SUPPORTING ON-LINE MATERIAL

MATERIALS AND METHODS

Subjects

12 right-handed healthy normal subjects participated in the experiment (Mean age 28.75, range: 21-49), of which 7 were female. The subjects were pre-assessed to exclude those with a prior history of neurological or psychiatric illness. All subjects gave informed consent and the study was approved by the Joint Ethics Committee of the National Hospital for Neurology and Neurosurgery. Subjects were asked to refrain from eating, or drinking sweet drinks for 6 hours prior to scanning.

Stimulus Presentation

The gustatory stimuli were contained in 6 50ml syringes (two for each fruit juice and the neutral taste) which were attached to an SP220I electronic syringe pump, positioned in the scanner control room and delivered to the subjects via three separate 6 metre long 3mm wide polythene tubes. The syringes were also attached to a computer controlled valve system which enabled the different tastes to be delivered independently along the tubing. The apparatus was controlled by the stimulus presentation computer positioned in the control room, which also received volume trigger pulses from the scanner, and the visual stimuli were presented on a projector screen positioned ~10cm away from the subject's face.

Task description

Each trial in the instrumental task involved the simultaneous presentation of one of two pairs of fractal stimuli. Each pair signified the onset of one of two distinct trial types: Reward and Neutral, whose occurrence was fully randomised throughout the experiment. The order of presentation of the instrumental and Pavlovian tasks, and the specific assignment of fractal pairs to a given trial type were fully counterbalanced across subjects. The subjects' task on each instrumental trial was to choose one of the two stimuli by selecting the fractal to the left or right of the fixation cross via a button box (using their right hand). Once a fractal had been selected, it increased in brightness, and was followed 3 seconds later by 0.75ml of either a juice reward, the same quantity of an affectively neutral control solution, or else by nothing.

In Reward trials, one of the stimuli was arbitrarily designated as a High Probability stimulus, such that if the subject chose that stimulus they would obtain juice reward with a probability of 0.6, otherwise they would obtain nothing. The other stimulus was designated as a Low Probability stimulus, as the juice reward was presented with probability of only 0.3. The stimuli in the Neutral trials were similarly assigned High (0.6) and Low (0.3) outcome probabilities, but in this case the outcome was affectively neutral (control tasteless solution). We predicted that in the Reward trials over the course of the experiment subjects would learn to choose with greater frequency the High Probability stimulus, as this was associated with a greater probability of obtaining the juice reward. On the other hand in Neutral trials, we predicted that subjects would not choose the high probability stimulus more frequently than the low probability stimulus, as the outcome was affectively neutral and would thus not serve to bias action choice.

The Pavlovian conditioning task was identical to the instrumental task in terms of presentation (though using four different fractal stimuli), except that in this case subjects had no choice about which stimulus they would select on a given trial. Instead, the computer selected one of the stimuli according to the selections made and feedback obtained by another subject while performing the instrumental task. Thus, each subject's Pavlovian conditioning task was yoked to the instrumental choices of another subject. Timing was matched as best as possible to the instrumental task. Each trial began with the presentation of two arbitrary neutral stimuli either side of a fixation cross. Unlike the instrumental task, 500 msecs into the trial one of the two stimuli spontaneously increased in brightness, indicating that the computer had chosen that stimulus. Once a stimulus had been chosen, but before feedback, subjects were instructed to make a response indicating whether the selected stimulus was on the left or right of the screen. This ensured that subjects attended to the relevant stimulus, as well as controlling for motor confounds when comparing activity evoked by trials from the instrumental task. Furthermore, subjects' reaction time was recorded to provide a behavioral index of learning.

Two different juices were used for each subject: peach and blackcurrant juice. One juice was used in the instrumental task and the other in the Pavlovian Conditioning task, counterbalanced across subjects. The rationale for alternating the juices between sessions was to control for the possibility that subjects' habituated to a specific juice over the course of a conditioning episode, or that the pleasantness would decrease due to the phenomenon of selective satiation (S1). The use of two different juices ensured that the juice reward was judged as highly pleasant by subjects throughout both sessions. The specific juice assigned to each task was counterbalanced across subjects

Imaging Procedure

The functional imaging was conducted by using a 2 Tesla Siemens Vision MRI scanner to acquire gradient echo T2* weighted echo-planar images (EPI) images with BOLD (blood oxygenation level dependent) contrast. We employed a special sequence designed to optimise functional sensitivity in OFC and medial temporal lobes (S2). This consisted of tilted acquistion in an oblique orientation at 30* to the AC-PC line, as well as application of a preparation pulse with a duration of 1 msec and amplitude of –2 mT/m in the slice selection direction. The sequence enabled 33 axial slices of 3 mm thickness and 3 mm in-plane resolution to be acquired with a TR of 2.31 seconds. Subjects were placed in a light head restraint within the scanner to limit head movement during acquisition. A T1-weighted structural image was also acquired for each subject. Functional imaging data was acquired in two separate ~15 minute (390 volume) sessions in each subject during performance of the Pavlovian and instrumental conditioning tasks. To detect transient head movements due to swallowing we attached a 1.5 cm long copper coil with a radius of 0.5 cm to the neck of each subject. Small movements of the coil induced a current in the magnetic field that could be detected when amplified using one channel of an EEG system positioned in the scanner room (National Hospital for Neurology and Neurosurgery, London, UK). This produced a time-series over the whole experiment in which signal changes represented transient head movement.

Temporal difference prediction error signal

The temporal difference (TD) learning model used to generate a prediction error signal is as described by Schultz et al.(S3). On each trial, the prediction $\hat{V}(t)$ of the value $V(t)$ at any time *t* within a trial is calculated as a linear product of the weights w_i and the presence or absence of a conditioned stimulus (CS) stimulus at time *t*, coded in the stimulus representation vector $x_i(t)$:

$$\hat{V}(t) = \sum_i w_i x_i(t)$$

Learning occurs by up dating the predicted value of each time-point *t* in the trial by comparing the value at time *t*+1 to that at time *t*, leading to a prediction error (PE) or $\delta(t)$:

$$\delta(t) = r(t) + \gamma\hat{V}(t+1) - \hat{V}(t) \tag{0.1}$$

where $r(t)$ is the reward at time *t*

The parameter γ is a discount factor, that determines the extent to which rewards that arrive earlier are more important than rewards that arrive later on. In the present study we set $\gamma = 0.99$. The weights w_i are then updated on a trial by trial basis according to the correlation between PE and the stimulus representation:

$$\Delta w_i = \alpha \sum_t x_i(t)\delta(t)$$

where α is the learning rate

In the TD model, we assigned 6 time-points to each trial, and used each subject's individual event-history as input. On each trial the CS was taken to be delivered at time-point 1, and the reward was delivered at time-point 3. The stimuli x_i corresponding to the presence of the CS, were represented as vectors in which the i^{th} component was = 1 and 0 elsewhere (as used by Schultz et al.). We note that other stimulus representations are possible, and these could lead to a different output from the model.

The existence of two stimuli between which either the subject or the computer chooses complicates the analysis of $\hat{V}(t)$ for both Pavlovian and instrumental conditioning. One might expect a rich temporal microstructure of the development of this value prediction as the subject assesses which stimulus to choose, or the probability with which the computer will choose one or other stimulus. The details of this microstructure may differ on a trial-by-trial basis, and it is in any case invisible in fMRI. We therefore considered various possibilities for $\hat{V}(t)$, which lead to similar, though not quite identical, results.

For the analysis of the Pavlovian conditioning data, reward prediction errors are calculated for the specific CS that was illuminated (i.e., $\hat{V}(1)$ was generated based on just one of the two stimuli shown), since there is no evident reason for subjects to generate a prediction before this point. To generate regressors corresponding to PE responses separately for CS and UCS (unconditioned stimulus) trial components for the SPM analysis (see below), δt_{CS} was sampled at time point 1 in the trial, and δt_{UCS} was sampled at time point 3.

For the analysis of the instrumental conditioning data, we employed advantage learning (S4-S6). In this, subjects are assumed to use a prediction error like $\delta(t)$ to learn an estimate $\hat{A}(t,a)$ of the advantage, $A(t,a)$, specific to action a. The advantage is defined as the difference between the future reward $Q(t,a)$ expected if action a is chosen (called the Q value, (S7)), and the average value of the state $V(t)$. The advantage error signal is

$$\delta^{A}(t) = r(t) + \gamma\hat{V}(t+1) - \hat{Q}(t,a) \tag{0.2}$$

based on estimated values of the various quantities. In the case of just two actions (a and b, one for each stimulus), the advantages can be used to determine the probability with which an action is chosen according to

$$p(t,a) = \sigma(\beta(\hat{A}(t,a) - \hat{A}(t,b))$$

where $\sigma(z) = 1/(1+\exp(-z))$ is the Luce choice rule (S8) or logistic sigmoid, and β is an inverse temperature that determines the ferocity of the competition. This probability can be used to define the value of the initial state at $t = 1$ as

$$\widehat{V}(1) = p(1,a)\widehat{Q}(1,a) + p(1,b)\widehat{Q}(1,b) .$$

For the instrumental conditioning analyses, we used $\delta^{A}(t)$ as the advantage learning prediction error signal, and $\delta(t)$ from equation 1.1 (using $\widehat{V}(1)$ as above) as the critic prediction error signal.

To ensure that our results do not depend on the different interpretations of the predictions made by the subjects at time $t = 1$ in instrumental and Pavlovian conditioning, we confirmed that the significant difference in activity reported in dorsal striatum between the instrumental and Pavlovian tasks remains present if the Pavlovian data is modeled using exactly the same signal at this time as the instrumental data (at p<0.001).

Various other forms of prediction error signal are also possible. For instance, the original actor-critic model would suggest that the actor learns stimulus-response quantities that are not directly related to the estimated values of executing actions at states (unlike the $\widehat{Q}(t,a)$ or $\hat{A}(t,a)$ values) together with a critic whose predictions do not depend directly on the probabilities of taking particular actions (S9-S11).

There is active debate about the nature of the actor-critic learning models that subjects employ in tasks like this (S12-S14). We therefore calculated (Fig. S1) the log likelihood fit of the actual choices made by subjects according to advantage learning and the original actor-critic, for a variety of learning rates and exploration/exploitation tradeoff parameters (β). Our results favour advantage learning, justifying its use to generate regressors for the fMRI data. Furthermore, for advantage learning, a range of learning rates between 0.1 and 0.6 provide the best fit to the behavioral data, provided the exploration/exploitation constant (β) is chosen appropriately. We report the results of using a learning rate within this range of 0.2, with β = 0.5. We also used a learning rate of 0.2 in the Pavlovian task, consistent with that used in a previous study of Pavlovian conditioning (S15).

Image Analysis

Image analysis was performed using SPM99 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, U.K.). To correct for subject motion, the images were realigned to the first volume, spatially normalized to a standard T2* template with a resampled vowel size of 3mm^3, and spatial smoothing was applied using a Gaussian kernel with a full width at half maximum (FWHM) of 8mm. Intensity normalisation and high pass temporal filtering (using a filter width of 128 secs) were also applied to the data (S16).

In the TD analysis, the subject specific δt_{CS} and δt_{UCS} components were convolved with a haemodynamic response function and fitted to the data for each single subject. The δt_{CS} and δt_{UCS} components were modeled separately in order to accommodate the possibility of a variation in magnitude between the responses elicited by the CS and UCS (arising, for instance, from the sorts of differences in the time courses of CS and UCS activation seen in the activity of dopamine cells (S3). In addition, the 6 scan to scan motion parameters produced during realignment were included to account for residual effects of movement. To account for transient head motion, produced by e.g. swallowing, we also included an additional motion regressor which comprised the output of the motion-detector coil, band-pass filtered appropriately and sub-sampled to the number of scans in the experiment. Linear contrasts of regressor coefficients were computed at the individual subject level to enable comparison of PE responses between the Reward and Neutral trials. The specific contrasts were: δt_{CS} [Reward] - δt_{CS} [Neutral] and δt_{UCS} [Reward] - δt_{UCS} [Neutral]. These were computed separately for the instrumental and Pavlovian conditioning tasks. The results from each subject were taken to a random effects level by including the contrast images for the CS and UCS comparisons, separately for the Instrumental and Pavlovian conditioning sessions from each single subject into a one way analysis of variance with no mean term. Sphericity correction was applied to correct for possible violation of the independence assumption between the contrasts. Conjunctions were performed at the second level to identify areas showing significant prediction error related activity at both the time of presentation of the CS and UCS.

The structural T1 images were co-registered to the mean functional EPI images for each subject and normalised using the parameters derived from the EPI images. Anatomical localisation was carried out by overlaying the t-maps on a normalised structural image averaged across subjects, and with reference to an anatomical atlas (S17).

Behavioral measures

Pleasantness ratings

Subjects provided ratings of the pleasantness of the fruit juice and control tasteless solutions before and after the Instrumental and Pavlovian conditioning sessions, using a scale ranging from -10 to +10 were -10 = very unpleasant, +10 = very pleasant and 0 = neutral. A two-way repeated measures Anova with one factor Stimulus Type: Juice or Neutral, and the other factor Time: Before and After, revealed a significant main effect of Stimulus Type for both tasks (Instrumental: F=35.2; df=1,11; $p<0.001$; Pavlovian: F=32.4, df=1,11;$p<0.001$), indicating a significant difference between the pleasantness ratings for the two stimuli. Post-hoc paired t-tests confirmed that subjects rated the juice reward as significantly more pleasant than the neutral stimulus (Pavlovian: t=5.7, df=11; $p<0.001$; Instrumental: t=5.9, df=11;$p<0.001$). As rating scale data can violate the normal distribution we also confirmed our analysis using Wilcoxon non-parametric paired sample tests, which also revealed significant differences in pleasantness ratings between the stimuli (at $p<0.01$).

Instrumental choices

To determine whether subjects showed a preference for the High probability over the Low probability action during the Instrumental task**,** a repeated measures two-way Anova was performed on the instrumental choice data with one factor Valence (Reward vs Neutral) and the other Action Outcome Probability (High vs Low) revealed a significant Valence x Outcome Probability interaction across subjects (F=4.82;df=1,11;$p<0.05$). Paired t-tests indicated a significant difference between the number of responses made to the High Probability stimulus and the Low Probability stimulus in Reward trials (t=2.536,df=11,p<0.05. There was no significant difference between the number of responses made to the High and Low Probability actions in Neutral trials (t=-1.243,df=11;p=.240), indicating that the affectively neutral control solution did not bias action choice in either direction.

Pavlovian reaction times

To provide behavioral evidence of learning in the Pavlovian task we tested whether subjects were faster to respond (to indicate which stimulus had been chosen by the computer) during Reward trials than Neutral trials. The session was subdivided into two phases (early and late) to test for differences in response times across learning. A two-way repeated measures Anova revealed a time x condition interaction just approaching significance (f=4.769,df=1,10; p=0.054) . Subsequent paired t-tests revealed a significant difference between reaction times during Reward and Neutral trials in the second phase of the experiment (t=-2.768,df=10,p<0.05 two-tailed), such that subjects were faster to respond to Rewarding than Neutral conditioned stimuli, whereas there was no significant difference between the two stimulus types in the early part of the experiment. Due to a technical problem, reaction time data was not obtained for one subject during the Pavlovian conditioning task.

Assessment of model fitness across early and late trials

Since we are explicitly interested in modeling the changes happening through the course of learning, it is important to confirm that the fit of the model to the imaging and behavioral data is not significantly different at different times during learning. To assess this for the imaging data, we plot the predicted (according to the model) against the adjusted fMRI data (which is the real fMRI signal adjusted for all other effects in the statistical model) derived from the peak voxel in this region at the individual subject level and then subsequently averaged across subjects in Fig. S2(A) separately for the first and second half of the experiment. This figure shows that the model provides a reasonable fit to the data in both the early and late trials. Furthermore, there are no significant differences in the regression slopes between the predictions and the data across the two time periods for either the CS or UCS regressors. We also show plots of residuals for two typical single subjects in Fig. S2(B). There is no clear structure in the residual plots, a further indication that the model provides a good fit to the data in both early and late phases of the experiment even at the single subject level.

Calibration analysis of model fit with fMRI data and behavior

Figures S3 and S4 show finer scale analyses over the course of learning for individual subjects of the degree to which the model's prediction are appropriately calibrated (S18) to imaging data (Fig. S3) and behavioral choices (Fig. S4).

Fig. S3 shows calibration points, relating the predicted values of the BOLD signal in the ventral striatum to the actual (adjusted) values as a function of time (left; out of 385 scans per subject) and by subject (right). The predicted values of the BOLD signal produced by the TD model were separated into bins, and the mean values of the adjusted BOLD signals for those scans for which the predicted signal lives within each bin were calculated. If the model and data were perfectly calibrated and there was no noise, then the mean BOLD signal would have the same value as the predicted BOLD signal, and so the points would all lie exactly on the solid straight line. In the left plot, we have made the further manipulation of separating the plotted points into segments of scans, from early (scans 1-77), later (78-134), etc in the trial shown by color (see legend). The sizes of the points relates to the number of data points in each joint bin. The main conclusion to be drawn from the left plot is that since the points with different colors lie on top of each other, there are no privileged time bins – the model is generally as well calibrated for the early parts of learning (blue) as the late parts (red). The black line with circles shows the overall calibration (for values of predicted signal with more than 0.5% of the maximum number) across subjects and scans. The plot is very close to the identity, showing good overall calibration. The right hand of Fig. S3 shows the same data, but now focusing on the question of whether some subjects carry the weight of the statistical relationships shown in the text. Only half the subjects are shown for clarity (the numbers are shown by colour). In this plot, the whole session is split into two intervals (solid lines for the first 192 scans; dashed lines for the remaining scans). The line color indicates subject identity; only points with more than 2 samples are shown. That the lines for the individual subjects are very similar suggests that there are no particular subjects that bear the weight of the correlation. These two plots together suggest that the relationship between the predicted and actual BOLD data is relatively uniformly spread across time and subjects.

Figure S4 shows similar calibration plots, but now for the behavior of the subjects rather than the BOLD signal. The actor-critic model was used to generate predictions about the probability of choosing the better stimulus as the next action, given the (subject-specific) history of choices and outcomes experienced. These are divided into ranges, and the mean values of the actual predictions within these ranges are plotted against the actual probabilities of choice.

The left figure shows these values for early (1-15), later (16-30), etc, trials (by color). Again, that the colors substantially overly suggests that there are no privileged epochs of learning at which the model is specially good. The right figure joins up the dots, now by subject (see color legend) rather than trial. Again, the whole period of learning is separated into early (solid) and late (dashed) trials. This suggests that there are no significant systematic biases in the data.

References and Notes

S1. B. J. Rolls, E. T. Rolls, E. A. Rowe, K. Sweeney, *Physiol Behav.* 27, 137 (1981).

S2. R. Deichmann, J. A. Gottfried, C. Hutton, R. Turner, *Neuroimage.* 19, 430 (2003).

S3. W. Schultz, P. Dayan, P. R. Montague, *Science* 275, 1593 (1997).

S4. L. C. Baird, "Advantage Updating" *Report No. WL-TR-93-1146* (Wright Patterson Air Force Base, Dayton, OH, 1993).

S5. P. Dayan, in *NIPS*, T. G. Dietterich, S. Becker, Z. Ghahramani, Eds. 2001).

S6. P. Dayan , B. W. Balleine, *Neuron* 36, 285 (2002).

S7. C. J. Watkins , P. Dayan, *Machine Learning* 8, 279 (1992).

S8. D. R. Luce, Response Times (Oxford University Press, New York, 2003).

S9. A. G. Barto, R. S. Sutton, C. W. Anderson, *IEEE Transactions on Sytems, Man and Cybernetics* 13, 835 (1983).

S10. A. G. Barto, in *Models of Information Processing in the Basal Ganglia*, J. C. Houk, J. L. Davis, B. G. Beiser, Eds. (MIT Press, Cambridge, MA, 1996) ,chap. 11.

S11. P. Dayan and L. F. Abbott, Theoretical Neuroscience (MIT Press, Cambridge, MA, 2001).

S12. D. M. Egelman, C. Person, P. R. Montague, *J.Cogn Neurosci.* 10, 623 (1998).

S13. L.P. Sugrue, A.E. Rorie, G.S. Corrado, W.T. Newsome, paper presented at the 33rd Annual Meeting of the Society for Neuroscience, New Orleans, LA, 10 November 2003.

S14. B. Lau, P.W. Glimcher, paper presented at 33rd Annual Meeting of the Society for Neuroscience, New Orleans, LA, 10 Novermber 2003.

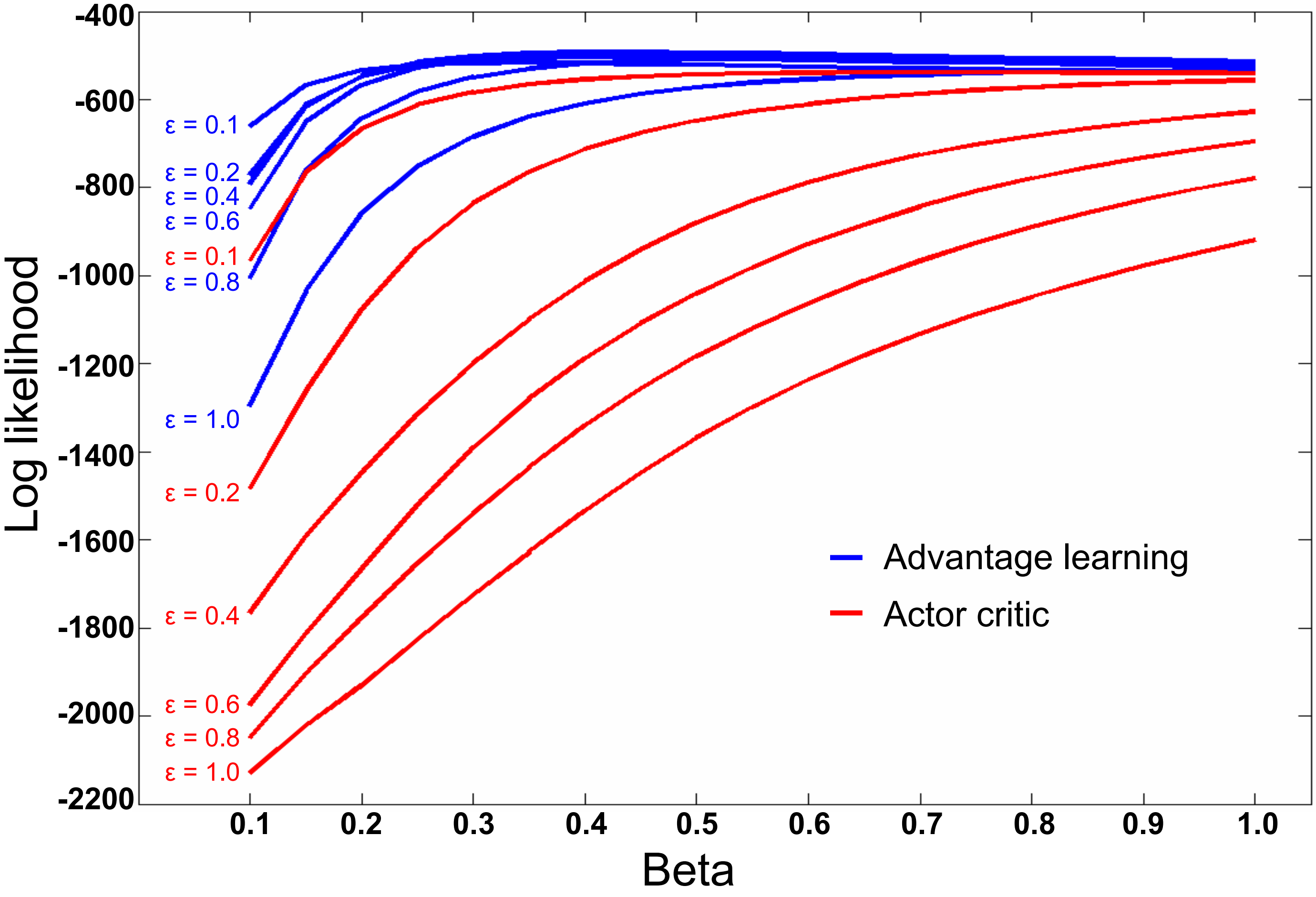
S15. J. P. O'Doherty, P. Dayan, K. Friston, H. Critchley, R. J. Dolan, *Neuron* 38, 329 (2003).

S16. K. Friston et al., *Hum.Brain Mapp.* 2, 165 (1995).

S17. H. M. Duvernoy, The Human Brain (Springer-Verlag, Vienna, 1999).

S18. A.P. Dawid. *The Annals of Statistics*, 13, 1251 (1985).

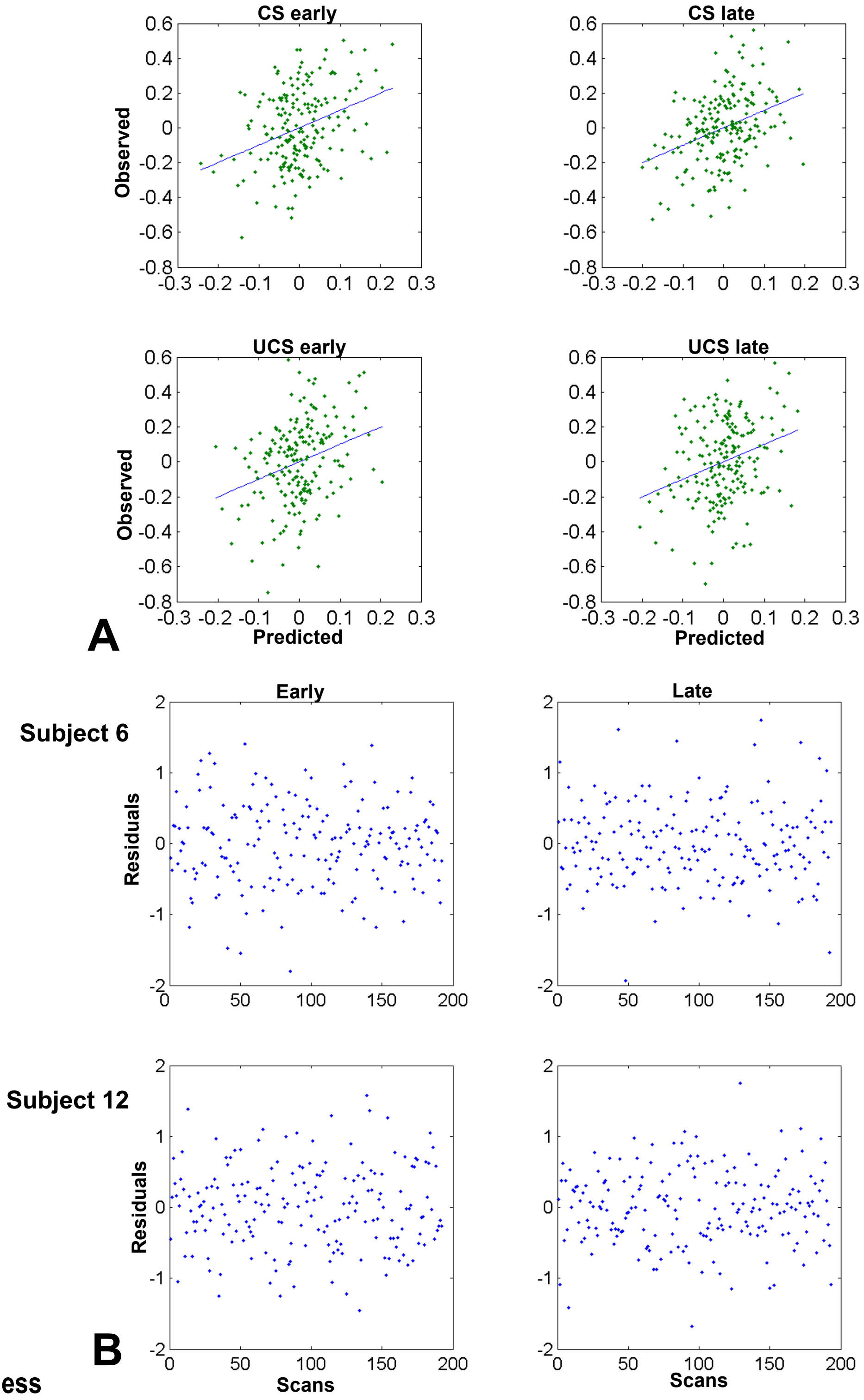
Figure S1



Model predictions vs behavioral choice data

Log likelihood of the actual choices made by all subjects according to the advantage learning model (in blue) and the original actor critic model (in red). The abscissa indicates the beta (β) parameter used to control the balance between exploration and exploitation. The ordinate shows the log-likelihood of the model fit. The higher the value the better the fit (i.e. the closer to the top of the graph). The different curves show different learning rates (ε) each separately annotated. The plot shows that the advantage learning model provides a much better fit to subjects' behavior than the actor-critic model. For the advantage model, learning rates between 0.1 and 0.6 provide a reasonable fit to the data with an appropriate beta value.
We used $\varepsilon = 0.2; \beta = 0.5$.

Figure S2

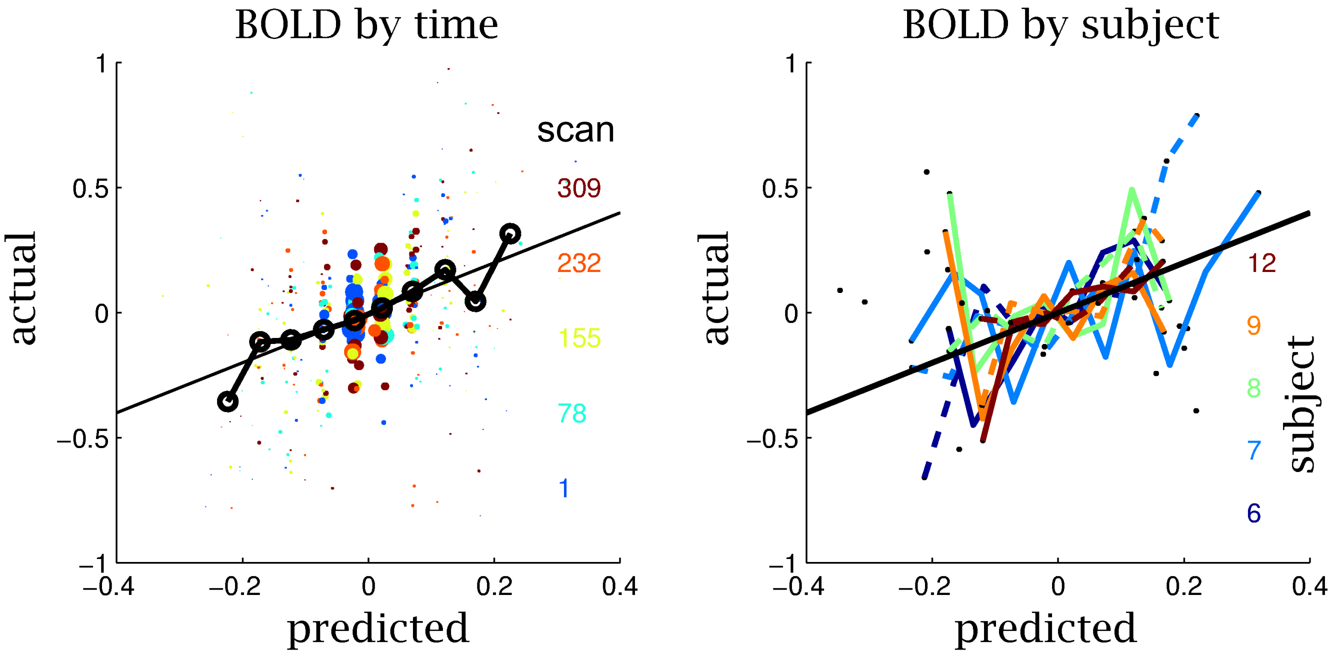


Assessment of model fitness

A. Plot of predicted (according to the advantage learning model) vs observed (adjusted) fMRI data in the ventral striatum during the instrumental task shown for CS and UCS regressors separately for early (scans 1 - 192) and late (scans 193 - 385) phases of the experiment. The solid blue line depicts y=x. The predicted data for each subject was rank ordered, and the same rank order was applied to the observed data. This is then averaged across subjects (for 9 out of 12 subjects who showed significant effects at p<0.01). There are no significant differences in the regressions slopes between early and late phases.

B. Residual plots of data from the ventral striatum for early and late phases of the experiment are shown for two typical single subjects. There is no obvious structure in the residuals, indicating that the model provides a good fit to the data at the single subject level.

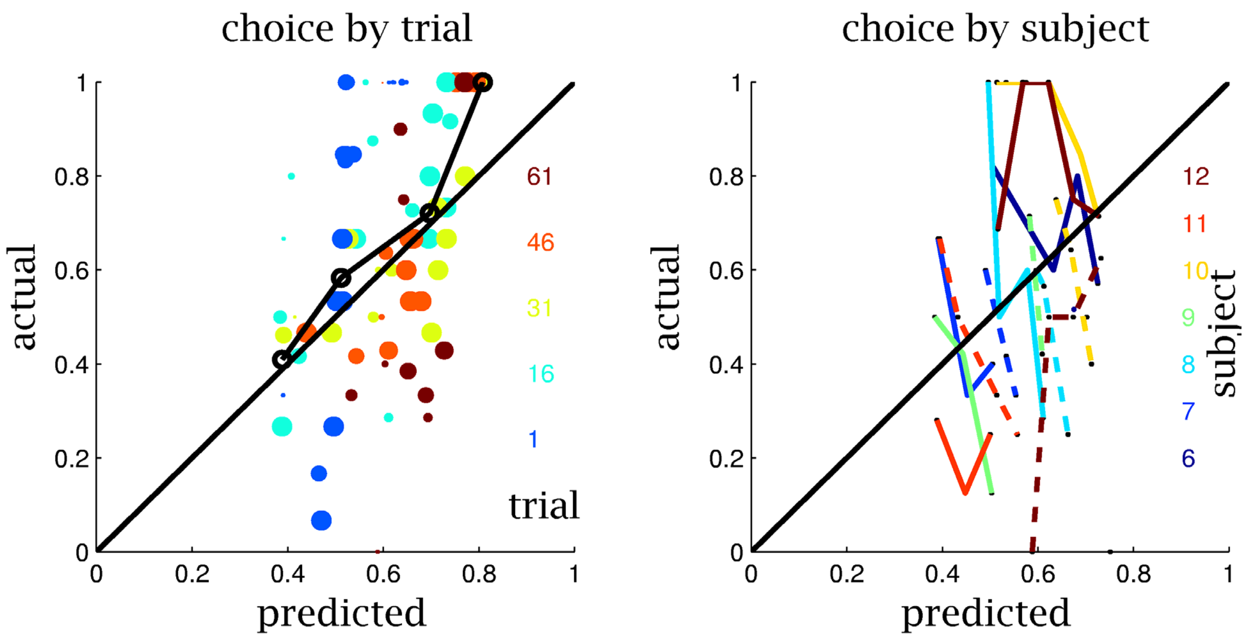
Figure S3



Calibration plot of advantage model predictions vs BOLD signal

Plot shows the relationship between the BOLD signal predicted by the advantage learning model and the observed BOLD signal in the ventral striatum (for the instrumental conditioning task). The left hand plot shows individual subject data from the peak voxels of 9/12 subjects with significant activations in this area (at p<0.01 or above) for five sets of scans (1-77; 78-154, etc., shown by color), together with the overall group averages (black line with circles) and the perfect calibration line (thin straight line). The right hand plot shows data for a few subjects (for clarity), shown by color, and for a division of the 385 scans into 1-192 (solid) and 193-385 (dashed).

Figure S4



Calibration plots of advantage model predictions vs behavior

Plot shows the calibration relationship between the advantage learning model predictions of choice probabilities across trials and subjects, and the actual choices they made. The left hand plot shows individual subject data for five sequential sets of trials (1-15; etc., shown by color), together with the overall group averages (black line with circles) and the perfect calibration line (thin straight line). The right hand plot shows data for a few subjects (for clarity), shown by color, and for a division of the trials into 1-35 (solid) and 36-end (dashed).