# Supporting Information

Burke et al. 10.1073/pnas.1003111107

## SI Materials and Methods

**Participants.** All participants were fluent speakers of English and had normal or corrected-to-normal vision in the scanner. Participants were preassessed to exclude previous histories of neurological or psychiatric illness. All participants gave informed consent, and the Local Research Ethics Committee of the Cambridgeshire Health Authority approved the study. To minimize error trials during scanning, participants learned the timings and sequence of task events (for 20 training trials per condition with stimuli not used in the experiment) no more than 7 d before scanning. During the training period, black and white portrait photographs were taken of each participant against a plain white background at a fixed distance of 2 m. The images were cropped to 100 × 100 pixels and adjusted to have equal luminance. During training, participants were instructed that they would be taking part in a social experiment with two players. They were instructed that they would be able to observe the behavior of another player but that the other player would not be able to observe them. When participants arrived at the scanner, an experimental confederate arrived a little later. The participants were gender matched to confederates. Confederates and participants sat together in the waiting area of the MR facility and went through the same procedures with regards to filling in forms, reading task instructions and being checked for metals. After these preliminary procedures, one research team member led the confederate into another room, where another computer was present. Another member of the research team led the true participant into the scanner. After scanning, the exit of the confederate from the facility was timed to coincide with the debriefing of the true participant (who was sat in the waiting area). The confederates never actually performed the task (except to familiarize themselves with the experiment), and the behavior of what the true participant believed to be the other player was controlled by a computer and kept constant across participants. There was very little difference in the performance of the computer over the trial types (Fig. S1), indicating that the differential participant performance according to trial type was a function of the amount of social information available.

**Participant Payment.** Participants were paid according to the total points accumulated during all sessions of the task, which were converted to British pounds sterling at a rate of 30 points to the pound. In accordance with local payment protocol, participants also received 20 pounds for participating regardless of task performance. The average participant payment was 52 pounds.

**Data Acquisition.** Scanning took place at the Medical Research Council's Cognition and Brain Sciences Unit (MRC-CBU), Cambridge, United Kingdom. The task was projected on a display, which participants viewed through a mirror fitted on top of the head coil. We acquired gradient echo T2*-weighted echo-planar images (EPIs) with blood-oxygen-level–dependent (BOLD) contrast on a Siemens Trio 3 Tesla scanner (slices/volume, 33; repetition time, 2 s). Depending on performance of participants, 280–350 volumes were collected in each session of the experiment, together with five "dummy" volumes at the start and end of each scanning session. Scan onset times varied randomly relative to stimulus onset times.

A T1-weighted MP-RAGE structural image was also acquired for each participant. Signal dropout in basal frontal and medial temporal structures resulting from susceptibility artifact was reduced by using a tilted plane of acquisition (30° to the anterior commissure-posterior commissure line, rostral > caudal). Imaging parameters were the following: echo time, 50 ms; field of view, 192 mm. The in-plane resolution was 3 × 3 mm, with a slice thickness of 2 mm and an interslice gap of 1 mm. High-resolution T1-weighted structural scans were coregistered to their mean EPIs and averaged together to permit anatomical localization of the functional activations at the group level.

**Image Analysis.** We used a standard rapid-event–related fMRI approach in which evoked hemodynamic responses to each event type are estimated separately by convolving a canonical hemodynamic response function with the onsets for each event and regressing these against the measured fMRI signal (1, 2). This approach makes use of the fact that the hemodynamic response function summates in an approximately linear manner over time (3). By presenting trials in strictly random order and using randomly varying intertribal intervals, it is possible to separate out fMRI responses to rapidly presented events without waiting for the hemodynamic response to reach baseline after each single trial (1, 2).

Statistical parametric mapping (SPM5; Functional Imaging Laboratory, University College London, available at www.fil.ion.ucl.ac.uk/spm/software/spm5) served to spatially realign functional data, normalize them to a standard EPI template and smooth them using an isometric Gaussian kernel with a full-width at half-maximum of 8 mm. Onsets of stimuli and outcomes were modeled as separate delta functions and convolved with a canonical hemodynamic response function. Participant-specific movement parameters were modeled as covariates of no interest. Linear contrasts of regression coefficients were computed at the individual subject level and then taken to group-level *t* tests.

**Computational Models.** We adapted a basic Q learning algorithm that has been previously shown to account for instrumental choice in probabilistic reward-learning tasks (4, 5). Generally, for a given binary choice between two stimuli (A and B), the standard Q learning model estimates the expected value of choosing A or B. Whenever an outcome is observed for choosing a particular stimulus at time *t*, a prediction error (δ) (corresponding to the realized minus the expected outcome) is computed. The Q value associated with that stimulus is updated accordingly by multiplying the prediction error by the learning rate (α). At the start of a session, the Q values associated with each stimulus were set to zero. If, for example, on the first trial the subject chose stimulus A and received an outcome (r) of 10 points, the prediction error δ would be given by $\delta(t) = r(t) - Qa(t)$. The value of stimulus A would then be updated according to $Qa(t+1) = Qa(t)+\alpha*\delta(t)$. The probability of the model subsequently selecting a stimulus was determined using the softmax function (6). The softmax function computes a probability of selecting a particular stimulus from a pair according to the ratio of the Q values associated with each stimulus and parameter β (the inverse temperature, which captures the degree of variability in choices). The softmax function has been shown to provide a good approximation of binary choice in previous experiments (4).

***Full observational learning.*** During full observational learning (i.e., when the action and outcome of the other player was observable), the standard Q learning algorithm was modified by incorporating a two-stage update process per trial (Fig. S3). The first update occurs during the "observation stage" and the second during the "action stage" (Fig. 1*A*). As such, the first update (after the observation stage) occurs at $t+0.5$, halfway through the trial. Upon observing an outcome received by the other player, the scanned

participant is assumed to experience an observational reward prediction error ($\delta^S$), according to the outcome received by the confederate ($r^S$) minus the Q value associated with that stimulus. This trial-by-trial observational outcome prediction error was entered as a parametric modulator at the onset of the other player's outcome. The scanned participant is able to learn from the reinforcement received by the other player by multiplying the social reward prediction error ($\delta^S$) by the observational learning rate ($\alpha^S$), capturing the degree to which participants are able to learn from outcomes that are not directly experienced. This update results in an observationally-updated Q value at time $t+0.5$ (denoted by $Q^S$ in Fig. S1 for display purposes).

At the choice of the participant (during the action stage), the probabilities of choosing a particular stimulus are modeled using the softmax function, taking the observationally updated Q values as arguments. Upon receipt of individual outcome ($r$), an individual reward prediction error ($\delta^i$) is computed by subtracting the previous observationally updated Q value ($Q^S$) from the individual outcome ($r$). These trial-by-trial values were entered into the general linear model as a parametric modulator at the onset of the participant's outcome. The Q values are then updated according to the standard algorithm by multiplying $\delta^i$ with the individual learning rate ($\alpha^i$).

***Action imitation learning.*** When only the action of the other player is observable, learning can be modeled with the incorporation of action prediction errors with the standard Q learning algorithm (Fig. S4). The protocol follows the two-stage update procedure outlined previously. At the start of the observation stage, the participant in the scanner has Q values associated with each stimulus on the screen, and therefore has some probability of choosing each stimulus according to the softmax function. For example, at the start of a session when no learning has occurred the ratio of the two stimulus values gives a 0.5 probability of choosing a particular stimulus. When the confederate goes on to make a choice, the action prediction error ($\delta^{action}$) is given by the actual choice minus the probability of choice associated with that stimulus from softmax. Because the actual choice is always 1 or 0, action prediction errors are always in the positive domain. For example, if at the start of the session the probability associated with choosing stimulus A is 0.5 (as no learning has occurred and the Q values for the stimuli are equal), the action prediction error would be [action (a) = 1] − [probability of choosing (a) = 0.5] = 0.5. The degree to which the participant incorporates this information in driving his subsequent choice behavior is controlled by an imitation factor ($\kappa$) analogous to the learning rate in the standard algorithm. Therefore, the probability of the participant subsequently choosing A is $P(a)_{(t+0.5)} = P(a)_{(t)} + \kappa^* \delta^{action}$. Conversely, the probability of choosing B is simply $1 - P(a)$. Although no actual outcome has been observed, after a number of trials the ratio of the stimulus values is inferred. This ratio of action probabilities drives the scanned participant's choice in an analogous fashion to softmax. When the scanned participant subsequently chooses and receives an outcome from a particular stimulus, the participant computes an individual reward prediction error and update the Q values according to the standard algorithm, thereby refining value estimations.

***Individual learning.*** On individual learning trials, the choice and outcome of the confederate player were not observable. On such trials, at the time of the confederate's choice during the observation stage both stimuli were surrounded by white rectangles, making it impossible for the scanned participant to determine which was chosen. At the time of the confederate's outcome, a scrambled image was displayed at the same location. On these trials, participants were required to learn from only their received reinforcements, and the standard Q-learning algorithm was used to model this.

To generate the regressors for the novel prediction error signals, the free parameters in each model ($\alpha^i$, $\alpha^S$, $\kappa$, and $\beta$) were adjusted to maximize the likelihood of observing each participant's choices given the respective model according to $L = \prod^N_{n=1}\prod^T_{t=1} P_{(choice,n,t)}$, where N is the number of participants; T is the number of trials per participant, and $P_{(choice,n,t)}$ is the likelihood of choice made by participant n at trial t given the model. MATLAB (Mathworks) was used to find the parameters maximizing the likelihood L, with values of the parameters searched in increments of 0.01 from 0 to 1, the results of which can be seen in Fig. S5. In such a manner, individual, session-specific regressors for theoretical social, individual, and action prediction errors were generated and subsequently tested for covariation with brain signals.

**Behavioral Results.** As noted in the main text, there was a significant increase in participants' performances with increasing amounts of observable information [ANOVA, $F(2,21) = 11.305$, $P < 0.001$]. For example, participants chose the correct stimulus at a significantly higher rate when they were able to observe the actions and outcomes of the confederate compared with when only confederate actions were observable ($P < 0.01$). In turn, the participants did significantly better when actions were observable compared with learning without any observable information ($P < 0.05$).

When the data were split according to gain and loss sessions (Fig. S2), the monotonic increase in performance with observable information was preserved ($P < 0.001$ in gain sessions, $P < 0.02$ in loss sessions). In both gain and loss scenarios, participants chose the correct stimulus at a significantly higher rate and received significantly more points in the fully observable condition compared with the individual learning baseline. However, in gain sessions, there was a significant difference between fully observable and action-only conditions for both correct choices and reward received ($P < 0.001$ and $P < 0.001$ respectively) but not between action-only and individual conditions. In loss session, this pattern was reversed (no significant differences between fully observable and action-only conditions for both performance and points received). However, a two-way ANOVA with sessions (gain/loss) and trial type (fully observable, action-only, and individual learning conditions) as factors failed to show a significant interaction ($P = 0.42$). In summary, it appears that participants were able to increase their performance by learning from the actions of the confederate more efficiently in loss sessions as opposed to gain sessions. Previous research has suggested that learning from positive and negative reinforcement may be mediated by opposing neural systems (7) and individual differences in the effectiveness of learning from rewards and punishments have been documented (8). One possibility underlying the differences we observed could be that learning through observing the actions of others is more effective in situations where an avoidance response is necessary, numerous examples of which exist in the literature (9, 10). Indeed, research in foraging theory predicts that observational learning should proceed more readily in resource-poor environments or for the learning of predator avoidance mechanisms (11).


1. Dale AM, Buckner RL (1997) Selective averaging of rapidly presented individual trials using fMRI. *Hum Brain Mapp* 5:329–340.
2. Josephs O, Henson RN (1999) Event-related functional magnetic resonance imaging: Modelling, inference and optimization. *Philos Trans R Soc Lond B Biol Sci* 354: 1215–1228.
3. Boynton GM, Engel SA, Glover GH, Heeger DJ (1996) Linear systems analysis of functional magnetic resonance imaging in human V1. *J Neurosci* 16:4207–4221.
4. Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 442: 1042–1045.
5. Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. *Science* 310:1337–1340.
6. Luce RD (1986) *Response Times: Their Role in Inferring Elementary Mental Organisation* (Oxford University Press, New York).

7. Daw ND, Kakade S, Dayan P (2002) Opponent interactions between serotonin and dopamine. *Neural Netw* 15:603–616.
8. Frank MJ, Woroch BS, Curran T (2005) Error-related negativity predicts reinforcement learning and conflict biases. *Neuron* 47:495–501.
9. Mineka S, Davidson M, Cook M, Keir R (1984) Observational conditioning of snake fear in rhesus monkeys. *J Abnorm Psychol* 93:355–372.
10. Olsson A, Phelps EA (2004) Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychol Sci* 15:822–828.
11. Laland KN (2004) Social learning strategies. *Learn Behav* 32:4–14.

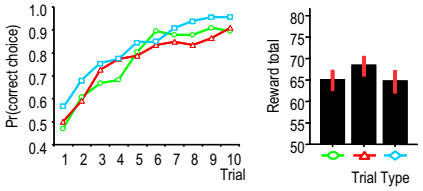


**Fig. S1.** Computer-controlled confederate's behavioral performance was constant across trial type.

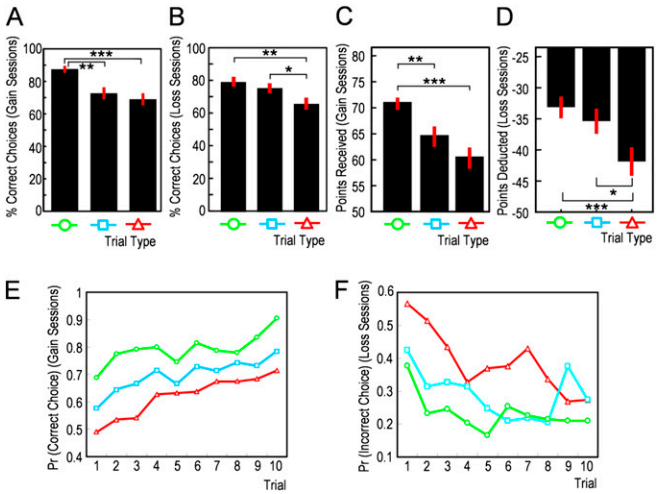


**Fig. S2.** Percentage of correct choices in gain (*A*) and loss (*B*) sessions and points scored by participants in gain (*C*) and loss sessions (*D*), separated according to trial type. (*E*) Probabilities of correct choices on a trial-by-trial basis in gain sessions. (*F*) Probabilities of correct choices on a trial-by-trial basis in loss sessions.

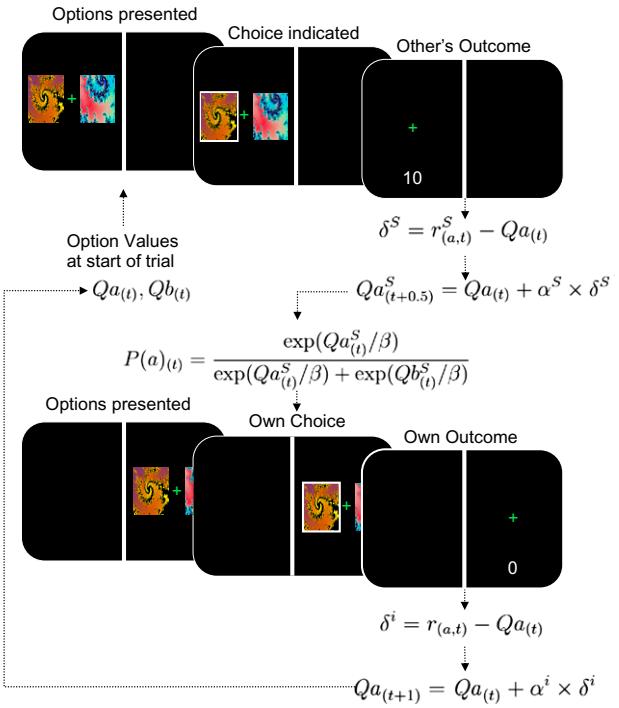


**Fig. S3.** Schematic diagram illustrating the learning model for when all social information is available (i.e., when the action and outcome of the confederate player is observable). The two-stage learning process is illustrated as if stimulus A has been chosen by both confederate and participant, denoted in the lowercase and subscript text. In this particular example, the confederate received 10 points and the participant 0 points for their choices.

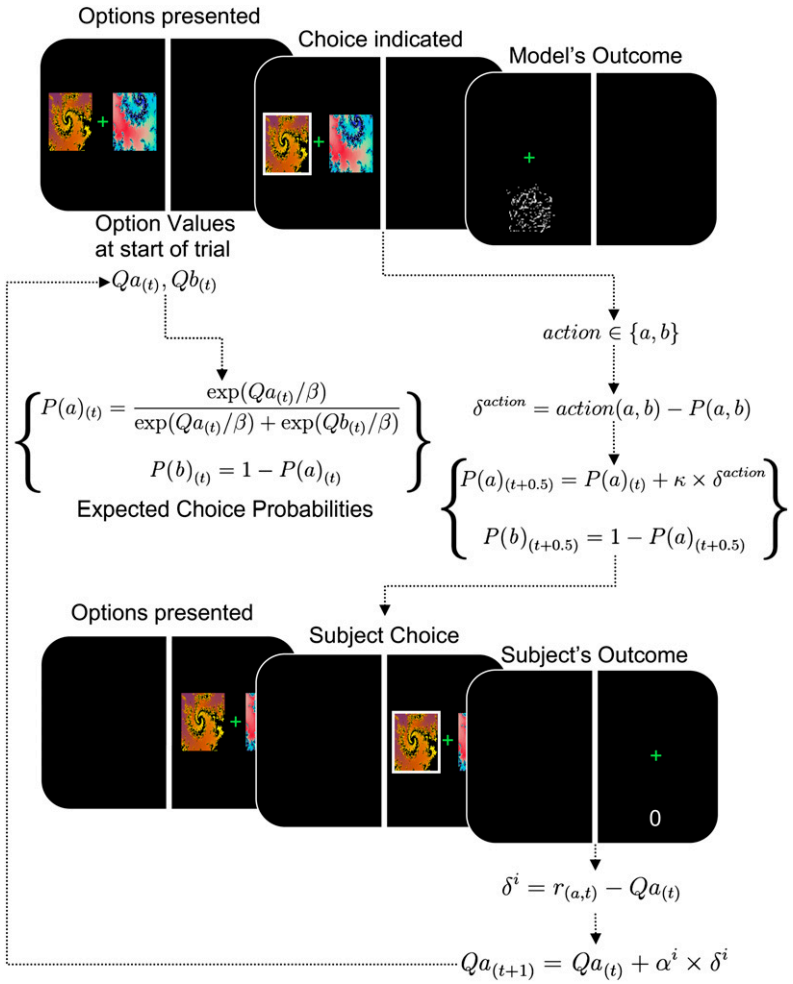


**Fig. S4.** Schematic illustrating the learning model for when only the actions of the other player are observable. In this particular example, both confederate and participant chose stimulus A, denoted in the lowercase and subscript text.

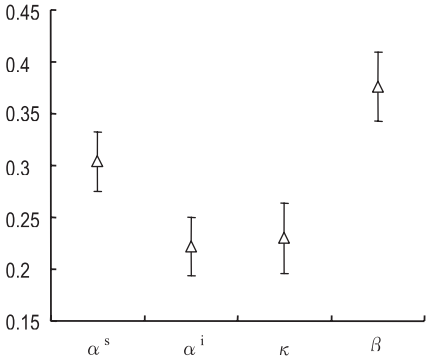


**Fig. S5.** Free parameter values that maximized the likelihood of observing participants' behavioral data given the social learning models.

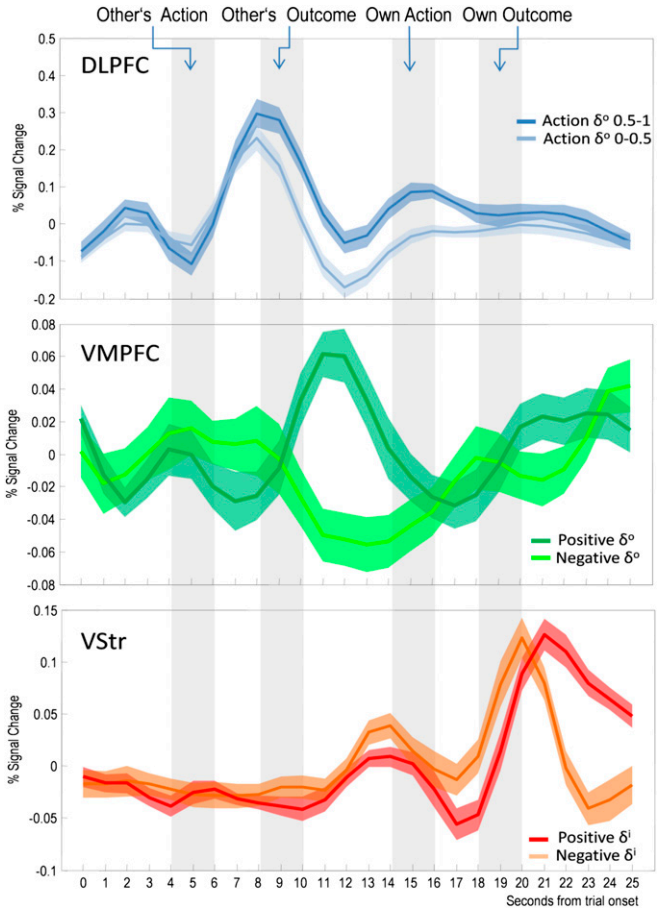


**Fig. S6.** Time course in the three regions of interest (DLPFC, VMPFC, and ventral striatum) over the course of a full trial.

**Table S1. Trial types summarized in terms of the availability of social information to the participant in the scanner**

| Learning type | Other's action | Other's outcome |
|---|---|---|
| Full observational learning | Observable | Observable |
| Action imitation learning | Observable | Hidden |
| Nonobservational learning | Hidden | Hidden |

Colored shapes are the same as those used in the main text.

**Table S2. Locations of significant activation clusters for the action learning, action + outcome learning, and individual learning conditions in a whole-brain analysis**

| Cluster location | MNI X (mm) | MNI Y (mm) | MNI Z (mm) | No. of voxels | Peak z score |
|---|---|---|---|---|---|
| Action learning condition | | | | | |
| R inferior temporal gyrus | 48 | −48 | −15 | 74 | 4.25 |
| R middle occipital gyrus | 39 | −72 | 21 | 27 | 3.98 |
| R inferior occipital gyrus | 39 | −72 | −15 | 16 | 3.52 |
| Full observational learning | | | | | |
| L insula | −42 | −6 | 3 | 12 | 2.93 |
| R medial orbitofrontal cortex | 12 | 45 | −12 | 3 | 2.56 |
| L medial orbitofrontal cortex | −3 | 39 | −15 | 2 | 2.48 |
| Individual learning | | | | | |
| L precuneus | −27 | −60 | 21 | 45 | 5.91 |
| R superior occipital gyrus | 30 | −78 | 0 | 43 | 5.80 |
| L angular gyrus | −48 | −72 | 36 | 35 | 5.60 |
| L middle temporal gyrus | −54 | −3 | −24 | 44 | 5.60 |
| L cerebellum | −21 | −57 | −45 | 13 | 5.54 |
| R cerebellum | 42 | −66 | −42 | 149 | 5.53 |
| R middle temporal gyrus | 60 | 0 | −15 | 30 | 5.30 |
| L paracentral Lobule | −18 | −9 | 66 | 28 | 4.13 |
| L rolandic opercularis | −57 | 15 | 12 | 20 | 3.90 |

Montreal Neurological Institute (MNI) coordinates denote the peak of each cluster. Activations at $P < 0.001$ uncorrected with an extent threshold of 10 voxels are listed. However, for the full observational learning condition, no other activations were observed at the usual threshold of $P < 0.001$ with an extent threshold of 10 voxels. As such, activations at $P < 0.01$ uncorrected with an extent threshold of 0 voxels are listed for this contrast. R, right; L, left.