Behavioral/Systems/Cognitive

Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain

Yael Niv,[1] Jeffrey A. Edlund,[2] Peter Dayan,[3] and John P. O'Doherty[2,4]

[1]Psychology Department and Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey 08540, [2]Division of Humanities and Social Sciences and Computation and Neural Systems Program, California Institute of Technology, Pasadena, California 91125, [3]Gatsby Computational Neuroscience Unit, University College London, London, WC1N 3AR, United Kingdom, and [4]Trinity College Institute of Neuroscience and School of Psychology, Trinity College Dublin, Dublin 2, Ireland

Humans and animals are exquisitely, though idiosyncratically, sensitive to risk or variance in the outcomes of their actions. Economic, psychological, and neural aspects of this are well studied when information about risk is provided explicitly. However, we must normally learn about outcomes from experience, through trial and error. Traditional models of such reinforcement learning focus on learning about the mean reward value of cues and ignore higher order moments such as variance. We used fMRI to test whether the neural correlates of human reinforcement learning are sensitive to experienced risk. Our analysis focused on anatomically delineated regions of a priori interest in the nucleus accumbens, where blood oxygenation level-dependent (BOLD) signals have been suggested as correlating with quantities derived from reinforcement learning. We first provide unbiased evidence that the raw BOLD signal in these regions corresponds closely to a reward prediction error. We then derive from this signal the learned values of cues that predict rewards of equal mean but different variance and show that these values are indeed modulated by experienced risk. Moreover, a close neurometric–psychometric coupling exists between the fluctuations of the experience-based evaluations of risky options that we measured neurally and the fluctuations in behavioral risk aversion. This suggests that risk sensitivity is integral to human learning, illuminating economic models of choice, neuroscientific models of affective learning, and the workings of the underlying neural mechanisms.

Introduction

As the recent economic downturn has shown, our attitude to risk during decision making can have major consequences. Sensitivity to risk (defined as the variance associated with an outcome) is a well-established phenomenon central to models of decision making in economics (Bernoulli, 1954), ethology (Kacelnik and Bateson, 1996), and neuroscience (Platt and Huettel, 2008). However, the mechanisms by which risk comes to influence the decision process are as yet unknown.

Much of our current understanding of the neural processes that give rise to risk sensitivity comes from studies that provided explicit information about the probabilities and magnitudes associated with a gamble (Elliott et al., 1999; Hsu et al., 2005, 2009; Huettel et al., 2005, 2006; Kuhnen and Knutson, 2005; Preuschoff et al., 2006, 2008; Tobler et al., 2007; Tom et al., 2007). However, we often do not have such knowledge a priori, but rather must learn about the different payoffs and their probabilities through experience (Hertwig et al., 2004; Hertwig and Erev, 2009). Although experiential decision making may depend on processes that are, at least partly, distinct from those invoked by explicit information (Jessup et al., 2008; Fitzgerald et al., 2010), the effects of risk on experiential learning have been less well studied. Indeed, traditional computational models of trial-and-error reinforcement learning (Sutton and Barto, 1998) that have proven invaluable in understanding the neural basis of human and animal experiential learning concentrate only on learning the mean (expected) value of different options, ignoring their variance.

Two interesting possibilities exist: Learning about means and variances may be separated in the brain, with the neural substrates of reinforcement learning concentrating only on learning means, as per the theory. Alternatively, reinforcement learning might itself be risk sensitive, due to nonlinear subjective utilities for outcomes (Bernoulli, 1954), nonlinear effects of unpredictable outcomes on the learning process (Mihatsch and Neuneier, 2002), or both.

Here, we concentrate on learning about, and choosing between, two stimuli associated with equal mean rewards but different variances: a "sure" stimulus associated with a 20¢ reward and a "risky" stimulus associated with 50% chance of receiving either 40¢ or 0¢. In our experiment, subjects were not told these payoffs, but had to experience them. Thus, we expected neural values of the stimuli to be learned through, and to fluctuate according to, an experiential reinforcement-learning process. In this situation, we asked whether or not these neural values are sensitive to the variance or risk associated with different options.

Received Oct. 20, 2010; revised Oct. 8, 2011; accepted Nov. 7, 2011.

Author contributions: Y.N., P.D., and J.P.O. designed research; Y.N. and J.A.E. performed research; Y.N., P.D., and J.P.O. analyzed data; Y.N., P.D., and J.P.O. wrote the paper.

This research was supported by a Hebrew University Rector's Fellowship, a Human Frontiers Science Program fellowship, and a Sloan Research Fellowship (Y.N.), the Gatsby Charitable Foundation (P.D.,Y.N.), the Gimbel Discovery Fund for Neuroscience (J.P.O.), and the Gordon and Betty Moore Foundation (J.P.O.). We are indebted to Laura deSouza for marking the anatomical ROIs and grateful to Timothy Behrens, Peter Bossaerts, Carlos Brody, Nathaniel Daw, Marie Monfils, Daniela Schiller, Geoffrey Schoenbaum, Ben Seymour, and John White for extremely helpful comments on previous versions of this manuscript.

Correspondence should be addressed to Yael Niv, Green Hall, Princeton University, Princeton, NJ 08540. E-mail: yael@princeton.edu.

DOI:10.1523/JNEUROSCI.5498-10.2012

 0270-6474/12/320551-12$15.00/0

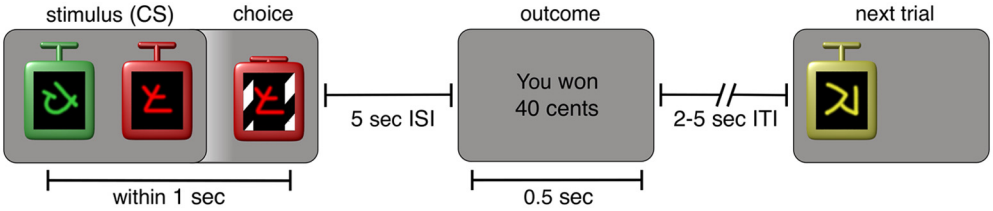

Figure 1. A reinforcement-learning task designed to assess the dynamic effects of risk on choice behavior and learning processes. In each trial, one or two slot machines differing in color and abstract symbol were presented on the left or right side of the screen. Subjects had to choose one of the two machines or indicate the location of the single machine. This triggered the rolling animation of the slot machine, followed by text describing the amount of money won.

We first show a close correspondence between fMRI blood oxygenation level-dependent (BOLD) signals in the anatomically delineated nucleus accumbens (NAC) and a hypothetical, model-derived reward prediction error signal. We then extract stimulus values from this signal and show a close neurometric–psychometric relationship between learned values and behavioral choice. This suggests that values learned through reinforcement learning are sensitive to risk and, furthermore, that these values affect decision making. Behavioral and neural analyses show that nonlinear utilities associated with sure outcomes do not fully account for this risk sensitivity, raising the possibility that the learning process is itself sensitive to variance.

Materials and Methods

Subjects

Sixteen subjects (four female; age, 18–34; mean, 24 years) participated in the study. An additional four subjects were scanned but were excluded from analysis, one due to color blindness and three due to failure to understand the task. Subjects first performed a practice session outside the scanner and then performed the task in the scanner. Subjects were remunerated according to the amount of money they won during the task (mean, $50.7; range, $42–$54; two subjects were also given a $5 compensation for showing up, which was later discontinued). All subjects gave informed consent and the study was approved by the Institutional Review Board of the California Institute of Technology.

Behavioral task

Subjects performed a mixed instrumental conditioning and classical conditioning task in which they chose between different visual stimuli displayed on a computer monitor to earn monetary rewards (Fig. 1). For each subject, five different-colored stimuli (portrayed as casino-style slot machines or bandits) were drawn randomly out of a total of six possible stimuli (yellow, orange, purple, green, red, or blue) and randomly allocated to five payoff schedules: sure 40¢, sure 20¢, two sure 0¢ stimuli, and one variable-payoff risky stimulus associated with equal probabilities of 0¢ or 40¢ payoffs (henceforth denoted the "0/40¢" stimulus). Subjects were never informed about the payoffs associated with the different stimuli, but had to estimate them from experience.

Two types of trials were intermixed randomly: choice trials and forced trials. In choice trials, two stimuli were displayed (left and right location randomized), and the subject was instructed to quickly select one of them by pressing either the left or right button on a button box. The chosen stimulus was then animated to emulate a spinning slot machine until 5 s had passed from the time of stimulus onset. The payoff associated with the chosen stimulus was then displayed for 500 ms. After a variable (uniformly distributed) intertrial interval of 2–5 s, the next trial began. In forced trials, only one stimulus was displayed on either the left or right side of the screen, and the subject had to indicate its location using the button box to cause the animation and obtain its associated outcome. Constant interstimulus intervals and variable intertrial intervals were chosen to allow for precisely timed predictions within a trial but not between trials. All button presses were timed out after 1 s, at which time the trial was aborted with a message indicating that the response was "too slow" and the intertrial interval commenced. Subjects were instructed to try to maximize their winnings and were paid according to their actual payoffs in the task.

Subjects were first familiarized with the task and provided with several observations of the stimulus–reward mapping in a training phase that included two subparts. The first part involved 10 randomly ordered forced trials (two per stimulus). The second part comprised 12 randomly ordered choice trials (two of each of six types of choice trials: 20¢ versus 0/40¢, 40¢ versus 0/40¢, 20¢ versus 40¢, 0¢ versus 0/40¢, 0¢ versus 20¢, and 0¢ versus 0¢). On-screen instructions for the task informed subjects that they would see five different slot machines, each associated with certain monetary rewards, and that they would play these machines with the goal of earning as much money as possible. They were also told explicitly that payoffs depended only on the slot machine chosen, not on its location or on their history of choices.

The experimental task was then performed inside an MRI scanner. On-screen instructions informed subjects that they would encounter the same stimuli as in the training phase. They were briefly reminded of the rules and encouraged to choose those machines that yielded the highest payoffs, as they would be paid their earnings in this part. The task consisted of 234 trials (three sessions of 78 trials each, with short breaks in between), with choice and forced trials randomly intermixed. The trials comprised (1) 30 "risk" choice trials involving a choice between the 20¢ stimulus and the 0/40¢ stimulus (from which we assessed subjects' behavioral risk sensitivity); (2) 20 "test" choice trials involving each of the pairs 40¢ versus 0/40¢, 20¢ versus 40¢, 0¢ versus 0/40¢, and 0¢ versus 20¢ (from which we assessed learning of the payoffs according to the frequency with which subjects chose the better option); (3) 24 forced trials involving each of the stimuli (16 only for each of the 0¢ stimuli), which were used to assess the neural representation of the value of the stimulus (see fMRI analysis below) uncontaminated by the presence of, or competition with, other stimuli; and (4) 20 trials in which subjects chose between the two 0¢ stimuli. These were originally designed as baseline trials; however, the analysis reported here does not make special use of them.

Trial order was randomized subject to the constraint that each half of the trials included exactly one half of all types of trials in the experiment. To minimize priming of responses in the critical trials that assessed risk sensitivity, trial order was further constrained such that a 20¢ versus 0/40¢ choice trial could not immediately follow a forced trial of one of its components. Payoffs for the 0/40¢ stimulus were counterbalanced in advance such that groups of eight consecutive payoffs included exactly four 40¢ payoffs ("wins") and four 0¢ payoffs ("losses"), and such that streaks of winning or losing were no longer than four. Although this preadjustment would make the "gambler's fallacy" true in the unlikely event that it was detected by the subjects, we used this counterbalancing method to eliminate between-subject differences due to potential variability in the actual experienced percentage of wins. All task events were controlled using Cogent (Wellcome Department of Imaging Neuroscience, London, UK).

Models of risk-sensitive choice

Temporal difference (TD) reinforcement learning (Sutton and Barto, 1998) offers a general framework for understanding trial-and-error learning and decision making in circumstances like this, and also for linking choice behavior to neurophysiological and fMRI signals (Barto,

1995; Montague et al., 1996; Schultz et al., 1997; O'Doherty et al., 2004; Li et al., 2006; Schönberg et al., 2007). Reinforcement learning postulates that subjects use past experience to estimate values for the different stimuli [denoted $V(S)$] corresponding to the expected payoffs to which these stimuli lead and, given a choice, pick between stimuli based on their values.

Three variants of TD learning that can give rise to risk-sensitive behavior are illustrated in Figure 3.

TD model. The TD model is a standard temporal difference learning model (Barto, 1995; Sutton, 1988; Sutton and Barto, 1998). In this model, a prediction error $\delta(t) = r(t) + V(t) - V(t-1)$ is computed at each of two consecutive time steps (t_{stimulus} and $t_{\text{outcome}} = t_{\text{stimulus}} + 1$), where $V(t)$ is the predicted value at time t, and $r(t)$ is the reward at time t. The prediction error at t_{outcome} is used to update $V(C)$, the value of the chosen stimulus, according to $V^{\text{new}}(C) = V^{\text{old}}(C) + \eta \cdot \delta(t_{\text{outcome}})$, with η being a learning rate or step-size parameter.

Utility model. The utility model is a TD learning model that incorporates nonlinear subjective utilities (Bernoulli, 1954) for the different amounts of reward. In this model, the prediction error is $\delta(t) = U(r(t)) + V(t) - V(t-1)$, where $U(r(t))$ is the subjective utility of the reward at time t, and the update rule is the same as in the TD model above: $V^{\text{new}}(C) = V^{\text{old}}(C) + \eta\, \delta(t_{\text{outcome}})$. To model the subjective utility of the reward, we assumed (without loss of generality) that $U(0) = 0$, $U(20) = 20$, and $U(40) = a \cdot 20$. Values of a that are smaller than 2 are thus consistent with a concave subjective utility curve, and $a > 2$ is consistent with a convex utility curve.

Risk-sensitive TD model. In the risk-sensitive TD (RSTD) model, positive and negative prediction errors have asymmetric effects on learning (Mihatsch and Neuneier, 2002). Prediction errors in this model are similar to those in the TD model above, $\delta(t) = r(t) + V(t) - V(t-1)$; however, there are separate update rules for positive and negative prediction errors [as in the study by Shapiro et al. (2001)]:

$$V^{\text{new}}(C) = \begin{cases} V^{\text{old}}(C) + \eta^{+} \cdot \delta(t_{\text{outcome}}) & \text{if } \delta(t_{\text{outcome}}) > 0, \\ V^{\text{old}}(C) + \eta^{-} \cdot \delta(t_{\text{outcome}}) & \text{if } \delta(t_{\text{outcome}}) < 0, \end{cases} \tag{1}$$

such that if $\eta^{+} < \eta^{-}$, the effect of negative prediction errors on learned values is larger than that of positive errors, leading to risk aversion, and vice versa if $\eta^{+} > \eta^{-}$.

Using a trial-based (finite horizon) reinforcement-learning scheme, we modeled prediction errors at two time points in each trial: the time at which the stimulus is presented (t_{stimulus}) and the time at which the outcome is delivered ($t_{\text{outcome}} = t_{\text{stimulus}} + 1$). At stimulus onset, in the absence of reward and assuming that the prior expectation $V(t_{\text{stimulus}} - 1)$ is zero (a common assumption based on recordings from dopaminergic neurons, which show that the neural activity at stimulus onset reflects a prediction error comparing the value of that stimulus to zero, rather than to the average reward obtainable in the experiment; Fiorillo et al., 2003; Tobler et al., 2005), the prediction error is $\delta(t) = V(t_{\text{stimulus}})$, that is, the predicted reward for this trial, which is the value of the chosen stimulus $V(C)$. At outcome onset, $r(t_{\text{outcome}})$ depends on the actual outcome presented to the subject, and we assumed that $V(t_{\text{outcome}})$, the expected reward in the intertrial interval, is zero. Thus, the prediction error $\delta(t_{\text{outcome}}) = r(t_{\text{outcome}}) - V(t_{\text{outcome}} - 1) = r(t_{\text{outcome}}) - V(C)$ is the difference between the obtained and the expected outcome [as in the study by Rescorla and Wagner (1972)].

Additionally, for all three models we assumed a softmax action selection function:

$$p(A) = \frac{e^{\beta V(A)}}{e^{\beta V(A)} + e^{\beta V(B)}} = \frac{1}{1 + e^{-\beta(V(A) - V(B))}}, \tag{2}$$

where $P(A)$ is the probability of choosing stimulus A, and β is an inverse temperature parameter.

Model fitting and model comparison

We used subjects' behavioral data from both training and test sessions to fit the free parameters of the three models: η and β for the TD model; η, a, and β for the utility model; and η^{+}, η^{-}, and β for the RSTD model. Model likelihoods were based on assigning probabilities to the 142 choice trials for each subject, according to the softmax function (Eq. 2). We modeled learning of values in all 256 training and test trials; thus, forced trials also contributed to the model likelihoods, albeit indirectly.

As we were interested specifically in intersubject differences, we did not pool data across subjects, but rather fit each subject's parameters separately. To facilitate this in the face of the multiplicative interaction between learning rates and softmax inverse temperatures in the model, we used regularizing priors that favored realistic values and maximum a posteriori (rather than maximum likelihood) fitting (Daw, 2011). This prevented any unreasonable individual fits. Thus, learning rates were constrained to the range $0 \leq \eta \leq 1$ with a Beta (2,2) prior distribution slightly favoring values that were in the middle of this range, and the inverse temperature was constrained to be positive with a Gamma (2,3) prior distribution that favored lower values. Additionally, the utility parameter was constrained to the range $1 \leq a \leq 30$ (with a uniform prior). Priors were implemented by adding to the data log likelihood, for each setting of the parameters, the log probability of the parameters given the prior distribution. Stimulus values were initialized to 0. Under a standard Kalman filter scheme, it would be optimal for learning rates to start high and decay as experience with the stimuli accumulates (Dayan et al., 2000). We approximated this effect by adding a decaying learning rate $\eta_d = 0.5/(1 + T_s)$ (with T_s being the number of trials in which the consequences of stimulus S had been experienced) to the constant (fit) learning rate(s). This predominantly affected learning in the training trials. [We could not simply fit different learning rates for the training and the test trials due to the small number of training trials; however, fitting the three models to test data only while initializing stimulus values to their correct values [0,0,20,40,20] (or, in the case of the utility model, their nonlinear counterparts) and eschewing the decaying learning rate, gave similar results.] We optimized model parameters by minimizing the negative log posterior of the data given different settings of the model parameters using the Matlab function fminunc. This function performs unconstrained linear optimization using a subspace trust-region method that approximates the objective function using a Taylor expansion around the current point and minimizes it within a subspace around this point. We used the "large-scale optimization" algorithm of fminunc; that is, we provided fminunc with a custom-built routine that computed both the negative log posterior probability of the data given any setting of the parameters and its analytical partial derivatives with respect to each of the parameters, evaluated at the current setting of the parameters. To facilitate finding the global maximum of the log posterior, we ran the routine multiple times for each subject, starting each run from random initial values for the parameters, and kept track of the highest value of the log posterior achieved.

We assessed the contribution of the extra parameters in the utility and RSTD models to the overall data likelihood using a likelihood ratio test for nested models and the maximum a posteriori parameters for each model. In this, twice the difference in log likelihoods of the RSTD or utility model and the TD model was compared to a χ^2 statistic with one degree of freedom (the number of extra parameters in the more complex model). Comparisons between the RSTD and utility models, which had the same number of parameters, were performed directly on the posterior probability of the models evaluated using the maximum a posteriori parameters.

fMRI data acquisition

Functional brain images were acquired using a Siemens 3.0 Tesla Trio MRI scanner. Gradient echo T2*-weighted echoplanar images (EPIs) with BOLD contrast were acquired at an oblique orientation of 30° to the anterior–posterior commissure line, using an eight-channel phased array coil. Each volume comprised 32 axial slices. Volumes were collected during the experiment in an interleaved ascending manner, with the following imaging parameters: echo time, 30 ms; field of view, 192 mm; in-plane resolution and slice thickness, 3 mm; repetition time, 2s. Whole-brain high-resolution T1-weighted structural scans ($1 \times 1 \times 1$ mm) were also acquired for all subjects and coregistered with their mean EPIs. These were used to map anatomical regions of interest in the right

and left nucleus accumbens for each subject, and were averaged together to permit anatomical localization of functional activations at the group level.

Preprocessing of the images and whole-brain image analysis were performed using SPM2 (statistical parametric mapping software; Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK; http://www.fil.ion.ucl.ac.uk/spm/spm2.html). Preprocessing of EPIs included temporal normalization (adjusting each slice to the middle of the scan), subject motion correction (rigid body realignment of all images to the first volume), and spatial normalization to a standard T2* template in Montreal Neurological Institute space. Anatomical images were also normalized to the same template, and anatomical regions of interest (ROIs) were marked for each subject using Analysis of Functional NeuroImages (Cox, 1996). Whole-brain images were then further preprocessed by spatially smoothing the images using a Gaussian kernel (with a full-width at half-maximum of 8 mm) to allow for statistical parametric mapping analysis.

ROI analysis

A wealth of evidence suggests that prediction error signals based on a diverse array of rewards can be seen in BOLD fMRI activity in areas such as the NAC in the ventral striatum (Breiter et al., 2001; McClure et al., 2003; O'Doherty et al., 2003, 2004; Abler et al., 2006; Li et al., 2006; Preuschoff et al., 2006; Hare et al., 2008). These prediction error signals have a precise computational interpretation as the momentary differences between expected and obtained outcomes (Niv and Schoenbaum, 2008). Prediction errors are not the same as value signals—value signals are prolonged, in our case spanning the interval between cue onset (which gives rise to the expected value) and receipt of the outcome, whereas prediction errors correspond to momentary differences in expected value (Niv and Schoenbaum, 2008). However, given a tailored experimental design such as we use here, neural values of cues [in our case, the value of a chosen stimulus $V(C)$] can be extracted from prediction error signals at the time of cue onset. We thus chose a priori to concentrate on BOLD signals in the NAC to compare different models for learning the values of sure versus risky stimuli.

The NAC was anatomically defined as the area bordered ventrally by the caudate nucleus, dorsally by the anterior commisure, laterally by the globus pallidus and putamen, and medially by the septum pellucidum. The border with the caudate was taken to be at the bottom of the lateral ventricle, and with the putamen at the thinnest part of gray matter. We considered the anteriormost border to be at the axial slice in which the caudate and putamen fully separated and the posterior border to be where the anterior commisure was fully attached between hemispheres. Voxels in functional space were taken that were wholly within the NAC as delineated in the higher-resolution anatomical space. This resulted in means of 36.75 and 40.62 voxels for the left and right NAC ROIs, respectively (ranges, 24–52 and 26–66 voxels, respectively). Data were then extracted for each of the two anatomical ROI and averaged per ROI using singular value decomposition. This resulted in two time course vectors of BOLD activity (with samples every 2 s) for each subject on which all additional ROI analyses were conducted. Parallel analyses performed on more inclusive ROIs that also incorporated voxels that were only partly within the NAC (means, 58.56 and 59.75; ranges, 37–82 and 43–79 voxels for left and right NAC ROIs, respectively) showed similar results and are not reported.

To analyze the ROI time courses, we first removed effects of no interest due to scanner drift and subject motion from the time course activity by estimating and subtracting from the data, for each session separately, a linear regression model that included six motion regressors (3D translation and rotation), two trend regressors (linear and quadratic), and a baseline. To compare the remaining signal to a theoretical model-derived prediction error signal, we then upsampled the raw, corrected time courses to 100 ms resolution using spline interpolation and averaged the time courses across the bilateral ROIs, all subjects, and similar trials (upsampling was done for the purpose of this comparison only; all subsequent analyses were performed on the original time course data). To verify statistically that the signal in each of the two ROIs corresponds to a prediction error signal $\delta(t) = V(t) + r(t) - V(t-1)$, we regressed against each of the ROI time courses a linear model that included in three different regressors the three necessary components of the prediction error in this one-step task: $V(t_{\text{stimulus}})$ at stimulus onset [which is $V(C)$, the current value of the chosen stimulus], $r(t_{\text{outcome}})$ at reward onset, and $-V(t_{\text{outcome}} - 1)$ [which is $-V(C)$] at reward onset. To conclude that the BOLD activity corresponds to a prediction error signal, we required that the ROI time series be significantly correlated with each of these three regressors at $p < 0.05$ across subjects and that these correlations be positive. Moreover, we tested for equality of the regression coefficients for $r(t_{\text{outcome}})$ and $-V(t_{\text{outcome}} - 1)$ using the method suggested by Rouder et al. (2009); that is, we computed Bayes factors measuring the odds for the hypothesis that these regressors are equal compared to the hypothesis that they are unequal, using both an uninformative and a unit-informative prior on the alternative hypothesis. When computed in this manner, Bayes factors >1 support the null hypothesis over the alternative, with the commonly used cutoff for support for one hypothesis versus another being at least three (Rouder et al., 2009). We did not test for similar equality of the regressor coefficients of $V(t_{\text{stimulus}})$ and $-V(t_{\text{outcome}} - 1)$, as these may differ due to discounting of future rewards, the degree of which we could not estimate using the current fixed-interstimulus-interval design.

We then used the ROI time courses to estimate the value of each stimulus. For this we estimated the mean BOLD response corresponding to the time of stimulus onset in forced trials by estimating two additional regression models. These designs included separate stimulus-onset stick-function regressors for each of the four types of forced trials (collapsing across the two 0¢ stimuli). In the first design, the onset regressors were modeled across the whole experiment (four regressors in total). In the second design, they were modeled for each session separately (12 regressors in total). Due to large between-subject differences in the initial experience of the risky 0/40¢ stimulus, the first two forced trials of this stimulus were not included in the stick regressors. [Including these trials does not change any of our results; however, with these trials included, the TD model also predicts a weakly significant relationship between behavioral risk aversion and the difference in values of the 20 and 0/40 stimuli, which is absent otherwise (supplemental Fig. 2, available at www.jneurosci.org as supplemental material).] To deconvolve from these signals of interest other NAC activations due to outcome presentation and due to the onset of choice trials, we also modeled temporal difference errors for all other events in a separate regressor (or in three regressors, one for each session, in the second design). To generate the prediction error for each subject at these times, we used the subject's actual experienced choices and rewards and maximum a posteriori parameters for the RSTD model (results using the TD or utility model instead were similar and are not shown). Following the convention in the literature (O'Doherty et al., 2004; Daw et al., 2006; Schönberg et al., 2007), we modeled the prediction error at stimulus onset as that corresponding to the to-be-chosen stimulus. All regressors were created as covariate regressors and convolved with a canonical hemodynamic response function before being entered into the design matrix. The β values for the stick-function regressors for each subject were then treated as estimates of the mean value of each stimulus in each session, and differences in values were correlated with behavioral risk sensitivity as measured by the proportion of choices of the 20¢ stimulus in the 20¢ versus 0/40¢ choice trials. All ROI model fitting and statistical analysis was done using software written in Matlab (MathWorks).

Whole-brain analysis

A supplemental whole-brain image analysis was performed using SPM2. In this, we searched for brain areas in which BOLD activity correlates with a prediction error signal. The design matrix comprised, for each of the three sessions, a regressor for prediction errors, two stick-function regressors for stimulus onsets and for outcome onsets, and nuisance covariate regressors for motion, linear and quadratic drift, and baseline. The prediction error regressor was created as a covariate regressor by convolving the punctate prediction errors as modeled via the RSTD model (using the maximum a posteriori parameters for each subject) with the canonical hemodynamic response function. Other stick regressors were convolved with the canonical hemodynamic response function

as is usual in SPM2. The six scan-to-scan motion parameters produced during realignment were used as nuisance motion regressors to account for residual effects of movement. This design matrix was entered into a regression analysis of the fMRI data of each subject. A linear contrast of regressor coefficients was then computed at the single-subject level for the temporal difference regressor. The results were analyzed as random effects at a second, between-subjects level by including the contrast images from each subject in a one-way ANOVA with no mean term.

Group-level activations were localized using the group-averaged structural scan, and functional ROIs were marked on the group-level statistical parametric map using xjView (http://www.alivelearn.net/xjview). We analyzed activations at a whole-brain familywise error-corrected threshold of $p < 0.05$ (corresponding to $t_{(15)} > 7.38$). For each of the five functional ROIs identified, we extracted time course activity for every subject, averaging over voxels within the ROI using singular value decomposition. As in the NAC analysis above, we first removed effects of no interest due to scanner drift and subject motion from the time course activity in each ROI by estimating and subtracting from the data, for each session separately, a linear regression model that included six motion regressors (3D translation and rotation), two trend regressors (linear and quadratic), and a baseline. We then tested whether each functional ROI corresponded to a prediction error signal using the same analysis applied to the anatomical ROI time course data. To conclude that BOLD activity in an ROI corresponded to a prediction error signal, we required that the ROI time series be significantly correlated with each of the three component regressors of a prediction error signal at $p < 0.05$ across subjects.

Results

We scanned human subjects using fMRI while they performed a learning and decision making task in which they chose between different stimuli to earn monetary rewards. Four stimuli were associated with fixed payoffs of 0¢, 0¢, 20¢, and 40¢, and one stimulus was associated with a risky 50% probability of receiving either 0¢ or 40¢ (called the 0/40¢ stimulus). Subjects were not informed about these payoffs, but had to learn them from experience. The task consisted of choice trials, in which subjects chose one of two stimuli with the aim of maximizing monetary earnings, and forced trials, in which only one stimulus was available and subjects earned its associated payoff (Fig. 1). We were especially interested in learning about, and choices between, the 20¢ and 0/40¢ stimuli, which had the same mean payoff but different variances, or risk.

Behavior

We first examined performance in choice trials in which one of the two options is objectively better than the other one (40¢ vs 0/40¢, 20¢ vs 40¢, 0¢ vs 0/40¢, and 0¢ vs 20¢). This showed that subjects had learned the payoffs associated with the different stimuli such that they could make the correct choice on most trials (78% correct choices in the first test session, 95% in the final test session; $t_{(15)} = 4.66$; $p < 0.0005$, paired one-tailed Student's t test). Individual learning curves further confirmed that all 16 subjects had mastered the task (supplemental Fig. 1, available at www.jneurosci.org as supplemental material).

In contrast to trials in which one stimulus was objectively better than the other, our main focus was on understanding those choices in which there was no "correct" answer, namely, the choices between the sure 20¢ and the risky 0/40¢ stimuli, and how these were influenced by fluctuating experience-based estimations of the value of the risky stimulus. Figure 2 shows each subject's preference in these risky choice trials for the sure 20¢ option in the three sessions. As expected (Huettel et al., 2006; Hayden and Platt, 2008), risk sensitivity varied widely among the subjects, with the majority preferring the sure option and only a few preferring the risky option. Moreover, risk preference was not stationary throughout the experimental sessions, even within individual subjects. We thus set out to explain this behavioral variability both computationally and neurally, by contrasting three possible explanations for risk-sensitive choice.

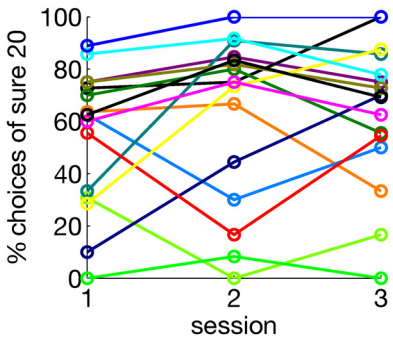

Figure 2. Risk sensitivity varied between subjects and across sessions within subjects. The percent choice of the sure 20¢ stimulus over the risky 0/40¢ stimulus in each of three experimental sessions is plotted for each subject (~10 choices per subject per session).

Model fits

To understand subjects' choice behavior, we used the framework of reinforcement learning (Sutton and Barto, 1998) that has been widely used to model trial-and-error learning and decision making (Niv, 2009). Basic reinforcement-learning models, and specifically temporal difference learning, concentrate on estimating only the mean outcome for each stimulus and do not explicitly take risk into account. However, risk-sensitive variants have been suggested. We compared three variants of TD learning, diagramed in Figure 3.

The first model we considered, basic TD learning, does not track variance. However, risk aversion arises implicitly in this model from biased sampling due to the interaction between choice and learning (March, 1996; Denrell and March, 2001; Niv et al., 2002; Hertwig et al., 2004; Denrell, 2007). This is illustrated in Figure 3*a* and arises because risky stimuli are, by definition, associated with outcomes that are larger or smaller than their means. Learned predictive values will thus fluctuate above and below the mean according to the specific sequence of past experienced rewards. When the value fluctuates below the mean, the risky option will be chosen less frequently, and thus its value will not be updated. As a result, the value of the risky option will be lower than its actual mean payoff more often than it will be higher than the mean payoff, and choices of the risky option will, on average, be suppressed. The higher the learning rate η, the larger the fluctuations in the estimated value, and so the greater the risk aversion (Niv et al., 2002). This was borne out by the model fits: across subjects, the fit learning rate tended to be higher for more risk-averse subjects (Fig. 3*d*; Pearson's correlation, $r = 0.38$). Thus, even the basic, risk-neutral, TD model might account for risk sensitivity, although we note that basic TD learning is, a priori, an unlikely explanation for behavior in our task because it cannot generate risk-seeking behavior, and because, by design, the effects of these biases were mitigated by the forced trials.

The second model we considered, the utility model (Fig. 3*b*), is the standard explanation for risk sensitivity from economics. This involves convex or concave subjective utilities for the outcomes (Bernoulli, 1954) that are incorporated into TD learning. Here, as in the previous model, outcome variance is not explicitly

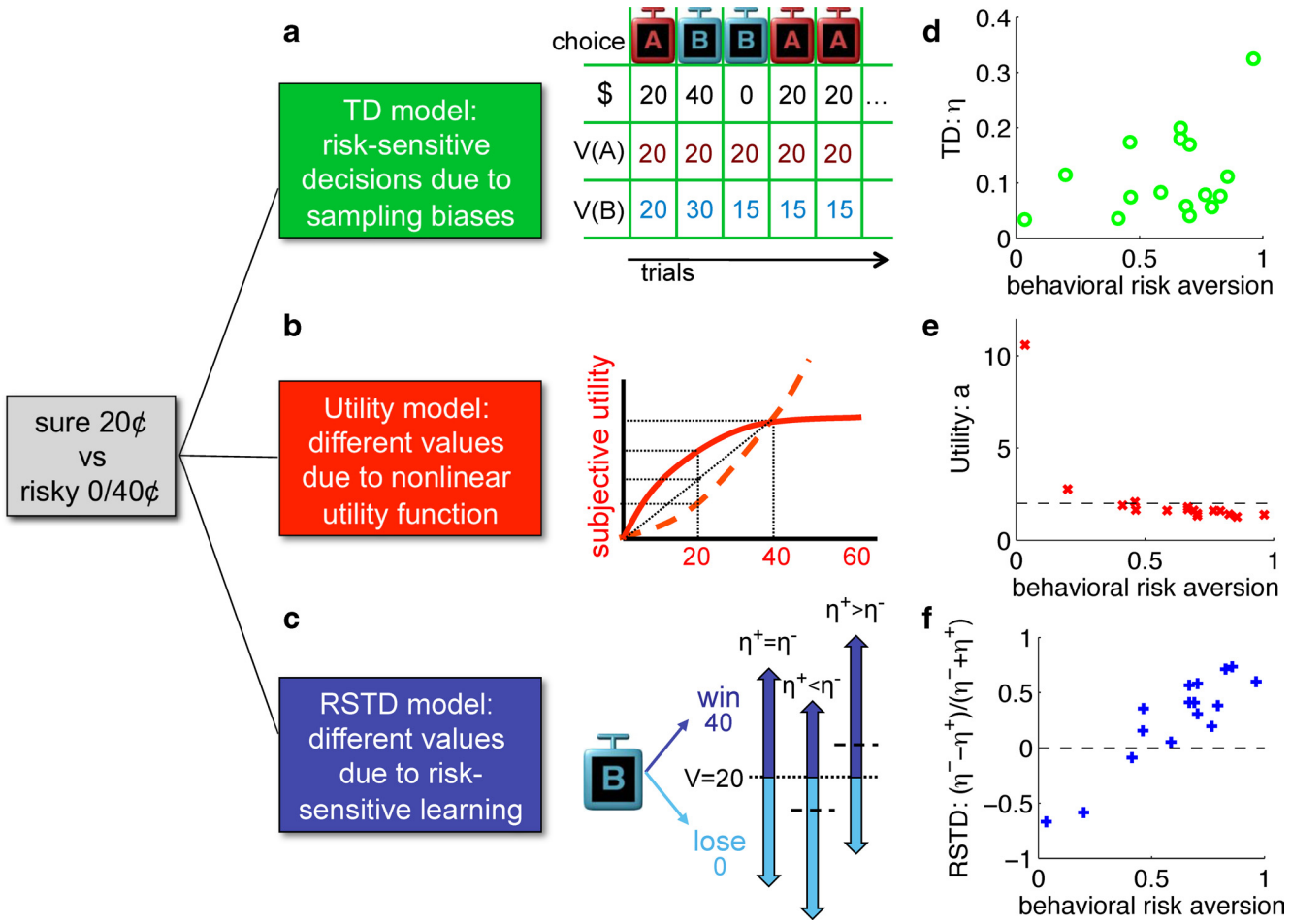

Figure 3. Three qualitatively different models for risk-sensitive choice. ***a***, TD model: biased sampling that tends to choose the better of two options implies that fluctuations in the prediction of risky options below their mean persist longer than fluctuations above their mean. Simulation of five choice trials, starting from equal predictions $V(A) = V(B) = 20$, with $\eta = 0.5$, and preferential choice of the stimulus with the higher predicted value are shown. The sure option, *A*, dominates. ***b***, Utility model: concave (solid) or convex (dashed) nonlinear subjective utility functions for different monetary rewards lead to risk-averse or risk-seeking behavior, respectively. ***c***, RSTD model: positive prediction errors $\delta > 0$ are weighted by η^+ during learning, while negative prediction errors are weighted by η^-. Whereas symmetric weighting ($\eta^+ = \eta^-$) results in an average predictive value of 20 for the 0/40¢ stimulus (left), losses loom larger if $\eta^+ < \eta^-$, leading to a value that is on average smaller than 20, and consequently to risk aversion (middle), and conversely when for $\eta^+ > \eta^-$ (right). ***d–f***, Parameter fits for three different models that can potentially explain the subjects' risk-sensitive choices. Each subject's best-fit parameter values are plotted against the fraction of their choices of the 20¢ stimulus over the alternative 0/40¢ stimulus throughout the whole experiment. ***d***, In the classic TD model, higher learning rates (η) account for more risk aversion (Niv et al., 2002). ***e***, In the utility model, risk-neutral subjects have values of *a* near 2 (dashed horizontal line), implying a rather linear utility function for this range of monetary rewards; *a* is lower than 2 for risk-averse subjects (consonant with a concave utility function), and higher than 2 for risk seekers (consonant with a convex utility function). ***f***, In the RSTD model, the normalized difference between η^- and η^+ is strongly correlated with risk sensitivity (Mihatsch and Neuneier, 2002).

taken into account. Instead, risk-sensitive preferences emerge because the two options that have objectively been designated as having the same mean payoff do not lead to an equal subjective mean reward. We modeled subject-specific nonlinearities using a parameter *a* as the ratio of the subjective utilities of 40¢ and 20¢ [$U(40) = a \cdot U(20)$; see Materials and Methods]. As expected, subject-specific parameter fits resulted in an overall monotonically decreasing relationship between *a* and behavioral risk aversion across subjects (Fig. 3*e*; Spearman's rank correlation, $\rho = -0.91$; $p < 10^{-6}$, one-tailed Student's *t* test).

In the third model (risk-sensitive TD learning; Fig. 3*c*), risk sensitivity arises from an explicitly risk-sensitive variant of TD learning that penalizes or favors outcome variance through asymmetric learning from positive and negative prediction errors (Shapiro et al., 2001; Mihatsch and Neuneier, 2002). If negative errors have the effect of decreasing V^{new} (stimulus) more than positive errors increase it, then the learned value will be lower than the mean nominal outcome, leading to risk aversion. Conversely, higher learning rates for positive compared to negative prediction errors will result in overestimation of the values of risky options and thus to risk seeking. This underestimation or overestimation of values only occurs for risky options, as these are, by definition, associated with persistent positive and negative prediction errors. We correlated the normalized difference between the two fit learning rates for each subject $(\eta^- - \eta^+)/(\eta^- + \eta^+)$ with behavioral risk aversion across subjects (Fig. 3*f*). As expected, this correlation was highly significant (Pearson's correlation, $r = 0.90$; $p < 10^{-6}$, one-tailed Student's *t* test).

Model comparison

We compared the posterior likelihoods of each subject's choice data according to each of the three models to evaluate which model provides the best fit for each subject. Likelihood ratio tests to assess the contribution of the extra parameter in the more complex RSTD or utility models (compared to the simpler TD model) showed that the additional parameter was justified in 14

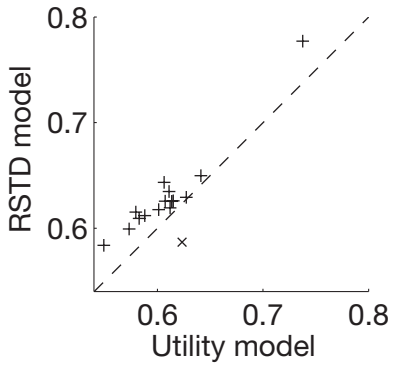

Figure 4. Direct comparison of the posterior probability per choice trial for the utility and the RSTD models shows that the RSTD model assigns a higher average probability per trial to the choices of 15 out of 16 subjects. The average probability assigned to a choice trial for each subject is the likelihood of the data divided by the number of choice trials. This provides a measure of the choice variance explained by the model (chance, 0.5). The one subject for which the utility model assigned a higher probability per choice is denoted by ×.

of 16 subjects for the RSTD model, but only in 6 of 16 for the utility model (χ^2 test with one degree of freedom, $p < 0.05$). As the utility and RSTD models have the same number of parameters, we compared their posterior likelihoods directly. This showed that the RSTD model provided a better fit that the utility model for all but one subject (Fig. 4).

Prediction error signaling in the nucleus accumbens

The above results showed that the behavioral data strongly favored the RSTD model. Another way to test which model best aligns with human learning in this task is by comparing the predictions of the three models to neural measurements of the learned values of the stimuli. Here, the three models make qualitatively different predictions (supplemental Fig. 2, available at www.jneurosci.org as supplemental material): The TD model predicts that in forced trials, in which choice biases are eliminated, the values of the 20¢ stimulus and the 0/40¢ stimulus would be similar. According to the utility model, if $a \neq 2$, the values of these two stimuli will be unequal even if only forced trials are taken into account. Moreover, the utility model predicts that nonlinearity in the values of the sure 40¢ and 20¢ stimuli should match the subjects' risk-sensitive behavior: risk-averse subjects should exhibit concavity [$2\,V(20) > V(40)$], and risk-seeking subjects convexity [$2\,V(20) < V(40)$]. Finally, the RSTD model also predicts a lawful relationship between the difference $V(20) - V(0/40)$ and risk-sensitive choice. However, unlike the utility model, according to the RSTD model, the evaluation of deterministic options should still be linear in the nominal outcomes, since they do not induce persistent prediction errors.

To test these predictions, we took advantage of the fact that under all three models the prediction error at the time of stimulus onset theoretically corresponds to the (learned) value of the stimulus. Thus, we could use fMRI BOLD signals shown previously to correlate with reward prediction errors to extract from activations at the time of stimulus onset the neural values of the different stimuli and determine which model is best supported by the neural data.

Numerous previous fMRI studies employing whole-brain analysis methods have shown that BOLD activity in the ventral striatum of humans engaged in classical or instrumental conditioning is correlated with prediction errors (McClure et al., 2003; O'Doherty et al., 2003, 2004; Abler et al., 2006; Li et al., 2006; Preuschoff et al., 2006; Tobler et al., 2006; Seymour et al., 2007), and that when monetary rewards are used, this signal is more specifically located in the nucleus accumbens (Kuhnen and Knutson, 2005; Knutson et al., 2005; Daw et al., 2006; Kim et al., 2006; Schönberg et al., 2007; Hare et al., 2008). Therefore, to determine prediction errors (and thence from them to extract the neural values of stimuli), we defined two a priori ROIs, one in the left and one in the right nucleus accumbens. Given the variable size of these nuclei across subjects and the unreliability of standard normalization of brain images when dealing with subcortical nuclei (D'Ardenne et al., 2008), we delineated these structures separately for each subject using anatomical criteria (see Materials and Methods). Indeed, individual ROIs in normalized Montreal Neurological Institute coordinates showed relatively weak between-subject overlap (Fig. 5*a*; supplemental Fig. 3, available at www.jneurosci.org as supplemental material).

We first assessed qualitatively whether the BOLD signal in these ROIs indeed reflected a temporal difference prediction error signal. For this we aligned raw BOLD signal time courses to the onset of the stimuli, correcting only for scanner drift, mean baseline, and motion artifacts (see Materials and Methods). Figure 5*b* shows the time course of the activations averaged over all subjects and both ROIs, and divided according to the chosen stimulus (the 0/40¢ trials were further divided according to the actual outcome on these trials; the two stimuli predicting 0¢ were combined). As a point of comparison, we considered what form these signals would take if they represented the theoretical prediction errors in our task. In all three models, prediction errors occur at two time points in each trial (Fig. 5*c*): at the time of the stimulus, corresponding to the expected value of the chosen option, and at the time of the outcome, corresponding to the difference between the reward and this expectation. Since we expected the BOLD signal to reflect subject-specific prediction errors during learning, we used the fit parameters for each subject in conjunction with his or her own sequence of choices and payoffs to create a personalized prediction error sequence. By convolving these prediction errors with the canonical hemodynamic response function (Fig. 5*d*) and averaging across subjects and trials, we obtained the expected signature for a grand average prediction error BOLD signal (Fig. 5*e*).

Visual comparison of Figure 5, *b* and *e*, shows that the averaged responses in the NAC ROIs were extremely similar to the activations we expected to see based on the TD prediction error, in both general form and ordinal relations. Notably, BOLD signals associated with stimulus onset roughly corresponded to the value of the chosen option, and in trials in which a sure option was chosen, no additional BOLD response was seen at the presentation of the (expected) payoff. For the 0/40¢ stimulus, however, a second BOLD response was above or below baseline in correspondence with positive and negative prediction errors due to the 40¢ or 0¢ outcomes, respectively. We also found that the activations in the left and right ROIs were very similar (supplemental Fig. 4, available at www.jneurosci.org as supplemental material). The one qualitative discrepancy between predicted and actual BOLD responses was a positive response associated with the onset of the sure 0¢ stimulus and a negative response at the time of the 0¢ outcome (yellow). This response pattern may have resulted from confusion or generalization between stimuli at the beginning of the experiment, as it disappeared after the first session (supplemental Fig. 5, available at www.jneurosci.org as supplemental material).

We verified these results statistically by modeling the time course in each ROI as a linear combination of the three necessary components of a prediction error signal, $\delta(t) = V(t) + r(t) -$

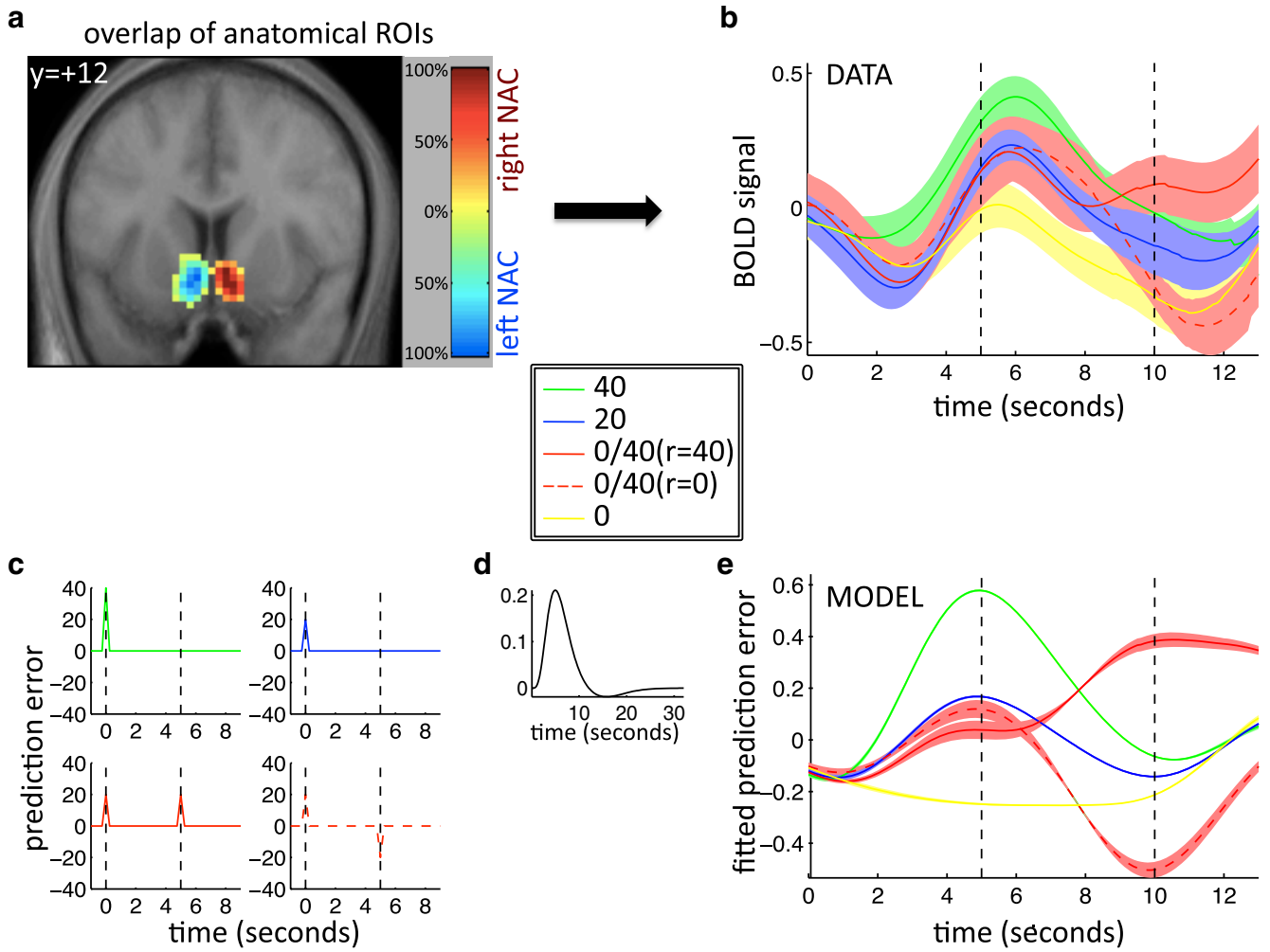

Figure 5. ***a***, Overlap between anatomical ROIs depicted on the average anatomical image of the subjects. Darker red (right NAC) and blue (left NAC) denote a higher degree of overlap between the ROIs of different subjects. ***b***, The raw BOLD signal (in arbitrary units) as extracted from the ROIs in ***a***, aligned on trial onset, averaged over all subjects and all trials, and separated according to chosen stimulus and payoff. Shading corresponds to the SEM for each trace. Compare with ***e***. ***c***, Illustration of the theoretical asymptotic prediction errors for each stimulus. Stimuli (presented at time $t = 0$) appear unpredictably and so induce prediction errors approximately equal to their mean values (40 and 20 for the sure stimuli and 20 for risky stimulus, ignoring subject-specific risk-related perturbations for purposes of illustration only). For the sure stimuli, the outcomes (at $t = 5$) are fully predicted, and thus induce no further prediction errors. For the 0/40¢ stimulus, the 40¢ outcome induces a positive prediction error (red solid line) of ~20; the 0¢ outcome induces a negative prediction error of approximately (−20) (red dashed line). The 0¢ stimulus (data not shown) is not expected to generate a prediction error at time of stimulus onset or payoff. ***d***, The canonical hemodynamic response function. ***e***, Model prediction errors at the time of the stimulus and outcome for every subject for each condition (using individual best-fit parameters for the RSTD model) were convolved with the hemodynamic response function (***d***) and averaged to predict the grand average BOLD signal. The hemodynamic lag adds 5 s to the times to peak (dashed black lines). The yellow trace corresponds to the 0¢ stimuli and is below baseline due to the residual dip from the hemodynamic response in the previous trial.

$V(t-1)$, in our task: (1) $V(t)$, the value of the chosen stimulus at time of stimulus onset, (2) $r(t)$ the magnitude of the reward at time of outcome, and (3) $-V(t-1)$ the negative of the expected value at time of outcome. All three regressors were significant at the $p < 0.05$ level in each of the ROIs (right NAC, value at stimulus, $p = 0.003$; reward at outcome, $p = 0.030$; negative value at outcome, $p = 0.003$; left NAC, value at stimulus, $p = 0.003$; reward at outcome, $p = 0.018$; negative value at outcome, $p = 0.026$), with positive average regression coefficients in all cases. Moreover, as predicted by the fact that at outcome the prediction error is $\delta(t_{outcome}) = r(t_{outcome}) - V(t_{outcome})$, there was a strong correlation across subjects between the regression coefficients for the reward and negative value regressors (Fig. 6). For both ROIs, Bayes factors comparing the odds ratio for the null hypothesis that the regression coefficients for $r(t_{outcome})$ and for $-V(t_{outcome})$ are equal versus the hypothesis that they are unequal supported the null hypothesis (Bayes factors for the right NAC were 4.02 and 3.09, and for left NAC were 5.29 and 4.12, using an uninformative prior and a unit-informative prior, respectively).

Relationship between learned values and risk preference

Together, the above ROI results confirmed that NAC BOLD signals correspond to a prediction error signal. They furthermore showed that the signals associated with the onset of the sure 20¢ and the risky 0/40¢ stimuli reflected similar subjective values, on average. The main aim of our study was to determine whether subject-specific behavioral risk sensitivity was related to differences in the evaluation of these stimuli. We thus used impulse function regressors aligned on stimulus onsets in forced trials to extract from the prediction error signals the mean values of the different stimuli for each subject in each session, while modeling away prediction errors for all other events using a separate regressor (see Materials and Methods). We only used forced trials to extract values for two reasons: First, choice trials are biased to-

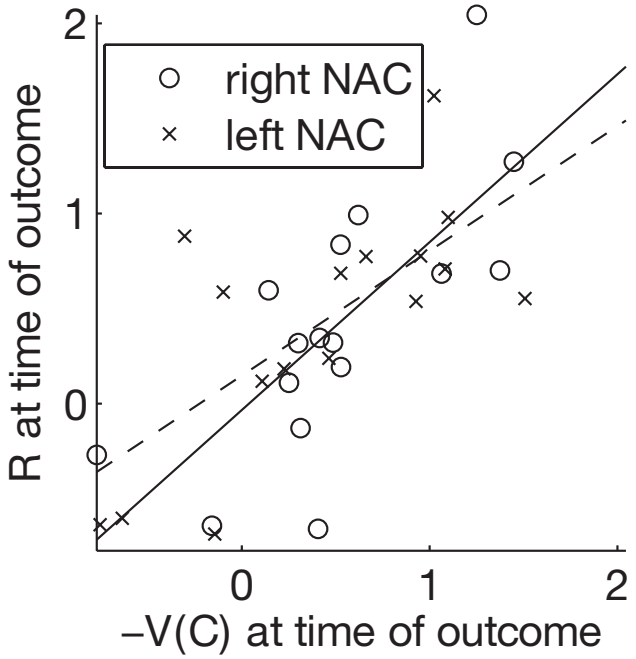

Figure 6. Regression coefficients for the obtained reward and the expected value at time of outcome in the two anatomically defined ROIs. Each data point represents one subject. In each ROI, the coefficients are strongly correlated (right NAC, dashed line, $\rho = 0.72$; $p = 0.002$; regression slope, 0.66; left NAC, solid line, $\rho = 0.70$; $p = 0.003$; regression slope, 0.88).

ward stimuli with higher values. As a result, all three models predict risk-related differences between the sure 20¢ and risky 0/40¢ stimulus values in choice trials, rendering any result we obtain inconclusive. By comparison, as forced trials allowed learning without a choice bias, the TD model predicted no difference between the sure 20¢ and risky 0/40¢ stimulus values for these trials (supplemental Fig. 2, available at www.jneurosci.org as supplemental material). Second, in choice trials there may be some ambiguity regarding which of the two stimuli the prediction error might reflect (Morris et al., 2006; Roesch et al., 2007). In forced trials there is no such ambiguity.

The correlation between subjects' behavioral risk preferences (preference for the sure 20¢ stimulus over the risky 0/40¢ stimulus) and the difference between the extracted neural values of the 20¢ and 0/40¢ stimuli was significantly positive when considering each of the left and right NAC ROIs as samples of the value differences (Fig. 7*a*; overall Pearson's $r = 0.60$, $p < 0.05$; right NAC only, Pearson's $r = 0.60$, $p < 0.05$; left NAC only, Pearson's $r = 0.44$, $p = 0.089$). This was also true when considering separately the value differences and risk sensitivity in each session of the experiment (Fig. 7*b*; overall Pearson's $r = 0.40$, $p = 0.005$; second session alone, Pearson's $r = 0.55$, $p < 0.05$, two tailed; third session alone, Pearson's $r = 0.36$, $p = 0.086$, marginally significant one tailed). These results argue against risk-neutral TD learning, in agreement with our analysis of choice behavior in the task.

Finally, to differentiate neurally between the utility and RSTD models, we relied on the fact that the former, but not the latter, predicts that the nonlinearity in the neural values of the sure 0¢, 20¢, and 40¢ stimuli (which we extracted from the BOLD signal in the same way) should correlate with behavioral risk sensitivity. In contrast to the previous results, we did not find any significant correlation between neural evaluations of sure outcomes and subjects' risk preferences. This was the case when we considered

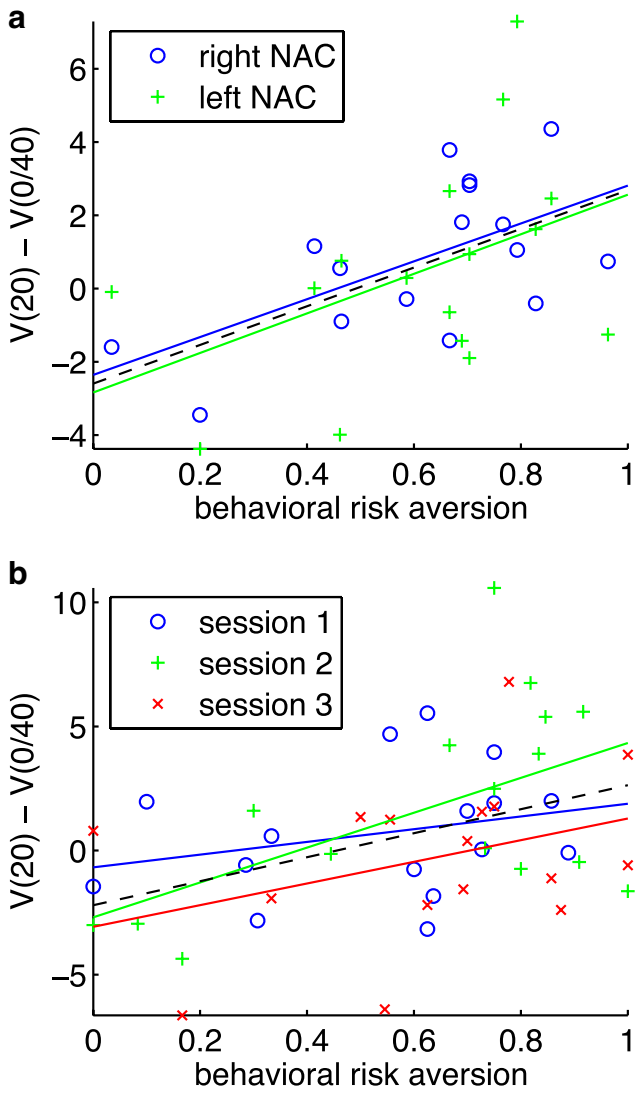

Figure 7. Risk sensitivity can be inferred from neural values extracted from prediction error signals. The neural values of the sure 20¢ stimulus and the risky 0/40¢ stimulus were extracted from the BOLD signal for each anatomical ROI and for each subject. ***a***, Across subjects, the difference between the neural values of these two stimuli correlated with behavioral risk aversion, with similar correlations apparent when considering each of the ROIs separately. ***b***, When considering each session separately (and averaging over both ROIs to reduce noise), the correlations between value differences and behavioral risk sensitivity were remarkably similar, despite the fact that behavioral risk aversion within subjects varied widely across sessions.

each ROI as a sample, or used each session as a sample, and for a variety of ways of assessing the presumed nonlinearity [using ratios, $V(40)/V(20)$ or $[V(40) - V(0)]/[V(20) - V(0)]$, or differences, $V(40) - V(20)$ or $[V(40) - V(20)] - [V(20) - V(0)]$; all p values >0.05, one sided; supplemental Figure 6, available at www.jneurosci.org as supplemental material]. That we did not find evidence for nonlinear utilities of 0¢, 20¢, and 40¢ payoffs in this task is maybe not surprising, as the monetary amounts involved were very small and likely in the linear range of the utility curve (see Discussion). In any case, together, these results suggest that reinforcement learning in the brain is influenced by the risk associated with different payoffs, for instance, as is implemented in the RSTD model.

Whole-brain analysis

Our analysis so far has concentrated on a priori, anatomically defined regions of interest in the nucleus accumbens. In a sup-

Table 1. All clusters that survived a whole-brain familywise error-corrected threshold of $p < 0.05$ in the random effects contrast for an RSTD prediction error signal

Anatomical location	Peak *x*, *y*, *z* (mm)	Cluster size	$t_{(15)}$	*V*(stimulus) (at stimulus)	Reward (at outcome)	−V(stimulus) (at outcome)
Left nucleus accumbens	−15, 6, −15	4	8.02	$p < 0.0001$**	$p = 0.0011$**	$p < 0.0001$**
Right tail of caudate	21, −12, 24	4	8.34	$p = 0.0436$*	$p = 0.3884$	$p = 0.0476$*
Right tail of caudate	24, −36, 3	3	10.73	$p = 0.0185$*	$p = 0.8693$	$p = 0.8979$
Tail of caudate combined		7		$p = 0.0098$**	$p = 0.6123$	$p = 0.2320$
Left orbitofrontal cortex	−15, 36, −18	6	10.13	$p = 0.0231$*	$p = 0.0350$	$p = 0.3969$
Left cerebellum	−29, −78, −36	4	8.04	$p = 0.1728$	$p = 0.0401$*	$p = 0.0198$*

Anatomical locations were determined through inspection with respect to the average anatomical image of all 16 subjects. "Tail of caudate combined" is a combined cluster including the two disjoint clusters in the right tail of the caudate.
*$p < 0.05$; ** significant after Bonferroni correction for 18 comparisons (post hoc analyses).

plemental analysis, we searched the whole brain for areas correlating with the RSTD prediction error regressor. We designated as functional ROIs all activations that survived a whole-brain-corrected threshold of $p < 0.05$, leading to five small clusters of activation: three in the striatum (left NAC and two noncontiguous activations in the left tail of the caudate), one in the orbitofrontal cortex, and one in the cerebellum (Table 1; for the raw time courses from these ROIs, see supplemental Fig. 7, available at www.jneurosci.org as supplemental material). To determine which of these areas indeed corresponded to a prediction error signal, in a post hoc (nonindependent) analysis, we asked whether each ROI was significantly correlated with each of the three necessary components of the prediction error signal in our task: (1) the value of the chosen stimulus at time of stimulus onset, (2) the magnitude of the reward at time of outcome, and (3) the negative of the expected value at time of outcome. As expected from our (fully independent) analysis of anatomical ROIs in the NAC, this analysis showed that the nucleus accumbens corresponds fully to a prediction error signal. However, this was not true for any of the other areas identified in the whole-brain analysis, where in each case activity failed to correlate significantly with one or more of the prediction-error components. This suggests that the NAC shows the most consistent evidence of neural prediction error coding out of all the areas implicated by the whole-brain analysis.

Discussion

We investigated the neural basis of risk-sensitive choice in an experiential learning task. We were particularly interested in neural representations associated with reinforcement learning, asking whether these were sensitive only to mean payoffs, as traditional theory suggests, or whether some of the effects of risk on choice may be realized through this form of learning. Using a priori anatomically defined ROIs in the NAC, we first confirmed that the dominant component of the raw BOLD signal in the NAC corresponds to a reward prediction error, a key component in reinforcement-learning models.

We then extracted from this prediction error signal the neural representations of the subjective values of different stimuli and showed that these correlated significantly with risk-sensitive choice: subjects for whom the extracted values of the sure 20¢ stimulus were greater than those of the risky 0/40¢ stimulus were more risk averse in their choices, and those whose neural signals favored the risky stimulus were behaviorally more risk seeking. Moreover, experience-related session-by-session fluctuations in the subjective evaluations of these two options correlated significantly with risk sensitivity in each session, testifying to the relationship between values estimated through reinforcement learning and behavioral choice.

Traditional TD reinforcement learning estimates the mean rewards associated with different options, ignoring their variance. The TD model therefore predicted similar values for the sure and risky options, and so was inconsistent with our results. Common accounts of risk sensitivity involving nonlinear utility functions (Bernoulli, 1954; Smallwood, 1996) were also not supported by the results of our analysis of the values of sure outcomes. This was evident from the poor behavioral fits, as well as the lack of neural support for the utility model. Although the latter was a null result and thus inconclusive, we note that this analysis had similar power to the analysis of differences between the values of the 0/40¢ and 20¢ stimuli, which did reach significance on all measures. Moreover, risk sensitivity in the domain of small payoffs (such as those we used here) is generally incompatible with accounts based on nonlinear utility (Rabin and Thaler, 2001). Thus, although it is possible, even likely, that large outcomes are subjectively valued in a nonlinear way, another source of sensitivity to risk is likely at play in causing the risk sensitivity that we observed in the domain of small outcomes.

Our results are consistent with risk-sensitive TD (RSTD) learning. In RSTD learning, nonlinearity is associated with the learning process rather than the evaluation of outcomes, with positive and negative prediction errors having asymmetric affects on changes in predictions (Fig. 3*c*). Applying a nonlinear transformation to prediction errors (as in RSTD learning) rather than to outcomes (as in the utility model) maintains the simplicity of recursive learning, as in traditional TD learning (Mihatsch and Neuneier, 2002). This form of learning predicts a precise relationship between amount of outcome variance and deviations from the risk-neutral mean value, as a function of the differences between η^- and η^+ for each subject. Because in the present experiment only one stimulus was associated with variable outcomes, we could not compare the effects of different degrees of variance, leaving this for future work. Another yet-unanswered question is whether the asymmetry is fixed, or adapts to the amount of risk in the task (or to other characteristics such as volatility or amount of training). For instance, Preuschoff et al. (2008) identify neural substrates correlating with signals that can be used to learn explicitly about the risk associated with different outcomes.

NAC BOLD signals have been implicated previously, directly and indirectly, in representing prediction errors. Early fMRI studies showed that the NAC is activated by monetary or other gains (Breiter et al., 2001; Knutson et al., 2005; Kuhnen and Knutson, 2005), while model-driven analyses have shown convincingly that BOLD signals in the NAC (as well as in other areas of the ventral striatum extending to the ventral putamen, and in some tasks to more dorsal striatal areas) correlate well with a prediction error in a variety of pavlovian or instrumental conditioning tasks (McClure et al., 2003; O'Doherty et al., 2003, 2004; Abler et al., 2006; Li et al., 2006; Tobler et al., 2006; Seymour et al., 2007). A recent study that used an experimental design aimed specifically at teasing apart prediction errors, outcome value, and decision value, showed that NAC BOLD responses correlate with

a prediction error signal and not other related value signals (Hare et al., 2008). Our supplemental analysis of other regions that were identified as potential prediction error correlates showed that signals there did not, in fact, satisfy all three defining features of a prediction error signal.

It is commonly believed that striatal BOLD signals reflect (at least appetitive) prediction errors because of dopaminergic input to this area (although other influences also weigh in) (Knutson et al., 2005; Knutson and Gibbs, 2007; Schott et al., 2008). Our results may thus have implications for the semantics of dopaminergic firing patterns and their downstream effects, either of which could be responsible for the asymmetric weighting of prediction errors in RSTD. Dopaminergic representations of negative and positive prediction errors are asymmetric because of the low baseline firing rate of these neurons (Fiorillo et al., 2003; Bayer and Glimcher, 2005); however, this need not necessarily result in asymmetric effects on downstream targets (Niv et al., 2005). Furthermore, at least in the BOLD signal, we did not find risk-sensitivity-related asymmetry in the prediction errors at the time of the outcome (Seymour et al., 2004). It is, however, straightforward to imagine that the inherently different synaptic processes of long-term potentiation and depression could result in asymmetric effects of increased or decreased levels of dopamine on downstream plasticity (Wickens et al., 2003; Shen et al., 2008).

Neither of the two previous studies that have associated NAC BOLD signals with risk involved learning. Matthews et al. (2004) showed selective activation in NAC during risk taking that was correlated with a harm-avoidance scale across subjects. Preuschoff et al. (2006) identified a correlate of anticipated risk in the NAC BOLD signal. This component is different from ours, since it appeared later in the trial, closer to the time of the outcome. It may be related to ramping dopaminergic signals that have sometimes been seen in single-unit recordings in a task with probabilistic rewards (Fiorillo et al., 2003), and might reflect nonlinearities in the coding (rather than the effect) of prediction errors (Niv et al., 2005).

Although we have discussed the proposal that risk is incorporated within learned values, it is conceivable that expected value and risk are represented and learned separately, with the two signals converging at the level of the NAC (or at least in the NAC BOLD signal). Risk and return are typically separated in financial theory (also related to Markowitz portfolio theory) (Markowitz, 1952; Weber et al., 2004; Weber and Johnson, 2008). Indeed, previous studies of risk-sensitive choice have implicated cortical areas such as insula, posterior cingulate cortex, the orbitofrontal cortex and medial prefrontal cortex in tracking and representing risk (Elliott et al., 1999; Huettel et al., 2005, 2006; Kuhnen and Knutson, 2005; McCoy and Platt, 2005; Knutson and Bossaerts, 2007; Tobler et al., 2007; Platt and Huettel, 2008; Preuschoff et al., 2008). However, these studies all involved explicit knowledge of the risks. This was also the case in a study of ambiguity aversion that showed that ambiguity influences striatal BOLD signals (albeit arguing that its effect on choice was mediated by an orbitofrontal representation) (Hsu et al., 2005). As a result, these studies, and the mechanism that they suggest for incorporating risk into choice, may be more relevant to model-based decision making. In the present study as well, a contrast between trials involving risk and all other trials showed prominent activations in bilateral anterior insula and supplementary motor area (supplemental Fig. 8, available at www.jneurosci.org as supplemental material), perhaps suggesting that model-based learning was taking place in parallel to model-free learning in the areas that we investigated (Daw et al., 2005).

In sum, our results provide evidence that risk sensitivity is indeed present in prediction error signaling in the nucleus accumbens, with a direct correlation between the risk-averse or risk-seeking choices of our subjects and the neural prediction error signals. This finding supports a reinforcement-learning-based account of risk sensitivity at least in cases in which payoffs must be learned from experience and suggests that risk sensitivity, as in RSTD learning, should be imported into computational models of human choice. Furthermore, it can potentially provide insight into the functions of related neural substrates such as dopamine. Thus, our study completes a full circle starting from computational theory, through the psychology of affective choice, to the neural basis in signals in the nucleus accumbens, and back again to influence and inform the computational theory.

References

Abler B, Walter H, Erk S, Kammerer H, Spitzer M (2006) Prediction error as a linear function of reward probability is coded in human nucleus accumbens. Neuroimage 31:790–795.

Barto, A. G. (1995) . Adaptive critic and the basal ganglia. In: Models of information processing in the basal ganglia (Houk JC, Davis JL, Beiser DG, eds), pp 215–232. Cambridge, MA: MIT.

Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129–141.

Bernoulli D (1954) Exposition of a new theory on the measurement of risk. Econometrica 22:23–36.

Breiter HC, Aharon I, Kahneman D, Dale A, Shizgal P (2001) Functional imaging of neural responses to expectancy and experience of monetary gains and losses. Neuron 30:619–639.

Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162–173.

D'Ardenne K, McClure SM, Nystrom LE, Cohen JD (2008) BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science 319:1264–1267.

Daw ND (2011) Trial by trial data analysis using computational models. In: Decision making, affect, and learning: attention and performance XXIII (Delgado MR, Phelps EA, Robbins TW, eds), pp 3–38. Oxford, UK: Oxford UP.

Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704–1711.

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876–879.

Dayan P, Kakade S, Montague PR (2000) Learning and selective attention. Nat Neurosci [Suppl] 3:1218–1223.

Denrell J (2007) Adaptive learning and risk taking. Psychol Rev 114:177–187.

Denrell J, March JG (2001) Adaptation as information restriction: the hot stove effect. Organ Sci 12:523–538.

Elliott R, Rees G, Dolan RJ (1999) Ventromedial prefrontal cortex mediates guessing. Neuropsychologia 37:403–411.

Fiorillo CD, Tobler PN, Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299:1898–1902.

Fitzgerald TH, Seymour B, Bach DR, Dolan RJ (2010) Differentiable neural substrates for learned and described value and risk. Curr Biol 20:1823–1829.

Hare TA, O'Doherty J, Camerer CF, Schultz W, Rangel A (2008) Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. J Neurosci 28:5623–5630.

Hayden B, Platt M (2008) Gambling for Gatorade: risk-sensitive decision making for fluid rewards in humans. Anim Cogn 12:201–207.

Hertwig R, Erev I (2009) The description-experience gap in risky choice. Trends Cogn Sci 13:517–523.

Hertwig R, Barron G, Weber EU, Erev I (2004) Decisions from experience and the effect of rare events in risky choice. Psychol Sci 15:534–539.

Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF (2005) Neural systems responding to degrees of uncertainty in human decision-making. Science 310:1680–1683.

Hsu M, Krajbich I, Zhao C, Camerer CF (2009) Neural responses to reward

anticipation under risk is nonlinear in probabilities. J Neurosci 29:2231–2237.
Huettel SA, Song AW, McCarthy G (2005) Decisions under uncertainty: probabilistic context influences activation of prefrontal and parietal cortices. J Neurosci 25:3304–3311.
Huettel SA, Stowe CJ, Gordon EM, Warner BT, Platt ML (2006) Neural signatures of economic preferences for risk and ambiguity. Neuron 49:765–775.
Jessup RK, Bishara AJ, Busemeyer JR (2008) Feedback produces divergence from prospect theory in descriptive choice. Psychol Sci 19:1015–1022.
Kacelnik A, Bateson M (1996) Risky theories—the effect of variance on foraging decisions. Am Zoologist 36:402–434.
Kim H, Shimojo S, O'Doherty JP (2006) Is avoiding an aversive outcome rewarding? neural substrates of avoidance learning in the human brain. PLoS Biol 4:e233.
Knutson B, Bossaerts P (2007) Neural antecedents of financial decisions. J Neurosci 27:8174–8177.
Knutson B, Gibbs SE (2007) Linking nucleus accumbens dopamine and blood oxygenation. Psychopharmacology (Berl) 191:813–822.
Knutson B, Taylor J, Kaufman M, Peterson R, Glover G (2005) Distributed neural representation of expected value. J Neurosci 25:4806–4812.
Kuhnen CM, Knutson B (2005) The neural basis of financial risk taking. Neuron 47:763–770.
Li J, McClure SM, King-Casas B, Montague PR (2006) Policy adjustment in a dynamic economic game. PLoS One 1:e103.
March JG (1996) Learning to be risk averse. Psychol Rev 103:309–319.
Markowitz H (1952) Portfolio selection. J Finance 7:77–91.
Matthews SC, Simmons AN, Lane SD, Paulus MP (2004) Selective activation of the nucleus accumbens during risk-taking decision making. Neuroreport 15:2123–2127.
McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. Neuron 38:339–346.
McCoy AN, Platt ML (2005) Risk-sensitive neurons in macaque posterior cingulate cortex. Nat Neurosci 8:1220–1227.
Mihatsch O, Neuneier R (2002) Risk-sensitive reinforcement learning. Machine Learn 49:267–290.
Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947.
Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H (2006) Midbrain dopamine neurons encode decisions for future action. Nat Neurosci 9:1057–1063.
Niv Y (2009) Reinforcement learning in the brain. J Math Psychol 53:139–154.
Niv Y, Schoenbaum G (2008) Dialogues on prediction errors. Trends Cogn Sci 12:265–272.
Niv Y, Joel D, Meilijson I, Ruppin E (2002) Evolution of reinforcement learning in uncertain environments: a simple explanation for complex foraging behaviors. Adapt Behav 10:5–24.
Niv Y, Duff MO, Dayan P (2005) Dopamine, uncertainty and TD learning. Behav Brain Funct 1:6.
O'Doherty J, Dayan P, Friston K, Critchley H, Dolan R (2003) Temporal difference learning model accounts for responses in human ventral striatum and orbitofrontal cortex during Pavlovian appetitive learning. Neuron 38:329–337.
O'Doherty JP, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452–454.
Platt ML, Huettel SA (2008) Risky business: the neuroeconomics of decision making under uncertainty. Nat Neurosci 11:398–403.
Preuschoff K, Bossaerts P, Quartz SR (2006) Neural differentiation of expected reward and risk in human subcortical structures. Neuron 51:381–390.
Preuschoff K, Quartz SR, Bossaerts P (2008) Human insula activation reflects risk prediction errors as well as risk. J Neurosci 28:2745–2752.
Rabin M, Thaler RH (2001) Anomalies: risk aversion. J Econom Perspect 15:219–232.
Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Classical conditioning II: current research and theory (Black AH, Prokasy WF, eds), pp 64–99. New York: Appleton-Century-Crofts.
Roesch MR, Calu DJ, Schoenbaum G (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat Neurosci 10:1615–1624.
Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian *t* tests for accepting and rejecting the null hypothesis. Psychon Bull Rev 16:225–237.
Schönberg T, Daw ND, Joel D, O'Doherty JP (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J Neurosci 27:12860–12867.
Schott BH, Minuzzi L, Krebs RM, Elmenhorst D, Lang M, Winz OH, Seidenbecher CI, Coenen HH, Heinze H-J, Zilles K, Düzel E, Bauer A (2008) Mesolimbic functional magnetic resonance imaging activations during reward anticipation correlate with reward-related ventral striatal dopamine release. J Neurosci 28:14311–14319.
Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599.
Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, Friston KJ, Frackowiak RS (2004) Temporal difference models describe higher-order learning in humans. Nature 429:664–667.
Seymour B, Daw N, Dayan P, Singer T, Dolan R (2007) Differential encoding of losses and gains in the human striatum. J Neurosci 27:4826–4831.
Shapiro MS, Couvillon PA, Bitterman ME (2001) Quantitative tests of an associative theory of risk-sensitivity in honeybees. J Exp Biol 204:565–573.
Shen W, Flajolet M, Greengard P, Surmeier DJ (2008) Dichotomous dopaminergic control of striatal synaptic plasticity. Science 321:848–851.
Smallwood PD (1996) An introduction to risk sensitivity: the use of Jensen's inequality to clarify evolutionary arguments of adaptation and constraint. Am Zoologist 36:392–401.
Sutton RS (1988) Learning to predict by the method of temporal difference. Machine Learning 3:9–44.
Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, MA: MIT.
Tobler PN, Fiorillo CD, Schultz W (2005) Adaptive coding of reward value by dopamine neurons. Science 307:1642–1645.
Tobler PN, O'Doherty JP, Dolan RJ, Schultz W (2006) Human neural learning depends on reward prediction errors in the blocking paradigm. J Neurophysiol 95:301–310.
Tobler PN, O'Doherty JP, Dolan RJ, Schultz W (2007) Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. J Neurophysiol 97:1621–1632.
Tom SM, Fox CR, Trepel C, Poldrack RA (2007) The neural basis of loss aversion in decision-making under risk. Science 315:515–518.
Weber EU, Johnson EJ (2008) Decisions under uncertainty: psychological, economic, and neuroeconomic explanations of risk preference. In: Neuroeconomics: Decision making and the brain (Glimcher PW, Camerer C, Fehr E, Poldrack R, eds), pp 127–144. London: Elsevier.
Weber EU, Shafir S, Blais, A-R (2004) Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. Psychol Rev 111:430–445.
Wickens JR, Reynolds JN, Hyland BI (2003) Neural mechanisms of reward-related motor learning. Curr Opin Neurobiol 13:685–690.