Published in final edited form as:
Soc Neurosci. 2010 ; 5(1): 76–91. doi:10.1080/17470910903135825.

Neural Regions that Underlie Reinforcement Learning Also Engage in Social Expectancy Violations

Lasana T. Harris, Ph.D.[*] and **Susan T. Fiske, Ph.D.**[†]

[*]Psychology Department, New York University, 6 Washington Place, Rm # 865, New York, New York, 10003

[†]Psychology Department, Princeton University, Green Hall, Rm # 2-N-14, Princeton, New Jersey, 08540

Abstract

Prediction error, the difference between an expected and actual outcome, serves as a learning signal that interacts with reward and punishment value to direct future behavior during reinforcement learning. We hypothesized that similar learning and valuation signals may underlie social expectancy violations. Here, we explore the neural correlates of social expectancy violation signals along the universal person-perception dimensions of trait warmth and competence. In this context, social learning may result from expectancy violations that occur when a target is inconsistent with an a priori schema. Expectancy violation may activate neural regions normally implicated in prediction error and valuation during appetitive and aversive conditioning. Using fMRI, we first gave perceivers warmth or competence behavioral information. Participants then saw pictures of people responsible for the behavior; they represented social groups either inconsistent (rated low on either warmth or competence) or consistent (rated high on either warmth or competence) with the behavior information. Warmth and competence expectancy violations activate striatal regions and frontal cortex respectively, areas that represent evaluative and prediction-error signals. These findings suggest that regions underlying reinforcement learning may be engaged in warmth and competence social expectancy violation, and illustrate the neural overlap between neuroeconomics and social neuroscience.

Keywords

stereotype content model; trait attribution; social expectancy violation; trait warmth and competence

Reinforcement learning describes a process whereby past and current information is used to guide future behavior. In classical conditioning, people and animals learn rewards and punishment contingent on arbitrary cues. Learning signals captured by the Rescorla-Wagner models result from a discrepancy between predicted and actual outcomes. This prediction

Corresponding Author: Lasana T. Harris, Ph.D., 6 Washington Place, C&P, RM# 865, NY, N.Y. 10003. Telephone—212-998-3860 (office), 212-995-4768 (fax). lasana@nyu.edu.

[2]Assessments of warmth and competence satisfy the components of the Rescorla-Wagner model:

$$V_{new} = V_{old} + \eta(R - V_{old})$$

Here, R is a scalar quantity that is an assessment of goodness or badness, consistent with warmth assessments that specify the valence, and competence assessments that specify magnitude or value as a function of warmth.

error is then used to update representations of the relationship between cues and outcomes that subsequently guide future behavior (see Niv & Schoenbaum, 2008 for a review).

Perhaps social learning can be described using this learning model. Stereotypes serve as cues to guide future behavior toward other people, and inconsistencies can cause a revision of the stereotype (see Fiske, 1999, for a review). Imagine a magazine blurb about a person who reads to sick children three times a week for hours, a sacrifice few people make. This person also reads to other sick people in the hospital, and has been reading to sick children for years. You then turn to the page to see that the person is a apparently an injection drug addict, pictured with eyes closed sitting on the floor, surrounded by heroin needles. Expectancy violation? Of course.

Behavioral information that leads to a dispositional attribution (e.g., warm, generous) creates a social expectancy. The behavior in this example (reading to children) activates a category of possible social targets consistent with stereotypically high-warm people, the elderly perhaps. However, a social expectancy can be violated if the behavioral information turns out to be inconsistent with the subsequently revealed social category. For instance, the drug addict in this example demonstrates a clear expectancy violation.

The current study examines the neural correlates of such expectancy violations. People attend to expectancy violations, and such people who contradict prior knowledge are salient (Jones & McGills, 1976). People abstract the most typical or central features of category members, and compare subsequent examples to this prototype, according to one plausible account (Hayes-Roth & Hayes-Roth, 1977; Posner & Keele, 1968, 1970; Reed, 1972); social categories develop in the same way (Fiske & Dyer, 1985). Therefore, one can assume that social expectancies derived from behavioral information link to prototypic representations of categories consistent with the stereotype implied by the behavior. For example, people believe that someone described as politically conscious and liberal who works as a bank teller is a feminist bank-teller rather than simply a bank-teller, even though bank-teller is more probable (Tversky & Kahneman, 1983). This suggests that people use stereotypes when generating schemas of people from dispositional information.

Like value and error signals, stereotypes and their violations have affective components and consequences for future behavior. Research on schema-triggered affect shows that people expecting to interact with a schizophrenic show more non-verbal signs of anxiety than if the expectancy was about a heart-patient (Neuberg & Fiske, 1987). Research shows that perceivers modify their behavior for negative expectancies such as hostile or cold targets (Bond, 1972; Ickes, Patterson, Rajecki, & Tanford, 1982; Shelton & Richeson, 2005; Swann & Snyder, 1980). Also, the expectancy remains in the perceivers' minds even if they cannot confirm the expectancy (Ickes et al., 1982).

Stereotype-relevant behavior is also primed by social group categories (Dijksterhuis & Bargh, 2001; Dijksterhuis & van Knippenberg, 1998; Wheeler & Petty, 2001). Perceivers often automatically assimilate their behavior to stereotypes (Dijksterhuis, Spears, & Lepinasse, 2001). Priming of stereotypes also results in stereotype-consistent behavior (Dijksterhuis, Aarts, Bargh, & van Knippenberg 2000; Dijksterhuis & van Knippenberg, 1999), and these automatic associations predict subtle forms of discriminatory behavior (Dovidio, Kawakami & Gaertner, 2002; Dovidio, Kawakami, Johnson, Johnson, & Howard, 1997; Word Zanna, & Cooper, 1974).

Trait Warmth and Competence

What types of expectancies do perceivers typically carry about other people? Trait dimensions of warmth and competence—respectively, perceived intent for good or ill and

the ability to enact those intentions—are the fundamental dimensions of person perception (see Fiske, Cuddy, & Glick, 2007, for review). Social groups are therefore perceived primarily along these two trait dimensions, and they fall into one of four quadrants created by low and high values on each dimension (Fiske, Cuddy, Glick, & Xu, 2002). That is, the Stereotype Content Model (SCM) predicts that groups perceived as high on both trait dimensions (e.g., middle-class) elicit the ingroup emotion pride, groups perceived as high in warmth but low in competence (e.g., the elderly) elicit the paternalistic emotion pity, groups perceived as low in warmth but high in competence (e.g., rich people) elicit the ambivalent emotion envy, and groups perceived as low on both dimensions (e.g., drug addicts) elicit the basic negative emotion disgust for perceived moral violations (Fiske, et al., 2002). Affect elicited by individuals also depends on perceptions of trait warmth and competence (Russell & Fiske, 2008).

Group stereotypes assist in trait attribution by categorizing people into different social groups with prescribed attributions of warmth and competence, which allow for rapid attributions during person perception (Fiske, Lin, & Neuberg, 1999). People are constantly adjusting their perception of other individuals along warmth and competence dimensions.

Expectancy violations can serve as a learning signal to guide future behavior, much like prediction error serves as a learning signal in appetitive and aversive conditioning. This suggests that social expectancy violation may depend on the same structures underlying prediction error. The inconsistency when expecting a warm target but perceiving a cold target we consider a *warmth expectancy violation signal (W_{EV})*, a prediction error signal. The same holds for competence: expecting an able target but perceiving an inept target, we consider this inconsistency a *competence expectancy violation signal (C_{EV})*.

Social targets have intentions and traits that predict their likelihood and ability to help or harm the perceiver. First, perceivers must infer the target's likely action. Warmth is person perception's initial dimension, an assessment of good or ill intent. Intent suggests whether the target's behavior toward the perceiver will bear potentially helpful or harmful outcomes for the perceiver. Therefore, perceived intent focuses on whether the target is a threat or not, an assessment of their possible behavior. Threat detection is an essential evolved ability, and even animals without a theory of mind (ToM)—the ability to infer mental states—assess threat to guide approach-avoid behavior (de Waal, 2005). This suggests that warmth expectancy violations may signal that an expected harmless target may be harmful (or vice versa). As such, warmth judgments and their violations concern the likely valence (positive-negative or help-harm) of intended actions. An expectancy violation signal underlying this most important social dimension may depend on structures involved in calculating prediction error (for example receiving a punishment when expecting a reward). Structures in the striatum (caudate, putamen, and nucleus accumbens), frontal cortex and the amygdala, has been implicated in calculating prediction error (Niv & Schoenbaum, 2008).

Along with inferring intended and therefore likely valence (goodness or badness) of action, perceivers also infer ability and therefore likelihood to take action. Competence assessments are judgments of ability, a "how-much" (more or less) judgment that describes the degree of "good-bad" appraisal. Inferences of ability may rely on neural networks implicated in tracking reward and punishment value, that is, its degree of goodness or badness. Neural areas that track value work in concert with prediction error signals during learning, and include frontal regions, specifically orbital frontal and medial frontal regions (see Montague, King-Casas, Cohen, 2006 for a review). Changes in competence valuation may moderate neural activity in frontal regions associated with reward and punishment value.

There are two kinds of prediction errors, positive and negative (see Schultz, Dayan, & Montague, 1997, for a review). Positive prediction errors occur when the animal received more reward (or punishment) than anticipated, and negative prediction errors occur when the animal receives less reward (or punishment) than anticipated. This distinct is mirrored by neural activity in the striatum; positive prediction error leads to increases in striatal activity, while negative prediction error leads to decreases in striatal activity. Both kinds of prediction errors may be present during social learning, but we make no a priori distinction about the specific kind of prediction error involved in trait warmth and competence violations and confirmations.

To determine the neural correlates of social learning, we had participants respond to information that led them to make either highly warm (moral) or highly competent (intelligent) attributions for behavior. When attributed to a person, the behavior allows a prediction of that person's future behavior because the predisposition to respond resides in the person. When attributed to the situation, the behavior is held constant but does not allow for the prediction of the person's future behavior. Dispositional attributions involve a neural network centered on medial prefrontal cortex (MPFC) and superior temporal sulcus (STS; Harris, Todorov, & Fiske, 2005). Areas of MPFC are tuned especially to valence differences in people, even more than objects (Harris, McClure, Van den Bos, Cohen, & Fiske, 2007; Van den Bos, McClure, Harris, Fiske, & Cohen, 2007). More importantly, dispositional attributions focus specifically on the actor of the behavior, not the action, or context. As a control in our current study, some behaviors were attributable to context, that is, a unique situation. This allowed us to hold constant the person and the situation described in the behavior, but cognitively focus participants either on the person or the situation in different attributional combinations. After presenting the dispositionally or situationally attributed behavior, we then showed participants a picture of the person who supposedly performed the behavior. This allows participants first to form expectancies about the person after a dispositional attribution but before receiving visual information that is either consistent or inconsistent with those expectancies. The situational attributions serve as a control.

Therefore, we predict that when viewing the pictured social targets after dispositional attributions, participants should show activation in (a) striatal regions associated with prediction error after warmth expectancy violations, and (b) frontal structures associated with reward or punishment value after competence expectancy violations.

Methods

Participants

Fifteen Princeton University undergraduates participated for course credit. Three participants' data were excluded because of either excessive movement or data recording errors, resulting in 12 participants' data averaged in the analyses. Participants reported no abnormal neurological conditions, were right-handed, and had suffered no incidence of head trauma or brain lesions. All participants had normal or corrected vision, were native English speakers, and provided informed consent. The mean age was 20.4 years, with 4 women, and 3 ethnic minorities.

Stimuli

The behavioral information established the judged warmth (valence) or competence (quantity) of likely behavior by each target. We attempted to capture a spontaneous response to social targets immediately after making an attribution for their behavior. Therefore, each stimulus comprised (a) behavior sentences relevant to warmth or competence, (b) additional information leading to the crucial dispositional or the control, situational attribution, and (c)

a photograph of a social target who presumably performed the behavior. We examine BOLD responses only to this picture, presented separately after the sentences.

As noted, participants first saw a target sentence describing a person's behavior. The 60 target sentences describing behavior came from a group of sentences rated on intelligence and moral goodness (see Skowronski & Carlston, 1987[1]). The sentences used in the experiment were the 30 rated highest on intelligence, a competence trait, and the 30 rated highest on positive moral behavior, a warmth trait. Sentences that described implausible behavior (e.g., won the Nobel Peace Prize) were replaced with the next ranked sentence.

Target sentences appeared with additional information about the behavior that encouraged a dispositional or situational attribution (Harris et al., 2005; Kelley, 1967; McArthur, 1972). Half the behaviors led to a dispositional attribution: low consensus information (*hardly any other* [target does this]), low distinctiveness information ([this target does this] *also… to every other entity*), and high consistency information (in the past… [this target] *would almost always* [do this]). The information combinations encouraging a situational attribution describe the other half of behaviors: high consensus information (*almost all other* [targets do this]), high distinctiveness information ([this target] *does not…to any other entity*), and high consistency information (in the past… [this target] *would almost always* [do this]).

Participants next saw one of 60 pictures of different people. Pictures were taken from a larger set already rated on warmth and competence (see Harris, 2007). Therefore, each pictured social target illustrates the warmth and competence interaction described in the SCM. There were 30 pictures per quadrant, high and low on warmth and competence. Each picture depicted a person who was from a group pretested as eliciting warmth expectancies as high (American hero [firefighter, police officer, astronaut, athlete], college student, elderly person, disabled person), or low (business person, rich person, homeless person, drug addict), and competence expectancies as high (American hero, college student, business person, rich person) or low (elderly person, disabled person, homeless person, drug addict). This results in 7–8 pictures in each cell of the 2 (warmth, competence information) X 2 (target is high [consistent] or low [inconsistent] on the dimension) X 2 (dispositional, situational focus of attribution) design.

Scanning Parameters

All fMRI scanning was conducted at Princeton's Center for the Study of Brain, Mind, and Behavior, which uses a 3.0 Tesla Siemens Allegra head-dedicated MR scanner. A Dell computer presented the stimuli projected to a screen mounted at the rear of the scanner bore. Stimuli reflected through a series of mirrors, which participants viewed while supine. Responses were recorded using bimanual fiber-optic response pads (Current Designs Inc. url: http://www.curdes.com/response). Prior to the functional echo planar image (EPI) acquisitions, subjects received a short series of structural MRI scans to allow for subsequent functional localization. These scans took approximately 12 minutes and included: 1) a brief scout for landmarking; 2) a high-resolution whole-brain MPRAGE sequence for later localization and intersubject registration. Functional imaging then proceeded using an EPI sequence that allowed for whole-brain coverage in a relatively short period of time (32 3mm axial slices; 1mm gap, TR: 2 sec; TE: 30 msec). In-plane resolutions were 3mm × 3mm (196mm FOV, 64×64 matrix).

[1]These are correlates of warmth and competence. Though sociability is a separate dimension of warmth (Leach, Ellemers, & Barreto, 2007), morality underlies the same dimension (Fiske et al., 2007). Therefore, we used behavioral sentences rated high on morality.

Procedure

The method is adapted from the Harris et al. (2005) paradigm (see Figure 1). Participants read a series of sentences that provided information about the behavior of different people. Each of the 60 warmth and competence sentences was paired with information suggesting dispositional and situational attributions. This resulted in 120 sentence-attribution combinations that were split evenly between two versions of the experiment, resulting in 60 stimulus epics per participant. No sentence repeated in any version, and led to only one kind of attribution in that version; half the combinations to a dispositional attribution, while the other half to a situational attribution. Similarly, half the sentences described warm behavior, and half described competent behavior in each version.

Each picture also appeared once per version, and was paired quasi-randomly with a sentence and the resulting attribution combination. Therefore, each of the 30 pictures per trait dimension was paired with a warmth or competence situation or dispositional attribution. Each subject completed one version of the experiment, with six completing the first version, and six completing the second version.

Participants practiced the task before scanning. The experimenter never explained what information combination led to which attributions for behavior, but they have pretested as intuitively obvious. No participants were allowed in the scanner until they made the standard attributions for behavior on one complete practice run. However, we did not exclude any participants for failure to make correct attributions since all of our participants did so during the practice run.

In the scanner, participants first saw a fixation cross for 4 seconds. The information screen then appeared and remained for 20 seconds following the fixation cross. This screen contained the target sentence and the consensus, distinctiveness, and consistency information, which presumably led participants to make either a dispositional or a situational attribution about a person. Pronouns “he” or “she” identified the person, consistent with either a male or female picture. A picture of the person whose behavior had just been described appeared after the information screen for 2 seconds[3]. A response screen appeared after the picture, during which participants attributed responsibility for the behavior to the person, the situation, or some combination of circumstances. This screen remained for 4 seconds, followed by a fixation cross. Each run contained 15 trials, and each participant completed 4 runs.

The order of attributions was random across participants. The stimuli appeared via the computer display program E-prime. After the scanning session, participants were probed for suspicion; none were suspicious. They were then thoroughly debriefed, given course credit, and thanked.

Preprocessing

Both image preprocessing and statistical analysis used Brain Voyager QX (www.brainvoyager.de). Before statistical analysis, image preprocessing consisted of: 1) slice acquisition order correction; 2) 3D rigid-body motion correction; 3) voxelwise linear detrending across time; 4) temporal bandpass filtering to remove low frequency (scanner

[3]We reverse the conventional order, presenting the behavior first then the social target, because of the nature of our independent variable, ANOVA-styled sentences leading to dispositional attributions. Previous work suggests that people make dispositional attributions using ANOVA-styled sentences 9–14 seconds after the sentences are presented (see Harris, Todorov, & Fiske, 2005). Therefore, we employ a block design in order to capture the attributions across the entire 20-second period, regardless of when they occurred. Given the variance in making the attribution, it would be very difficult to estimate precisely when the violation occurred if the order were reversed. By the time the social target is present, the participant has made an attribution. There is a cleaner, more precise response to a picture as an isolated event in a stream of sentences.

and physiology related) noise. We later add Fourier predictors (2 cycles) to correct for high frequency noise associated with scanner drift. Distortions of EPI images were corrected with a simple affine transformation. Functional images were registered to the structural images and interpolated to cubic voxels. After coregistering participants' structural images to a standard image using a 12-parameter spatial transformation, their functional data were similarly transformed, along with a standard moderate degree of spatial smoothing (Gaussian 8 mm FWHM).

Data Analysis

Data analysis used the general linear model available on the Brain Voyager QX software package. We first computed a GLM focusing just on the two seconds when the pictures were displayed because this was the period of expectancy violation. We computed series of regressors to examine BOLD brain activity, as well as contrast maps. For all neuroimaging analyses, we report the average signal change value of all the clusters of voxels that overlay the neural region of interest, and provide the coordinates at the center of this cluster, not maximum values. Random effects analyses were performed on all imaging data. All data are presented with their coordinates based on a standard system (Talairarch & Tournoux, 1988).

Additionally, we conducted region of interest (ROI) analyses. That is, we extracted the average signal change for each participant within a cluster of voxels active in the exploratory analyses. This is a measure neural activity to each information combination and picture interaction in regions identified in the exploratory analysis. We computed 2 (*target warmth*) X 2 (*target competence*) X 2 (*trait*) X 2 (*focus of attribution*) repeated measures ANOVAs on each cluster.

Results and Discussion

Recall our main predictions: When viewing the pictured social targets after dispositional attributions, participants should show activation (a) striatal regions associated with prediction error after warmth expectancy violations, and (b) frontal structures associated with reward or punishment value after competence expectancy violations. All reported neural areas contain at least 10 contiguous voxels, and are significant after correction for multiple comparisons at $p < .001$[4]. All follow-up region-of-interest (ROI) analyses are significant at $p < .05$.

Warmth Expectancy Violation

We first performed a whole-brain analysis, contrasting high- versus low-warmth targets engaged in warm behavior after dispositional attributions for the behavior.[5] Areas *less*[6] active in this contrast are to dispositional attributions to social targets after *inconsistent* warmth behavior (that is, initially warm behavior revealed to come from a social target not expected to be warm). This suggests that these neural regions are involved in calculating expectancy violation along the warmth trait dimension, expecting a warm target but perceiving a cold target. We consider this a *warmth expectancy violation signal (W_{EV})*.

[4]We used significance criteria of 10 contiguous voxels and *p*-value < .001after correction for multiple comparisons, instead of just a set *p*-value for determining which regions were active in our paradigm. We do not report specific *p*-values, instead, we report effect sizes for all significant results, indicated by partial η^2, and observed power (for ROI analyses), indicated by Ω, not specific *p*-values, because partial η^2 is a measure of effect magnitude independent of *N*, unlike *p*.

[5]We employ this strategy, using just some of our data to define the ROIs, to allow unbiased comparisons within our ROIs.

[6]The nature of our contrasts makes it difficult to determine whether we are reporting positive or negative prediction errors. For instance, given our paradigm, one might predict that negative prediction errors should result from warmth expectancy violations. However, the low warmth social groups include both high and low competence groups. Therefore, it is possible that the high competence groups could lead to a positive warmth prediction error. It is difficult to make either case our contrasts do not allow for independent exploration of positive and negative prediction errors.

Consistent with our hypothesis, we find activity in the lentiform nucleus of the right putamen, t (11) = −3.39, at x = 24, y = 16, z = 3, 23 voxels, partial η^2 = .51 (see Figure 2a). We also find activity in middle frontal gyrus (see Table 1 for Talairarch coordinates and statistics)[7].

We performed a follow-up region of interest (ROI) analysis that includes all of our data, looking at the percent signal change for each possible kind of attribution that created expectancies before perceiving the social targets. Because our ROIs are based on just some of the data in the a priori contrast, these ROI analyses are unbiased. This strategy allows us to compare, in the same subject in the same paradigm, the neural responses to expectancy violation (warmth attributions but low-warmth targets) against warmth attributions to the situation (not about the target), and to the orthogonal person-perception trait dimension (in this case, competence). Therefore, we perform a four-way *competence of social target* (high vs. low) X *warmth of social target* (high vs. low) X *behavior trait* (competence vs. warmth) X *focus of attribution* (dispositional vs. situational) repeated measures ANOVA. We predict a significant three-way (high vs. low) *warmth* X (warmth vs. competence) *trait* X (dispositional vs. situational) *focus of attribution* interactions.

The predicted three-way interaction is not significant in the putamen. However, there is a significant *warmth* main effect, F (1, 11) = 7.23, partial η^2 = .40, Ω = .69, such that there is more activation to low warmth than high warmth targets after dispositional attributions (see Figure 2b for means). We also find a significant two-way *competence* X *trait* interaction, F (1, 11) = 4.95, partial η^2 = .31, Ω = .53, such that warmth attributions to high competence social targets elicit less activity than warmth attributions to low competence social targets and competence attributions to high and low competence social targets (see Figure 2c for means). Finally, in the putamen there is also a significant three-way *competence* X *warmth* X *trait* interaction, F (1, 11) = 8.02, partial η^2 = .42, Ω = .73, such that high warmth high competence social targets elicit less activity after warmth trait attributions than after competence trait attributions. There are no differences between the warmth and competence attributions for any of the three remaining competence X warmth social targets (see Figure 2d for means). There is also a marginally significant *competence* X *focus* two-way interaction, F (1, 11) = 3.91, p = .07.

These findings suggest that an area of the striatum activates to expectancy violations for behavior of low warmth people. However, the putamen responds not only to the warmth trait dimension, but to competence as well. This suggests that this area is indeed calculating an expectancy violation in the social domain. The putamen responds particularly to distinctions along the warmth dimension of the target's social group, and to violations involving competent behavior.

Competence Expectancy Violation

We next performed a whole-brain analysis, contrasting competence attributions to low- versus high-competence targets after dispositional attribution. Areas *less* active in this contrast are to social targets after inconsistent competence behavioral information. This suggests that these neural regions are involved in an expectation violation for competence

[7]This contrast also reveals an area of middle orbitofrontal cortex responding to warmth expectancy violations. The predicted three-way interaction is not significant in OFC. However, there is a significant two-way *warmth* X *focus of attribution* interaction, F (1, 11) = 4.89, p < .05, partial η^2 = .31, Ω = .52, showing more activation to high than low warmth targets for situational attributions. The two-way *trait* X *focus of attribution* interaction is also significant, F (1, 11) = 5.36, p < .04, partial η^2 = .33, Ω = .56, showing more activation to warmth than competence information for situational attributions, and the inverse for dispositional attributions. This pattern of responding is unlike the striatum.

information, expecting a competent target but perceiving an inept target. We consider this a *competence expectancy violation signal (CEV)*.

We find a subgenual cingulate area in frontal cortex underlying C_{EV}, t (11) = 3.49, at x = 8, y = −6, z = 21, 170 voxels, partial η^2 = .58 (see Figure 4a). We perform follow-up ROI analyses as described above for warmth expectancy violations. Again, we predict a significant three-way interaction. Instead, there is a significant two-way *competence* X *trait* interaction, F (1, 11) = 4.72, partial η^2 = .30, Ω = .51, showing more activation to low than high competence targets for dispositional attributions, and more activation to high than low competence targets for situational attributions (see Figure 5b for means). This cross-over interaction is partially consistent with our hypotheses and suggests that this brain region responds uniquely to expectancy violations along the competence trait dimensions, though not differentiating between the trait conveyed by the behavior.

Consistency Signals

Different neural patterns emerge when the social target is consistent with the expectancy. Areas *more* active in the warm expectancy violation contrast respond to social targets after consistent warmth behavioral information. These neural regions respond consistent with expectations for warmth information, expecting a warm target and perceiving a warm target. We consider this a *warmth consistency signal (W_C)*. We find areas of hippocampus, precuneus, and superior temporal gyrus underlying W_C (see Table 1 for statistics).

Areas *more* active in the competent expectancy violation contrast respond to social targets after consistent competence behavioral information. These neural regions respond consistent with expectations for competence information, expecting a competent target and perceiving a competent target. We consider this a *competence consistency signal (C_C)*. We find a number of regions underlying C_C, including precuneus, bi-lateral areas of frontal cortex, parietal cortex, thalamus, and cerebellum (see Table 2 for statistics).

Conclusion

Consistent with our hypotheses, we find striatal regions underlying expectancy violation along the trait warmth dimension, and valuation areas underlying expectancy violation along the trait competence dimension. Furthermore, these activations are specific to each trait dimension, and specific to dispositional attributions, that is, specific to the target as the focus of the violation. This suggests that these regions may be used to calculate learning signals for social information. In particular, the putamen does not distinguish the expectancies induced by the narratives, while the caudate is insensitive to the character of the target picture.

These findings suggest that the focus of the attribution matters. Social psychological research demonstrates that the stimulus context can influence stereotype activation and subsequent implicit prejudice (Dasgupta & Greenwald, 2001; Karpinsky & Hilton, 2001; Wittenbrink, Judd, & Park, 2001), and social neuroscience agrees (Harris & Fiske, 2007; Wheeler & Fiske, 2005). However, the significant interactions for this study in the trait warmth domain suggest that situational influences on attributions may be specific to the trait competence domain. Moreover, the data also suggest that these learning signals are specific for each trait dimension.

These data are initial evidence that structures underlying reinforcement learning are also involved in social learning. Therefore, this study enriches the understanding of the valuation neural network by extending it to the social domain. Researchers often show that the punishment is associated with a faster decay in the BOLD signal than the reward (Delgado,

2007). We cannot directly test this feature of neural networks of prediction error because our task is not the same kind of learning task often employed in those studies. However, the findings in the social domain described in this manuscript may demonstrate a prediction error signal as it is defined by neuroeconomists. This suggests that social learning may occur via a similar neural mechanism engaged in reinforcement learning, and that the independent exploration of warmth and competence (as opposed to studying the interaction between the two traits that spontaneously occurs during person perception) has revealed more complexities about a social learning signal. In either case, further research must be conducted to delineate the nuances of this most important learning signal.

Finally, previous research using event-related potentials (ERPs) has shown a positivity waveform to expectancy violation after 300 milliseconds (Bartholow, Fabiani, Gratton, & Bettencourt, 2001), presumably localized to anterior cingulate cortex (ACC; Oliveira, McDonald, & Goodman, 2007; van Veen, Holroyd, Cohen, Stenger, & Carter, 2004). Imaging research also implicated the dorsal ACC in expectancy violation (Somerville, Heatherton, & Kelley, 2006). However, we did not find areas of ACC more active for either warmth or competence violations at our a priori thresholds.

Because the same neural areas engaged in prediction error and valuation during reinforcement learning are engaged for social learning, then it suggests that conditioning strategies used to modify instrumental action could be used to modify person perception. Most stereotype change research focuses on changing the perception of social groups by changing the affective response to the group. Perhaps dopaminergic[8] agonists may be useful tools that influence social learning and could change existing stereotypes. Therefore, these results suggest that strategies commonly practiced in behavioral neuroscience and emotion-learning research may be used to modify stereotypes and prejudices.

Acknowledgments

We thank the Center for Brain Mind and Behavior at Princeton University for funding this research and technical support. We also thank Bruce Barcelow, Jian Li, and Samuel McClure for feedback on earlier versions of the manuscript, and Mina Cikara and Lulu Kuang for assistance in collecting the imaging data.

References

Asch SE. Forming impressions of personality. Journal of Abnormal and Social Psychology. 1946; 41:258–290.

Bargh JA, Chen M, Burrows L. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. Journal of Personality and Social Psychology. 71:230–244. (199). [PubMed: 8765481]

Bartholow BD, Fabiani M, Gratton G, Bettencourt BA. A psychophysiological examination of cognitive processing of and affective responses to social expectancy violations. Psychological Science. 2001; 12:197–204. [PubMed: 11437301]

Chen M, Bargh JA. Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. Personality and Social Psychology Bulletin. 1999; 23:215–224.

Cunningham WA, Van Bavel JJ, Johnsen IR. Affective Flexibility: Evaluative Processing Goals Shape Amygdala Activity. Psychological Science. 2008; 19:152–160. [PubMed: 18271863]

Dasgupta N, Greenwald AG. On the malleability of automatic attitudes: Combining automatic prejudice with images of admired and disliked individuals. Journal of Personality and Social Psychology. 2001; 36:316–328.

[8]Though it is a highly contentious debate as to the exact role of dopamine in reward, specifically when present in striatal regions, considering these regions receive afferent inputs from other regions beside the substantia nigra of the basal ganglia, we raise this possibility as a potential subsequent study, not a definitive statement about the role of dopamine in social learning.

de Waal FBM. How animals do business. Scientific American. 2005; 292:72–79. [PubMed: 15882024]

Delgado MR. Reward-related responses in the human striatum. Annals of the New York Academy of Science. 2007; 1104:70–88.

Delgado MR, Frank RH, Phelps EA. Perceptions of moral character modulate the neural systems of reward during the trust game. Nature Neuroscience. 2005; 8:1611–1618.

Delgado MR, Locke HM, Stenger VA, Fiez JA. Dorsal striatum responses to reward and punishment: Effects of valence and magnitude manipulations. Cognitive, Affective & Behavioral Neuroscience. 2003; 3:27–38.

Dijksterhuis A, Aarts H, Bargh JA, van Knippenberg A. On the relation between associative strength and automatic behavior. Journal of Experimental Social Psychology. 2000; 36:531–544.

Dijksterhuis, A.; Bargh, JA. The perception-behavior expressway: Automatic effects of social perception on social behavior. In: Zanna, MP., editor. Advances in experimental social psychology. Vol. 33. New York: Academic Press; 2001. p. 1-40.

Dijksterhuis A, Spears R, Lepinasse V. Reflecting and deflecting stereotypes: Assimilation and contrast in impression formation formation and automatic behavior. Journal of Experimental Social Psychology. 2001; 37:286–299.

Dijksterhuis A, van Knippenberg AV. The relationship between perception and behavior, or how to win a game of Trivial Pursuit. Journal of Personality and Social Psychology. 1998; 74:865–877. [PubMed: 9569649]

Dijksterhuis A, van Knippenberg AV. On the parameters of associative strength: Central tendency and variability as determinants of stereotype accessibility. Personality and Social Psychology Bulletin. 1999; 25:527–536.

Dijksterhuis A, van Knippenberg AV. Behavioral indecision: Effects of self-focus on automatic behavior. Social Cognition. 2000; 18:55–74.

Dovidio JF, Kawakami K, Gaertner SL. Implicit and explicit prejudice and interracial interaction. Journal of Personality and Social Psychology. 2002; 82:62–68. [PubMed: 11811635]

Dovidio JF, Kawakami K, Johnson C, Johnson B, Howard A. On the nature of prejudice: Automatic and controlled processes. Journal of Experimental Social Psychology. 1997; 33:510–540.

Fiske ST, Cuddy AJ, Glick P, Xu J. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology. 2002; 82:878–902. [PubMed: 12051578]

Fiske ST, Dyer LM. Structure and development of social schemata: Evidence from positive and negative transfer effects. Journal of Personality and Social Psychology. 1985; 48:839–852.

Fiske, ST.; Taylor, SE. Social cognition: From brains to culture. New York: McGraw-Hill; 2008.

Frith U, Frith C. The biological basis of social interaction. Current Directions in Psychological Science. 2001; 10:151–155.

Gallagher HL, Frith CD. Functional imaging of ‘theory of mind’. Trends in Cognitive Sciences. 2002; 7:77–83. [PubMed: 12584026]

Harris LT, McClure SM, Van den Bos W, Cohen JD, Fiske ST. Regions of the MPFC differentially tuned to social and nonsocial affective evaluation. Cognitive and Behavioral Neuroscience. 2007; 7:309–316.

Harris LT, Todorov A, Fiske ST. Attributions on the brain: Neuro-imaging dispositional inferences beyond theory of mind. Neuroimage. 2005; 28:763–769. [PubMed: 16046148]

Heberlein SA, Adolphs R, Tranel D, Kemmerer D, Anderson S, Damasio AR. Impaired attribution of social meanings to abstract dynamic geometric patterns following damage to the amygdala. Society for Neuroscience Abstracts. 1998; 24:1176.

Heider, F. The Psychology of Interpersonal Relations. Wiley; New York: 1958.

Heider F, Simmel M. An experimental study of apparent behavior. American Journal of Psychology. 1944; 57:243–259.

Jones EE. The rocky road from acts to dispositions. American Psychologist. 1979; 34:107–117. [PubMed: 484919]

Karpinsky A, Hilton JL. Attitudes and the implicit associations test. Journal of Personality and Social Psychology. 2001; 81:774–788. [PubMed: 11708556]

Kelley, HH. Attribution in social interaction. In: Jones, EE.; Kanouse, DE.; Kelley, HH.; Nisbett, RE.; Valins, S.; Weiner, B., editors. Attribution: Perceiving the cause of behavior. Lawrence Elbaum & Associates; Hillsdale, NJ: 1972. p. 1-26.

Knutson B, Fong GW, Bennett SM, Adams CS, Hommer D. A region of medial prefrontal cortex tracks monetarily rewarding outcomes: Characterization with rapid event-related fMRI. NeuroImage. 2003; 18:263–272. [PubMed: 12595181]

Leach CW, Ellemers N, Barreto M. Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. Journal of Personality and Social Psychology. 2007; 93:234–249. [PubMed: 17645397]

McArthur LA. The how and what of why: Some determinants and consequences of causal attribution. Journal of Personality and Social Psychology. 1972; 72:171–193.

McClure SM, Laibson DI, Loewenstein G, Cohen JD. Separate neural systems value immediate and delayed monetary rewards. Science. 2004; 306:503–507. [PubMed: 15486304]

Montague PR, King-Casas B, Cohen JD. Imaging valuation models in human choice. Annual Review of Neuroscience. 2006; 29:417–448.

Niv Y, Schoenbaum G. Dialogues on prediction errors. Trends in Cognitive Sciences. 2008; 12:265–272. [PubMed: 18567531]

Oliveira FT, McDonald JJ, Goodman D. Performance monitoring in the anterior cingulate cortex is not all error related: Expectancy deviation and the representation of action-outcome associations. Journal of Cognitive Neuroscience. 2007; 19:1994–2004. [PubMed: 17892382]

Phelps, EA. Emotion, Learning, and the Brain: From Classical Conditioning to Cultural Bias. In: Baltes, P.; Reuter-Lorenz, P.; Rosler, F., editors. Lifespan development and the brain: The perspective of biocultural co-constructivism. NY: Cambridge University Press; 2006. p. 200-216.

Russell AM, Fiske ST. It's all relative: Social position and interpersonal perception. European Journal of Social Psychology. in press.

Saxe R, Wexler A. Making sense of another mind: The role of the right temporo-parietal junction. Neuropsychologia. 2005; 43:1391–1399. [PubMed: 15936784]

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275:1593–1599. [PubMed: 9054347]

Sears DO. The person-positivity bias. Journal of Personality and Social Psychology. 1983; 44:233–250.

Skowronski JJ, Carlston DE. Social judgment and social memory: The role of cue diagnosticity in negativity, positivity and extremity biases. Journal of Personality and Social Psychology. 1987; 52:689–699.

Somerville LH, Heatherton TF, Kelley WM. Anterior cingulate cortex responds differentially to expectancy violation and social rejection. Nature Neuroscience. 2006; 9:1007–1008.

Talairach, J.; Tournoux, P. Co-planar stereotaxic atlas of the human brain. New York: Thieme; 1988.

Tversky A, Kahneman D. Extension versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review. 1983; 90:293–315.

Van den Bos W, McClure SM, Harris LT, Fiske ST, Cohen JD. Dissociating affective evaluation and social cognitive processes in the ventral medial prefrontal cortex. Cognitive, Affective, and Behavioral Neuroscience. 2007; 7:337–346.

van Veen V, Holroyd CB, Cohen JD, Stenger VA, Carter CS. Errors without conflict: Implications for performance monitoring theories of anterior cingulate cortex. Brain and Cognition. 2004; 56:267–276. [PubMed: 15518940]

Wheeler SC, Petty RE. The effects of stereotype activation on behavior: A review of possible mechanisms. Psychological Bulletin. 2001; 127:797–826. [PubMed: 11726072]

Wheeler ME, Fiske ST. Controlling racial prejudice: Social-cognitive goals affect amygdala and stereotype activation. Psychological Science. 2005; 16(1):56–63. [PubMed: 15660852]

Winston JS, Strange BA, O'Doherty J, Dolan RJ. Automatic and intentional brain responses during evaluation of trustworthiness of faces. Nature Neuroscience. 2002; 5:277–283.

Wittenbrink B, Judd CM, Park B. Spontaneous prejudice in context: Variability in automatically activated attitudes. Journal of Personality and Social Psychology. 2001; 81:815–827. [PubMed: 11708559]

Word CO, Zanna MP, Cooper J. The nonverbal mediation of self-fulfilling prophecies in interracial interaction. Journal of Experimental Social Psychology. 1974; 10:109–120.

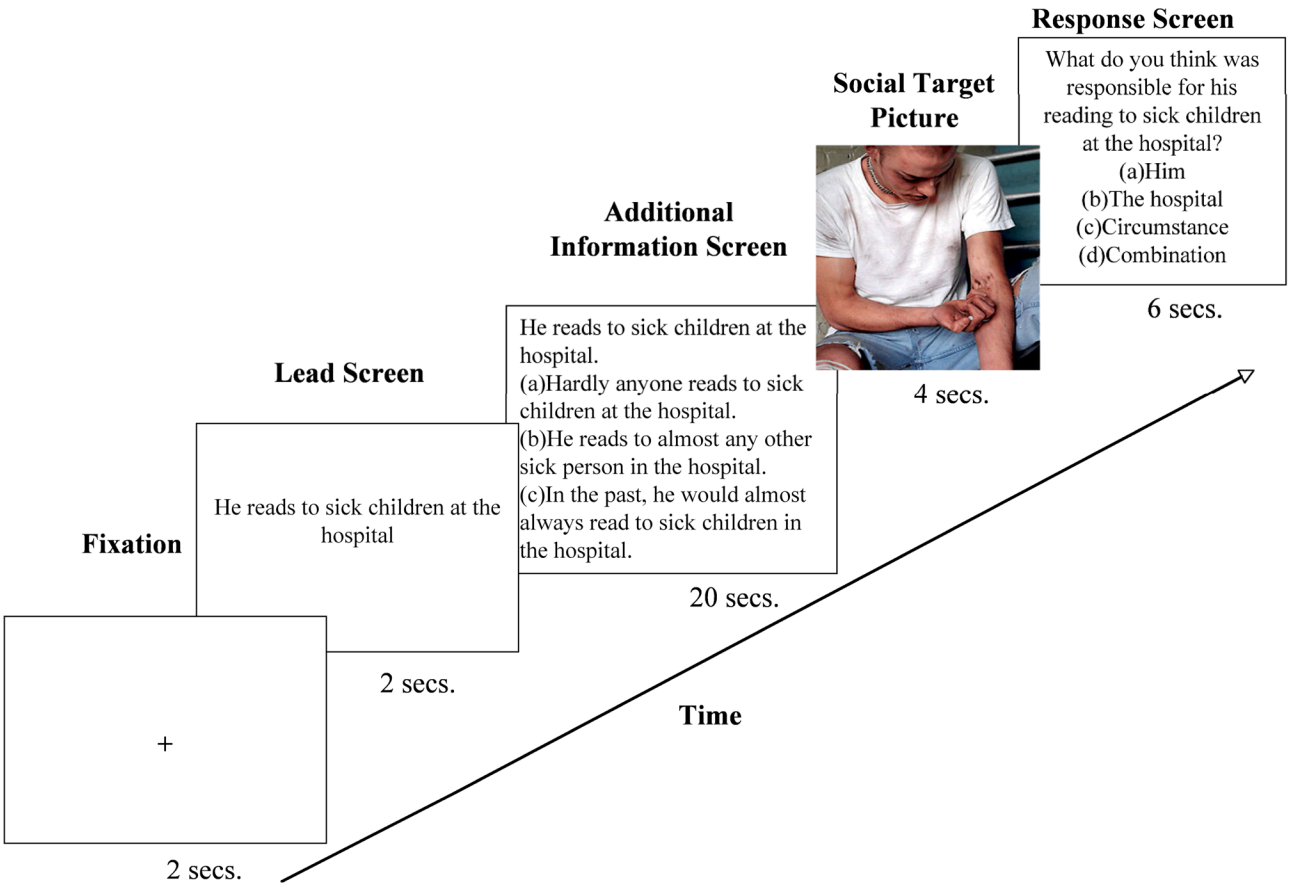

Figure 1. Expectancy Violation Attribution Paradigm

The schematic describes the timecourse of the paradigm. Participants first considered the behavior, then the behavior paired with information that created dispositional or situational attribution for the behavior. The participants, presumably with this expectancy in mind, then saw a picture of the person responsible for the behavior. The participants then indicate responsibility for the behavior.

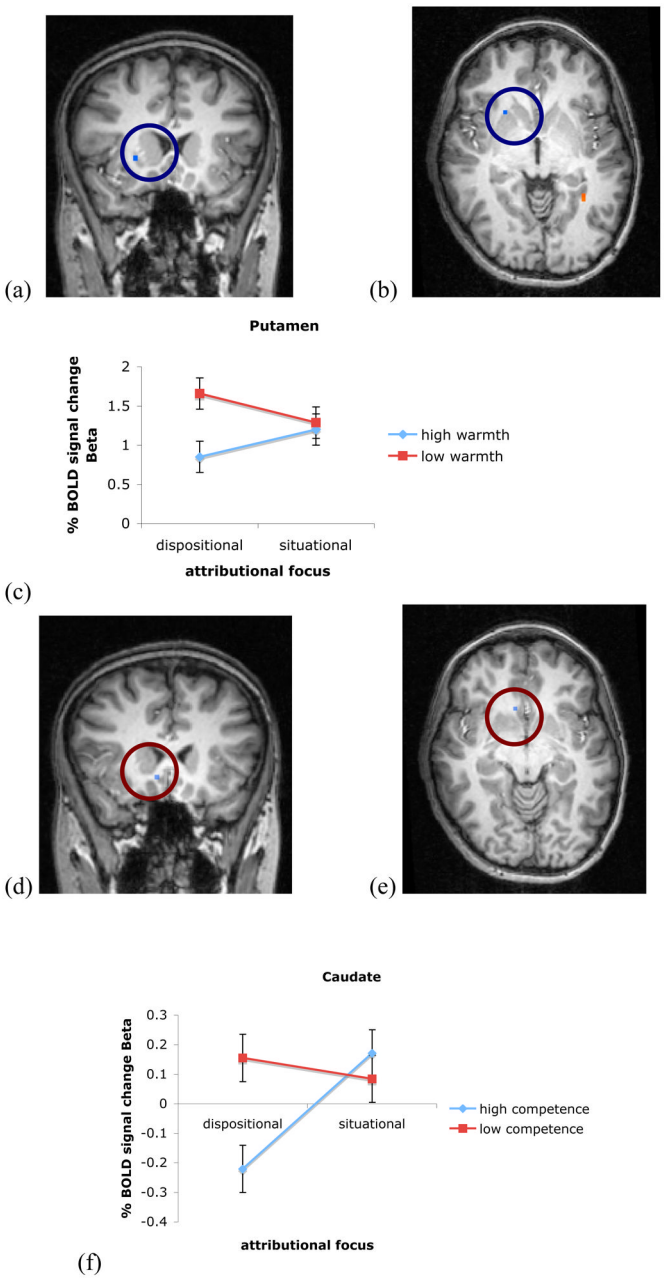

Figure 2. Striatal regions active for social expectancy violations

A mask showing the cluster of voxels more active in the putamen to warmth expectancy violations. The line graph depicts the mean % signal change to each of the four conditions, showing a significant interaction between focus of attribution and (high or low) warmth of the target.

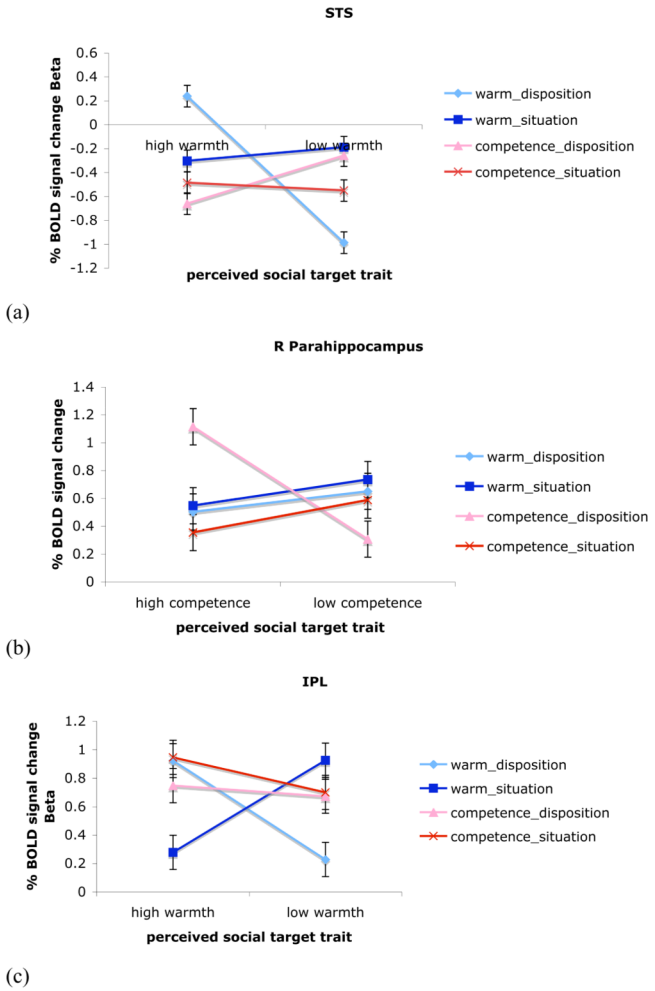

Figure 3. Three-way *warmth/competence* X *trait* X *focus* interactions

A mask showing the cluster of voxels more active in the subgenual cingulate in frontal cortex to competence expectancy violations. The line graph depicts the mean % signal change to each of the four conditions, showing a significant interaction between (high or low) competence of the target and trait information described in the behavior.

Table 1

Regions more active in the warmth expectancy inconsistent versus expectancy consistent contrast.

Neural Regions Active to Warmth Expectancy Violation Social Targets

Brain Region	Talairach Coordinates (*x*, *y*, *z*)	Cluster Size	*t*-value	Partial η^2
R, Middle Frontal Gyrus, (BA) 9	41, 11, 28	72	4.44	0.64
R, Lentiform Nucleus, Putamen	24, 16, 3	23	3.39	0.51

Neural Regions Active to Warmth Expectancy Consistent Social Targets

Brain Region	Talairach Coordinates (*x*, *y*, *z*)	Cluster Size	*t*-value	Partial η^2
R, Occipital Lobe, Precuneus	22, −76, 26	13	2.88	0.43
L, Hippocampus	−33, −44, 3	117	4.20	0.62
R, Superior Temporal Gyrus (BA 38)	24, 17, −32	17	3.36	0.51

Table 2

Regions more active in the competence expectancy inconsistent versus expectancy consistent contrast.

Neural Regions Active to Competence Expectancy Violation Social Targets

Brain Region	Talairach Coordinates (*x*, *y*, *z*)	Cluster Size	*t*-value	Partial η^2
L, Subgenual cingulate	8, 21, −6	51	3.50	0.58
R, Ventrolateral Prefrontal Gyrus	49, 27, −8	11	3.96	0.59

Neural Regions Active to Competence Expectancy Consistent Social Targets

Brain Region	Talairach Coordinates (*x*, *y*, *z*)	Cluster Size	*t*-value	Partial η^2
R, Occipital Lobe, Precuneus (BA) 31	10, −49, 30	275	4.89	0.68
L, Medial Frontal Gyrus, (BA) 6	−11, 5, 55	16	4.51	0.65
R, Middle Frontal Gyrus, (BA) 8	28, 39, 42	144	4.22	0.62
L, Middle Frontal Gyrus, (BA) 8	−25, 41, 46	46	4.38	0.64
L, Middle Frontal Gyrus, (BA) 9	−40, 32, 37	67	4.96	0.69
L, Middle Frontal Gyrus, (BA) 10	−41, 45, 18	133	4.29	0.63
R, Superior Frontal Gyrus (BA) 6	24, 10, 45	23	3.32	0.50
L, Superior Frontal Gyrus (BA) 6	−22, 9, 39	27	4.87	0.68
L, Parietal Lobe, Postcentral Gyrus, (BA) 2	−48, −26, 30	285	5.01	0.69
R, Parahippocampal Gyrus, (BA) 34	22, 5, −16	332	5.67	0.75
R, Cerebellum, Culmen, Anterior Lobe	39, −38, −22	64	5.08	0.70
L, Thalamus, Ventral Posterior Lateral Nucleus	−19, −16, 8	31	4.27	0.62