OXFORD

Social Cognitive and Affective Neuroscience, 2016, 1–12

doi: 10.1093/scan/nsv145
Advance Access Publication Date: 8 December 2015
Original article

Getting to know you: general and specific neural computations for learning about people

Damian A. Stanley

Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA

Correspondence should be addressed to Damian A. Stanley, California Institute of Technology, MC 228-77, 1200 East California Boulevard, Pasadena, CA 91125, USA. E-mail: dstanley@caltech.edu.

Abstract

Learning about other peoples' attributes, e.g. whether an individual is generous or selfish, is central to human social cognition. It is well documented that a network of cortical regions is reliably activated when we engage social processes. However, little is known about the specific computations performed by these regions or whether such processing is specialized for the social domain. We investigated these questions using a task in which participants ($N = 26$) learned about four peoples' generosity by watching them choose to share money with third party partners, or not. In a non-social control condition, participants learned the win/loss rates of four lotteries. fMRI analysis revealed learning-related general (social + non-social) prediction error signals in the dorsomedial and dorsolateral prefrontal cortices (bilaterally), and in the right lateral parietal cortex. Socially specific (social > non-social) prediction error signals were found in the precuneus. Interestingly, the region that exhibited social prediction errors was a distinct subregion of the area in the precuneus and posterior cingulate cortex that exhibited a commonly reported main effect of higher overall activity for social *vs* non-social stimuli. These findings elucidate the domain—general and—specific computations underlying learning about other people and demonstrate the increased explanatory power of computational approaches to social cognition.

Key words: fMRI; computational neuroscience; trait learning; prediction error; precuneus

Introduction

Learning about other peoples' behavioral dispositions and in particular their intentions toward others is crucial for survival in our social world. This ability requires that we maintain representations of other individuals that encode these characteristics and update them when we receive novel information. A network of cortical brain regions, including the temporal parietal junction (TPJ), precuneus/posterior cingulate cortex (Pc/PCC), dorsomedial prefrontal cortex (dmPFC) and the temporal poles (TP), has been consistently implicated by studies involving the representation of others' beliefs, preferences and intentions (for reviews see Frith and Frith, 2006; Behrens *et al.*, 2009; Van Overwalle, 2009; Mar, 2011; Kennedy and Adolphs, 2012; Olson *et al.*, 2013). Disruption of components of this network results in impairments of social cognition (Todorov and Olson, 2008; Krajbich *et al.*, 2009; Young *et al.*, 2010; Olson *et al.*, 2013); abnormal functioning may underlie social impairments associated with autism spectrum disorder (Castelli *et al.*, 2002; Kennedy *et al.*, 2006; Kennedy and Courchesne, 2008; Kana *et al.*, 2009); and gray matter volume in regions of this network reflects social network size (Lewis *et al.*, 2011; Sallet *et al.*, 2011). However, much remains unknown about the specific computations performed by this network's constituent components.

Most research on the neuroscience of representing other people has focused on the assignment of specific subprocesses (e.g. the representation of beliefs *vs* preferences) to specific brain regions. The paradigms used often examine the representation of other people in isolated, static social situations, with little ambiguity (but see Jenkins and Mitchell, 2010) and no requirement for maintenance or updating through experience. This is in stark contrast to the real-world, in which representations of other people are uncertain and dynamic, evolving over time as we observe their behavior and revise our understanding.

Received: 14 April 2015; **Revised:** 20 November 2015; **Accepted:** 28 November 2015

Two notable exceptions are the literatures on impression formation and computational modeling of social learning.[1,2]

Neuroimaging studies of impression formation have shown that Blood-Oxygen Level Dependent (BOLD) activity in regions of the social cognition network, as well as dorsolateral prefrontal and dorsolateral parietal cortices (dlPFC and dlPC, respectively), is greater when presented with impression-inconsistent (compared with impression-consistent) information about a social target (Cloutier *et al.*, 2011a,b; Ma *et al.*, 2012; Bhanji and Beer, 2013; Mende-Siedlecki *et al.*, 2013)—(but see Ames and Fiske, 2013). In one elegant study, Schiller *et al.* (2009) demonstrated that BOLD activity in the PCC was higher for pieces of information that were more heavily weighted by participants during subsequent ratings of whether or not they 'liked' a social target and also scaled with the magnitude of the participant's ratings. Although these data provide compelling evidence for the general involvement of these regions in impression formation, for the most part (with the possible exceptions of Bhanji and Beer, 2013 and Schiller *et al.*, 2009) they cannot speak to the specific computations being performed. In addition, none of them include a non-social control condition and therefore none of them is able to address domain specificity.

The few studies that have directly investigated the neural computations underlying social learning (Behrens *et al.*, 2008; Hampton *et al.*, 2008; Yoshida *et al.*, 2010; Suzuki *et al.*, 2012; Boorman *et al.*, 2013) have focused on two computational signals necessary for learning; estimates of the probability that another person will perform a given action (e.g. how predictable they are in a given situation), and estimates of prediction errors (PEs), i.e. how surprising an individual's behavior is, given previous estimates of their predictability. To identify neural signals, such as these, it is crucial to employ paradigms that require participants to maintain and revise their representations of other people given new information. These studies have tied BOLD correlates of social predictability and PE to a number of regions including the medial prefrontal cortex (mPFC; dorsal and ventral), superior temporal sulcus (STS)/TPJ and the right TP. Unfortunately, as with the impression formation literature, the majority of these studies have lacked non-social learning conditions and cannot speak to social specificity (but see Boorman *et al.*, 2013). A further concern is that these studies have generally confounded learning about reward to oneself with learning about the traits of another person, making it difficult to distinguish whether any putative neural signal is related to reward likelihood or the other person's character.

In this study, to identify neural signals specific for social learning we used a paradigm in which participants learned about other real people ('Gifters') by observing them make generous or selfish decisions concerning real third party individuals (Figure 1a). We focused on generosity because of the ubiquity of altruistic behavior in human societies (Henrich *et al.*, 2001; Camerer, 2003), suggesting that generosity may be a fundamental attribute that we evaluate in others. Critically, participants also completed a computationally matched, non-social learning condition and were not rewarded during learning.

1 Author's Note: At the time of the design and implementation of this study (Spring, 2010), much of the literature described here did not exist. Because of this, we provide a review of the field as it currently stands, however, we omit what would be *post hoc* predictions about expected patterns of BOLD activity. Instead, we reserve discussion of our results in the context of this literature for the discussion.

2 For the purposes of this article, we use the term 'social learning' to specifically refer to learning about the traits, beliefs and intentions of another person.

Materials and methods

MRI participants

Thirty participants (median age = 23.5 years, range = 19–37; all female to match the gender of the 'Gifters') took part in the fMRI study, four were excluded from the final analysis (1 for excessive head motion and 3 because they did not meet behavioral criterion). All participants were right-handed and had normal or corrected-to-normal vision. Participants were recruited through the subject pool of the Social Sciences Experimental Laboratory at the California Institute of Technology (which included participants from Pasadena City College and the surrounding area) and were paid $40/h as well as earnings from the experiment (up to $40 additional; see Procedure). All experimental procedures were undertaken with the understanding and written consent of each participant and were approved by the California Institute of Technology Institutional Review Board.

Procedure

Participants were scanned on two separate days (intersession interval median = 2 days, range = 1–4) so as to limit the duration of a single Magnetic Resonance Imaging (MRI) session to <1.5 h. On the first day, participants were consented, briefed concerning the nature of the experiment (details later), and then completed 1 structural and 3 functional MRIs. On the second day, participants again completed 1 structural and 3 functional MRIs, continuing the experiment where they left off on the first day.

During the briefing, MRI participants were informed that in an earlier phase of the experiment (see later), four real female[3] 'Gifters' had made a series of economic decisions to share or keep all of $10 with 48 distinct real partners (i.e. a dictator game; Forsythe *et al.*, 1994; Kahneman *et al.*, 1986). The MRI participants' task was to observe these Gifters' decisions (presented in random order) and form an estimate of each Gifter's overall share (or keep; counterbalanced) percentage. They were aware that all the interactions were real and had actual consequences for the Gifters and their partners. Finally, they were told they would also be estimating the overall win (or loss; counterbalanced) percentage of four Lotteries (represented by pictures of fractals) that generated outcomes for the same 48 partners (again with real consequences).

Each of the six functional runs contained 64 randomly intermixed trials (8×4 Gifter and 8×4 Lottery). Participants learned about the same Gifters and Lotteries throughout the experiment, always continuing from where they left off in previous run. On each day, once MRI data collection was complete, participants provided a final estimate for each Gifter and Lottery. To incentivize participants to learn, they were told that at the end of the experiment they would be rewarded $2.50 for each of these final estimates that was within 5% of the true share/keep, or win/loss, percentage. Participants instructed to estimate share/win percentage on day 1 were instructed to estimate keep/loss percentage on day 2 and vice versa.

Trial design

On each Gifter trial, participants saw a color photograph representing one of the four Gifters (identity-to-behavior pairings were randomly assigned across participants) and had up to 4 s to estimate the overall percentage of the time (0–100% in increments of 10%) that particular Gifter chose to share $10 with their partners

3 The gender of participants and Gifters was matched (i.e. all female) to avoid potential cross gender effects influencing participant learning.

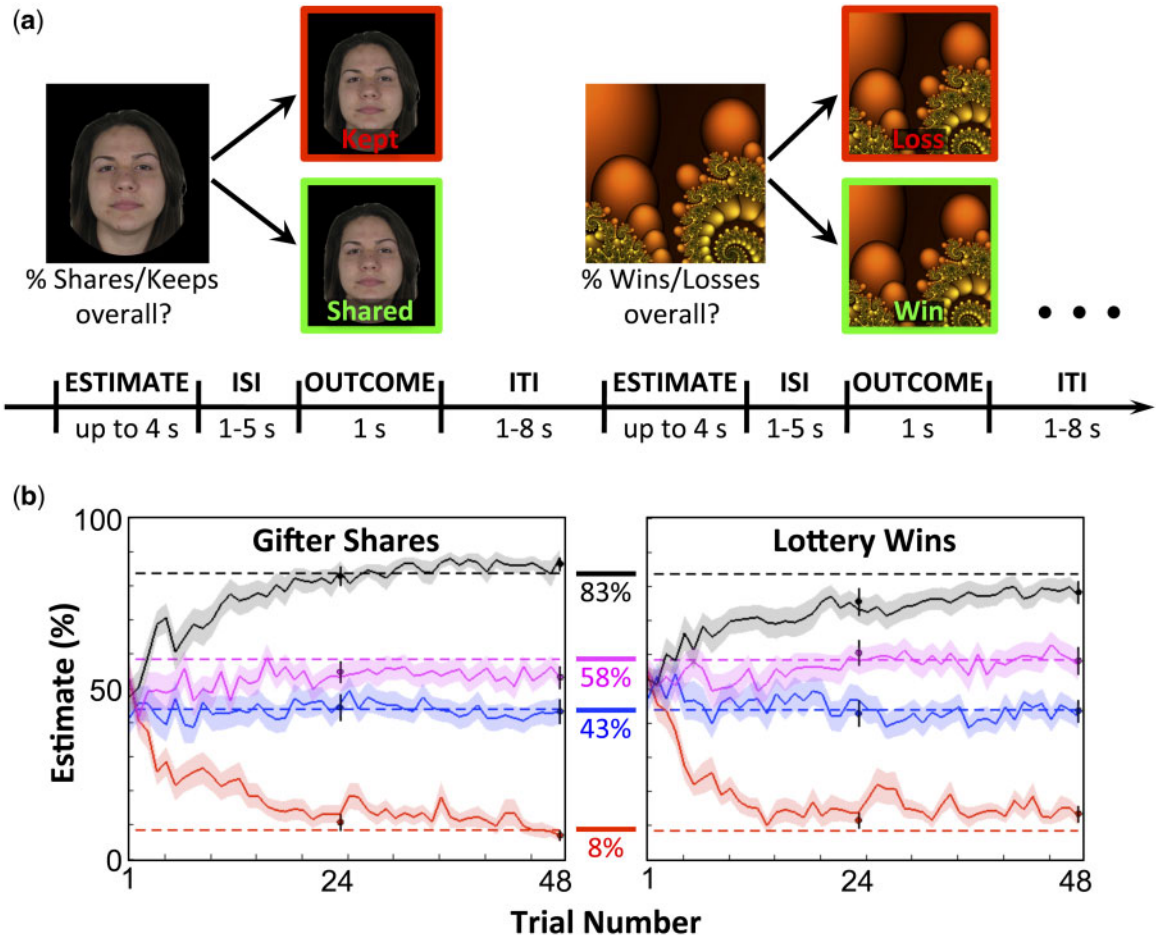

Fig. 1. (a) Schematic of experimental design. Participants learned about the generosity of four distinct 'Gifters' as each made single-shot decisions to 'Share' or 'Keep' $10 with 48 different partners [a.k.a. a 'Dictator' game (Kahneman *et al.*, 1986; Forsythe *et al.*, 1994)]. On each trial, participants first estimated the overall percentage of the time that the current Gifter shared (or kept) $10 with their partners (0–100% in increments of 10%; Gifters were preselected to have a range of sharing rates: 8%, 43%, 58% and 83%). Participants then saw the actual outcome (shared/kept) for that trial. To enable the identification of neural mechanisms specific for social learning, participants also learned about the overall percentage that four matched Lotteries (represented by fractals) generated 'Wins' or 'Losses' for the same partners (non-social control condition). Participants were incentivized to learn, but did not themselves receive rewards during the experiment (see Methods). (b) Average participant estimates for the four Gifters (left) and the corresponding Lotteries (right). Solid colored lines indicate mean participant estimate of each Gifter Share, and Lottery Win, percentage over the course of the experiment with the surrounding transparency indicating standard error across participants ($N = 26$). Dotted lines indicate actual Gifter Share and Lottery Win percentages. Circles indicate participant final estimates (outside the scanner) and standard errors on day 1 (following trial 24) and day 2 (following trial 48) of the experiment. No significant differences in learning for Gifters compared with Lotteries were identified (see Results).

(Figure 1). To enter their response, participants first adjusted the percentage number underneath the Gifter's photo up or down (initial percentage value was randomized on each trial) using the index and middle finger of one hand (right/left counterbalanced across subjects) and then signaled their final answer by pressing a button with the index finger of the other hand. Once a response was finalized, the Gifter's photo disappeared and only the fixation point remained for the remainder of estimate period. This was followed by a 1–5 s (randomized, uniform distribution) inter-stimulus-interval and then an outcome screen (duration = 1 s) that displayed the Gifter's actual decision ('shared' or 'kept') for that specific partner in green ('shared') or red ('kept'). Finally, trials were separated by a 1–8 s (randomized, uniform distribution) inter-trial-interval. The procedure for Lottery trials was exactly the same.

Bayesian learner analysis

To assess learning performance, each participant's behavioral data were compared with that of an ideal Bayesian learner. For each of the four Gifters (and four Lotteries) a Bayesian learner with a flat prior described by a Beta distribution [i.e. prior = Beta(1, 1)] was provided the participant-specific outcome history (shared = 1, kept = 0). For these learners, the mean of the posterior distribution on a given trial t (i.e. the current generosity estimate) is given by the formula:

$$\text{Estimate}_t = \left((1 + \text{num}_{\text{shares}_t}) / \left[(1 + \text{num}_{\text{shares}_t}) + \left(1 + \text{num}_{\text{keeps}_t}\right)\right]\right),$$

in which Estimate_t is the generosity estimate following the observation of outcome_t and the variables num_shares_t and num_keeps_t refer to the total number of times including trial t that the Gifter has shared or kept, respectively. Once estimates for each Gifter and Lottery were generated, they were combined and ordered by trial number. The resulting idiosyncratic series of estimates were correlated with the participant's actual estimates.

Gifter behavior data collection

Prior to the fMRI experiment, descriptive information (age, gender, years of higher education, city of birth and a movie they

would highly recommend) was collected from 48 partners (20 female). Partners were paid $5 each for their participation. Subsequently, a non-overlapping pool of eight female Gifters viewed the descriptive information of each of the 48 partners while making a decision to share ($4:$6, $5:$5 or $6:$4 splits) with the partner or keep the whole $10 for themselves [a.k.a. a dictator game (Kahneman *et al.*, 1986; Forsythe *et al.*, 1994)]. Gifters were paid $10 for participation. When the experiment was over, one random trial was selected and both the Gifter and the partner from that trial received the actual monetary outcome of the trial. Gifter choice data were subsequently analyzed and 4/8 Gifters with a range of sharing rates (8%, 43%, 58% and 83%) were selected to be the Gifters that fMRI participants learned about.

MRI data acquisition

Imaging data were collected at the Caltech Brain Imaging Center using a Siemens 3T Trio scanner and a Siemens 8-channel phased array head coil. Imaging sessions began with a T1-weighted MPRAGE anatomical scan collected in the sagittal plane (176 slices, Repetition Time (TR) = 1.5 s, Echo Time (TE) = 3.05 ms, slice thickness = 1 mm, inplane resolution = 1 × 1 mm, flip angle = 10°, Field of view (FOV) = 256 mm^2, number of averages = 2). Following this, T2*-weighted gradient-echo echo-planar images (EPI) with BOLD contrast were acquired in three scans (duration ~ 13 m 48 s, TR = 2.75 s, TE = 25 ms, flip angle = 80°, slice thickness = 3 mm, inplane resolution = 3 × 3 mm, FOV = 192 mm). We collected 44 slices with an oblique orientation of 30° to the anterior commissure-posterior commissure line. Slice acquisition order was interleaved with no gap and the first two acquisitions of each functional run were discarded. Foam inserts were used to restrict participant head motion. Stimuli presented using MATLAB (The MathWorks, Natick, MA) and PsychToolBox-3 (Brainard, 1997; Pelli, 1997; Kleiner *et al.*, 2007), were projected onto a screen at the back of the MRI machine and viewed through a mirror. Responses were collected using two 2-button response boxes (Current Designs, Philadelphia, PA).

Imaging data preprocessing

Imaging data were preprocessed and analyzed using SPM8 (Wellcome Trust Centre for Neuroimaging, University College London, UK). Functional data from each scan were corrected for slice acquisition time, then motion-corrected (3d affine transformation) to the first image of the scan. Following this, the data underwent normalization to the Montreal Neurological Institute's standard EPI template and spatial smoothing (3d Gaussian filter, 8 mm Full width at half maximum). Finally, we applied a high-pass temporal filter (width = 128 s) to the data to remove low frequency noise associated with scanner drift.

fMRI data analysis

For each participant, to identify regions with differential responses to Gifters and Lotteries we estimated a GLM with AR(1) and the following regressors of interest:

R1) a boxcar indicator function for the Gifter estimation screen (duration = reaction time = time between trial onset and the first key press).

R2) a parametric modulator function of estimated Gifter predictability: a transformation of participants' estimates to a V-shaped function that increased with distance from the point at which Gifter behavior was minimally predictable, P(share) = 0.5. Thus,

$$\mathrm{Pred}_{(t)} = \mathrm{abs}(P(\mathrm{share}_{(t)}) - 0.5)$$

In which $\mathrm{Pred}_{(t)}$ is the estimated predictability on trial t and $P(\mathrm{share}_{(t)})$ is the participants' estimated probability the Gifter is a sharer on trial t.

R3) a boxcar indicator function for the outcome screen (duration = 1 s).

R4) a parametric modulator function of participant state prediction error (SPE) (Gläscher *et al.*, 2010) at outcome.

$$\mathrm{SPE}_{(t)} = \mathrm{abs}(\mathrm{Outcome} - P(\mathrm{share}_{(t)}))$$

R5–R8) The equivalent predictors for Lottery trials (with win/loss substituted for share/keep).

In addition, the General Linear Model (GLM) included 6 head-motion regressors, 6 constant regressors (1 per scan), a regressor for response-related finger movements (a boxcar covering the period between the first response-related button press and the button press indicating the final response) and a regressor for missed trials (median = 0.2% of all trials; range = 0–9.4%). The regressors of interest, motor movements and missed trials were convolved with a canonical double-gamma hemodynamic response function (SPM8).

Because we were interested in determining whether there were regions selectively involved in the processing of Gifter attributes, we computed each of the following single-subject contrasts between the two conditions: the main effect contrast of Gifter indicators > Lottery indicators during the estimate (R1 > R5) and outcome (R3 > R7) periods and the parametric effects contrasts of Gifter > Lottery for the modulators of Predictability (R2 > R6) and SPE (R4 > R8).

Second-level group statistics were calculated using one-sample t-tests of the beta weights from the first-level contrasts. For inference purposes, we applied a voxel-wise statistical threshold of $P < 0.005$, and then applied a whole-brain cluster-correction (threshold: $P < 0.05$). For subcortical regions-of-interest (e.g. the caudate), we used small-volume-correction based on anatomical ROIs taken from the automated anatomical labeling (AAL) Atlas (Tzourio-Mazoyer *et al.*, 2002). Finally, to identify regions commonly activated by both Gifter and Lottery trials we used minimum statistic conjunction analyses (Nichols *et al.*, 2005), and set cluster thresholds to the larger of those estimated for the two statistical maps contributing to the conjunction (thereby remaining conservative in our inference).

Precuneus ROI analysis

When considering contrasts of parametric regressors (i.e. Gifter SPE > Lottery SPE), different underlying patterns of BOLD response can lead to voxels/clusters being deemed significant. To examine the response patterns that gave rise to significant clusters in the parametric contrast of [Gifter SPE > Lottery SPE], we first defined individual functional ROIs for each participant using a leave-one-out procedure. For each participant, the parametric contrast of [Gifter SPE > Lottery SPE] was calculated at the group-level for the remaining 25 participants. That statistical parametric map was then used to define a functional Region-of-Interest (ROI) for that participant. A separate GLM was estimated for each participant in which trials of low, medium and high predictability and SPE were modeled as distinct regressors. To remain unbiased with regards

to the relationship between behavioral measures (i.e. predictability and SPE) and the distribution of BOLD responses, the data were first split into approximate tertiles to ensure that there were a similar number of trials in each bin (because the data were discrete, tertile split points were restricted to being between SPE values). For predictability, the resulting mapping of low, medium and high predictability to participant share likelihood estimates was: low = [0.4, 0.5, 0.6] (average number of trials per participant = 125.4); medium = [0.2, 0.3, 0.7, 0.8] (average number of trials per participant = 143.2); high = [0, 0.1, 0.9, 1] (average number of trials per participant = 112.6). For SPE, the resulting mapping of low, medium and high SPE to participant SPEs was: low = [0, 0.1] (average number of trials per participant = 119.7); medium = [0.2, 0.3, 0.4, 0.5] (average number of trials per participant = 146.2); high = [0.6, 0.7, 0.8, 0.9, 1] (average number of trials per participant = 115.23). For each participant a GLM with AR(1) and six regressors-of-interest (low/medium/high × predictability/SPE) as well as regressors for motion, missed trials and response-related finger movement was estimated. Then, for each participant and ROI the beta values for each regressor of interest in each voxel were extracted and averaged together. Finally, data for each ROI and regressor of interest were averaged across participants (Figure 3; top right panel).

Results

Behavior

To assess learning, we compared each participant's estimates to those of a purely Bayesian learner exposed to the same history of outcomes (see Methods for details). Three participants whose estimates across the whole experiment had a correlation of <0.2 (Pearson's r) with those of the Bayesian learner were classified as poor learners and were excluded from further analysis. For the remaining participants ($N = 26$), the behavioral data indicated that participants reliably learned the overall share/win rates in both Gifters and Lottery conditions. Figure 1b displays the mean participant estimates over the course of the experiment for each of the four Gifters (left panel) and the corresponding Lotteries (right panel). Note that average participant estimates (solid lines) converge toward the true share/win rate (dotted lines) for each of the four Gifters and the corresponding Lotteries. The filled circles indicate the mean of the participants' final incentivized estimates (outside the scanner) for each Gifter and Lottery after trial 24 (day 1) and trial 48 (day 2). To assess performance, for each participant we calculated the difference between the participant's final incentivized estimates and the actual share/win rates for each Gifter and Lottery. Statistical testing revealed that none of the average differences (across participants) between participant estimate and actual share/win rates was significant (at $P < 0.05$, Bonferroni-corrected, one-sample t-tests).

We also assessed whether there were any significant systematic differences between learning in the Gifter and Lottery conditions. For each participant, the signed difference between the participant's actual estimates on each trial and those of the ideal Bayesian observer was calculated for each Gifter and Lottery and summed across trials, providing a single number for each participant, Gifter and Lottery. These numbers were entered into a 2 (Condition: Gifter/Lottery) × 4 (Share/Win Rate: 8%, 43%, 58%, 83%) repeated-measures ANOVA with Participant as a random factor. This analysis revealed that there was no main effect of Condition [$F(1,75) = 0.12$] though the main effect of Rate was trending [$F(3,75) = 2.32$, $P = 0.08$] —reflecting the fact that for both extreme Rates (8 and 83%) the difference between the ideal Bayesian estimates and participant estimates were of equal but opposite magnitude (i.e. Bayesian observer estimates were quicker to asymptote than those of participants in both conditions). Importantly, there was no significant interaction between Condition and Rate [$F(3,75) = 1.22$, $P = 0.31$]. To ensure that there were no significant systematic learning differences in early trials, we repeated this analysis for trials in the first half and quarter of the experiment. In both cases, there was no main effect of Condition [$F_{first\_half}(1,75) = 0.04$, $P = 0.84$; $F_{first\_quarter}(1,75) = 0.29$, $P = 0.60$] but the main effect of Rate was stronger [$F_{first\_half}(3,75) = 3.84$, $P = 0.01$; $F_{first\_quarter}(3,75) = 7.35$, $P < 0.001$]. Importantly, there was no interaction between Condition and Rate [$F_{first\_half}(3,75) = 0.57$, $P = 0.64$; $F_{first\_quarter}(1,75) = 0.3$, $P = 0.82$]. These analyses indicate that observed differences in BOLD signal between the two conditions was not likely due to discrepancies in task difficulty.

Neural correlates of social processing

Main effects. To identify regions in which there was a main effect of social processing, we compared the average response on Gifter trials to that on Lottery trials, both during the estimate period (Table 1), and during the outcome period (Table 2; see also Figure 3, top left and bottom panels). Consistent with the large body of work on mentalizing and social processing (e.g. Behrens *et al.*, 2009; Van Overwalle, 2009; Kennedy and Adolphs, 2012), these contrasts identified BOLD activity in a network of regions previously implicated in social cognition and face perception (including: Pc, TPJ, STS, rFFA, Anterior Temporal Lobe (ATL), mPFC). The only region to show greater activity for Lotteries than Gifters at both time points was a swath of ventral visual cortex (medial to, and not including, the Fusiform Face Area).

As there were observable differences in which of set of clusters were deemed 'significant' at estimate and at outcome, we further investigated this question using a contrast of contrasts ($[Gifter_{Estimate} > Lottery_{Estimate}] - [Gifter_{Outcome} > Lottery_{Outcome}]$. This analysis revealed that BOLD responses in both the left [peak Montreal Neurological Institute (MNI) XYZ = −27 −1 −20; 32 voxels, P(cluster) = 0.009, small-volume corrected (SVC) for bilateral amygdalae] and right [peak MNI XYZ = 27 −7 −20; 18 voxels, P(cluster) = 0.021, SVC for bilateral amygdalae] amygdalae were significantly greater for Gifter (compared with Lottery) outcomes but did not distinguish between the two conditions during estimates. Interestingly, this effect was primarily driven by a reduction/inversion of the response to Lottery outcomes. As the amygdalae are known be involved in the processing of social stimuli, coding of saliency and learning (e.g. Adolphs, 2010), this could be indicative of a decreased role for the amygdala in non-social outcome processing. However, given the lack of evidence for parametric modulation in the amygdalae in either condition, we refrain from drawing any strong conclusions. There were also two clusters in which responses were higher for the Lottery condition. In a large swath of occipital-parietal cortex [peak MNI XYZ = 45 −79 7; 3148 voxels, P(cluster) < 0.001, whole-brain corrected], the BOLD response showed higher responses to Lotteries than Gifters during outcomes but not during estimates. In bilateral ventral temporal cortex, BOLD responses were higher for Lotteries than Gifters at both estimate and outcome but significantly more so at estimate [peak MNI XYZ = 30 −61 −5; 389 voxels, P(cluster) < 0.001, whole-brain corrected]. The diffuseness of these responses (they cover many brain regions known to perform distinct functions), and the lack of parametric modulation, suggests they may be related to some large-scale modulatory

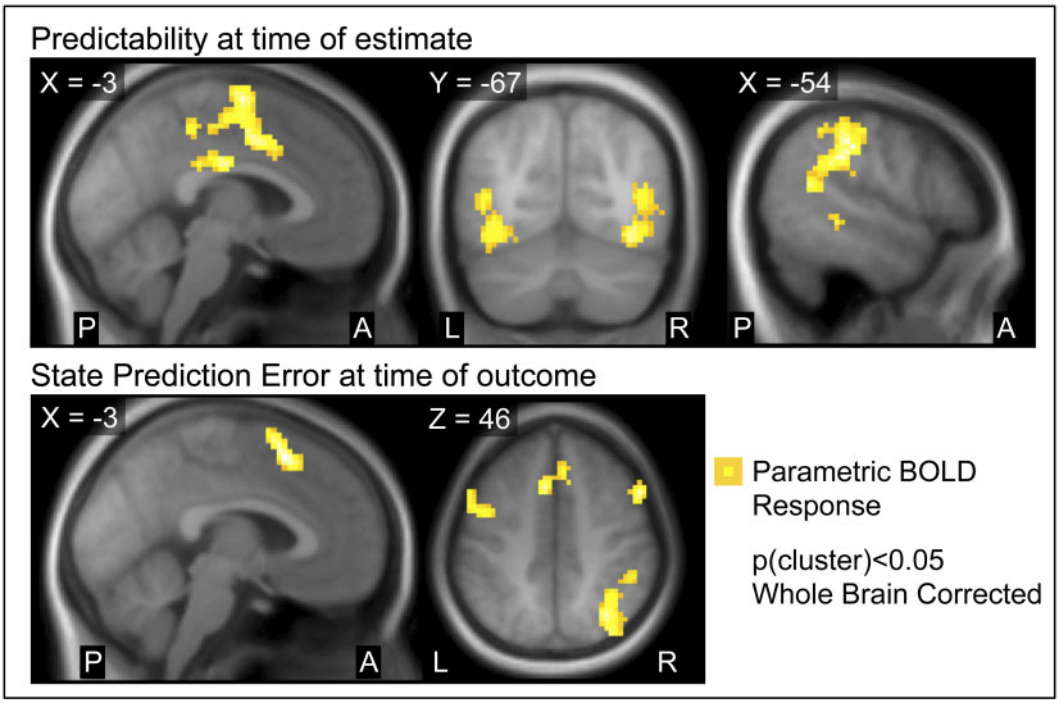

Fig. 2. BOLD signals reflecting shared learning computations for social and non-social targets. Top: Brain regions in which BOLD activity at the time of estimate correlated with trial-to-trial variation in participant estimates of the predictability of Gifter and Lottery behavior—i.e. the conjunction [$Gifter_{predictability}$ & $Lottery_{predictability}$]. Bottom: Brain regions in which BOLD activity at the time of outcome correlated with trial-to-trial variation in participant SPE for both Gifter and Lottery outcomes—i.e. the conjunction of [$Gifter_{SPE}$ & $Lottery_{SPE}$]. Statistical maps were thresholded at $P < 0.005$ (voxel-wise) and whole-brain cluster-correction was applied ($P < 0.05$; see Methods).

process (e.g. attention) but makes their interpretation difficult. Because we focus on specific computations related to learning, we mention these main effect findings here for completeness, but do not discuss them further.

Parametric effects. We were specifically interested in identifying BOLD activity that varied parametrically with participants' trial-to-trial learning about Gifter behavior. With this in mind, we searched for parametric signals that reflected how predictable participants believed Gifters' generosity to be at the time of estimation (i.e. predictability), and the magnitude of participant surprise at Gifter behavior when the Gifter's choice was revealed at outcome (i.e. state or action PE; Gläscher *et al.*, 2010; Suzuki *et al.*, 2012).[4] Importantly, our Lottery control condition allowed us to identify regions in which signals reflecting predictability and SPE were specific for learning about people.

Neural correlates of predictability. At the time when participants entered their estimate, we were specifically interested in identifying brain regions with signals that reflected participants' beliefs about how predictable Gifters' behavior and Lottery payoffs were, i.e. Gifter/Lottery predictability (see Methods). This signal was high when participants believed Gifter and Lottery behavior was highly deterministic (i.e. when their estimate of overall Gifter/Lottery outcome probability was closer to either 0 or 100%), and lowest when they believed Gifter and Lottery behavior was purely random (i.e. when their estimate was 50%).[5] We first examined the conjunction of Gifter and Lottery predictability to identify regions in which neural activity reflected domain general estimations. We found a number of regions in which BOLD activity was positively correlated with both Gifter and Lottery predictability (Figure 2 and Table 3) —i.e. the more predictable the participant believed a Gifter or Lottery to be, the higher the BOLD signal. We next investigated whether BOLD activity in any region selectively represented predictability for Gifters and not Lotteries (and vice versa). No clusters of BOLD activity survived our whole-brain-corrected cluster threshold for either contrast ([$Gifter_{predictability} > Lottery_{predictability}$] or [$Lottery_{predictability} > Gifter_{predictability}$]). Finally, no regions were found to correlate negatively with predictability in either condition.

Neural correlates of SPE. During the outcome period, we looked for signals that could be used to update a neural representation of Gifters' generosity and Lotteries' payoff likelihood, given the trial outcome and the participant's estimate at the beginning of the trial, i.e. a PE or surprise signal. Because our participants were estimating stimulus-outcome transition probabilities and did not receive rewards from outcomes themselves, we used

4 Because participants were estimating the likelihood that Gifters were associated with one outcome or another, and the outcomes had no inherent positive or negative value for the participants, we used SPEs and predictability (a V-shaped function of estimate) rather than reward PEs and raw estimates. We also investigated whether there were any regions in which BOLD activity directly reflected participants' estimates of generosity (in the place of predictability) or signed PEs often found in studies of reward learning (in the place of SPE). The only significant cluster of activity identified was for the contrast of Gifter PE > Lottery PE and was located in early visual cortex. Given the large body of evidence on the function of early visual cortex, as well as the focus of this study on learning in the absence of reward to self, we do not discuss this result further.

5 Note: predictability need not correspond to what participants would report were they asked how confident they were in their estimates.

Table 1. Main effects at time of estimate

	Pk vox MNI				
nVox	X	Y	Z	Peak vox region (nVox)	Other AAL regions > 5 voxels
Main effect of Gifters > Lotteries at time of estimate					
2475	42	−52	−20	Fusiform_R(143)	Temporal_Mid_R(735), Angular_R(355), Temporal_Inf_R(209), Temporal_Sup_R(179), Parietal_Inf_R(115), Occipital_Inf_R(104), Occipital_Mid_R(51), Temporal_Pole_Mid_R(48), Cerebelum_Crus1_R(46), Cerebelum_6_R(39), SupraMarginal_R(25), Temporal_Pole_Sup_R(6)
1809	6	50	37	Frontal_Sup_Medial_R(388)	Frontal_Sup_Medial_L(331), Frontal_Med_Orb_R(155), Frontal_Med_Orb_L(141), Frontal_Sup_R(132), Cingulum_Ant_R(107), Cingulum_Ant_L(101), Rectus_L(97), Frontal_Sup_L(77), Rectus_R(56)
1567	−45	−64	19	Temporal_Mid_L(651)	Angular_L(173), Occipital_Mid_L(137), Fusiform_L(59), Temporal_Sup_L(52), Occipital_Inf_L(48), SupraMarginal_L(25), Cerebelum_Crus1_L(24), Temporal_Inf_L(6)
1028	33	35	−26	Frontal_Inf_Tri_R(234)[a]	Frontal_Inf_Orb_R(215), Frontal_Inf_Oper_R(124), Frontal_Mid_R(124), Precentral_R(80), Temporal_Pole_Sup_R(58), Temporal_Pole_Mid_R(32), Insula_R(19), Frontal_Mid_Orb_R(7)
893	0	−52	34	Precuneus_L(268)	Precuneus_R(411), Cingulum_Post_L(71), Cingulum_Mid_L(40), Cingulum_Mid_R(31), Cingulum_Post_R(20), Cuneus_L(14), Cuneus_R(13)
235	−39	14	−20	Temporal_Pole_Sup_L(75)	Frontal_Inf_Orb_L(78), Temporal_Inf_L(30), Frontal_Inf_Tri_L(14), Insula_L(12), Temporal_Pole_Mid_L(8)
156	−33	8	49	Frontal_Mid_L(53)	Precentral_L(88), Frontal_Mid_L(53)
Main effect of Lotteries > Gifters at time of estimate					
3089	−27	−55	−11	Fusiform_L(222)	Lingual_R(308), Occipital_Mid_L(290), Lingual_L(282), Occipital_Mid_R(278), Fusiform_R(277), Calcarine_L(219), Cerebelum_6_L(134), Cerebelum_6_R(112), Calcarine_R(95), Cerebelum_4_5_R(79), Occipital_Sup_R(67), Cerebelum_Crus1_L(55), ParaHippocampal_R(50), Occipital_Sup_L(49), Occipital_Inf_L(43), Cerebelum_4_5_L(41), Cerebelum_Crus1_R(40), ParaHippocampal_L(13), Cuneus_R(9), Occipital_Inf_R(7)

[a]The peak voxels of these clusters were in regions undefined by AAL (e.g. white matter), so the largest contributing AAL region is reported instead

the SPE (Gläscher *et al.*, 2010; see also action PE, Suzuki *et al.*, 2012) which, in our paradigm, is the absolute value of the difference between the outcome value and estimate value. This SPE is high when the outcome is unexpected (e.g. if a Gifter's decision is out of character) and low when it is not.

We first looked for common regions of SPE-related activity across the Gifter and Lottery conditions ([$Gifter_{SPE}$ & $Lottery_{SPE}$]; Figure 2 and Table 3). This conjunction analysis identified four clusters—dlPFC (bilaterally), dmPFC and right lateral parietal cortex (left lateral parietal cortex was also present but did not pass cluster threshold) —in which BOLD activity reflected SPE for both Gifters and Lotteries. It is noteworthy that Gläscher *et al.* (2010) found a similar pattern of activity (see also Suzuki *et al.*, 2012) that was positively correlated with SPE in their model-based state-learning condition, which is highly similar in structure to our Lottery condition.

We were most interested in determining whether there were any regions in which Gifter SPE was represented and Lottery SPE was not. We therefore looked for regions in which BOLD activity was correlated specifically with SPE for Gifters and not with SPE for Lotteries. The contrast of [$Gifter_{SPE} > Lottery_{SPE}$] identified robust activity in the Pc [Figure 3, top right and bottom panel; peak MNI XYZ = 6 −67 31; 394 voxels, P(cluster) < 0.001, whole-brain corrected] as well as a single cluster covering portions of the right thalamus [peak MNI XYZ = 12 −16 13; 49 voxels, P(cluster) = 0.012, SVC for bilateral thalamus] and the right caudate [peak MNI XYZ = 21 −16 22; 23 voxels, P(cluster) = 0.058, SVC for bilateral caudate]. Inspection of the pattern of response within the Pc cluster on low (L), medium (M) and high (H), Gifter and Lottery SPE trials (Figure 3, top left panel; see Methods for details), verified that it selectively reflected SPE for Gifters and *not* Lotteries (note: this was not the case for the sub-cortical ROIs). These data suggest a specific role for the Pc in representing when another person's behavior is surprising given our prior beliefs about them, a calculation critical for social learning.

Parametric effects related to social outcomes in the Pc are in a distinct subregion. Our computational approach enabled us to investigate the extent to which the parametrically varying learning-related BOLD activity in the Pc (i.e. Gifter SPE) occurred in a spatially distinct region from commonly found BOLD activity in the Pc/PCC reflecting the main effect of social processing (i.e. the contrast of [Gifters > Lotteries]) at the time of outcome. This analysis revealed that the SPE ROI was in a distinct subregion of the main effect ROI, with the former located entirely within the Pc, dorsal and posterior to the latter, which spanned the border into the PCC (Figure 3, bottom panel).

Discussion

This study used a learning paradigm to both identify BOLD signals that reflect neural computations for learning about other people and assess their specificity for social learning compared with learning about non-social contingencies. To answer these questions, our paradigm directly contrasted learning about other people to learning in a computationally well-matched non-social condition. Furthermore, to isolate learning about other peoples' generosity from learning about rewards,

Table 2. Main effects at time of outcome

	Pk vox MNI				
nVox	X	Y	Z	Peak vox region (nVox)	Other AAL regions > 5 voxels
Main effect of Gifters > Lotteries at time of outcome					
7100	51	−58	28	Angular_R(341)	Temporal_Mid_R(760), Frontal_Sup_Medial_R(524), Frontal_Sup_Medial_L(409), Frontal_Inf_Orb_R(353), Frontal_Sup_R(328), Temporal_Inf_R(303), Frontal_Inf_Tri_R(296), Temporal_Pole_Mid_R(234), Frontal_Mid_R(225), Temporal_Sup_R(206), Frontal_Med_Orb_R(160), Rectus_L(157), Temporal_Pole_Sup_R(145), Frontal_Sup_L(143), Frontal_Med_Orb_L(141), Parietal_Inf_R(131), Frontal_Inf_Oper_R(125), Cingulum_Ant_R(118), Rectus_R(106), Cingulum_Ant_L(79), Precentral_R(72), Hippocampus_R(68), Amygdala_R(61), Thalamus_R(58), Insula_R(45), Supp_Motor_Area_R(35), ParaHippocampal_R(29), Olfactory_R(21), Frontal_Sup_Orb_L(21), Thalamus_L(20), SupraMarginal_R(19), Occipital_Mid_R(18), Fusiform_R(14), Putamen_R(14), Frontal_Mid_Orb_R(14), Frontal_Sup_Orb_R(11), Lingual_R(7)
2495	−60	−55	16	Temporal_Mid_L(909)	Angular_L(184), Temporal_Inf_L(160), Frontal_Inf_Orb_L(147), Temporal_Pole_Sup_L(124), Frontal_Inf_Tri_L(85), Temporal_Pole_Mid_L(72), Insula_L(42), Amygdala_L(42), Temporal_Sup_L(35), Hippocampus_L(24), SupraMarginal_L(15), Putamen_L(9), Olfactory_L(6)
858	6	−55	40	Precuneus_R(302)	Precuneus_L(226), Cingulum_Mid_R(90), Cingulum_Post_L(76), Cingulum_Mid_L(59), Cingulum_Post_R(48)
210	48	−85	−14	Occipital_Inf_R(97)[a]	Lingual_R(9)
174	−30	−82	−41	Cerebelum_Crus2_L(131)	
173	42	−49	−20	Fusiform_R(66)	Temporal_Inf_R(60), Cerebelum_Crus1_R(27), Cerebelum_6_R(20)
Main effect of Lotteries > Gifters at time of outcome					
6554	−24	−79	−14	Lingual_L(445)[b]	Occipital_Mid_L(516), Calcarine_L(460), Lingual_R(438), Occipital_Mid_R(424), Calcarine_R(385), Occipital_Sup_R(345), Occipital_Sup_L(307), Cuneus_R(306), Cuneus_L(273), Fusiform_R(268), Fusiform_L(219), Cerebelum_6_R(201), Cerebelum_6_L(195), Parietal_Sup_L(149), Parietal_Sup_R(132), Cerebelum_4_5_R(79), Cerebelum_4_5_L(71), Precuneus_L(64), Cerebelum_Crus1_R(61), ParaHippocampal_R(42), Occipital_Inf_L(39), Precuneus_R(27), Cerebelum_Crus1_L(23), ParaHippocampal_L(18), Vermis_4_5(18), Occipital_Inf_R(17), Angular_R(16), Parietal_Inf_L(9)

[a]The peak voxels of these clusters were in regions undefined by AAL (e.g. white matter), so the largest contributing AAL region is reported instead.
[b]The reader may note that the large cluster identified by the contrast of Lotteries > Gifters contained a small amount of activity in the Pc. This was at the edge of the cluster, which was primarily located in early visual cortex, and likely resulted from spatial smoothing. Visual inspection of the regions verified that there was no overlap with the region of the Pc that was parametrically modulated by Gifter SPE.

participants in this study did not receive rewards during learning. Our results demonstrate that while there is considerable overlap between regions of the brain exhibiting learning-related signals for learning about people and non-social stimuli, there is also at least one region, the Pc, in which learning-related signals are restricted to the social domain. Specifically, the BOLD signal in the Pc at the time of outcome reflected SPEs (i.e. how surprising participants found the others' actions; Figure 3) for people, and *not* for Lotteries. These findings provide clear evidence that there may be something unique about the neural computations underlying learning when social entities are involved.

The question of whether there exist brain regions that are specifically recruited for the processing of social stimuli has been extensively debated. One recurrent issue is that previous work has generally compared average BOLD activity in two or more conditions (e.g. thinking about people *vs* non-social control stimuli), leaving them open to the criticism that the social conditions may simply be more salient or require more processing than the non-social conditions. A major strength of our study is that we probed for computational signals that parametrically varied with learning. Specifically, we demonstrated that the BOLD signal in many brain regions (Table 3) reflected trial-by-trial estimates of predictability and/or SPE but *did not* distinguish between Gifter and Lottery trials. In other words, the parametric BOLD signal indicates they were performing similar computations in both cases. This suggests that these regions perform general computations related to learning of stimulus-outcome contingencies (i.e. they are not specialized for social cognition). In contrast, activity in the Pc reflected specific parametric variation for learning about people and *not* for learning about non-social targets, suggesting that these computations are specific to social learning.

The robust and selective SPE signal we found for Gifters in the Pc suggests this region may play a unique role in signaling the need to update one's mental representation of another person's character. Previous work has implicated the Pc and adjacent PCC in components of social processing such as updating impressions (Schiller *et al.*, 2009; Ma *et al.*, 2011, 2012; Mende-Siedlecki *et al.*, 2013) and sensitivity to social outcomes (Rilling *et al.*, 2004; Delgado *et al.*, 2005; Tomlin *et al.*, 2006). This study extends these findings by identifying a parametrically varying, socially specific, PE signal in the Pc. The Pc has been shown to subserve a wide range of cognitive processes, including mental imagery, episodic memory retrieval and self-processing (for review see Cavanna and Trimble, 2006) and is the hub of the default mode network (Raichle *et al.*, 2001; Buckner *et al.*, 2008). More recently, it has

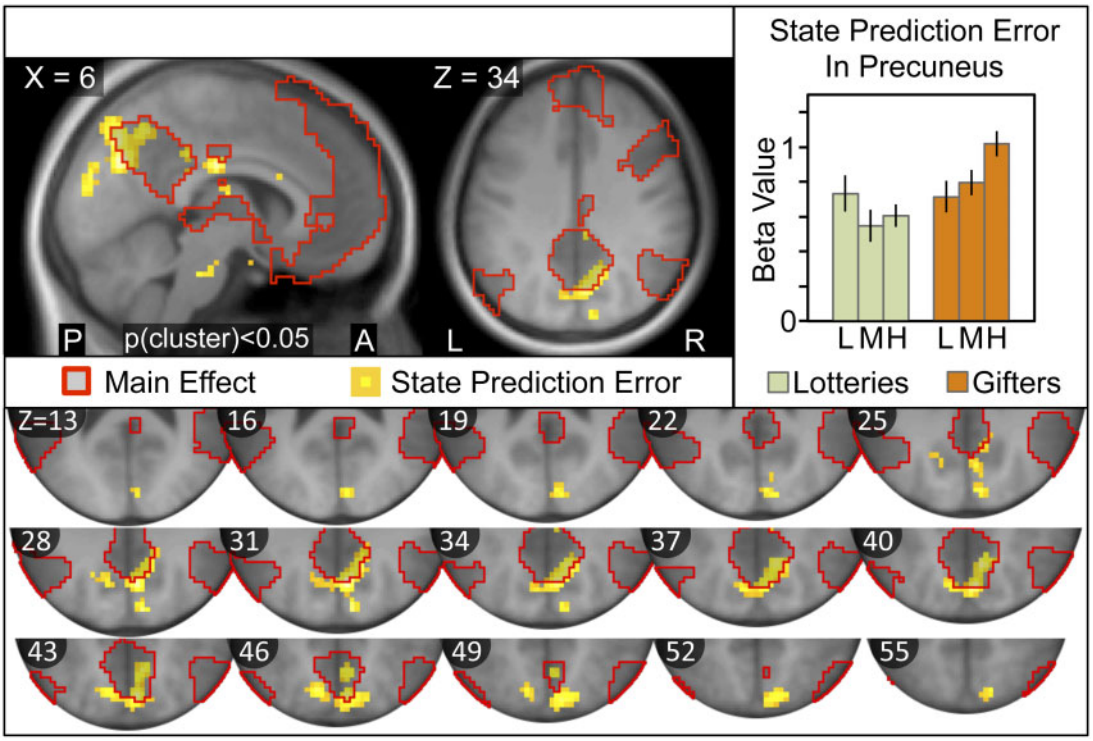

Fig. 3. BOLD signals specific for learning about people. Top Left: BOLD activity in the Pc [top panel; 394 voxels, P(cluster) < 0.001, whole-brain corrected, peak MNI XYZ = 6 −67 31] was selectively correlated with SPE (i.e. surprise) for Gifters and *not* SPE for Lotteries. The red-bordered overlay indicates brain regions that showed a main effect of social outcomes—i.e. that were more active overall for $Gifter_{outcomes}$ compared with $Lottery_{outcomes}$. Statistical maps were thresholded at $P < 0.005$ (voxel-wise) and whole-brain cluster-correction was applied ($P < 0.05$; see Methods). Top Right: The bar graph shows the mean beta value for trials with low (L), medium (M) and high (H) SPEs in the Pc (participant ROIs were independently created using a leave-one-out procedure—see Methods). Bottom: Posterior axial sections showing the full extent of brain regions in which the BOLD signal reflected the parametric (SPE) and main effects for the contrast of [Gifters > Lotteries]. It is noteworthy that the parametric response occurs in a distinct subregion of the area in which there is a main effect (see Results).

Table 3. Conjunctions [Gifters & Lotteries] of parametric effects

	Pk vox MNI				
nVox	X	Y	Z	Peak vox region (nVox)	Other AAL regions > 5 voxels
Clusters where BOLD signals positively correlated with social and non-social predictability					
553	−60	−31	37	SupraMarginal_L(171)	Parietal_Inf_L(84), Temporal_Mid_L(80), Postcentral_L(73), Parietal_Sup_L(44), Temporal_Sup_L(27)
537	−3	−7	64	Supp_Motor_Area_L(126)	Cingulum_Mid_L(164), Cingulum_Mid_R(99), Supp_Motor_Area_R(74), Paracentral_Lobule_L(11)
289	36	−64	−8	Occipital_Inf_R(65)	Temporal_Mid_R(113), Temporal_Inf_R(35), Fusiform_R(34), Occipital_Mid_R(18)
276	−39	−70	−8	Occipital_Inf_L(45)	Fusiform_L(115), Occipital_Inf_L(45), Temporal_Mid_L(29), Temporal_Inf_L(12), Cerebelum_6_L(7)
194	21	−19	64	Precentral_R(92)	Postcentral_R(74)
Clusters where BOLD signals positively correlated with social and non-social SPE					
457	48	8	31	Precentral_R(40)	Frontal_Mid_R(169), Frontal_Inf_Oper_R(138), Frontal_Inf_Tri_R(56)
344	36	−55	52	Angular_R(110)	Parietal_Inf_R(85), Parietal_Sup_R(82), Occipital_Sup_R(12), SupraMarginal_R(8), Occipital_Mid_R(6)
303	−51	17	34	Frontal_Inf_Oper_L(50)	Precentral_L(132), Frontal_Mid_L(56), Frontal_Inf_Tri_L(31)
181	−6	14	49	Supp_Motor_Area_L(92)	Supp_Motor_Area_R(50), Frontal_Sup_Medial_R(31)

been suggested that, the PCC (which is highly interconnected with the Pc; Margulies *et al.*, 2009), subserves change detection for adapting to a changing environment (Pearson *et al.*, 2011). Our results would suggest that the adjacent Pc may play a similar role, with specific emphasis on the social environment.

It is notable that the region of Pc in which we find the socially specific SPE signal is a distinct subregion of the Pc/PCC region in which we found a main effect of processing outcomes related to social *vs* non-social targets. That the Pc is comprised of distinct subregions with connections to distinct networks in

the brain is quite well documented both anatomically and functionally (Cavanna and Trimble, 2006; Margulies *et al.*, 2009; Zhang and Li, 2012), as is the interconnectivity between the Pc and PCC. Margulies *et al.* (2009) identified four distinct regions and labeled them on the basis of patterns of functional connectivity arising from seeds placed around the Pc; sensorimotor, visual, cognitive and limbic. An informal comparison of the regions we find to those of Margulies *et al.* (2009) suggests that the region in which brain activity correlated with Gifter SPE likely corresponds to their 'cognitive' region, whereas the region in the PCC showing the main effect of [Gifters > Lotteries] but no parametric modulation by Gifter SPE likely corresponds to their 'Limbic' region.

One possible mechanistic explanation for the socially specific SPE signal in the Pc is that when learning about social targets, the information about SPE in lateral fronto-parietal regions (where we find SPE signals for both social and non-social targets, and part of Margulies *et al.*'s 'cognitive' functional connectivity signature) is forwarded to the Pc. Once there, SPE information can be integrated with information from limbic systems and/or forwarded to other regions of the social cognition network. Another possibility is that coactivation of the limbic network and cognitive networks resulting from the presence of novel information about a social target creates the conditions for SPE to be represented in the Pc. These are but two possibilities of many that future studies will need to arbitrate between.

While this article was under review, Hackel *et al.* (2015) published a study with a similar design and goal, namely to identify neural computations that subserve learning about others' generosity. In contrast to the results reported here, the authors did not find any evidence for socially specific neural computations related to generosity learning (compared with a computationally matched non-social control condition). Two salient differences in the studies may account for this discrepancy in findings. First, Hackel *et al.* (2015) had participants learn whether partners were rewarding and/or generous by receiving money (i.e. reward) as a direct consequence of their partners' actions. Critically, while generosity in Hackel *et al.* (2015) varied orthogonally to reward amount, it still indicated how rewarding a given partner could be for the participant, and as such could be considered information about each partner's reward potential. It may be that information about self-relevant reward potential triggers the involvement of subcortical structures important for reward learning, and does so equally for social and non-social entities. In contrast, our study specifically examined learning about others in the absence of experienced reward, to disentangle social processes from basic reward learning. One possibility is that by having participants learn associations between cues and behaviors, rather than cues and rewards, we encouraged the use of more model-based forms of learning (e.g. Gläscher *et al.*, 2010; Doll *et al.*, 2012). Consistent with this idea, while Hackel *et al.* observed a ventral striatal learning signal (characteristic of studies that involve learning from experienced rewards), we do not (for either state or reward PEs), instead we observed PE signals in regions associated with model-based learning. A second distinction is that participants in Hackel *et al.* (2015) were provided with more specific information about the value and context of an agent's decisions than our participants (who only received binary information about Gifter/Lottery behaviors). This may have induced their participants to engage in value-based computations, which in turn may lead to the engagement of value-sensitive neural systems. Future work will need to investigate these questions directly.

There are some important limitations to acknowledge. First, by design, 'social learning' in our task is only distinguished from non-social learning by the presence of a face and participant knowledge that they were viewing the behavior of real individuals.[6] As such, the social condition was minimally social, requiring only simple computations of stimulus-outcome associativity with little-to-none of the abstraction or complexity found in real-world social behavior. In light of this, it is noteworthy that we find robust main and parametric effects that are present during social learning only, and suggests that these systems may engage automatically when a social target is present. A natural question to ask is whether more complex and/or abstract outcome behaviors engage the same regions? We definitely believe so, and indeed, there is evidence that surprise at more abstract social outcomes does engage the Pc (e.g. Schiller *et al.*, 2009; Cloutier *et al.*, 2011a,b; Mende-Siedlecki *et al.*, 2013). Our study supports and furthers these findings by demonstrating that responses in this region are socially specific and scale with the magnitude of surprise. Future work should incorporate the abstract nature of real-world social behaviors (e.g. Anita helped Noah finish his homework) into more quantitative computational models. What we present here is a first step in a series of many that we hope will help to bring a more nuanced and mechanistic understanding of computations underlying human person perception.

A second limitation is that we find no regions that reflect predictability for Gifters and not for Lotteries. Previous studies have found signals in the dmPFC reflect the expected reward of an action in an economic game given the other player's likely actions (Hampton *et al.*, 2008) as well as other agents' behavioral ambiguity [i.e. the inverse of certainty (Yoshida *et al.*, 2010); see also (Jenkins and Mitchell, 2010)]. Additionally, signals in dmPFC and other regions (e.g. TPJ, ATL) have been found when participants are accessing trait representations both intentionally and spontaneously (e.g. Ma *et al.*, 2011; Hassabis *et al.*, 2014; Welborn and Lieberman, 2015). We do note that we found a small region of the left TP in which predictability for Gifters, but not Lotteries, was represented. Recent work suggests the TP play a role in storing associations between social targets and related concepts (Todorov and Olson, 2008; Ross and Olson, 2010; Olson *et al.*, 2013). However, because the signal we found did not survive whole-brain cluster-correction and we did not optimize our fMRI data collection to identify signals in the TP (where susceptibility artifacts are strong), we only note this as an avenue of future interest. Finally, it is also interesting to note that we did not identify any person-specific learning signals in the TPJ, a region consistently implicated in theory of mind processes (Saxe, 2010) and found in other studies investigation the neural computations of social learning (e.g. Behrens *et al.*, 2008; Hampton *et al.*, 2008; Boorman *et al.*, 2013). This may be because we did not require participants to actively consider the internal thought processes of Gifters (i.e. 'mentalizing'), only that they learn the association between each Gifter and outcomes. It is possible the TPJ is only engaged when participants are actively 'mentalizing'.

Stepping back, our findings provide strong evidence that learning about other people may recruit a set of specifically social neural computations distinct from those that subserve general learning about stimulus-outcome contingencies. That these social signals are complex and reflect computations necessary for learning, suggests that they are not simply the result of increased attention or depth of processing. Rather, these

6 Participants were recruited through the Center for Experimental Social Science at the California Institute of Technology, which has a strict and explicit rule against deceiving participants.

social neural signals may result from unique demands that the social world places on the brain. One possibility is that representing the complex, ever-changing, multidimensionality of another individual's character may require systems that can quickly and flexibly assign semantic associations (another putative role of the anterior temporal lobes; Olson *et al.*, 2013) to different agents and social groups. In contrast to the social contingencies of other minds, the non-social contingencies in our world are relatively stable and mostly beholden to a generalizable set of rules. This necessity for flexibility and high dimensionality may have led us to develop specialized cortical systems capable of handling such problems.

Acknowledgements

The author thanks Antonio Rangel for contributions to study design and data analysis as well as Ralph Adolphs, Hanah Chapman, Shabnam Hakimi, Catherine Hartley, Cendri Hutcherson, Peter Sokol-Hessner and Bob Spunt for comments on the article.

Funding

This work was supported by the National Institute of Mental Health at the National Institutes of Health (grant number K01MH099343 to D.A.S.) and by a grant from the Gordon and Betty Moore Foundation to Antonio Rangel. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health. The author declares no competing financial interests.

Conflict of interest. None declared.

References

Adolphs, R. (2010). Annals of the New York Academy of Sciences: *The year in Cognitive Neuroscience*, **1191**, 42–61.

Ames, D.L., Fiske, S.T. (2013). Outcome dependency alters the neural substrates of impression formation. *NeuroImage*, **83**, 599–608.

Behrens, T.E.J., Hunt, L.T., Rushworth, M.F.S. (2009). The computation of social behavior. *Science*, **324**, 1160–4.

Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., Rushworth, M.F.S. (2008). Associative learning of social value. *Nature*, **456**, 245–9.

Bhanji, J.P., Beer, J.S. (2013). Dissociable neural modulation underlying lasting first impressions, changing your mind for the better, and changing it for the worse. *Journal of Neuroscience*, **33**(22), 9337–44.

Boorman, E.D., O'Doherty, J.P., Adolphs, R., Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron*, **80**(6), 1558–71.

Brainard, D.H. (1997). The psychophysics toolbox. *Spatial Vision*, **10**, 433–6.

Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, **1124**, 1–38.

Camerer, C.F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Castelli, F., Frith, C.D., Happé, F., Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, **125**, 1839–49.

Cavanna, A.E., Trimble, M.R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, **129**, 564–83.

Cloutier, J., Gabrieli, J.D.E., O'Young, D., Ambady, N. (2011a). An fMRI study of violations of social expectations: when people are not who we expect them to be. *NeuroImage*, **57**, 583–8.

Cloutier, J., Kelley, W.M., Heatherton, T.F. (2011b). The influence of perceptual and knowledge-based familiarity on the neural substrates of face perception. *Social Neuroscience*, **6**, 63–75.

Delgado, M.R., Frank, R.H., Phelps, E.A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, **8**, 1611–8.

Doll, B.B., Simon, D.A., Daw, N.D.D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, **22**(6), 1075–81.

Forsythe, R., Horowitz, J.L., Savin, N.E., Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, **6**(3), 347–69.

Frith, C.D., Frith, U. (2006). How we predict what other people are going to do. *Brain Research*, **1079**, 36–46.

Gläscher, J.P., Daw, N.D., Dayan, P., O'Doherty, J.P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, **66**, 585–95.

Hackel, L.M., Doll, B.B., Amodio, D.M. (2015). Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience*, **18**, 1233–5.

Hampton, A.N., Bossaerts, P., O'Doherty, J.P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 6741–6.

Hassabis, D., Sprend, R.N., Rusu, A.A., Robbins, C.A., Mar, R.A., Schacter, D.L. (2014). Imagine all the people: how the brain creates and uses personality models to predict behavior. *Cerebral Cortex*, **24**, 1979–87.

Henrich, J., Boyd, R., Bowles, S., *et al.* (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *The American Economic Review*, **91**, 73–8.

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

Kahneman, D., Knetsch, J.L., Thaler, R.H. (1986). Fairness as a Constraint on Profit Seeking: Entitlements in the Market. *American Economic Review*, **76**(4), 728–41.

Kana, R.K., Keller, T.A., Cherkassky, V.L., Minshew, N.J., Just, M.A. (2009). Atypical frontal-posterior synchronization of theory of mind regions in autism during mental state attribution. *Social Neuroscience*, **4**, 135–52.

Kennedy, D.P., Adolphs, R. (2012). The social brain in psychiatric and neurological disorders. *Trends in Cognitive Sciences*, **16**, 559–72.

Kennedy, D.P., Courchesne, E. (2008). The intrinsic functional organization of the brain is altered in autism. *NeuroImage*, **39**, 1877–85.

Kennedy, D.P., Redcay, E., Courchesne, E. (2006). Failing to deactivate: resting functional abnormalities in autism. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 8275–80.

Kleiner, M., Brainard, D.H., Pelli, D.G. (2007). What's new in Psychtoolbox-3? *Perception* **36**(14), 1.

Krajbich, I., Adolphs, R., Tranel, D., Denburg, N.L., Camerer, C.F. (2009). Economic games quantify diminished sense of guilt in patients with damage to the prefrontal cortex. *The Journal of Neuroscience*, **29**, 2188–92.

Lewis, P.A., Rezaie, R., Brown, R., Roberts, N., Dunbar, R.I.M. (2011). Ventromedial prefrontal volume predicts understanding of others and social network size. *NeuroImage*, **57**, 1624–9.

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**, 937–50.

Ma, N., Vandekerckhove, M., Van Overwalle, F., Seurinck, R., Fias, W. (2011). Spontaneous and intentional trait inferences recruit a common mentalizing network to a different degree: spontaneous inferences activate only its core areas. *Social Neuroscience*, **6**, 123–38.

Mar, R.A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology*, **62**, 103–34.

Margulies, D.S., Vincent, J.L., Kelly, C., *et al.* (2009). Precuneus shares intrinsic functional architecture in humans and monkeys. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 20069–74.

Mende-Siedlecki, P., Cai, Y., Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, **8**(6), 623–31.

Nichols, T., Brett, M., Andersson, J., Wager, T., Poline, J-B. (2005). Valid conjunction inference with the minimum statistic. *Neuroimage*, **25**(3), 653–60.

Olson, I.R., McCoy, D., Klobusicky, E., Ross, L.A. (2013). Social cognition and the anterior temporal lobes: a review and theoretical framework. *Social Cognitive and Affective Neuroscience*, **8**, 123–33.

Pearson, J.M., Heilbronner, S.R., Barack, D.L., Hayden, B.Y., Platt, M.L. (2011). Posterior cingulate cortex: adapting behavior to a changing world. *Trends in Cognitive Sciences*, **15**, 143–51.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, **10**, 437–42.

Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., Shulman, G.L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 676–82.

Rilling, J.K., Sanfey, A.G., Aronson, J.A., Nystrom, L.E., Cohen, J.D. (2004). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage*, **22**, 1694–703.

Ross, L.A., Olson, I.R. (2010). Social cognition and the anterior temporal lobes. *NeuroImage*, **49**, 3452–62.

Sallet, J., Mars, R.B., Noonan, M.P., *et al.* (2011). Social network size affects neural circuits in macaques. *Science*, **334**, 697–700.

Saxe, R.R. (2010). The right temporo-parietal junction: a specific brain region for thinking about thoughts. In: Leslie, A., German, T., editors. *Handbook of Theory of Mind*. p. 1–35. Taylor and Francis, UK: Psychology Press.

Schiller, D., Freeman, J.B., Mitchell, J.P., Uleman, J.S., Phelps, E.A. (2009). A neural mechanism of first impressions. *Nature Neuroscience*, **12**, 508–14.

Suzuki, S., Harasawa, N., Ueno, K., *et al.* (2012). Learning to simulate others' decisions. *Neuron*, **74**, 1125–37.

Todorov, A., Olson, I.R. (2008). Robust learning of affective trait associations with faces when the hippocampus is damaged, but not when the amygdala and temporal pole are damaged. *Social Cognitive and Affective Neuroscience*, **3**, 195–203.

Tomlin, D., Kayali, M.A., King-Casas, B., *et al.* (2006). Agent-specific responses in the cingulate cortex during economic exchanges. *Science*, **312**, 1047–50.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., *et al.* (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, **15**, 273–89.

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

Welborn, B.L., Lieberman, M.D. (2015). Person-specific theory of mind in medial pFC. *Journal of Cognitive Neuroscience*, **27**, 1–12.

Yoshida, W., Seymour, B., Friston, K.J., Dolan, R.J. (2010). Neural mechanisms of belief inference during cooperative games. *The Journal of Neuroscience*, **30**, 10744–51.

Young, L.L., Camprodon, J.A., Hauser, M., Pascual-Leone, A., Saxe, R.R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 6753–8.

Zhang, S., Li, C.S. (2012). Functional connectivity mapping of the human precuneus by resting state fMRI. *NeuroImage*, **59**(4), 3548–62.