−4	10	3.91	4.94	22
Agent-mental: unexpected > expected							
Inferior frontal gyrus, orbital part	Left	−39	35	−20	4.37	5.84	74
Object-behavior: unexpected > expected							
Middle frontal gyrus	Left	−42	29	25	4.13	5.36	98
Middle frontal gyrus	Left	−33	11	40	3.87	4.86	51

effect of ROI ($P > 0.10$). Finally, we observed both an ROI × expectedness interaction [$F(3,15) = 5.256$, $P = 0.011$, $\eta_p^2 = 0.512$] and an ROI × story type interaction [$F(6,102) = 2.220$, $P = 0.047$, $\eta_p^2 = 0.116$], indicating that the dimensions of expectedness and story type are processed differently across ToM ROIs; however, the expectedness × story type and three-way ROI × expectedness × story type interactions did not reach significance (P's > 0.50).

We performed a series of planned comparisons analyzing the dimension of expectedness across story types in each of the ROIs with separate 2 (expectedness) × 3 (story type) repeated measures ANOVAs. In PC, we observed a main effect of story type [$F(2,40) = 5.719$, $P = 0.007$, $\eta_p^2 = 0.222$] in the absence of a main effect of expectedness ($P > 0.60$). Social stories elicited greater activity than object-behavior stories (agent-behavior greater than object-behavior, $t(20) = 3.087$, $P = 0.006$; $d = 0.69$; agent-mental marginally greater than object-behavior, $t(20) = 1.908$, $P = 0.071$; $d = 0.42$), with no difference between the two social stories [$t(20) = 1.673$, $P = 0.110$; $d = 0.37$]. We observed the opposite pattern of results in DMPFC: a main effect of expectedness [$F(1,18) = 7.147$, $P = 0.016$, $\eta_p^2 = 0.284$] in the absence of a main effect of story type ($P > 0.25$). Across stories, unexpected outcomes elicited greater activity than expected outcomes [$t(18) = 2.673$, $P = 0.016$; $d = 0.61$]; however, this difference reached significance only for social stories [agent-behavior: $t(18) = 3.14$, $P = 0.006$; $d = 0.72$; agent-mental: $t(18) = 2.40$, $P = 0.027$; $d = 0.55$; object-behavior: $P > 0.25$]. Finally, in TPJ bilaterally we observed main effects of both expectedness [RTPJ: $F(1,23) = 4.917$, $P = 0.037$, $\eta_p^2 = 0.176$; LTPJ: $F(1,22) = 11.632$, $P = 0.003$, $\eta_p^2 = 0.346$] and story type [RTPJ: $F(2,46) = 9.741$, $P < 0.001$, $\eta_p^2 = 0.298$; LTPJ: $F(2,44) = 4.859$, $P = 0.012$, $\eta_p^2 = 0.181$]. For both LTPJ and RTPJ, unexpected outcomes elicited greater activity than expected outcomes [RTPJ: $t(23) = 2.217$, $P = 0.037$; $d = 0.48$; LTPJ: $t(22) = 3.411$, $P = 0.003$; $d = 0.71$]. This effect was largely driven by a significant difference in the agent-behavior condition [RTPJ: $t(23) = 2.603$, $P = 0.016$; $d = 0.54$; LTPJ: $t(22) = 3.132$, $P = 0.005$; $d = 0.65$], but not in the agent-mental or object-behavior conditions (all P's > 0.15). When comparing story types, RTPJ responded more to agent-behavior stories than agent-mental stories [$t(23) = 2.284$, $P = 0.032$; $d = 0.47$] and more to agent-mental stories than object-behavior stories [$t(23) = 3.349$, $P = 0.003$; $d = 0.69$; Figure 2]. LTPJ responded most to agent-behavior stories [agent-behavior greater than agent-mental, $t(22) = 3.204$, $P = 0.004$; $d = 0.68$], with no difference between agent-mental and object-behavior stories ($P > 0.25$). Note though that pairwise comparisons should be interpreted with caution given the absence of significant expectedness × story type interactions across all ROIs (P's > 0.15).

Given the lack of interactions between outcome expectedness and story type in analyses of average response magnitude, we turned to ROI-based MVPA, reported in Table 4. MVPA enables an independent test of the predicted difference between conditions over and above conventional univariate analyses (Kok *et al.*, 2012). In particular, we investigated which ROIs could discriminate unexpected from expected outcomes within each story type (classification accuracy exceeding chance in a one-tailed *t*-test; Table 4). Mirroring the difference observed in average response magnitude, the pattern of activity within DMPFC distinguished unexpected from expected outcomes for both social stories (agent-mental stories: $P = 0.001$; agent-behavior stories: $P = 0.019$), but not non-social stories (object-behavior: $P > 0.70$). Additionally, unexpected outcomes could be discriminated from expected outcomes for only agent-behavior stories in RTPJ ($P = 0.035$) and agent-mental stories in LTPJ ($P = 0.047$). Unexpected outcomes could not be discriminated from any expected outcomes in PC (P's > 0.06; see supplementary material for pattern discrimination between story types in each ROI).

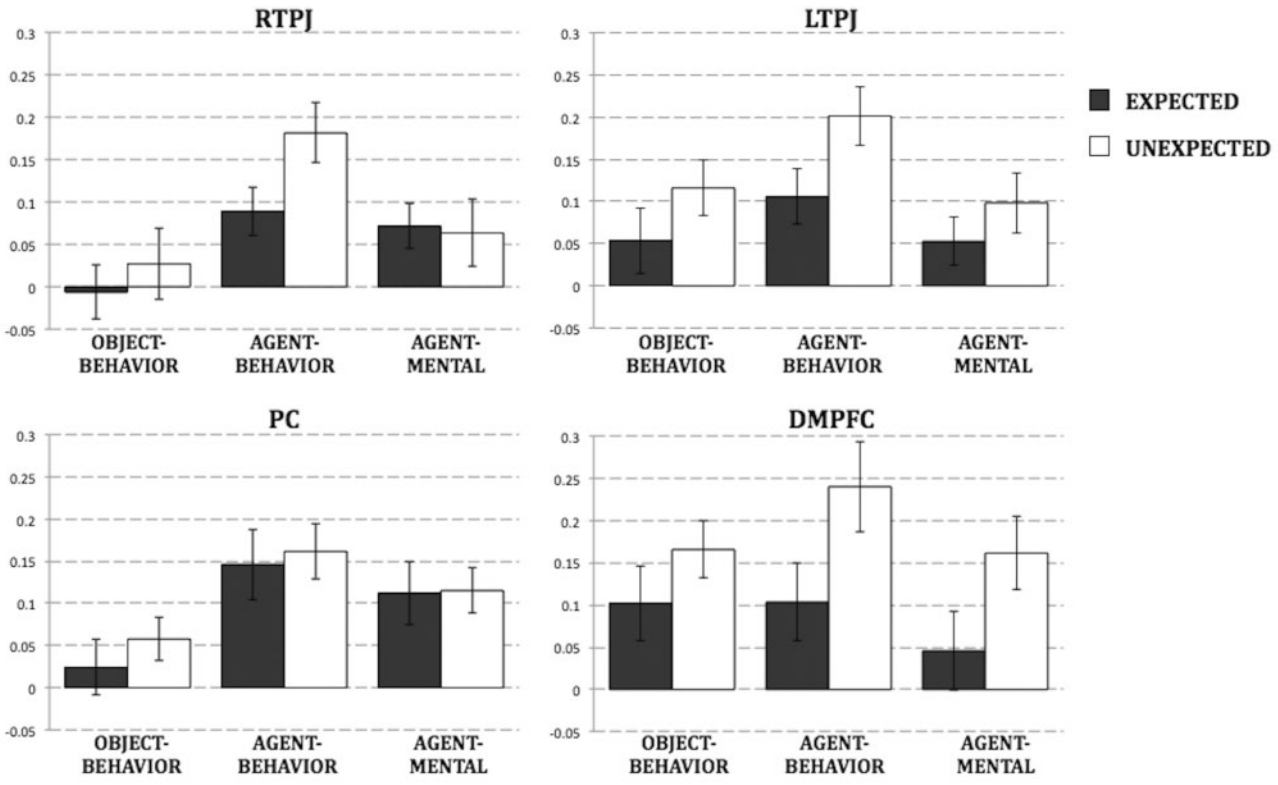

Fig. 2. PSC in ROIs across conditions during the Outcome segment. Error bars show standard error.

Table 4. MVPA results showing average classification accuracy for unexpected *vs* expected outcomes within each story type (error given in standard deviation)

	ROI			
Comparisons between expected and unexpected outcomes	RTPJ	LTPJ	PC	DMPFC
Object-behavior	55.2 ± 3.5	55.5 ± 3.7	49.6 ± 4.2	47.5 ± 4.0
Agent-behavior	**59.0 ± 4.8***	57.5 ± 4.5	48.8 ± 4.7	**61.1 ± 5.0***
Agent-mental	50.7 ± 3.9	**57.6 ± 4.3***	56.3 ± 3.9	**64.8 ± 4.3****

Significance for classification accuracies are one-tailed tests:
*$P < 0.05$;
**$P < 0.01$.

Correlation with autistic-like traits

We computed the difference between the average response to unexpected *vs* expected outcomes within each story type (object-behavior, agent-behavior, agent-mental) and investigated how this difference score within each story type correlates with the presence of autistic-like traits across individuals, as measured by the AQ (Baron-Cohen *et al.*, 2001). Activation to unexpected *vs* expected outcomes in RTPJ correlated with AQ scores only in the agent-behavior condition [$r(22) = -0.471$, $P = 0.020$; 95% CI: [−0.08, −0.73]; Figure 3]—not in the agent-mental or object-behavior condition (P's > 0.40).[2] No other correlations were significant between AQ scores and activation in any of the ToM ROIs for any condition (P's > 0.05).

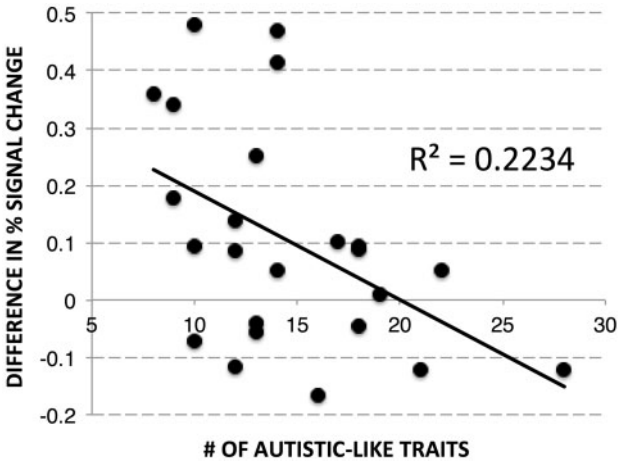

Fig. 3. Scores on the AQ correlate with the difference in RTPJ response (PSC) to unexpected *vs* expected outcomes based on an agent's behavior.

Discussion

The present study adds to our understanding of the precise role of ToM by investigating how ToM supports the processing of unexpected events across contexts. In DMPFC, unexpected outcomes elicited greater activity than expected outcomes, particularly for social *vs* non-social contexts. PC was preferentially recruited for reasoning about social *vs* non-social contexts, regardless of whether behavior was unexpected or not.

2 Pattern discrimination in RTPJ for unexpected agent-behavior *vs* expected agent-behavior shows a similar but non-significant correlation with AQ [$r(22) = -0.32$, $P = 0.12$].

Finally, TPJ bilaterally responded more to social *vs* non-social stories and to unexpected *vs* expected outcomes. At the broadest level, these results demonstrate that unexpected outcomes are processed differently across contexts and brain regions for ToM.

The convergent evidence from analyses of average response magnitude and pattern classification demonstrating a difference between unexpected and expected outcomes in DMPFC is consistent with recent work suggesting that DMPFC plays a general role in monitoring violations of expectations (Amodio and Frith, 2006; Venkatraman and Huettel, 2012; Bzdok *et al.*, 2013). For example, in one study (Desmet *et al.*, 2014), DMPFC responded to unexpected behavior of both humans and machines. Although unexpected outcomes in the current study tended to elicit greater activity than expected outcomes for all story types, it is worth noting that this difference reached significance only for social stories. Similarly, the pattern of activation within DMPFC could discriminate between expected and unexpected outcomes for both 'social' story types (agent-behavior and agent-mental), but not non-social (object-behavior) stories. Although activity in DMPFC likely reflects a general mechanism for monitoring and detecting unexpected events across many contexts, the present findings tentatively suggest that this mechanism may be especially attuned to monitoring events in the social domain.

Why did unexpected outcomes in the agent-mental condition not elicit greater RTPJ activity than expected outcomes, as might have been predicted by prior work? Two points are worth noting. First, prior work featured unexpected events containing explicit mental state information (e.g. a Democrat wants a smaller government, a married man would find it fun if his wife had extramarital relations; Saxe and Wexler, 2005; Cloutier *et al.*, 2011). In contrast, outcomes in the present study (the target of our primary analyses) depicted physical behavior in all conditions, e.g. belief information in the agent-mental condition was presented only in the Background segment as the basis of participants' expectations. Second, the finding that these outcomes did not elicit robust activity in RTPJ is consistent with previous studies showing that the ToM network may be recruited primarily to reason about unexpected behavior when people are explicitly instructed to attend to agents' intentions (de Lange *et al.*, 2008; Ampe *et al.*, 2014). In many contexts, people may therefore refrain from spontaneously engaging in ToM to resolve inconsistencies between present behavior and prior information about mental states.

Nevertheless, the neural pattern within RTPJ, as revealed by MVPA, distinguished unexpected outcomes from expected outcomes, but only given an agent's past behavior—not an object's behavior or, notably, an agent's mental states. We suggest that ToM capacities may have evolved primarily to explain and integrate information about behaviors (in terms of mental state inferences) rather than to process explicit information about mental states. Behavior in general elicits spontaneous ToM insofar as people must often infer abstract mental states to explain their observations (Spunt *et al.*, 2010), especially when behavior is unexpected (Brass *et al.*, 2007; Van Overwalle, 2009; Van Overwalle and Baetens, 2009; Young and Saxe, 2009). Thus, people may rely on ToM particularly in contexts where an agent's behaviors are internally inconsistent in order to construct a coherent representation of the agent's mind.

The finding that neurotypical participants exhibiting more autistic-like traits show a decreased RTPJ response to unexpected *vs* expected outcomes provides empirical support for recent theoretical accounts suggesting that individuals with ASD have difficulty forming predictions (Lawson *et al.*, 2014; Sinha *et al.*, 2014; Van de Cruys *et al.*, 2014). However, we did not find a relationship between responses to unexpected events and the presence of autistic-like traits in all conditions, but rather specifically in the agent-behavior condition. Future research should investigate whether difficulty with forming adequate predictions about complex social behavior is a cause or consequence of decreased motivation to engage with social stimuli (Chevallier *et al.*, 2012). Notably, children with ASD prefer non-social toys (e.g. cars) to social toys (e.g. interactive animals), but this difference is diminished when the social toys are made more predictable (Ferrara and Hill, 1980). Although strong conclusions cannot be drawn from our small sample of correlational data, the present findings emphasize the importance of investigating how unexpected events are processed across a broader range of participants, including both neurotypical individuals and individuals on the autism spectrum.

Interestingly, contrasting unexpected *vs* expected outcomes revealed activity in caudate nucleus. This region of dorsal striatum plays a key role in processing prediction errors—when actual outcomes violate expectations (Rescorla and Wagner, 1972; O'Doherty *et al.*, 2004; Balleine *et al.*, 2007). Although the present study made targeted predictions about ToM ROIs, future work should directly investigate the relationship between brain regions for ToM and brain regions that process prediction errors more generally (Harris and Fiske, 2010). Understanding how midbrain regions underlying domain-general learning interact with domain-specific regions for social cognition will be crucial for determining the unique contribution of the ToM network for social learning.

Finally, we acknowledge potential limitations of the current study. First, although a key feature of the present paradigm was to test how brain regions for ToM respond to unexpected outcomes across distinct contexts, the large number of conditions of interest limited events per condition to 10. To increase power, future studies may benefit from focusing on pairs of contexts, particularly the understudied comparison between behaviors that are unexpected based on an agent's past behavior *vs* an agent's mental states. Second, despite the attempt to remove the influence of participants' pre-existing beliefs and attitudes on the formation of their expectations, some items may have contained information that was tied to social conventions or evoked personality attributions. However, this incidental information was typically only part of the broader context of an item and not central to the judgment being made, compared with the primary information presented in the Background segment. Furthermore, it is unlikely that the amount of personality information differed systematically across conditions (e.g. between stories describing an agent's behavior *vs* an agent's mental states). Thus, the systematic differences that emerged across conditions and brain regions seem robust to noise introduced by the wide range of stimuli used. Nevertheless, it is possible that there are other dimensions relevant for social cognition that cut across the conditions defined here. Future research should continue to examine neural responses across different contexts to further characterize the precise role of ToM.

In sum, the current work reveals important differences in how unexpected events are processed across contexts and brain regions in the ToM network. Specifically, key regions for ToM may be especially sensitive to actions that are unexpected based on previous knowledge of an agent's past behavior. These results indicate a unique role for the ToM network in constructing coherent models of agents' minds by integrating information about their past and present behaviors.

Funding

This work was supported by an NSF Graduate Research Fellowship awarded to J.A.D. and a grant from the Dana Foundation awarded to L.Y.

Supplementary data

Supplementary data are available at *SCAN* online.

Conflict of interest. None declared.

Acknowledgements

We thank Fiery Cushman, Elizabeth Kensinger, Jorie Koster-Hale, Jim Russell, Rebecca Saxe and the Boston College Morality Lab.

References

Amodio, D.M., Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, **7**(4), 268–77.

Ampe, L., Ma, N., Van Hoeck, N., Vandekerckhove, M., Van Overwalle, F. (2014). Unusual actions do not always trigger the mentalizing network. *Neurocase*, **20**(2),144–9.

Balleine, B.W., Delgado, M.R., Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *The Journal of Neuroscience*, **27**(31), 8161–5.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., Clubley, E. (2001). The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, **31**(1), 5–17.

Baron-Cohen, S., Ashwin, E., Ashwin, C., Tavassoli, T., Chakrabarti, B. (2009). Talent in autism: hyper-systemizing, hyper-attention to detail and sensory hypersensitivity. *Philosophical Transactions of the Royal Society B*, **364**, 1377–83.

Best, C.S., Moffat, V.J., Power, M.J., Owens, D.G., Johnstone, E.C. (2008). The boundaries of the cognitive phenotype of autism: theory of mind, central coherence and ambiguous figure perception in young people with autistic traits. *Journal of Autism and Developmental Disorders*, **38**(5), 840–7.

Blakemore, S.J., Winston, J., Frith, U. (2004). Social cognitive neuroscience: where are we heading? *Trends in Cognitive Sciences*, **8**(5), 216–22.

Brass, M., Schmitt, R.M., Spengler, S., Gergely, G. (2007). Investigating action understanding: inferential processes versus action simulation. *Current Biology*, **17**(24), 2117–21.

Bzdok, D., Langner, R., Schilbach, L., *et al.* (2013). Segregation of the human medial prefrontal cortex in social cognition. *Frontiers in Human Neuroscience*, **7**, 1–17.

Carrington, S.J., Bailey, A.J. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Human Brain Mapping*, **30**(8), 2313–35.

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E.S., Schultz, R.T. (2012). The social motivation theory of autism. *Trends in Cognitive Sciences*, **16**(4), 231–9.

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**(2), 583–8.

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), 580–93.

Desmet, C., Deschrijver, E., Brass, M. (2014). How social is error observation? The neural mechanisms underlying the observation of human and machine errors. *Social Cognitive and Affective Neuroscience*, **9**(4), 427–35.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., Saxe, R. (2011). fMRI item analysis in a theory of mind task. *Neuroimage*, **55**(2), 705–12.

Ferrara, C., Hill, S.D. (1980). The responsiveness of autistic children to the predictability of social and nonsocial toys. *Journal of Autism and Developmental Disorders*, **10**(1), 51–7.

Fletcher, P., Happe, F., Frith, U., *et al.* (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition*, **57**(2), 109–28.

Gobbini, M.I., Koralek, A.C., Bryan, R.E., Montgomery, K.J., Haxby, J.V. (2007). Two takes on the social brain: a comparison of theory of mind tasks. *Journal of Cognitive Neuroscience*, **19**(11), 1803–14.

Haffey, A., Press, C., O'Connell, G., Chakrabarti, B. (2013). Autistic traits modulate mimicry of social but not nonsocial rewards. *Autism Research*, **6**(6), 614–20.

Harris, L.T., Fiske, S.T. (2010). Neural regions that underlie reinforcement learning are also active for social expectancy violations. *Social Neuroscience*, **5**(1), 76–91.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, **293**(5539), 2425–30.

Herry, C., Bach, D.R., Esposito, F., *et al.* (2007). Processing of temporal unpredictability in human and animal amygdala. *The Journal of Neuroscience*, **27**(22), 5958–66.

Ingersoll, B., Hopwood, C.J., Wainer, A., Donnellan, M.B. (2011). A comparison of three self-report measures of the broader autism phenotype in a non-clinical sample. *Journal of Autism and Developmental Disorders*, **41**(12), 1646–57.

Jenkins, A.C., Mitchell, J.P. (2010). Mentalizing under uncertainty: dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex*, **20**(2), 404–10.

Kobayashi, C., Glover, G.H., Temple, E. (2007). Children's and adults' neural bases of verbal and nonverbal 'theory of mind'. *Neuropsychologia*, **45**(7), 1522–32.

Kok, P., Jehee, J.F., de Lange, F.P. (2012). Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, **75**(2), 265–70.

Koster-Hale, J., Saxe, R. (2013). Theory of mind: a neural prediction problem. *Neuron*, **79**(5), 836–48.

de Lange, F.P., Spronk, M., Willems, R.M., Toni, I., Bekkering, H. (2008). Complementary systems for understanding action intentions. *Current Biology*, **18**(6), 454–7.

Lawson, R.P., Rees, G., Friston, K.J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, **8**, 302, doi: 10.3389/fnhum.2014.00302.

Ma, N., Vandekerckhove, M., Baetens, K., Van Overwalle, F., Seurinck, R., Fias, W. (2012). Inconsistencies in spontaneous and intentional trait inferences. *Social Cognitive and Affective Neuroscience*, **7**(8), 937–50.

Mitchell, J.P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex*, **18**(2), 262–71.

Niv, Y., Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, **12**(7), 265–72.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, **304**(5669), 452–4.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, **10**(9), 424–430.

Rescorla, R. A., Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, **2**, 64–99.

Ruzich, E., Allison, C., Smith, P., *et al.* (2015). Measuring autistic traits in the general population: a systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. *Molecular Autism*, **6**(1), 2–13.

Saxe, R., Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". *Neuroimage*, **19**(4),1835–42.

Saxe, R., Powell, L.J. (2006). It's the thought that counts specific brain regions for one component of theory of mind. *Psychological Science*, **17**(8), 692–9.

Saxe, R., Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*, **43**(10), 1391–9.

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E.N., Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS One*, **4**(3), e4869.

Schultz, W., Dayan, P., Montague, P.R. (1997). A neural substrate of prediction and reward. *Science*, **275**(5306), 1593–9.

Sinha, P., Kjelgaard, M.M., Gandhi, T.K., *et al.* (2014). Autism as a disorder of prediction. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(42), 15220–5.

Slotnick, S.D., Moo, L.R., Segal, J.B., Hart, J. (2003). Distinct prefrontal cortex activity associated with item memory and source memory for visual shapes. *Cognitive Brain Research*, **17**(1), 75–82.

Somerville, L.H., Heatherton, T.F., Kelley, W.M. (2006). Anterior cingulate cortex responds differentially to expectancy violation and social rejection. *Nature Neuroscience*, **9**(8),1007–8.

Spunt, R.P., Falk, E.B., Lieberman, M.D. (2010). Dissociable neural systems support retrieval of how and why action knowledge. *Psychological Science*, **21**(11), 1593–8.

Van de Cruys, S., Evers, K., Van der Hallen, R., *et al.* (2014). Precise minds in uncertain worlds: predictive coding in autism. *Psychological Review*, **121**(4), 649.

Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human Brain Mapping*, **30**(3), 829–58.

Van Overwalle, F., Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage*, **48**(3), 564–84.

Venkatraman, V., Huettel, S.A. (2012). Strategic control in decision-making under uncertainty. *European Journal of Neuroscience*, **35**(7),1075–82.

Young, L., Dodell-Feder, D., Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia*, **48**(9), 2658–64.

Young, L., Saxe, R. (2009). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, **21**(7), 1396–405.