Brain and Cognition 88 (2014) 55–64

Contents lists available at ScienceDirect

Brain and Cognition

journal homepage: www.elsevier.com/locate/b&c

Neural correlates of social perception on response bias

Yeon Soon Shin, Hye-young Kim, Sanghoon Han *

Department of Psychology, Yonsei University, Seoul 120-749, Republic of Korea

ARTICLE INFO

Article history:
Accepted 27 April 2014
Available online 22 May 2014

Keywords:
Response bias
Social perception
Feedback-based learning
Caudate nucleus
DLPFC
fMRI

ABSTRACT

Accurate person perception is crucial in social decision-making. One of the central elements in successful social perception is the ability to understand another's response bias; this is because the same behavior can represent different inner states depending on whether other people are yea-sayers or naysayers. In the present study, we have tried to investigate how the internal biases of others are perceived. Using a multi-trial learning paradigm, perceivers made predictions about a target's responses to various suggested activities and then received feedback for each prediction trial-by-trial. Our hypotheses were that (1) the internal decision criterion of the targets would be realized through repeated experiences, and (2) due to positive–negative asymmetry, yea-sayers would be recognized more gradually than naysayers through the probabilistic integration of repeated experiences. To find neural evidence that tracks probabilistic integration when forming person knowledge on response biases, we employed a model-based fMRI with a State-Space Model. We discovered that person knowledge about yea-sayers modulated several brain regions, including caudate nucleus, DLPFC, hippocampus, etc. Moreover, when person knowledge was updated with incorrect performance feedback, brain regions including the caudate nucleus, DLPFC, dmPFC, and TPJ were also involved. There were overlapping regions for both processes, caudate nucleus and DLPFC, suggesting that these regions take crucial roles in forming person knowledge with repeated feedback, while reflecting acquired information up to the current prediction.

1. Introduction

A substantial part of our lives consists of meeting other people and getting to know them better. To thrive in our social life, we spend considerable time speculating on what others would think and do (Dunbar, 2003). That is because having accurate knowledge about other individuals is a key to success in this area (for a review, see Zaki & Ochsner, 2011). In this sense, it is crucial to know the degree to which people's behavior genuinely represents their mind. Response biases are closely related to this representational discrepancy between observable behavior and internal states of mind. According to classical decision theories (Green, 1966; Yonelinas, 2002), the criterion is an important factor that determines behavior, because a decision results from an interaction between the criterion and evidence (i.e. the strength of stimuli). A response bias entails that the decision criterion is biased, and in this manner, evidence falls above the criterion with either a high or a low probability. Thus, those who have a liberal criterion (*yea-sayers*) are likely to give positive responses, while those with a more conservative criterion (*naysayers*) are less likely to do so. Interestingly, response biases are stable across different contexts (Berg, 1955; Couch & Keniston, 1960; Furnham, 1986), and reflect underlying personality traits (e.g., acquiescence, agreement, and social desirability) (Couch & Keniston, 1960; Furnham, 1986). Thus, response biases carry valuable information required for understanding others' current behavior and making correct predictions about their future behaviors as well. Despite its importance in successful social cognition, however, little is known about how we perceive others' response biases in social interactions and which neural regions are involved in this process.

The key question that needs to be answered first is how we come to realize others' response biases. A growing body of research have shed light on the neural underpinnings for person impression formation and the update process with inconsistent information which violates the first impression (Bhanji & Beer, 2013; Ma et al., 2012; Mende-Siedlecki, Baron, & Todorov, 2013; Mende-Siedlecki, Cai, & Todorov, 2013). Although these studies neatly showed how rapidly-formed impression is updated with inconsistent evidence, in the current study, we explored further to investigate how impression is gradually formed over the multiple encounter with the opposite person. Unlike in a spontaneous trait inference, a perceiver should integrate a number of incidents in order to read

* Corresponding author.
E-mail addresses: ys.cecilia.shin@gmail.com (Y.S. Shin), hyeyoung.astin.kim@gmail.com (H.-y. Kim), sanghoon.han@yonsei.ac.kr (S. Han).

http://dx.doi.org/10.1016/j.bandc.2014.04.012
0278-2626/

how frequently the person answered positively/negatively. Thus, the crux of this problem lies in the multiple experiences with the person and the way we attribute those experiences. A single response observed only once is not attributable to either a situational factor or an internal response bias, and this fact necessitates multiple observations. Moreover, it is still impossible to infer a response bias if attribution is made to an external context. Consequently, we should focus more attention on the internal state of the person. In this manner, the multiple responses of a person can be generalized and integrated, and thus, can contribute to refining our knowledge about a single response criterion. In the repeated experiences, spontaneous probabilistic computation is likely to take place, like in typical multi-trial feedback-based learning (for a review, see Niv, 2009). That is, response observation would not necessarily explicitly bring up the concept of response bias, but the experiences will instead be calculated probabilistically and stored for future use. To quantify the amount of such probabilistic knowledge (i.e., learning state), a learning model named State-Space Model (SSM, Smith et al., 2004) is used. Previous studies on SSM showed that this model is more sensitive than other Reinforcement Learning (RL) models in capturing hidden learning performance (i.e., the degree to which knowledge is formed). This is because the model assumes an ideal observer who knows the entire trials when fitting observed data into a hidden learning equation (Kakade & Dayan, 2002; Smith et al., 2004) while other RL model only considers the trials up to the current observation. Therefore, SSM has strong validity in that the estimated amount of accumulated knowledge that is obtained from the model is well tracked at a neural level (Kumaran, Summerfield, Hassabis, & Maguire, 2009; Smith et al., 2004; Solomon, Smith, Frank, Ly, & Carter, 2011).

In addition to accumulated knowledge, knowledge-updating process itself is worth examining as well. Incorrect performance feedback is especially important here, because it elicits internal expectation violation, and so guides alternative correct predictions (Holroyd & Coles, 2002; Zanolie, Van Leijenhorst, Rombouts, & Crone, 2008). Although learners may capitalize on both correct and incorrect performance feedback, correct feedback conveys no more information than has already been accrued. In this sense, negative outcomes (i.e., "wrong") in feedback-based gradual learning have greater informational value than positive outcomes (i.e., "right").

Positivity and negativity of the biases is another critical issue. Given that yea-sayers have a higher probability of giving *positive* responses, while naysayers have a higher probability of giving *negative* ones, the positivity and negativity of responses would affect the way repeated experiences are generalized into knowledge about criterion. If positivity and negativity exhibit an asymmetrical influence upon the generalization process, this can have two possible consequences. The first is that positive responses are more readily generalizable and so serve as a better means of highlighting the underlying response criterion. In this way, observing a "yes" response would contribute more to person knowledge (asymmetrical integration – positivity dominance). The second possibility is that it is easier to recognize and integrate knowledge about a person's decision criterion from their negative responses (asymmetrical integration – negative dominance). On the other hand, if positivity and negativity do not asymmetrically influence the generalization process, observing a "yes" or a "no" would equally develop into adequate person knowledge about yea-sayers and naysayers (symmetrical integration).

In line with potential asymmetrical integration – in particular, the positivity dominance hypothesis – a substantial body of literature has shed light on positive–negative asymmetry in a range of diverse cognitive domains, such as valuation (Kahneman & Tversky, 1979), mood (Forgas, 1998), and episodic memory (Kensinger & Schacter, 2006; Ochsner, 2000). For example, it was discovered from mood-induced processing differences (Bless, Mackie, & Schwarz, 1992; Bless et al., 1996) that a perceiver is more likely to commit fundamental attribution errors when they are in a good mood, so that they tend to attribute the behavior of others to dispositional factors, while neglecting the role of situations (Forgas, 1998). More importantly, positivity itself also plays a role in the degree to which representation is generalized. As Tolstoy observes in a famous statement from *Anna Karenina*, "Happy families are all alike; every unhappy family is unhappy in its own way" (quoted in Unkelbach, Fiedler, Bayer, Stegmuller, & Danner, 2008), there seems to be much less variety in positivity than there is in negativity. Supporting this observation, positive objects have denser semantic nodes and a more homogeneous representation, while negative objects have a more heterogeneous representation (Unkelbach et al., 2008). Similarly, there are lines of research that suggest positivity induces broader and more generalized cognitive processing. Positive mood expands attentional breadth (Fredrickson & Branigan, 2005) and increases exploration behavior (Fredrickson, 2001). Moreover, positivity also plays a role in memory. Memories about positive stimuli are less accurate (Ochsner, 2000), while, in contrast, those about negative stimuli are more detailed and accurate (Kensinger & Schacter, 2006). For example, Ochsner (2000) found that individuals respond that they "know", but do not "remember", the positive item, suggesting that people have a less detailed memory about the positive items they encounter. On the other hand, Kensinger and Schacter (2006) have demonstrated that our memories about negative episodes are formed in a more detailed manner.

In the current study, we sought to investigate how multiple responses are generalized into the concept of a response criterion and how we learn about positive and negative response biases to different degrees. To examine this process, we employed a feedback-based learning paradigm (Gluck, Shohamy, & Myers, 2002; Knowlton, Mangels, & Squire, 1996; Maddox, Ashby, Ing, & Pickering, 2004). In our experimental paradigm, participants made a prediction and observed other people's responses to various suggested activities. A respondent's answers were expected to serve as a cognitive feedback for the observer, who will then accumulate this information and generalize it in order to make predictions. Although similar to traditional weather prediction tasks (Knowlton et al., 1996), our paradigm is distinct in that the objects (i.e., the activity that a responder was asked to perform) varied trial-by-trial, with a target person and question ("Would she perform the activity?") fixed. By doing so, we focused on inducing generalized knowledge rather than activity-reaction associations. With functional magnetic resonance imaging, we further aimed to explore the neural correlates of both the representation of probabilistic knowledge in the brain and the knowledge update process. Furthermore, by using a conjunction analysis, we sought to locate the knowledge-updating regions of the brain that are modulated by previously acquired knowledge. An information-sensitive caudate nucleus and DLPFC were hypothesized as providing the means by which knowledge about response biases was updated while being modulated by the amount of information.

2. Experiment 1: behavioral study

We first conducted a behavioral experiment in which participants made a prediction about a responder's reaction (i.e. "yes" or "no") and received feedback from multiple cases. With this feedback-based learning paradigm, we aimed to investigate if learning occurs in line with the responder's actual response tendency, and if the learning performances for yea-saying and naysaying are potentially asymmetrical.

2.1. Materials and method

2.1.1. Participants

Eighteen participants took part in the study in return for course credit or monetary compensation (10,000 KRW/h). Informed consent was obtained from all participants in accordance with the guidelines of the Yonsei University Institutional Review Board.

2.1.2. Stimuli

For the response prediction task, each target responder's face and name, as well as the particular activity they had to perform, were all used as prediction cues. In order to minimize possible gender effects in perceiving response biases, we confined all target responders to female. Three faces with neutral facial expressions were pooled from the Korea University Facial Expression Collection (KUFEC, Lee, Lee, Lee, Choi, & Kim, 2006) and peudorandomly assigned to Yea-sayer, Naysayer, and Neutral conditions. The photos had gray backgrounds and their size was set to 180 × 200 pixels. The faces assigned to yea-sayer and naysayer conditions, which were the conditions of interest, did not differ in terms of attractiveness, masculinity, or extroversion (attractiveness $t(17) = .08$, $p = .94$, *n.s.*, masculinity $t(17) = .09$, $p = .93$, *n.s.*, extroversion $t(17) = .-21$, $p = 83$, *n.s.*) when tested with pilot ratings from seven independent raters. The target responders had a predetermined dominant response of saying "yes" or "no" Specifically, a yea-sayer target responded "yes" for 80% of the trials while a naysayer target responded "no" for 80% of them. A neutral target responder with no response bias responded "yes" and "no" in equal probabilities (50% each). As a cue to assist learning, names appeared with faces. The set of activities, about which participants were asked to predict whether the target responder would be willing to perform them or not, consisted of 50 everyday activities including hobbies, sports activities, etc.

2.1.3. Procedure

Prior to the main prediction task, a familiarization process took place so that participants could focus on relevant information (i.e., face and activity) in the learning phase. Participants were instructed to remember the face-name pairs of six people, including the target responders, for 1 min and to then recall the name when each face was presented separately. After they were exposed to all the target responders, participants were given the main prediction task.

The response prediction task required participants to guess whether the target responder would agree or disagree to perform a suggested activity (Fig. 1). The goal was to have as many correct guesses as possible, which required participants to learn the general response tendency of the target responder trial-by-trial. There were five blocks in total, with each block consisting of thirty learning trials, followed by probe trials. The cue periods and feedback periods were repeated for each learning trial. For cue periods, the face and name of a target responder, as well as the activity in question, were presented on the screen for 3000 ms. The suggested activities for a target responder varied across the trials, while the set of activities was the same for all targets. The question asked to the participants (whether the target responder would perform a certain activity) appeared underneath the target's face. Responses were made by pressing the "G" and "H" keys on a keyboard for "yes" and "no", respectively. The response duration was limited and fixed to 3000 ms. When participants provided an answer, an asterisk appeared for the remaining response duration above the option selected. Feedback showing the target responders' actual answer ("yes" or "no") and the correctness of the prediction was given for 3000 ms after the prediction period. The correctness was indicated by the font color of the responder's answer: green for correct predictions and red for incorrect ones.

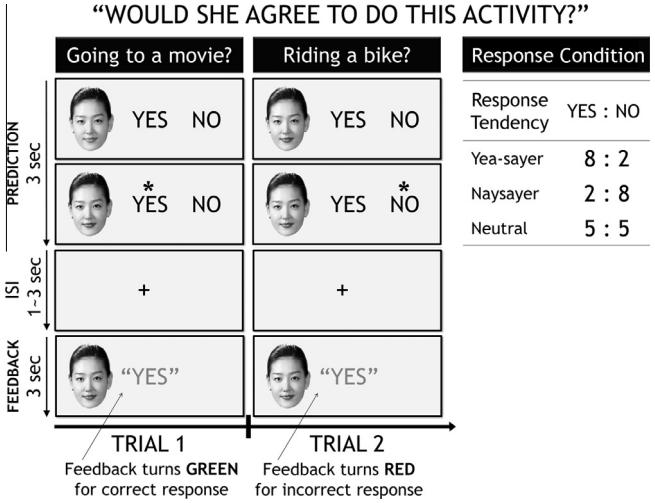

Fig. 1. Response prediction task. In a response prediction task, participants were asked to make a prediction on whether the target would agree to engage in an activity (3000 ms). Feedback indicating a response of target and the correctness of prediction followed (3000 ms). Correct performance feedback was indicated by green color whereas incorrect feedback appeared in red color. There were three conditions where targets answered yea for 80% (YEA), 20% (NAY), and 50% (NTR), respectively.

In-between the prediction and feedback periods a fixation cross appeared for a variable period (somewhere between 1000 and 3000 ms, but averaging at 2000 ms). Probe trials were given at the end of each block, which asked the participant to estimate the probability that each responder would allow the participants to check whether they had acquired knowledge properly. Participants provided their estimation of the targets' general response tendency, and the degree of confidence they felt in their estimation, by using a 5-point scale. Responses were self-paced and made by manipulating the"←" and "→" buttons, marked by a sticker on the keyboard, in order to go left and right respectively. A third button, between the two arrow keys, was used to finalize their responses. Each condition was pseudorandomized, keeping the response ratio (i.e., 8:2, 2:8, and 5:5) constant within each block. Each block consisted of trials on all targets that are intermixed with one another. Participants made predictions and received feedback 50 times for each target responder over five blocks, and there were 150 learning trials in total. After the experiment, participants completed personality inventories on the Barratt Impulsivity Scale 11th edition (Patton, Stanford, & Barratt, 1995) and the Social Interaction Anxiety Scale (Mattick & Clarke, 1998). Upon completion, all participants were debriefed.

2.2. Results and discussion

To check if participants were able to make a prediction in line with a target responders' response bias, we first checked the probability of correctness with a one-sample *t*-test. The probability of obtaining a correct answer was significantly higher than chance level (.5) for yea-sayer learning (*mean probability correct* = .75, $t(17) = 7.385$), $p < .01$) and naysayer learning (*mean probability correct* = .63, $t(17) = 2.83$, $p < .05$), but not for neutral target learning (*mean probability correct* = .56, $t(17) = 1.78$, $p > .05$, *n.s.*).

Of interest was the evident asymmetry between the way in which participants learned about and identified yea-sayers as opposed to naysayers. A paired *t*-test comparing the learning performances for yea-sayer and naysayer targets revealed that the accumulated probability of getting a correct answer under the yea-sayer condition was significantly higher than under the

naysayer condition (*mean difference* = .12, $t(17) = 2.914$, $p < .01$; Fig. 2). Interestingly, differences in performance success between learning conditions were not uniform across test blocks. Although the interaction of an ANOVA conducted between the bias direction and test blocks did not reach statistical significance ($F(4,14) = .73$, $p > .05$, *n.s.*), post hoc *t*-tests on bias conditions for each learning block showed that, for the first and second learning blocks, learning performances were significantly different between yea-sayer targets and naysayer targets (*mean difference*: first block $M = .18$, $t(17) = 2.5$, $p < .05$, second block $M = .17$, $t(17) = 2.35$, $p < .05$), but that these differences gradually lessened as the learning process continued (third block, $M = .07$, $t(17) = 1.29$, $p = .21$, fourth block, $M = .09$, $t(17) = 1.49$, $p = .15$, fifth block, $M = .09$, $t(17) = 1.49$, $p = .15$).

The results showed that predictions were made in line with each target's response bias, indicating that learning had occurred during the trials. Moreover, learning performance was better for yea-sayer targets than for naysayer targets, suggesting a discrete learning process. This also supports the hypothesis that positive responses are perceived as the evidence of a responder's internal judgment criterion (i.e., response bias) rather than as the evidence of his or her preference to engage in a specific activity that is attributable to the situation of being suggested to do such activity. To further test this hypothesis and to investigate whether knowledge about response biases is formed by integrating multiple experiences probabilistically, it is necessary to examine the neural correlates of probabilistic knowledge about response biases. In addition, it is possible that, when the number of experiences increases, knowledge on response biases could reach a similar level, despite the heterogeneity of the learning process. To answer these questions, Experiment 2 examined the neural correlates associated with the feedback-based learning of response biases, with increased number of learning trials.

3. Experiment 2: fMRI study

We conducted a functional Magnetic Resonance Imaging (fMRI) study to identify the neural correlates of forming integrated knowledge about other people's response biases. First, we sought to elucidate the neural underpinnings by means of which knowledge is updated through incorrect performance feedback (which has high informational value). Second, we aimed to locate the brain regions that reflect the amount of probabilistic knowledge about response biases asymmetrically for yea-sayer and naysayer. Third, in order to examine the common neural substrates in the two functional processes, we tried to find the overlapped regions by conducting a conjunction analysis. Finally, to support that response biases matter in social perception, we further examined individual difference according to perceivers' own response tendencies.

3.1. Materials and methods

3.1.1. Participants

Sixteen participants' functional neuroimaging data were acquired. All fMRI participants (10 women, mean age = 22.68 - years, range 19–27) provided informed consent according to the protocols approved by the Department Review Committee of Yonsei University, had normal or corrected-to-normal vision, and were right-handed and screened for magnetic imaging risk factors. One participant's imaging data were not included in the analysis due to severe signal artifacts.

3.1.2. Stimuli

As in Experiment 1, each target responder's face, name and activity were used as prediction cues for the prediction task. All target responders were female. Three faces with neutral facial expressions from Korea University Facial Expression Collection (KUFEC) were randomly assigned to three experimental conditions. The faces were presented in 180×200 pixels with gray backgrounds. A post hoc analysis of the pilot ratings given by seven independent raters revealed that the faces assigned to the positive and negative response bias conditions were not significantly different in terms of attractiveness, masculinity, or extroversion (attractiveness $t(15) = -.41$, $p = .69$, *n.s.*, masculinity $t(15) = .57$, $p = .57$, *n.s.*, extroversion $t(15) = -.88$, $p = .39$, *n.s.*). Yea-sayers and naysayers would predominantly answer "Yes" or "No" respectively for 80% of the trials. That is, 80% of trials were congruent with responders' internal response biases. A neutral target had no response bias. Names were shown below faces as cues. The number of given activities was increased to 80 as there was an increased number of learning trials in Experiment 2.

3.1.3. Behavioral procedure

The behavioral procedure was identical to Experiment 1 except for the number of trials, which was increased to 80 trials per target responder. This was done in order to test whether the learning performance for positive and negative response biases would ultimately become similar as experience accumulates (based on the results of Experiment 1, where the difference gradually decreased). Before proceeding to the response prediction task participants were first familiarized with the target responders by remembering their names and faces. In the multi-trial feedback-based learning

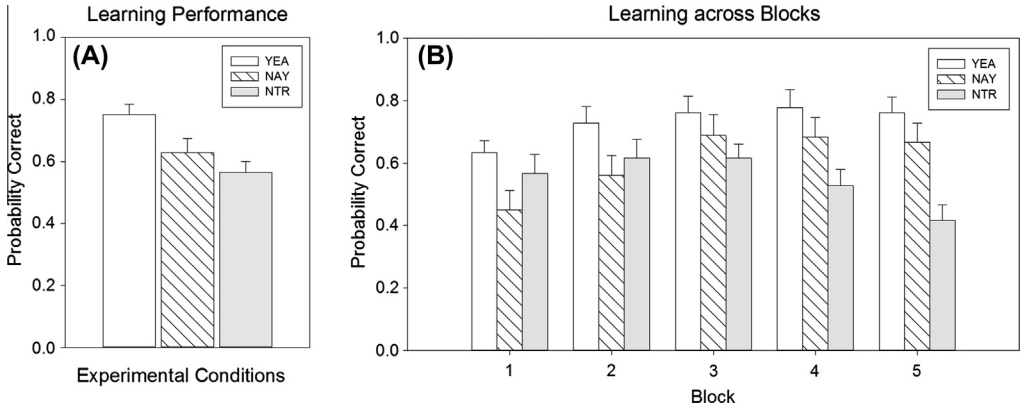

Fig. 2. Behavioral results. (A) Probability of correct prediction across all trials. Learning performance for positive bias was higher than that of negative bias, suggesting that positive response is taken as homogeneous concept of judgment criterion (i.e., response bias). (B) Probability of correct prediction for each learning block. Difference for YEA and NAY conditions were significant at the first two blocks.

paradigm, participants aimed to have as many correct guesses as possible. The prediction task was scanned in eight runs, each of which was comprised of thirty learning trials. Conditions were pseudorandomized within each block. After scanning each run, self-paced probe trials asked participants to identify the target responders' general tendency in agreements. After all eight runs, participants answered personality inventories on the Barratt Impulsivity Scale 11th edition (Patton et al., 1995) and the Social Interaction Anxiety Scale (Mattick & Clarke, 1998).

3.1.4. fMRI data acquisition and analysis methods

Functional magnetic resonance imaging (fMRI) data was acquired using an ISOL 3.0 Tesla forte MRI system (ISOL Tech, Oxford OR63). After the acquisition of high-resolution T1-weighted anatomical images for visualization, $T2^*$-weighted echo planar images (EPI) were obtained (TR = 2000 ms, TE = 31 ms, 25 axial slices parallel to the AC–PC plane, slice thickness = 5 mm, no gap, matrix size = 64 × 64 mm, FOV = 220 mm, interleaved collection).

fMRI data were analyzed using the statistical parametric mapping, SPM8 (Wellcome Department of Cognitive Neurology, London, UK). The first five images from each run were discarded for T1 equilibration. For slice timing correction, imaging data were temporally corrected by resampling all slices in time relative to the middle slice. Then, for motion correction, images were realigned to the first image of each run. To enable the analysis of any random effects, imaging data were normalized to match an echo planar imaging (EPI) template by using a 12-parameter affine and nonlinear cosine transformation. Normalized images were written in a functional voxel size of 3 × 3 × 3 mm and spatially smoothed using an 8 mm full-width-at-half-maximum (FWHM) isotropic Gaussian kernel. After preprocessing, fMRI data were analyzed with two models: a general linear model and a parametric regression model with state-space parameters. In the general linear model, all eight runs were concatenated in order to analyze runs with an insufficient number of trials in conditions of interest. In addition, head movement parameters were modeled as regressors of no interest and covaried out in order to minimize head movement artifacts. For a parametric regression model analysis, each run was rescaled for the mean global signal to be 100 across the volumes. Volumes were modeled by convolving a canonical hemodynamic response function and its temporal derivative with a boxcar function.

3.1.4.1. Knowledge update with incorrect performance feedback: general linear model analysis. To find brain regions that update knowledge on response biases with incorrect performance feedback, we first analyzed neuroimaging data with a general linear model. Specifically, congruent trials were modeled separately for yea-sayer and naysayer conditions that are then divided into correct and incorrect trials, resulting in four regressors of interest: "yea-sayer_correct", "yea-sayer_incorrect", "naysayer_correct", "naysayer_incorrect". Incongruent trials, where targets gave opposite responses to their biases, were modeled as a separate regressor and covaried out for the main contrast. The neutral target condition was modeled as a combination of four regressors of right and wrong predictions for "yes" and "no" responses, respectively. We used boxcar functions convolved with the hemodynamic response function, whose onsets were marked with feedback presentation. The cue period of all conditions was modeled and covaried out as a nuisance regressor. As our interest was focused on the contribution that incorrect performance feedback makes to the updating of knowledge, incorrect feedback was contrasted with correct feedback for both response bias conditions. Unless stated otherwise, the statistical threshold for the general linear model was corrected for multiple comparisons to $p < .05$ by using Monte Carlo simulations (Slotnick, Moo, Segal, & Hart, 2003).

3.1.4.2. The asymmetrical formation of person knowledge: parametric regression analysis with the State-Space Model. To ascertain the neural correlates of integrated probabilistic knowledge for yea-sayers that combine individual experiences, the estimated amount of knowledge was taken into account with parametric regression analysis. The accumulated amount of knowledge at each trial, which is expressed by the probability of obtaining a correct answer, was estimated by using the learning curve analysis software package (available at www.neurostat.mit.edu) in order to generate a state-space model parameter. This model employs a smoothing algorithm that estimates the hidden learning process based on the observed binary response sequence (assuming an ideal observer who knows the whole response to the end at each point of estimation). This smoothing algorithm has its advantage in that it detects and represents the learning process at the neural level better than traditional models such as found in the Reinforcement Learning (RL) model (Smith et al., 2004).

For estimation, this model contains two equations: an *observation equation* and a *state equation*. An *observation equation* is in a Bernoulli probability mass function of the earned data. It defines the probability of observing a particular response n_k (correct = 1, incorrect = 0) given a hidden learning probability x_k, when the probability of a correct response, p_k, is constrained between 0 and 1:

$$P_r(n_k|p_k, x_k) = p_k^{nk}(1-p_k)^{1-nk}$$

The state equation is in a Gaussian random-walk model to represent the hidden learning process, where a Gaussian random variable is defined from the distribution of the mean 0 and the variance σ_e^2:

$$x_k = x_{k-1} + \varepsilon_k$$

Using an expectation maximization (EM) algorithm, the state-space model finds the best fit for the derived value, p_k, from the observation equation and the state equation. Thus, it estimates the learning curve by showing the probability of a correct answer with the maximum likelihood of the observed data. As this learning curve is believed to accurately represent the learning process at the neural level (Smith et al., 2004), the obtained learning curve was used to determine the parameters that covary with the hemodynamic response function (see Fig. 4). The state-space parametric modulation regressors of biased response conditions for both cue and feedback periods were created in the SPM design matrix. The canonical hemodynamic response function was convolved with a boxcar function starting at the onset of each regressor. To ascertain where integrated knowledge formation about yea-sayer targets occurs, the degree to which knowledge modulates neural activation in the yea-sayer target condition was contrasted to that in naysayer target condition.

3.1.4.3. Overlapping regions for incorrect feedback processing and the estimated knowledge reflection: conjunction analysis. To find the overlapping regions that are more sensitive to any updated knowledge the more knowledge has been acquired about a target yea-sayer, a conjunction map was constructed. The map identified activation in both the incorrect feedback contrast map and the state-space model-based parametric modulation map.

3.1.4.4. The influence of perceiver's own response bias: regression analysis. To assess the influence of a perceiver's own social desirability bias in perceiving the response of other people, we ran a regression analysis that used the Marlowe-Crown Social Desirability Scale (Crowne & Marlowe, 1960), high scores of which indicate that respondents are biased toward providing socially desirable answers. Regardless of feedback correctness, we focused

on perceiving whether or not the target said they would perform certain activities, and looked for the regions with an increased contrast between negative and positive response perception as the social desirability bias increased. Beta estimates for medial-temporal lobe ROIs (Region of Interest), which we defined through the state-space model analysis reflecting knowledge accumulation, were extracted using the MarsBar Toolbox for SPM8 (available at http://marsbar.sourceforge.net/) and regressed with the MCSDS score.

3.2. Results and discussion

3.2.1. Behavioral data

Replicating the results of Experiment 1, one-sample *t*-test revealed that the probability of obtaining a correct prediction was significantly higher for both positive response bias conditions (*mean probability correct* = .71, $t(15) = 5.35$, $p < .01$) and negative response bias conditions (*mean probability correct* = .66, $t(15) = 2.80$, $p < .05$), but not for the control condition in which there was no response bias (*mean probability correct* = .51, $t(15) = 1.18$, $p = .86$, *n.s.*), suggesting that participants learned about the response tendency of other people in line with the actual response bias. The learning performance during positive response bias trials was significantly higher than for negative response bias trials on 50 accumulated experiences (*mean difference* = .08, $t(15) = 2.27$, $p < .05$), which was also found in Experiment 1, but when all 80 trials had been experienced, the difference in learning performance between yea-sayer and naysayer targets was no longer evident (*mean difference* = .06, $t(15) = .175$, $p = .10$). The results suggest that, when experiences accrue enough, knowledge about response biases could reach a similar level despite the heterogeneity of the learning process.

State-space model learning parameter, ideal observer certainty value, was also compared between the two experimental conditions. We conducted Areas Under the Curve (AUC) analysis to examine if there is greater certainty (i.e., less variability) in learning yea-sayer target. Paired *t*-test revealed that AUC is greater for yea-sayer learning certainty (mean AUC = 63.65) than naysayer learning certainty (mean AUC = 51.98), $t(15) = 2.44$, $p = .03$. The results suggest that certainty in learning process is significantly greater for yea-sayer target, supporting the hypothesis that representation of positivity is less variable.

3.2.2. fMRI data

3.2.2.1. Knowledge update with incorrect performance feedback. To examine how the knowledge is updated by incorrect performance feedback during learning, we contrasted it with positive feedback, regardless of target's actual bias direction. Contrasting incorrect and correct feedback revealed that the following brain regions all played a significant role: the caudate nucleus, bilateral dorsolateral prefrontal cortex (DLPFC), right temporo-parietal junction (rTPJ), dorsomedial prefrontal cortex (dmPFC), hippocampus, etc. (see Fig. 3 and Table 1). We did not find any gray matter voxels that were more involved in correct feedback at a very liberal threshold ($p < .01$, 5 contiguous voxel extent thresholds), thus supporting the idea that error feedback has the dominant role in the updating process. The results imply that the DLPFC and the caudate nucleus are involved in the knowledge updating process, especially in the production of the incorrect performance feedback that is used in prediction.

In addition, the involvement of the dmPFC and the rTPJ suggests that a social cognitive process that updates impressions occurs while learning about other people's response biases. These regions have been implicated in person perception, especially when a perceiver reads descriptions about others that are incongruent to their group identity, and in turn, prior expectation (Cloutier, Gabrieli, O'Young, & Ambady, 2011). Also, they are reliably implicated in impression formation and mentalizing tasks (Baron, Gobbini, Engell, & Todorov, 2011; Cloutier, Gabrieli, O'Young, & Ambady, 2011; Saxe & Baron-Cohen, 2006; Saxe & Kanwisher, 2003; Saxe & Wexler, 2005; Saxe, Whitfield-Gabrieli, Scholz, & Pelphrey, 2009; Schiller, Freeman, Mitchell, Uleman, & Phelps, 2009). Therefore, we believe that this activation pattern provides evidence of the social cognitive process beyond simple learning.

3.2.2.2. The asymmetrical formation of person knowledge. More importantly, we were primarily interested in finding which brain regions were modulated by variations in the amount of knowledge accrued. A parametric modulation analysis with state-space model parameters was conducted to detect the neural correlates involved in forming probabilistic knowledge about yea-sayers. If positive responses are perceived and integrated as an internal judgment criterion, rather than as the outcome of situational factors, there would be neural correlates modulated by the amount of acquired knowledge up to the trial. For a stricter analysis when extracting state-space parameters, data from three non-learners, whose learning performance did not reach a chance level (.5) for either of the response bias conditions, were excluded from the analysis. Restricting the analysis only to the feedback periods of those who demonstrated successful learning behavior, the statistical threshold was set at $p < .005$, uncorrected, with an extent threshold of 5 contiguous 3 mm isotropic voxels. Brain regions for learning, including the caudate nucleus, DLPFC, hippocampus, ventromedial prefrontal cortex (vmPFC), and the intraparietal lobe (IPL), were all found to track the amount of knowledge that had accumulated about a particular response bias (see Fig. 4 and Table 2). The results suggest that learning-related regions, such as the caudate nucleus and DLPFC, are not only involved in updating knowledge but also reflect the amount of accrued knowledge. Furthermore, in line with previous studies which showed that the hippocampus and vmPFC track conceptual knowledge (Kumaran et al., 2009; Schnyer et al., 2009; Zeithamova, Dominick, & Preston, 2012; Zeithamova, Schlichting, & Preston, 2012), these regions were found to track probabilistic knowledge about yea-sayers as well. We also found IPL involvement, a region that is thought to process numerical information (Chochon, Cohen, van de Moortele, & Dehaene, 1999), and which is essential in probabilistic reasoning.

3.2.2.3. Overlapping regions for incorrect feedback processing and the estimated knowledge reflection. Since we found the common regions that process incorrect performance feedback and reflect the amount of estimated knowledge that has accrued, we next sought to identify any overlapping regions. Regions that survived both analyses were examined by creating binary masks for the each statistical map can calculating overlapping areas. The analysis demonstrated that the caudate nucleus (MNI coordinates = [16,2,20]) and DLPFC (MNI coordinates = [−45,14,37]) were commonly involved in both processes (see the inset in Fig. 3 and notes in Table 1 and Table 2). This suggests that these regions are involved more sensitively when there is greater amount of knowledge on yea-sayer target.

3.2.2.4. The influence of perceiver's own response bias. To examine how a social desirability bias modulates the perception of negative responses, we looked for regions showing greater contrasts between negative versus positive response perception as a function of the social desirability bias. A regression analysis revealed that the MTL region is modulated by a perceiver's own social desirability bias to a greater degree when processing negative answers, than when processing positive answers, $p < .001$, 18 contiguous voxel extent threshold, corrected for multiple comparisons using Monte Carlo Simulation (Slotnick et al., 2003). Regression analysis

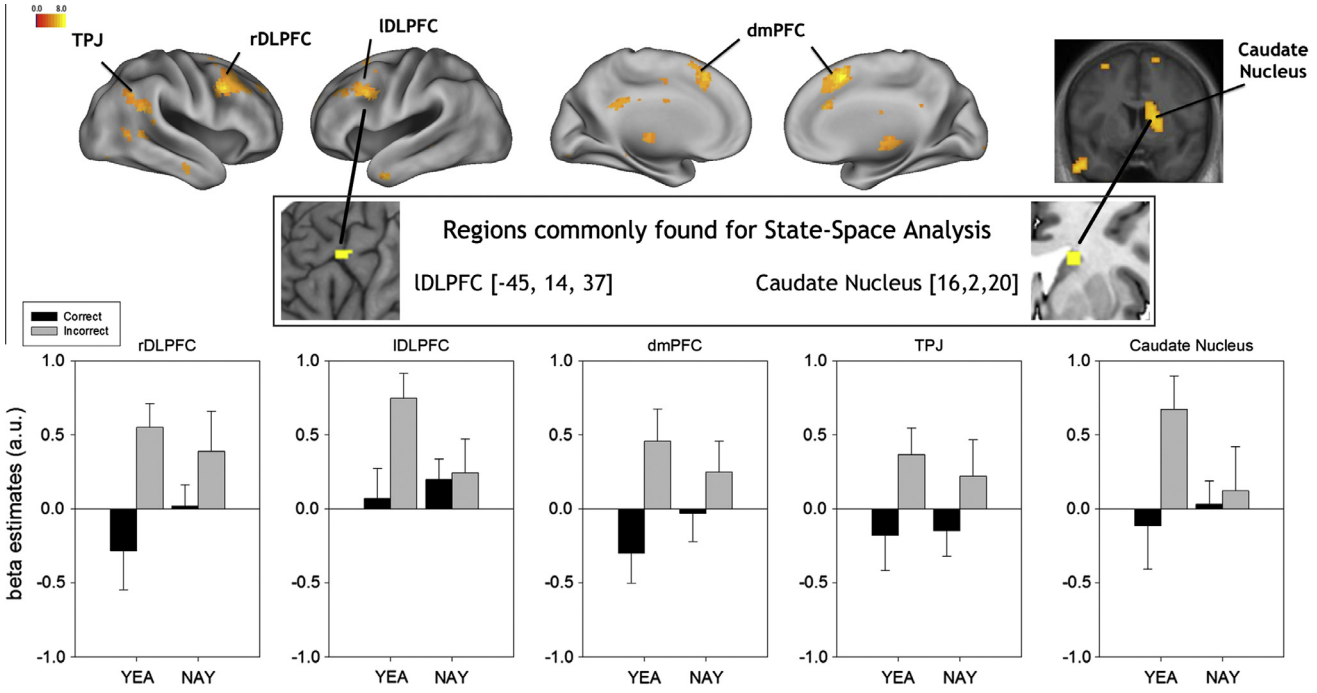

Fig. 3. Regions involved in knowledge update with incorrect performance feedback. Brain regions including caudate nucleus, bilateral DLPFC (dorsolateral prefrontal cortex), dmPFC, TPJ (temporo-parietal junction) were activated when receiving incorrect performance feedback ($p < .001$, 5 contiguous voxel extent thresholds, uncorrected for display purposes). Indicated on an inset are the results of conjunction analysis. The caudate nucleus (MNI coordinates = [16,2,20]) and DLPFC (MNI coordinates = [−45, 14, 37]) were commonly involved in both processing incorrect feedback and reflecting the amount of accumulated knowledge.

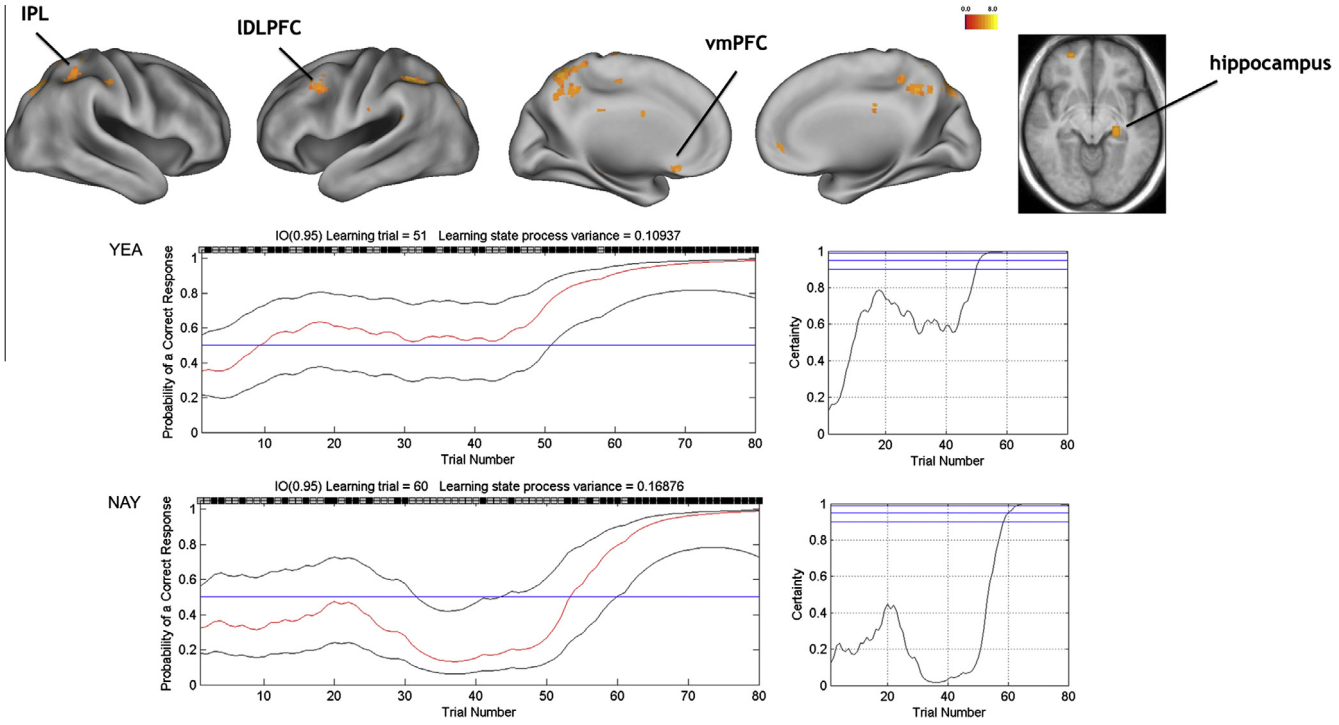

Fig. 4. Regions reflecting the amount of person knowledge on response biases. (A) Brain regions including IPL (inferior parietal lobule), DLPFC (dorsolateral prefrontal cortex), caudate nucleus, hippocampus reflected the probabilistic knowledge on response biases ($p < .005$, 5 contiguous voxel extent thresholds, uncorrected). (B) Individual sample graph for State-Space model parameters. The learning curve estimates are represented in the red lines. The upper and below lines indicate 90% confidence intervals. The obtained parameter was used as parametric regressor in fMRI analysis. Probability of correct responses reached higher-than-chance level faster and remained more stable for yea-sayer than naysayer targets. Moreover, certainty level that the ideal observer has was higher for yea-sayer learning, suggesting that perception of negative responses involves greater variability.

showed that the Social Desirability score (Crowne & Marlowe, 1960) predicted the degree to which the MTL region processes negative responses compared to positive ones (Fig. 5). Although our experimental design did not allow us to directly examine explicit memory for each item, the neuroimaging data provides preliminary evidence that, depending on a perceiver's personality,

Table 1
Regions processing negative versus positive performance feedback. Below are the regions that demonstrate greater activation for incorrect performance feedback versus correct performance feedback.

Regions	Lat.	x	y	z	z-Score
Dorsomedial prefrontal cortex (dmPFC)	L	−6	26	46	3.85
	R	6	20	49	4.72
		9	23	40	3.92
Dorsolateral prefrontal cortex (DLPFC)	L	−36	41	34	3.79
		−39[a]	17	34	3.77
		−42	32	31	3.29
	R	42	14	37	4.31
		36	26	25	3.9
		39	26	34	3.84
Temporo-parietal junction (TPJ)	R	42	14	37	4.31
		36	26	25	3.9
		39	26	34	3.84
Insula	R	45	11	1	3.28
Inferior temporal gyrus	L	−51	5	−38	3.87
Posterior cingulate cortex (PCC)	L	−3	−43	31	3.55
Caudate nucleus	R	12[a]	−1	22	3.31
Hippocampus/parahippocampal gyrus	L	−15	−31	−11	3.2
		−18	−22	−11	3.18
	R	3	−10	−14	3.37

Note: x, y, z corresponds to MNI coordinates of the maximum peak voxel.
[a] Clusters identified in conjunction analysis.

Table 2
Regions tracking the State-space parameter. Below are the regions that reflect the amount of knowledge estimated with the State-Space Model (SSM).

Regions	Lat.	x	y	z	z-Score
Inferior Parietal Lobule (IPL)	L	−36	−46	43	4.33
	R	24	−55	64	4.11
		15	−64	52	4.01
Dorsolateral prefrontal cortex (DLPFC)	L	−36	17	49	3.5
		−45[a]	14	40	3.17
Temporo-parietal junction (TPJ)	R	54	−28	43	3.4
Caudate nucleus	L	−21	−19	28	3.18
		−12	2	25	2.77
	R	18[a]	5	19	2.63
Ventromedial prefrontal cortex (vmPFC)	L	−9	23	−17	3.2
Hippocampus/parahippocampal gyrus	L	−21	−28	−11	2.73
		−3	−10	−23	2.64

Note: x, y, z corresponds to MNI coordinates of the maximum peak voxel.
[a] Clusters identified in conjunction analysis.

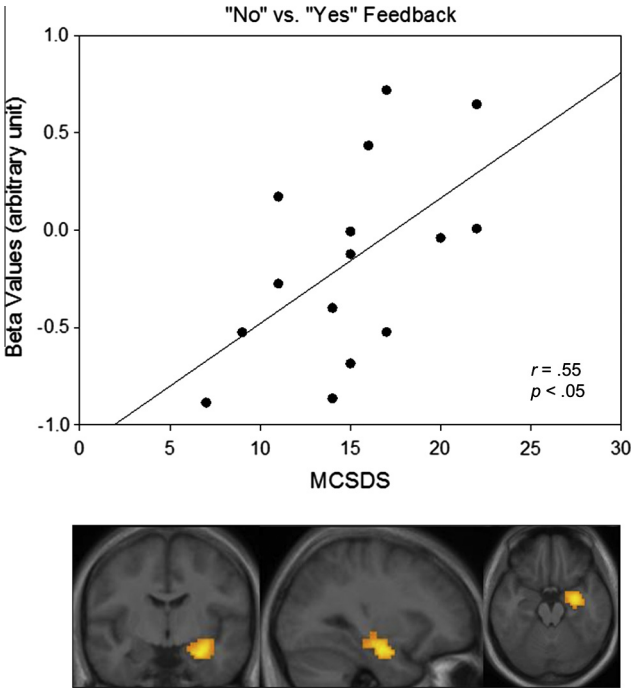

Fig. 5. MTL involvement in processing "No" feedback as a function of social desirability score. The map shows MTL ROI that is used for the analysis. Beta estimates in MTL (medial temporal lobe) ROI were greater as a function of Marlowe-Crown Social Desirability Scale (MCSDS, Crowne & Marlowe, 1960), suggesting that a perceiver's own response bias has an influence on perception of others' biases.

the "no" remarks of other people can have a more specific representation than their "yes" remarks. There are two possibilities for explaining this: either a perceiver's own response bias led them to treat the negative response as being more unique, or a perceiver who tries to behave in a socially desirable way will regard a "no" response as being socially undesirable, resulting in more sensitive representation. It is necessary to further disambiguate explicit memory for each item and the personality's role in forming global and specific representations.

4. General discussion

In the current study, we examined the neural correlates of response bias learning, a procedure that helps us to accurately predict the future behavior of other people. We hypothesized that the positivity or negativity of each response would asymmetrically influence the formation of knowledge from interactions with a peer. In order to test this hypothesis, we conducted behavioral and neuroimaging experiments with a feedback-based learning paradigm in which participants were required to, although not explicitly asked to, form generalized knowledge about response biases in order to make better predictions. The participants made guesses and received feedback on whether the targets would agree or disagree to perform suggested activities. We found that participants learned about their targets' biases and made gradually improved predictions about their targets' answers in the direction of their bias. With a limited number of interactions (Experiment 1), learning performance was better with yea-sayer targets than with naysayer targets, although this difference disappeared with an increase in the number of interactions (Experiment 2). This different learning pattern suggests that there is a fundamental asymmetry in the way that we learn about positive/negative response biases. The fMRI results also support the idea of there being a dissimilar learning process. Brain regions, including the caudate nucleus and DLPFC, reflected the degree to which positive answers aggregated into evidence for yea-sayers, compared to the integration of negative answers into evidence for naysayers. The results suggest that a perceiver regards a target's positive answer as evidence, as this indicates what the target's internal judgment criterion is. That is, individual experiences of seeing the target answer "yes" are accumulated to form a generalized form of person knowledge. This person knowledge then helps future prediction about the target's answers to novel questions. Additionally, our data also suggests that negative responses will have a more specific representation, especially when modulated by perceiver characteristics. Those who are more biased to behave in socially desirable ways showed greater levels of neural processing when dealing with negative answers.

Some might argue that the integrated learning of positive response biases is opposed to intuition, since, in general, negatively valenced events have a stronger impact than positively valenced events (for a review, see Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001). Anderson (1982) proposed an information-integration model in which negative information

outweighs positive information when integrated into knowledge, and Gottman's positive-to-negative ratio (Gottman & Levenson, 1992, 1999) provides empirical evidence that positive interaction is needed five times more than negative interaction for a successful relationship. In other words, one bad interaction may ruin five good interactions. This significance of negatively valenced events makes sense in terms of survival value. It is crucial in preventing bad consequences to remember a person to avoid. However, it is important to note that saying "no" is not necessarily a negative characteristic of the responder. Moreover, observing a negative answer is not necessarily a negatively valenced event to a perceiver. The primary information conveyed by "no" is less a decision against the perceiver, than a preference of the responder. If participants choose to suggest that a target perform certain activities instead of merely observing answers, they may more readily recognize naysayers because "no" is more emotionally charged and negatively valenced in this case. This is a possible independent research topic on its own. Nevertheless, in the present experiment, the positive/negative answer of the target was information whose *content* was positive or negative without direct emotional relevance to the perceiver. Thus, positive answers, with a lower variation in the kind of positivity (Unkelbach et al., 2008), can be integrated into homogeneous person knowledge, as found in our results. Moreover, our results are well suited to Gottman's positive to negative ratio (Gottman & Levenson, 1992), so that not doing what the person hates is more important than doing what she likes. To be recognized and avoided appropriately, activities that elicit negative response should have representations that are more specific. On the contrary, it is more efficient to recognize the positive person, not the activity per se, when there is nothing to be avoided.

Concerning the neural correlates involved in learning about yea-sayers, we discovered that, when analyzed in terms of how knowledge is updated through incorrect performance feedback and the modulation of integrated knowledge, among the most important brain regions were the caudate nucleus and DLPFC. Although these regions have been found to process feedback and help future behavioral adjustment (Zanolie et al., 2008), to our knowledge, this was the first study to determine the degree to which they respond to incorrect performance feedback and to also show that this response varied according to the amount of knowledge which had accrued. It is worth noting that these regions were not equally activated for every incorrect performance feedback, but were modulated by previously accrued information. This pattern implies that these regions possess the ability to utilize prior information when updating knowledge with newly acquired information.

The involvement of the caudate nucleus, uncovered in the current study, is also in accordance with previous research, which found that this part of the dorsal striatum was involved in instrumental learning (Cooper, Dunne, Furey, & O'Doherty, 2012). When there is a response-outcome contingency, which requires that we update action values for future positive outcomes, the caudate nucleus functions by updating action values and making decision according to the outcome. For example, Tricomi, Delgado, McCandliss, McClelland, and Fiez (2006) found that performance feedback elicited bilateral caudate nucleus activation in a phonological learning task, and Cooper, Dunne, Furey, and O'Doherty (2012) discovered the caudate nucleus's role in instrumental observational learning. Moreover, in a comparative single cell recording study, Samejima, Ueda, Doya, and Kimura (2005) showed that neurons in the caudate nucleus of monkeys encoded action values. Although there is inconsistency between our result and most of the previous findings which indicated that a positive feedback drives caudate nucleus activation (Delgado, Miller, Inati, & Phelps, 2005; Tricomi et al., 2006), Tricomi and Fiez (2012) recently showed that the caudate nucleus is activated when receiving incorrect performance feedback. In their study, the incorrect feedback was manipulated to provide greater informative value and, thus, could be interpreted as an intrinsic reward that in the long run helps goal attainment (Tricomi & Fiez, 2012). Moreover, Han, Huettel, Raposo, Adcock, and Dobbins (2010) showed that striatal activation is related to goal attainment in episodic memory, as manipulating a goal in episodic retrieval (i.e. a hit and correct rejection) heightened the importance of information that pertained to goal attainment. Likewise, in the present study, incorrect performance feedback indicated an informational gap between the predicted and actual response. Subjective evidence for prediction was gradually increased as targets were repeatedly presented, and when the prediction was incorrect, there was a larger gap to be updated and the feedback conveyed greater information. Therefore, we reasoned that it is the informational value of incorrect feedback that drives caudate nucleus activation. Together with this, the finding that caudate nucleus activation was modulated by the amount of knowledge (i.e., previously acquired information) gives further support to informational processing in the caudate nucleus.

In addition to striatal activation, we found that feedback processing hired the DLPFC, and that the region was modulated by probabilistic knowledge about the response bias. This result is in line with findings that the DLPFC plays a role in adjusting behaviors through acquired information. In a study by Zanolie et al. (2008), the DLPFC was activated by informational error feedback that helped behavioral adjustment, but unexpectedly incorrect feedback, without any informational value, did not activate the DLPFC. Moreover, adults hired the DLPFC in feedback processing while children, who have not fully developed the ability to utilize performance feedback to adjust future behavior, did not show feedback sensitivity in the DLPFC (van Duijvenvoorde, Zanolie, Rombouts, Raijmakers, & Crone, 2008). These findings contribute to the conclusion that the DLPFC helps to exert goal-directed behavior through utilizing error feedback. Although the current study's gradual learning paradigm was different from the feedback-based rule learning paradigm in those studies, the DLPFC's involvement was still found in the current study, expanding its role to include feedback utilization in gradual learning. Additionally, the amount of knowledge that accrued, which is accumulated up to the point of prediction, was found to modulate how incorrect feedback is utilized to make an accurate prediction, further supporting an informational account. That is, the DLPFC showed a higher level of involvement when there was a greater amount of information to update.

In summary, we found that learning about yea-sayers hires the caudate nucleus and DLPFC for feedback processing. In these regions, "yes" feedback elicited a probabilistic integration into the homogeneous concept of a yea-sayer. Further, the greater the amount of previously acquired knowledge about the response bias, the more updates with feedback occurred in these regions. This result implies that these regions utilize newly acquired information to form generalized knowledge while being modulated by previously acquired information. These findings suggest that the caudate nucleus and DLPFC are information-processing regions that store and update knowledge about response biases, thus enabling us to make better predictions about the behavior of other people.

Acknowledgments

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0003882 to Sanghoon Han).

References

Anderson, N. H. (1982). *Methods of information integration theory*. New York: Academic Press.

Baron, S. G., Gobbini, M. I., Engell, A. D., & Todorov, A. (2011). Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions. *Social Cognitive and Affective Neuroscience, 6*(5), 572–581.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370.

Berg, I. A. (1955). Response bias and personality: The deviation hypothesis. *Journal of Psychology: Interdisciplinary and Applied, 40*(1), 61–72.

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), 9337–9344.

Bless, H., Mackie, D. M., & Schwarz, N. (1992). Mood effects on attitude judgments: Independent effects of mood before and after message elaboration. *Journal of Personality and Social Psychology, 63*(4), 585–595.

Bless, H., Schwarz, N., Clore, G. L., Golisano, V., Rabe, C., & Wolk, M. (1996). Mood and the use of scripts: Does a happy mood really lead to mindlessness? *Journal of Personality and Social Psychology, 71*(4), 665–679.

Chochon, F., Cohen, L., van de Moortele, P. F., & Dehaene, S. (1999). Differential contributions of the left and right inferior parietal lobules to number processing. *Journal of Cognitive Neuroscience, 11*(6), 617–630.

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), 583–588.

Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2012). Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. *Journal of Cognitive Neuroscience, 24*(1), 106–118.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology, 60*, 151–174.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349–354.

Delgado, M. R., Miller, M. M., Inati, S., & Phelps, E. A. (2005). An fMRI study of reward-related probability learning. *Neuroimage, 24*(3), 862–873.

Dunbar, R. I. M. (2003). The social brain: Mind, language, and society in evolutionary perspective. *Annual Review of Anthropology, 32*, 163–181.

Forgas, J. P. (1998). On being happy and mistaken: Mood effects on the fundamental attribution error. *Journal of Personality and Social Psychology, 75*(2), 318–331.

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology. The broaden-and-build theory of positive emotions. *The American Psychologist, 56*(3), 218–226.

Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition & Emotion, 19*(3), 313–332.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*.

Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the "weather prediction" task?: Individual variability in strategies for probabilistic category learning. *Learning & Memory, 9*(6), 408–418.

Gottman, J. M., & Levenson, R. W. (1992). Marital processes predictive of later dissolution: Behavior, physiology, and health. *Journal of Personality and Social Psychology, 63*(2), 221–233.

Gottman, J. M., & Levenson, R. W. (1999). What predicts change in marital interaction over time? A study of alternative models. *Family Process, 38*(2), 143–158.

Green, D. M. (1966). *Signal detection theory and psychophysics*. Huntington, NY: R.E. Krieger Pub. Co.

Han, S., Huettel, S. A., Raposo, A., Adcock, R. A., & Dobbins, I. G. (2010). Functional significance of striatal responses during episodic decisions: Recovery or goal attainment? *Journal of Neuroscience, 30*(13), 4767–4775.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109*(4), 679–709.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291.

Kakade, S., & Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psychological Review, 109*(3), 533–544.

Kensinger, E. A., & Schacter, D. L. (2006). When the Red Sox shocked the Yankees: Comparing negative and positive memories. *Psychonomic Bulletin & Review, 13*(5), 757–763.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science, 273*(5280), 1399–1402.

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the emergence of conceptual knowledge during human decision making. *Neuron, 63*(6), 889–901.

Lee, T. H., Lee, K. Y., Lee, K., Choi, J. S., & Kim, H. T. (2006). *The Korea University facial expression collection: KUEFC*. Laboratory of Behavioral Neuroscience, Department of Psychology, Korea University). Seoul, South Korea.

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), 937–950.

Maddox, W. T., Ashby, F. G., Ing, A. D., & Pickering, A. D. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition, 32*(4), 582–591.

Mattick, R. P., & Clarke, J. C. (1998). Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behaviour Research and Therapy, 36*(4), 455–470.

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *Journal of Neuroscience, 33*(50), 19406–19415.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience, 8*(6), 623–631.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology, 53*(3), 139–154.

Ochsner, K. N. (2000). Are affective events richly recollected or simply familiar? The experience and process of recognizing feelings past. *Journal of Experimental Psychology: General, 129*(2), 242–261.

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology, 51*(51), 768–774.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4).

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science, 310*(5752), 1337–1340.

Saxe, R., & Baron-Cohen, S. (2006). The neuroscience of theory of mind. *Social Neuroscience, 1*(3–4), i–ix.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage, 19*(4), 1835–1842.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia, 43*(10), 1391–1399.

Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J., & Pelphrey, K. A. (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child Development, 80*(4), 1197–1209.

Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., & Phelps, E. A. (2009). A neural mechanism of first impressions. *Nature Neuroscience, 12*(4), 508–514.

Schnyer, D. M., Maddox, W. T., Ell, S., Davis, S., Pacheco, J., & Verfaellie, M. (2009). Prefrontal contributions to rule-based and information-integration category learning. *Neuropsychologia, 47*(13), 2995–3006.

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

Smith, A. C., Frank, L. M., Wirth, S., Yanike, M., Hu, D., Kubota, Y., et al. (2004). Dynamic analysis of learning in behavioral experiments. *Journal of Neuroscience, 24*(2).

Solomon, M., Smith, A. C., Frank, M. J., Ly, S., & Carter, C. S. (2011). Probabilistic reinforcement learning in adults with autism spectrum disorders. *Autism Research, 4*(2), 109–120.

Tricomi, E., Delgado, M. R., McCandliss, B. D., McClelland, J. L., & Fiez, J. A. (2006). Performance feedback drives caudate activation in a phonological learning task. *Journal of Cognitive Neuroscience, 18*(6), 1029–1043.

Tricomi, E., & Fiez, J. A. (2012). Information content and reward processing in the human striatum during performance of a declarative memory task. *Cognitive Affective & Behavioral Neuroscience, 12*(2), 361–372.

Unkelbach, C., Fiedler, K., Bayer, M., Stegmuller, M., & Danner, D. (2008). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology, 95*(1), 36–49.

van Duijvenvoorde, A. C., Zanolie, K., Rombouts, S. A., Raijmakers, M. E., & Crone, E. A. (2008). Evaluating the negative or valuing the positive? Neural mechanisms supporting feedback-based learning across development. *Journal of Neuroscience, 28*(38), 9495–9503.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*(3), 441–517.

Zaki, J., & Ochsner, K. (2011). Reintegrating the study of accuracy into social cognition research. *Psychological Inquiry, 22*(3), 159–182.

Zanolie, K., Van Leijenhorst, L., Rombouts, S. A., & Crone, E. A. (2008). Separable neural mechanisms contribute to feedback processing in a rule-learning task. *Neuropsychologia, 46*(1), 117–126.

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron, 75*(1), 168–179.

Zeithamova, D., Schlichting, M. L., & Preston, A. R. (2012). The hippocampus and inferential reasoning: Building memories to navigate future decisions. *Frontiers in Human Neuroscience, 6*, 70.