Neural regions that underlie reinforcement learning are also active for social expectancy violations

Lasana T. Harris
New York University, New York, NY, USA

Susan T. Fiske
Princeton University, Princeton, NJ, USA

Prediction error, the difference between an expected and an 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 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 high warmth or competence behavioral information that led to dispositional or situational attributions for the behavior. 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 that represent evaluative and prediction error signals. Social cognition regions underlie consistent expectations. These findings suggest that regions underlying reinforcement learning may work in concert with social cognition regions in warmth and competence social expectancy. This study illustrates the neural overlap between neuroeconomics and social neuroscience.

Keywords: Trait attribution; Social expectancy violation; Trait warmth and competence; Striatum; Reinforcement learning.

INTRODUCTION

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 model result from a discrepancy between predicted and actual outcomes. This prediction 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

Correspondence should be addressed to: Lasana T. Harris, 6 Washington Place, C&P, RM# 865, New York, NY 10003, USA. E-mail: lasana@nyu.edu

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.

www.psypress.com/socialneuroscience DOI:10.1080/17470910903135825

(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 apparently is 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 social targets who contradict prior knowledge are salient (Jones & McGill, 1967). 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; Richeson & Shelton, 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 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, allowing for rapid attributions during person perception (Fiske, Lin, & Neuberg, 1999), which can be useful for predicting their behavior. People are constantly adjusting their perception of other individuals along warmth and competence dimensions.[1]

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 complex 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).

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). Structures in the striatum (caudate, putamen, and nucleus accumbens) and the amygdala have also been implicated in calculating prediction error and reward value (Niv & Schoenbaum, 2008).

There are two kinds of prediction error, positive and negative (see Schultz, Dayan, & Montague, 1997, for a review). Positive prediction errors occur when the animal receives more reward (or punishment) than anticipated, and negative prediction errors occur when the animal receives less reward (or punishment) than anticipated. This distinction 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.

There is a vast literature using event-related potentials (ERPs) demonstrating a positivity waveform to expectancy violation after 300 ms (P3, P300, or late positive potential (LPP); Bartholow, Fabiani, Gratton, & Bettencourt, 2001) to both social (traits) and non-social (fruits) stimuli (Cacioppo, Crites, Berntson, & Coles, 1993). This response is insensitive to accuracy (Crites, Cacioppo, Gardner, & Berntson, 1995), suggesting that the P3 responds to the categorization process, not verbal report. Research has also implicated the P3 in spontaneous trait inferences (Van Duynslaeger, Sterken, Van Overwalle, & Verstraeten, 2008), a kind of categorization process involving binding social targets to trait and affective categories. Finally, like other cognitive processes, drugs such as alcohol modulate this response (Bartholow, Pearson, Gratton, & Fabiani, 2003).

The P3 is usually observed over centro-parietal parts of the scalp (Bartholow et al., 2001; Cacioppo et al., 1993; Cacioppo, Crites, Gardner, & Berntson, 1994; Crites et al., 1995; Van Duynslaeger et al., 2008). Research suggests this P3 response may originate from the locus coeruleus-norepinephrine (LC-NE) system distributed throughout the brain (for review, see Nieuwenhuis,

[1] Assessments of warmth and competence satisfy the components of the Rescorla-Wagner model: $V_{\text{new}} = V_{\text{old}} + \eta(R - V_{\text{old}})$. 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.

Aston-Jones, & Cohen, 2005). Imaging research also implicates dorsal anterior cingulate cortex (ACC) in expectancy, specifically outcome worse than expected (Somerville, Heatherton, & Kelley, 2006). In sum, there is a reliable neural response to stimuli inconsistent with expectancies independent of the social nature of the stimuli.

Predictions

To determine the neural correlates of social expectancy violation, we had participants respond to information that led them to make either highly warm (moral) or highly competent (intelligent) attributions for behavior.[2] 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. Dispositional attributions focus specifically on the actor that engages in the behavior, not the action, or context. As a control in our current study, behaviors were also attributable to context, that is, a unique situation. When attributed to the situation, the behavior is held constant but does not allow for the prediction of the person's future behavior. 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.

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 striatal 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 normal vision, were native English speakers, had lived in the U.S. for at least five years (suggesting they were aware of the cultural stereotypes about the social groups), and provided informed consent. The mean age was 20.4 years, with four women, and three members of ethnic minorities.

Stimuli

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 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). 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, Todorov, & Fiske, 2005; Kelley, 1972; McArthur, 1972). Half the behaviors led to a dispositional

[2] 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.

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]). Therefore, only information about consensus and distinctiveness differentiated dispositional from situational attributions.

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 15 pictures per quadrant, high and low on warmth and competence. Each picture depicted a person who was from a group pretested as eliciting high warmth expectancies (American hero [firefighter, police officer, astronaut, athlete], college student, elderly person, disabled person), or low warmth expectancies (business person, rich person, homeless person, drug addict), and high competence expectancies (American hero, college student, business person, rich person) or low competence expectancies (elderly person, disabled person, homeless person, drug addict). This results in pictures in each cell of the 2 (Target Warmth) × 2 (Target Competence) × 2 (Behavior Trait) × 2 (Focus of Attribution) design.

Scanning parameters

All fMRI scanning was conducted at Princeton's Center for the Study of Brain, Mind, and Behavior, using a 3.0 T 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.: www.curdes.com/responsepads.html). 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 min 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 3 mm axial slices; 1 mm gap, repetition time (TR): 2 s; TE: 30 ms). In-plane resolutions were 3 mm × 3 mm (196 mm FOV, 64 × 64 matrix).

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 a sentence led to only one kind of attribution in that version: half the combinations to a dispositional attribution; 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 attributions for behavior on one complete practice run. We did not exclude any participants for failure to make correct attributions.

In the scanner, participants first saw a fixation cross for 4 s. Next, the behavior sentence was presented for 2 s. The information screen then appeared and remained for 20 s following the

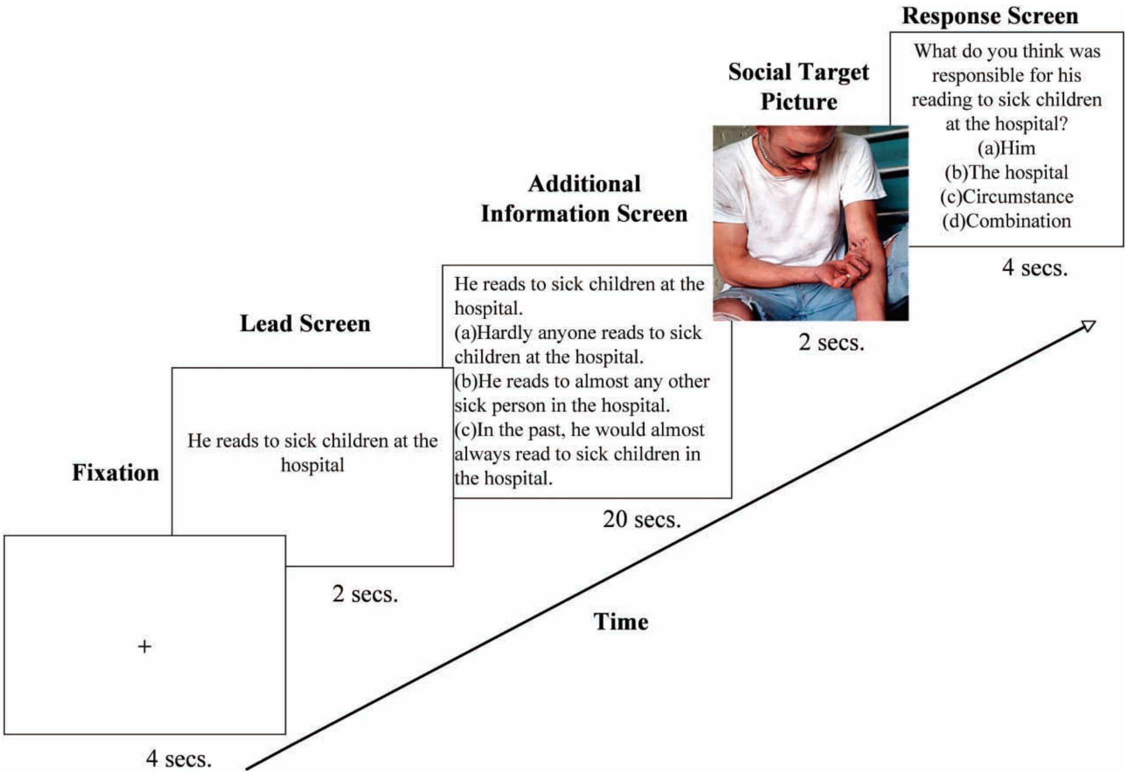
Response Screen
What do you think was responsible for his reading to sick children at the hospital?
(a)Him
(b)The hospital
(c)Circumstance
(d)Combination
4 secs.
Social Target Picture
2 secs.
Additional Information Screen
He reads to sick children at the hospital.
(a)Hardly anyone reads to sick children at the hospital.
(b)He reads to almost any other sick person in the hospital.
(c)In the past, he would almost always read to sick children in the hospital.
20 secs.
Lead Screen
He reads to sick children at the hospital
2 secs.
Fixation
+
4 secs.
Time

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 attribution for the behavior, or situational attribution. The participants, presumably with this expectancy in mind, then saw a picture of the person responsible for the behavior. The participant then indicated responsibility for the behavior.

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 for 2 s after the information screen.[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 s, followed by a fixation cross. Each run contained 15 trials, and each participant completed four 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) voxel-wise linear detrending across time; (4) temporal bandpass filtering to remove low frequency (scanner and physiology related) noise. We later

[3] We reverse the conventional order, presenting the behavior first and 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 s after the sentences are presented (see Harris et al., 2005). Therefore, we could employ a block design in order to capture the attributions across the entire 20-s period, regardless of when they occurred. However, 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 presented, 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.

add Fourier predictors (two 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 full width at half maximum (FWHM)).

Data analysis

Data analysis used the general linear model (GLM) available on the Brain Voyager QX software package. We first computed a GLM focusing just on the 2 s when the pictures were displayed because this was the occurrence of expectancy violations. We computed series of regressors to examine blood oxygen level-dependent (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 of the peak voxel. We present cluster sizes at the same resolution at which the data was analyzed. Random effects analyses were performed on all imaging data. All data are presented with their coordinates based on a standard system (Talairach & Tournoux, 1988).

Additionally, we conducted region of interest (ROI) analyses. That is, we extracted the mean betas representing percent BOLD signal change for each participant within a cluster of voxels active in the exploratory analyses. This is a measure of neural activity to each information combination and picture interaction in regions identified in the exploratory analysis. We computed 2 (Target Warmth) × 2 (Target Competence) × 2 (Trait) × 2 (Focus of Attribution) repeated measures ANOVAs on the mean betas extracted from each cluster.

RESULTS AND DISCUSSION

Recall our main predictions: 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 striatal 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 ROI analyses are significant at $p < .05$.

Unpacking of a four-way interaction

Recall from the task that there are three separate components together making up the expectation and outcome. The first is the behavioral trait, that is, whether the behavior described is warm or competent. This is the first main effect: *trait*. Neural regions sensitive to this main effect are influenced by which trait is presented 22 s before, and displayed until, perceiving the social target. Then, the participant receives additional information that leads to the second component: the attribution for the behavior, either to the social target or to the situation. This is the second main effect: *focus*. This information must be inferred, and occurs temporally after the behavioral trait. Neural regions sensitive to this main effect are responding to an attribution that requires mental calculation after the trait is presented, but before perceiving the social target. Therefore, these two components capture the expectancy. Neural regions responding to interactions between these expectancy components are responding to information before the social target is presented, that is, before the TR of analysis.

The third component is the social target: the outcome. This component is itself an interaction between two perceived trait dimensions inherent in the social target's social group: *warmth* and *competence*, the last two main effects. Neural regions that differentiate along these main effects are responding to the physical stimulus present at the TR of analysis. Neural regions that interact with either or both of these two outcome main effects and one or both of the expectancy main effects are integrating information: expectancy and outcome. Hence, our a priori predictions are either of two significant three-way interactions between each of the three

[4] We use just some of our data to define the ROIs, to allow unbiased comparisons within our ROIs. Therefore, we expect the biased simple effect to be significant within the larger ANOVA model. However, no other simple effect, main effect, or interaction is biased by this ROI selection strategy.

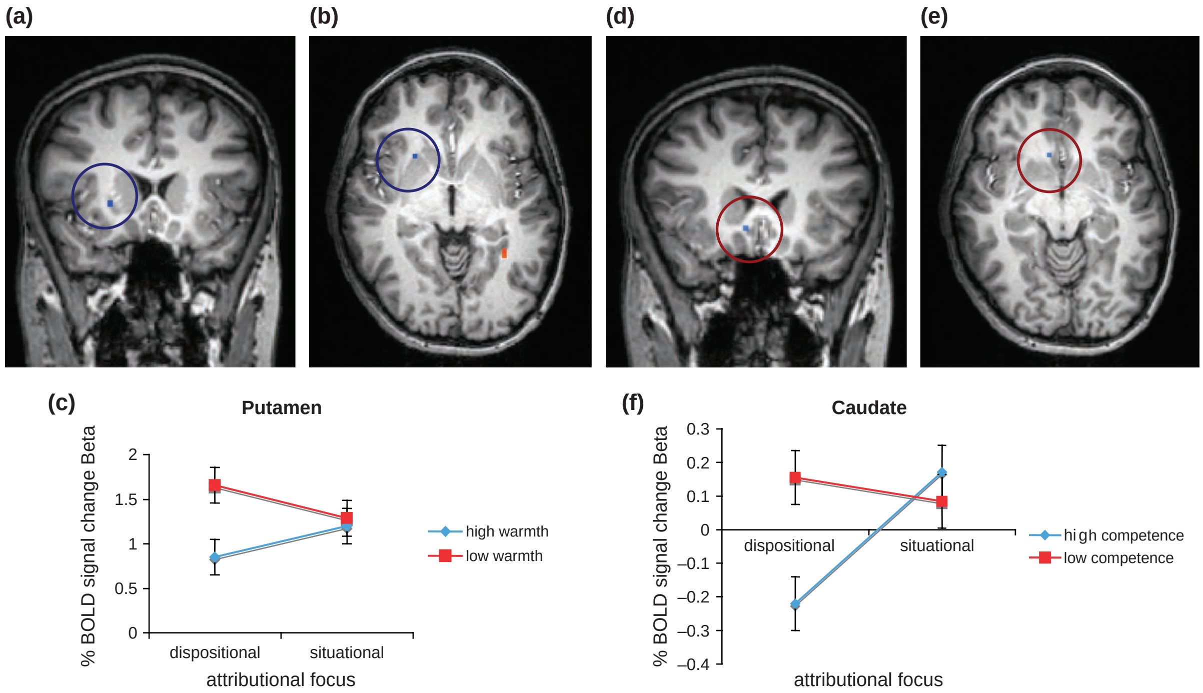

Figure 2. Masks depicting activation in striatal regions to expectancy violation: (a) putamen, coronal view; (b) putamen, axial view; (c) line graphs showing putamen ROI betas and *SE* of the mean, lines refer to levels of *warmth* main effect; (d) caudate, coronal view; (e) caudate, axial view; (f) line graphs showing ROI betas and *SE* of the mean, lines refer to levels of *competence* main effect. Blue circles on brain images signify ROI results from high vs. low warmth trait localizer; red circles on brain images signify ROI results from high vs. low competence trait localizer.

components that comprise expectancy and outcome, Warmth × Trait × Focus or Competence × Trait × Focus. We make no specific prediction about the four-way Warmth × Competence × Trait × Focus interaction.

Warmth expectancy violation localizer

We first performed a whole-brain analysis, contrasting high- versus low-warmth targets engaged in warm behavior after dispositional attributions for the behavior. Areas *less*[5] 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}).

Consistent with our hypothesis, we find activity in the lentiform nucleus of the right putamen, $t(11) = -4.38$, $p < .001$, at $x = 25$, $y = 18$, $z = 2$, 27 voxels, partial $\eta_p^2 = 0.64$ (see Figure 2a and 2b). We also find activity in right inferior frontal gyrus (see Table 1 for Talairach coordinates and statistics).

Warmth expectancy violation ROI analysis

We performed a follow-up region of interest (ROI) analysis 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

[5] 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 as our contrasts do not allow for independent exploration of positive and negative prediction errors.

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) × Warmth of Social Target (high vs. low) × Behavior Trait (competence vs. warmth) × Focus of Attribution (dispositional vs. situational) repeated measures ANOVA.

The predicted three-way interaction is not significant in the putamen. However, there is a significant Warmth main effect, $F(1,\ 11) = 8.38$, $p < .015$, $\eta_p{}^2 = 0.43$, $\Omega = 0.75$, such that there is more activation to low warmth than high warmth targets.

We also find a significant two-way Competence × Trait interaction, $F(1, 11) = 5.46$, $p < .039$, $\eta_p{}^2 = 0.39$, $\Omega = 0.59$. This suggests that the putamen is also sensitive to violations of trait, that is, expecting warm, but perceiving competent social targets. The trait main effect marginally significantly interacts with warmth in a two-way Warmth × Trait interaction, $F(1,\ 11) = 3.84$, $p = .076$, and three-way Competence × Warmth × Trait interaction, $F(1, 11) = 4.48, p = .058$. There is a marginally significant Warmth × Focus two-way interaction, $F(1,\ 11) = 3.33$, $p = .095$, showing more activation to low than to high warmth targets for dispositional attributions (see Figure 2c).

These findings suggest that an area of the striatum activates to expectancy violations for behavior along the warmth trait dimension. The putamen responds to a range of violations along the warmth behavior trait, differentiating low from high warmth social targets, and low from high competence social targets. Moreover, though the focus of the attribution may matter only for social targets along the perceived warmth trait dimension, the putamen generally does not distinguish whether the behavioral attribution is to the person or to the situation. This suggests that this area is calculating an expectancy violation in the social domain, specific to moral, high warmth behavior.

Competence expectancy violation localizer

We next performed a whole-brain analysis, contrasting competence attributions to high versus low 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 information, expecting a competent target but perceiving an inept target. We consider this a *competence expectancy violation signal (C_{EV})*. We find an area at the head of the caudate underlying C_{EV}, $t(11) = -4.46$, at $x = 9$, $y = 20$, $z = -5$, $p < .001$, 27 voxels, $\eta_p{}^2 = 0.64$ (see Figure 2d, 2e, and Table 2).

Competence expectancy violation ROI analyses

We performed follow-up ROI analyses as described above for warmth expectancy violations. Again, we predicted a significant three-way interaction. Instead, there was marginally significant two-way Competence × Focus interaction, $F(1,\ 11) = 3.94$, $p = .073$, showing more activation to low than high competence targets for dispositional attributions (see Figure 2f). This interaction is partially consistent with our hypotheses and suggests that this brain region implicated in reward value responds more to high than to low competent social targets after dispositional attribution. This pattern of the means is similar to the pattern in the putamen to warmth, suggesting that different areas of striatum perform the same computation for different traits.

Consistency localizers and ROI analyses

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; see Table 1 for all significant neural regions). We find left parahippocampal gyrus, right inferior parietal lobule (IPL), and areas of the person perception neural network underlying W_C, including right superior temporal sulcus (STS), $t(11) = 4.34$, $p < .001$, at $x = 61$, $y = -6$, $z = -1$, 27 voxels, $\eta_p{}^2 = .63$, and the lentiform nucleus of right precuneus, $t(11) = 4.42$, $p < .001$, at $x = 22$, $y = -75$, $z = 26$, 27

TABLE 1
Regions more active in the warmth expectancy violation and expectancy consistent contrast

Brain region	*Talairach coordinates (x, y, z)*	*Cluster size*	*t-value*	*p-value*	η_p^2
Neural regions active to warmth expectancy violation social targets					
R, Inferior frontal gyrus	43, 12, 29	81	−4.81	.0005	0.68
R, Lentiform nucleus, putamen	25, 18, 2	27	−4.38	.0010	0.64
Neural regions active to warmth expectancy consistent social targets					
R, Superior temporal gyrus (BA 21)	61, −6, −1	27	4.34	.0012	0.63
R, Inferior parietal lobule	52, −27, 47	27	4.43	.0010	0.64
R, Preuneus, occipital lobe (BA 18)	22, −75, 26	27	4.42	.0010	0.64
L, Parahippocampal gyrus (BA 19)	−33, −45, 1	54	4.78	.0006	0.68

voxels, $\eta_p{}^2 = 0.64$. Subsequent ROI analyses reveal a significant Warmth × Trait two-way interaction, $F(1, 11) = 5.28$, $p < .042$, $\eta_p{}^2 = 0.32$, $\Omega = 0.55$, a significant three-way Warmth × Trait × Focus interaction, $F(1, 11) = 8.33$, $p < .015$, $\eta_p{}^2 = 0.43$, $\Omega = 0.75$, and a marginally significant Warmth × Focus two-way interaction, $F(1, 11) = 3.63$, $p = .083$ in the STS. The precuneus reveals marginally significant Warmth × Focus, $F(1, 11) = 3.32$, $p = .096$, and Competence × Warmth × Focus interactions, $F(1, 11) = 4.38$, $p = .060$. This suggests that parts of the person perception neural network respond when the expectancy is consistent. Moreover, these neural regions are sensitive to the perceived trait warmth of the social target, and to the focus of the attribution.

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; see Table 2 for all significant neural regions). We find a region of right parahippocampus and left postcentral gyrus of parietal lobe underlying C_C.

We predicted significant three-way Warmth/Competence × Trait × Focus interactions. Right parahippocampal gyrus underlying C_C and both right STS and right IPL underlying W_C produced the significant three-way interaction (see Figure 3), while neither areas underlying W_{EV} nor C_{EV} revealed this interaction. Instead, they revealed significant two-way interactions of the perceived social target trait main effect (*warmth* or *competence*) with either the *trait* implied by the behavior or the *focus* of the attribution. This suggests that areas underlying social expectation violation are sensitive to the *trait* implied by the behavior regardless of the attribution focus, or the attribution *focus*, regardless of the behavioral trait. Conversely, areas underlying social expectation confirmation in both trait domains react to the interaction of implied behavioral *trait*, attribution *focus*, and perceived social target trait (*warmth* or *competence*) as predicted. Perhaps areas underlying consistency integrate all this information during social expectation, while separate areas underlie violation and represent some dimensions, not others, but in concert update future social expectations.

Further unpacking of a four-way interaction

Our results (see Table 3) reveal which neural regions differentiate which components of our task. No area differentiates neither the *trait* nor the *focus* component, though the left post-central gyrus significantly and the right parahippocampal gyrus marginally differentiate the *interaction*, that is, both regions respond to information presented before the physical stimulus: the social target. The inferior frontal gyrus is the only region that differentiates the physical stimuli, showing a significant *warmth* main effect and no interactions. All regions identified, except this region, show significant interactions between the main effects present before the TR of analysis and the main effects present at TR of analysis (social target presentation). This suggests that all these regions interpret the social targets in the context of the prior information. In particular, areas of right STS, right IPL, and right parahippocampal

TABLE 2
Regions more active in the competence expectancy violation and expectancy consistent contrast

Brain region	*Talairach coordinates (x, y, z)*	*Cluster size*	*t-value*	*p-value*	η_p^2
Neural regions active to competence expectancy violation social targets					
R, Caudate, head	10, 21, −4	27	−4.45	.0010	0.64
Neural regions active to competence expectancy consistent social targets					
R, Parahippocampal gyrus (BA 34)	25, 6, −16	54	4.65	.0007	0.66
L, Postcentral gyrus, parietal lobe (BA 2)	−47, −26, 32	108	4.81	.0005	0.68

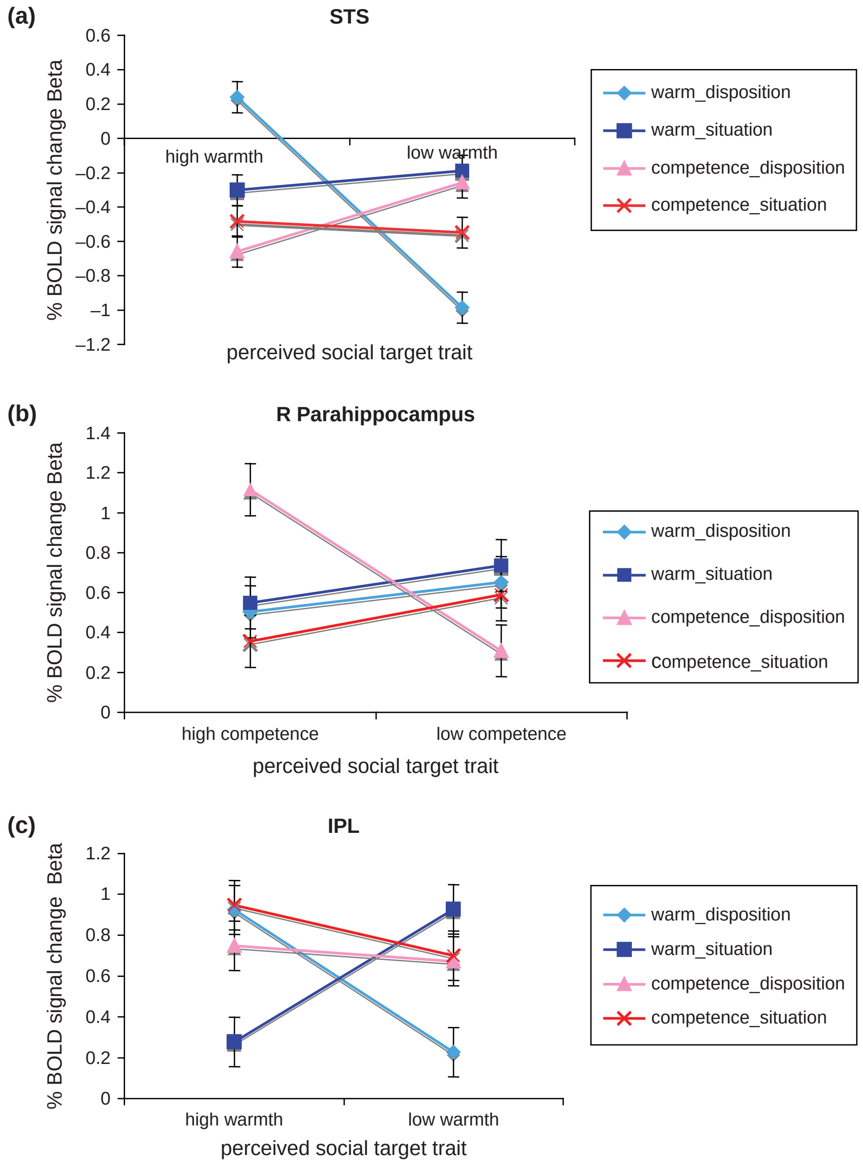

Figure 3. Graphs showing three-way Warmth/Competence × Trait × Focus interactions: (a) STS ROI betas and *SE* of the mean, lines refer to cell of Trait × Focus interaction; (b) parahippocampus ROI betas and *SE* of the mean, lines refer to cell of Trait × Focus interaction; (c) IPL ROI betas and *SE* of the mean, lines refer to cell of Trait × Focus interaction. Blue lines refer to warmth, red lines refer to competence, light lines refer to dispositional attributions, dark lines refer to situational attributions.

TABLE 3

Region of Interest (ROI) Analyses main effects, interactions, and effect sizes

Neural region (Brodman area; Talairach coordinates)	*ROI localizer trait & expectancy*	*ROI main effect and interaction analyses* (η_p^2)													
		W	***C***	***T***	*F*	*W x C*	*W x T*	*W x F*	*C x T*	*C x F*	*T x F*	*W x C x T*	*W x C x F*	*W x T x F*	*C x T x F*
R, Putamen (25, 18, 2)	W_{EV}	**.43**					*.26*	*.23*	**.33**			*.29*			
R, Inferior Frontal Gyrus (BA 9; 43, 12, 29)	W_{EV}	**.38**													
R, Caudate Head (9, 20, –5)	C_{EV}									*.26*					
R, Superior Temporal Gyrus (BA 21; 61, –6, –1)	W_C						**.32**	*.25*						**.43**	
R, Precuneus, Lentiform Nucleus (BA 31; 22, –75, 26)	W_C							*.23*					*.29*		
R, Inferior Parietal Lobule (BA 40; 52, –27, 47)	W_C		*.28*					*.23*					**.34**	**.30**	
L, Parahippocampal Gyrus (BA 19; –33, –45, 1)	W_C		**.32**				**.36**	**.41**	*.24*						
L, Parietal Lobe, Postcentral Gyrus (BA 2; –48, –25, 31)	C_C		**.55**								**.31**		**.30**		
R, Parahippocampal Gyrus (BA 34; 24, 5, –17)	C_C							*.25*		*.23*	*.23*				*.23*

W = social target warmth; C = social target competence; T = behavioral trait; F = focus of attribution; W_{EV} = warmth expectancy violation; C_{EV} = competence expectancy violation; W_C = warmth consistency; C_C = competence consistency; BA = Brodmann Area; **blue** = expectation; **yellow** = outcome; **light green** = expectation outcome interaction; **dark green** = predicted three-way interaction. Decimal numbers refer to effect size as indicated by η_p^2; **bold** = significant at $p < .05$; *italic* = marginal at $p < .10$. Blank cell means the main effect or interaction was not statistically or marginally significant in the neural region.

gyrus show the predicted three-way interactions, suggesting that they integrate each component of the task.

CONCLUSION

Consistent with our hypotheses, we find striatal regions putamen and caudate underlying respectively warmth and competence expectancy violation. In particular, the putamen integrates warmth expectancies induced by the narratives with outcomes, while the caudate integrates competent expectancies induced by the narratives with outcomes. The putamen responds to both trait dimensions of the social target, differentiating behavior traits and attributional focus. The caudate is specific to competence, and activations are specific to dispositional attributions, that is, specific to the social target as the focus of the violation. This suggests that these regions may be used to calculate learning signals for social information.

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). The significant interactions with this main effect suggest that dispositional attribution, a form of social cognition, modulates activity in neural regions that differentiate trait warmth and competence.

Dispositional attributions involve a neural network centered on medial prefrontal cortex (MPFC) and STS (Harris et al., 2005). These results reveal activity in STS underlying consistent expectations. Moreover, this region also activates in concert with other person perception and memory regions to show the predicted three-way interactions. These data are initial evidence that structures underlying reinforcement learning may interact with structures involved in social cognition to direct social learning. Therefore, this study enriches the understanding of the valuation neural network by extending it to the social domain.

Researchers often show that punishment is associated with a faster decay in the BOLD signal than 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 paper may demonstrate a prediction error signal as it is defined by neuroeconomists. Moreover, 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.

Previous research suggests that the ACC (Somerville et al., 2006) or the LC-NE system (Nieuwenhuis et al., 2005) may underlie expectancy violations. We did not find areas of ACC more active for either warmth or competence violations at our a priori thresholds, perhaps because our task did not involve errors *per se*, rather attributions and expectations were consistently informed or not. However, the neural generators of the P3 are still debated. The LC-NE system is distributed throughout the brain, and may underlie the activations we find in this paradigm. Nevertheless, future research may reveal that these brain regions do play a role in social expectancy violations.

The fact that the same neural areas engaged in prediction error and valuation during reinforcement learning are engaged for social learning 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[6] 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.

Manuscript received 28 October 2008
Manuscript accepted 16 June 2009

[6] There 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.

REFERENCES

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, *41*, 258–290.

Bargh, J. A., Chen, M., & Burrows, L. (199). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244.

Bartholow, B. D., Fabiani, M., Gratton, G., & Bettencourt, B. A. (2001). A psychophysiological examination of cognitive processing of and affective responses to social expectancy violations. *Psychological Science*, *12*, 197–204.

Bartholow, B. D., Pearson, M. A., Gratton, G., & Fabiani, M. (2003). Effects of alcohol on person perception: A social cognitive neuroscience approach. *Journal of Personality and Social Psychology*, *85*, 627–638.

Bond, M. H. (1972). Effect of an impression set on subsequent behavior. *Journal of Personality and Social Psychology, 24*, 301–305.

Cacioppo, J. T., Crites, S. L., Berntson, G. G., & Coles, M. G. H. (1993). If attitudes affect how stimuli are processed, should they not affect the event-related brain potential? *Psychological Science*, *4*, 108–112.

Cacioppo, J. T., Crites, S. L., Gardner, W. L., & Berntson, G. G. (1994). Bioelectrical echoes from evaluative categorizations: I. A late positive brain potential that varies as a function of trait negativity and extremity. *Journal of Personality and Social Psychology, 67*, 115–125.

Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin*, *23*, 215–224.

Crites, S. J., Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1995). Bioelectrical echoes from evaluative categorization: II. A late positive brain potential that varies as a function of attitude registration rather than attitude report. *Journal of Personality and Social Psychology*, *68*, 997–1013.

Cunningham, W. A., Van Bavel, J. J., & Johnsen, I. R. (2008). Affective flexibility: evaluative processing goals shape amygdala activity. *Psychological Science*, *19*, 152–160.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combining automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *36*, 316–328.

Delgado, M. R. (2007). Reward-related responses in the human striatum. *Annals of the New York Academy of Science*, *1104*, 70–88.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*, 1611–1618.

Delgado, M. R., Locke, H. M., Stenger, V. A., & Fiez, J. A. (2003). Dorsal striatum responses to reward and punishment: Effects of valence and magnitude manipulations. *Cognitive. Affective & Behavioral Neuroscience*, *3*, 27–38.

de Waal, F. B. M. (2005). How animals do business. *Scientific American*, *292*, 72–79.

Dijksterhuis, A., Aarts, H., Bargh, J. A., & van Knippenberg, A. (2000). On the relation between associative strength and automatic behavior. *Journal of Experimental Social Psychology*, *36*, 531–544.

Dijksterhuis, A., & Bargh, J. A. (2001). The perception–behavior expressway: Automatic effects of social perception on social behavior. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 33, pp. 1–40). New York: Academic Press.

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

Dijksterhuis, A., & van Knippenberg, A.V. (1998). The relationship between perception and behavior, or how to win a game of Trivial Pursuit. *Journal of Personality and Social Psychology*, *74*, 865–877.

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

Dijksterhuis, A., & van Knippenberg, A. V. (2000). Behavioral indecision: Effects of self-focus on automatic behavior. *Social Cognition*, *18*, 55–74.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*, 62–68.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology*, *33*, 510–540.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 357–411). New York: McGraw Hill.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*, 77–83.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*, 878–902.

Fiske, S. T., & Dyer, L. M. (1985). Structure and development of social schemata: Evidence from positive and negative transfer effects. *Journal of Personality and Social Psychology*, *48*, 839–852.

Fiske, S. T., Lin, M., & Neuberg, S. L. (1999). The continuum model: Ten years later. In S. Chaiken, & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 231–254). New York: Guilford Press.

Fiske, S. T., & Taylor, S. E. (2008). *Social cognition: From brains to culture*. New York: McGraw-Hill.

Frith, U., & Frith, C. (2001). The biological basis of social interaction. *Current Directions in Psychological Science*, *10*, 151–155.

Gallagher, H. L., & Frith, C. D. (2002). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences*, *7*, 77–83.

Harris, L. T. (2007). *Dehumanized perception fails to represent the contents of a social target's mind*. Unpublished PhD thesis, Princeton University, Princeton, NJ.

Harris, L. T., & Fiske, S. T. (2007). Social groups that elicit disgust are differentially processed in the mPFC. *Social Cognitive Affective Neuroscience*, *2*, 45–51.

Harris, L. T., Todorov, A., & Fiske, S. T. (2005). Attributions on the brain: Neuro-imaging dispositional inferences beyond theory of mind. *NeuroImage*, *28*, 763–769.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior, 16*, 321–338.

Heberlein, S. A., Adolphs, R., Tranel, D., Kemmerer, D., Anderson, S., & Damasio, A. R. (1998). Impaired attribution of social meanings to abstract dynamic geometric patterns following damage to the amygdala. *Society for Neuroscience Abstracts*, *24*, 1176.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*, 243–259.

Ickes, W., Patterson, M. L., Rajecki, D. W., & Tanford, S. (1982). Behavioral and cognitive consequences of reciprocal versus compensatory responses to preinteraction expectancies. *Social Cognition, 1*, 160–190.

Jones, A., & McGill, D. (1967). The homeostatic character of information drive in humans. *Journal of Experimental Research in Personality, 2*, 25–31.

Jones, E. E. (1979). The rocky road from acts to dispositions. *American Psychologist*, *34*, 107–117.

Karpinsky, A., & Hilton, J. L. (2001). Attitudes and the implicit associations test. *Journal of Personality and Social Psychology*, *81*, 774–788.

Kelley, H. H. (1972). Attribution in social interaction. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the cause of behavior* (pp. 1–26). Hillsdale, NJ: Lawrence Erlbaum Associates.

Knutson, B., Fong, G. W., Bennett, S. M., Adams, C. S., & Hommer, D. (2003). A region of medial prefrontal cortex tracks monetarily rewarding outcomes: Characterization with rapid event-related fMRI. *NeuroImage*, *18*, 263–272.

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology*, *93*, 234–249.

McArthur, L. A. (1972). The how and what of why: Some determinants and consequences of causal attribution. *Journal of Personality and Social Psychology*, *72*, 171–193.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, *306*, 503–507.

Montague, P. R., King-Casas, B., & Cohen, J. D. (2006). Imaging valuation models in human choice. *Annual Review of Neuroscience*, *29*, 417–448.

Neuberg, S. L., & Fiske, S. T. (1987). Motivational influences on impression formation: Outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology, 53*, 431–444.

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus–norepinephrine system. *Psychological Bulletin*, *131*, 510–532.

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, *12*, 265–272.

Phelps, E. A. (2006). Emotion, learning, and the brain: From classical conditioning to cultural bias. In P. Baltes, P. Reuter-Lorenz, & F. Rosler (Eds.), *Lifespan development and the brain: The perspective of biocultural co-constructivism* (pp. 200–216). New York: Cambridge University Press.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353–363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83*, 304–308.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382–407.

Richeson, J. A., & Shelton. J. N. (2005). Thin slices of racial bias. *Journal of Nonverbal Behavior, 29*, 75–86.

Russell, A. M., & Fiske, S. T. (2008). It's all relative: Social position and interpersonal perception. *European Journal of Social Psychology*, *38*, 1193–1201.

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

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Sears, D. O. (1983). The person-positivity bias. *Journal of Personality and Social Psychology*, *44*, 233–250.

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity and extremity biases. *Journal of Personality and Social Psychology*, *52*, 689–699.

Somerville, L. H., Heatherton, T. F., & Kelley, W. M. (2006). Anterior cingulate cortex responds differentially to expectancy violation and social rejection. *Nature Neuroscience*, *9*, 1007–1008.

Swann, W. B., & Snyder, M. (1980). On translating beliefs into action: Theories of ability and their application in an instructional setting. *Journal of Personality and Social Psychology, 36*, 879–888.

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

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

Van Duynslaeger, M., Sterken, C., Van Overwalle, F., & Verstraeten, E. (2008). EEG components of spontaneous trait inferences. *Social Neuroscience*, *3*, 164–177.

Wheeler, M. E., & Fiske, S. T. (2005). Controlling racial prejudice: Social-cognitive goals affect amygdala and stereotype activation. *Psychological Science*, *16*(1), 56–63.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin*, *127*, 797–826.

Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, *5*, 277–283.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology*, *81*, 815–827.

Word, C. O., Zanna, M. P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology*, *10*, 109–120.