Behavioral/Systems/Cognitive

Dissociable Brain Systems Mediate Vicarious Learning of Stimulus–Response and Action–Outcome Contingencies

Mimi Liljeholm,[1,2] Ciara J. Molloy,[1] and John P. O'Doherty[1,2]

[1]Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland, and [2]Division of the Humanities and Social Sciences and Computation and Neural Systems Program, California Institute of Technology, Pasadena, California 91125

Two distinct strategies have been suggested to support action selection in humans and other animals on the basis of experiential learning: a goal-directed strategy that generates decisions based on the value and causal antecedents of action outcomes, and a habitual strategy that relies on the automatic elicitation of actions by environmental stimuli. In the present study, we investigated whether a similar dichotomy exists for actions that are acquired vicariously, through observation of other individuals rather than through direct experience, and assessed whether these strategies are mediated by distinct brain regions. We scanned participants with functional magnetic resonance imaging while they performed an observational learning task designed to encourage either goal-directed encoding of the consequences of observed actions, or a mapping of observed actions to conditional discriminative cues. Activity in different parts of the action observation network discriminated between the two conditions during observational learning and correlated with the degree of insensitivity to outcome devaluation in subsequent performance. Our findings suggest that, in striking parallel to experiential learning, neural systems mediating the observational acquisition of actions may be dissociated into distinct components: a goal-directed, outcome-sensitive component and a less flexible stimulus–response component.

Introduction

The ability to acquire novel behaviors through observation of other individuals is at the core of a vast array of human skills, including tool-use, language, and hygiene. But what exactly is it that we learn when we observe others perform, and reap the rewards of, instrumental actions? Psychological theories of behavioral control distinguish between goal-directed learning, characterized by representations of action–outcome contingencies and outcome value, and habit formation, through which actions come to be rigidly and automatically elicited by their stimulus environment. Although substantial behavioral and neural evidence for this distinction has been demonstrated in rodent and human subjects learning from direct experience (Balleine and Dickinson, 1998; Tricomi et al., 2009), very little is known about whether an analogous dichotomy exists for vicarious learning. In particular, it is not clear whether habits can actually be acquired through observation. To test this hypothesis, we developed a novel observational learning task in which the structure of the observed environment encouraged either encoding the specific consequences of alternative actions or a mapping of actions to antecedent conditional cues.

Received Feb. 5, 2012; revised April 26, 2012; accepted May 30, 2012.

Author contributions: M.L. and J.P.O. designed research; M.L. and C.J.M. performed research; M.L. analyzed data; M.L. and J.P.O. wrote the paper.

This work was supported by a grant from the Wellcome Trust and by an NIH grant (DA033077-01) to J.P.O. We thank Sean B. Ostlund for his helpful comments on the manuscript.

Correspondence should be addressed to Mimi Liljeholm, Division of the Humanities and Social Sciences, Computation and Neural Systems Program, MC 228-77, California Institute of Technology, Pasadena, CA 91125. E-mail: mlil@hss.caltech.edu.

DOI:10.1523/JNEUROSCI.0548-12.2012

 0270-6474/12/329878-09$15.00/0

Participants learned, through observation, how to regulate a system of four fluid-filled beakers using a set of four instrumental actions. The beakers were graphically represented on the screen, as were the actions performed and any points gained or lost, by an ostensible observee (for details, see Fig. 1*A*). As long as all beakers had sufficient liquid, the system remained balanced and points corresponding to monetary reward were continuously gained. However, on each trial, the fluid in one of the beakers would drop below a required threshold, and points would be continually lost until that beaker was refilled. The below-threshold drop in a beaker's fluid was always preceded by the onset of one of four stimulus patterns (i.e., cues), appearing at the center of the screen. There were two conditions in the experiment. In the Response–Outcome (R-O) condition, each instrumental action refilled a particular beaker regardless of which cue was presented, so that identification of the relevant intermediate outcome (e.g., refilling beaker 1), combined with knowledge about specific action–outcome contingencies (i.e., action 1 refills beaker 1), indicated which action would restore system balance, thus yielding reward, on any given trial. Conversely, in the Cue–Response (C-R) condition, the identity of the antecedent cue determined which of the four actions would refill the emptied beaker, regardless of which particular beaker had lost its fluid.

Habitual control has been proposed to depend on the incremental formation of stimulus–response associations, void of any representation of specific outcome features (Dickinson et al., 1995; Daw et al., 2005). We hypothesized that, in the C-R condition, by decorrelating actions from specific beaker outcomes, while conditioning monetary reward on the mapping of actions to antecedent cues, we would bias the observer toward habitual performance following observational learning, as assessed by

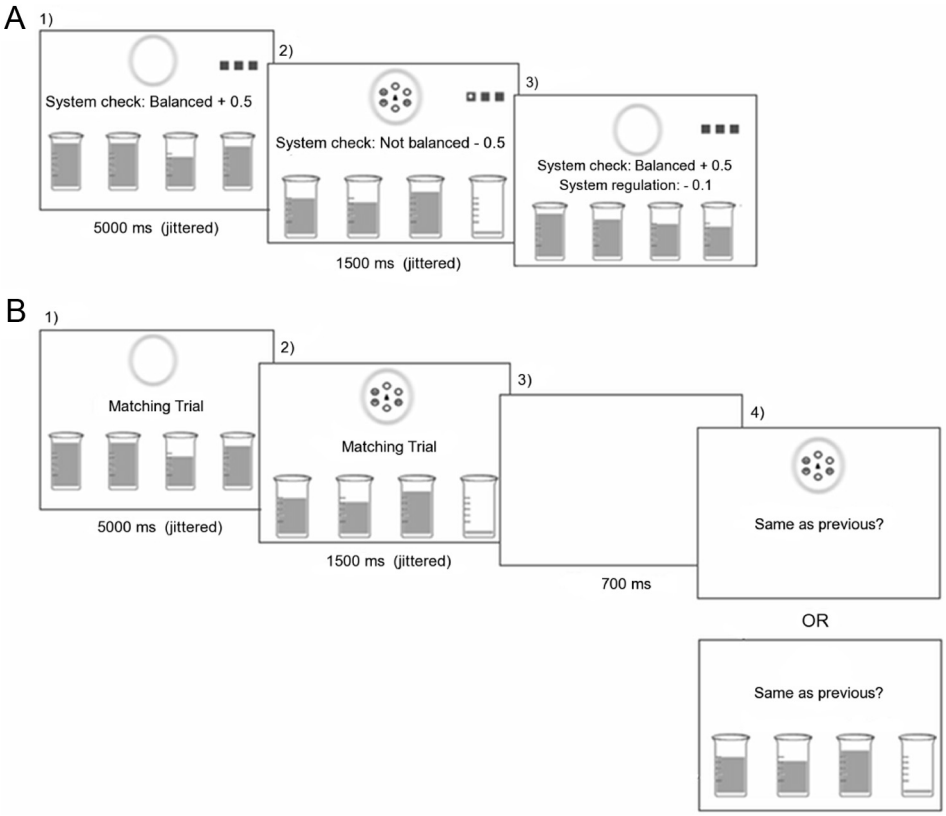

Figure 1. Illustration of observational learning trials. ***A***, Instrumental trial. Participants passively view the screen as an observee is ostensibly regulating the beaker system using the four different actions. The actions performed by the observee (each a 3-press sequence) were graphically represented by a white dot moving across three gray squares in the top right corner of the screen, and any points gained or lost by the observee were indicated by text messages. During the intertrial interval (1), the liquid in the beakers continually fluctuated but remained high, and balance checks occurring at brief random intervals yielded points for system balance. At the trial onset (2), one of four abstract cues appeared, the liquid in one of the beakers dropped to the bottom, and balance checks begin to indicate a loss of points due to system imbalance. After a short time interval, the observee performed the action required to refill the emptied beaker (as indicated by the response graphics). Following completion of the observed action (3), the abstract cue disappeared, the beaker is refilled, a small fee is charged for regulating the system, and balance checks again yield points for system balance. ***B***, Matching trial. The intertrial interval (1) and trial-onset (2) were as in the instrumental task and were followed by a blank screen with a 700 ms duration (3). The final screen (4) showed a matching/not matching cue (top) or set of beakers (bottom) in the C-R and R-O conditions, respectively, together with a query about the match.

outcome devaluation. To investigate differences in neural responses across our two observational learning conditions, we scanned participants with functional magnetic resonance imaging (fMRI) as they passively viewed the instrumental regulation of the beaker system.

Materials and Methods

Participants

Nineteen healthy normal volunteers (11 males and 8 females), recruited locally from the city of Dublin, Ireland, participated in the study. One participant was unsure about whether or not responding for the devalued beaker would lead to a gain in points, and was therefore excluded. An additional three participants (two in the C-R condition and one in the R-O condition) failed to acquire the task after extensive observational training, and were excluded on this basis, leaving a sample size of 15. Written informed consent was obtained from all participants and the study was approved by the Trinity College School of Psychology Research Ethics Committee.

Task and procedure

Each subject participated in both the R-O and C-R condition, with a novel set of four instrumental actions being used in each condition, and with the order of conditions counterbalanced across subjects. Each session (condition) included a response pretraining phase, an observational learning phase, a probe test, a devaluation phase, and a final extinction test, as described below (for a diagram illustrating the chronology of the procedure, see Fig. 2). The response-training phase of the first condition was conducted outside the scanner in a separate testing room. Once performance in this phase reached criterion, participants were transferred to the scanner and remained there throughout all subsequent stages of the experiment. The entire experiment lasted for ~2 h, with 1.5 h being spent in the scanner, and with ~60 min of active scanning during the observational learning phases and devaluation tests in each condition.

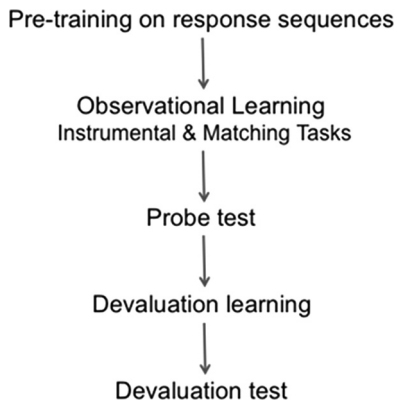

Figure 2. Diagram of procedure, showing each phase in one condition of the experiment. The entire procedure was repeated, with novel response sequences, for the alternate, within-subjects condition.

General instructions. At the start of the experiment, participants were presented with a cover story describing the beaker system and the task.

Briefly, participants were informed that monetary points could be earned as long as each of four beakers were filled up with liquid but that, whenever one of four abstract cues appeared, one of the beakers would be emptied and points would be lost until that beaker was refilled using one of four instrumental actions. They were further told that they would learn about the relationships between the cues, actions, and beakers (i.e., exactly how the system worked) by observing "someone else" perform the task, and that although they would not win or lose any points during this observational learning phase, they would eventually be given the opportunity to regulate the system themselves, for personal monetary gain. Finally, participants were instructed that they would be in one of two possible conditions—one in which each instrumental action refilled a particular beaker regardless of which cue was presented, and another in which the identity of the cue determined which of the four actions was required to refill an emptied beaker, regardless of the identity of that beaker. They were told that part of their task was to determine which of the two conditions they were in (piloting indicated that, without explicit instructions about these two possible structures, participants were unable to acquire the task within the time limits of the experiment).

Response pretraining. Before the observational learning phase, participants received pretraining on the four instrumental actions (each of which consisted of a three-press sequence). During this training, keypress sequences were represented by a white dot moving across three gray squares horizontally aligned at the center of the screen. Initially, participants viewed and then immediately attempted to replicate each response sequence, with feedback (i.e., correct/incorrect) given on each trial. After a total of five correct replications of each response sequence, they proceeded to a retrieval phase, in which they had to generate each unique sequence at least five times without any visual aids, again with feedback given at the completion of each three-press sequence. Participants were allowed to repeat these two phases as many times as they wanted to, knowing that they would have to use the actions to earn monetary reward in a subsequent phase.

Observational learning phase. The instrumental task was as described in the Introduction, above, and is illustrated in Figure 1*A*. Note that, in addition to the increase or decrease in monetary points based on system balance, there was a small cost for regulating the system. This response cost was included to ensure that, during test, participants would not respond simply based on any reinforcement intrinsic to executing the correct response. Critically, the stimulus materials presented in Figure 1*A* were identical across the two conditions: our manipulation consisted entirely of differences in the contingencies between cues, actions, and beaker outcomes. Specifically, in the R-O condition, each observed instrumental action (i.e., response sequence) was paired, across trials, with the refilling of a particular beaker but was decorrelated from the various cue identities. Conversely, in the C-R condition, each instrumental action was paired across trials with a particular antecedent cue but decorrelated from the refilling of any particular beaker. The rationale behind this manipulation was that the R-O condition would encourage encoding of the relationships between actions and specific outcomes, thought to mediate goal-directed performance, while the C-R condition would force participants to rely on the formation of stimulus-response associations (i.e., the incremental mapping of actions to discriminative cues) frequently argued to support habitual performance.

Of course, while the actual features of the visual display were identical across conditions, the relevant features (i.e., those to which participants were encouraged to attend) were quite different. To control for visual processes involved in attending to the abstract cues versus the beakers, a matching task was block-interleaved with the instrumental task (Fig. 1*B*) during observational learning. In matching blocks, the intertrial intervals and trial onsets were exactly as in the system balance task, except that there were no text messages indicating points for system balance and no response-key graphics; instead, the words "Matching Trial" were continuously displayed at the center of the screen. Following the appearance of the abstract cue and emptying of the relevant beaker, a white screen was displayed, followed by a depiction of either an abstract cue (C-R condition) or a set of beakers (R-O condition), together with a query about whether the currently shown cue/beaker set matched that on the previous screen. On 70% of these trials, participants were instructed to merely observe as a yes/no response to the matching query was indicated on the screen. However, on the remaining 30% of trials, they had to provide the answer themselves. Critically, participants did not know whether they would observe or provide the response until the matching query had appeared, ensuring that the to-be-matched display (Fig. 1*B*, second screen) was attended on all matching trials. The observational learning phase consisted of four blocks of trials, with each block being further divided into one block of 24 instrumental trials and a second block of eight matching trials (Fig. 1*A*,*B* respectively). Instrumental and matching blocks were separated by screens indicating the type of upcoming block; the order of trials within each type of block was randomized.

Probe test. A probe test, consisting of four trials with each action for a total of 16 randomly ordered trials, was administered immediately following observational learning to assess acquisition. The probe trials were identical to the observational learning trials, except that the participant was now performing the actions themselves. If, on any given trial, a participant failed to perform the action required to refill the currently empty beaker, the system would regulate itself (i.e., a variable interval with mean = 6000 ms). Participants were able to discern whether they had successfully refilled the emptied beaker using the correct action, or whether the system had regulated itself, based on the response cost indicated on the screen; the cost was only incurred on trials in which the participants action refilled the beaker. They were also informed that, just as during observational learning, they would not be actually gaining or losing any of the points displayed during the probe test, but that the purpose of the test was simply to determine how well they had mastered the task.

Devaluation phase. The distinction between goal-directed and habitual performance is most commonly demonstrated by changing the value of a particular action outcome. For example, in an animal conditioning paradigm, if rats have learned that one action results in sucrose pellets while another results in grain, and if sucrose pellets are subsequently devalued by pairing them with an aversive event or by feeding the animal on them to satiety, response rates decrease for sucrose but not for grain. The selective decrease indicates that behavior is sensitive to the subjective value of the anticipated outcome, as well as to the action–outcome contingency, and thus that performance is goal-directed. Conversely, the persistent execution of an action after its outcome has been devalued is a defining feature of habitual performance.

In the current study, following initial observational learning, we devalued one of the four beakers by degrading its relationship to the ultimate goal of gaining monetary reward. If participants were indeed relying largely on a stimulus–response strategy in the C-R condition, this change in beaker value should have a significantly lesser influence on subsequent instrumental performance in this condition than in the R-O condition. Specifically, in the devaluation phase, participants were instructed that the system had changed such that one of the beakers was no longer relevant for system balance, which would be maintained, and continue to yield points, even when the liquid in this beaker dropped below threshold. They then observed as the system regulated itself (i.e., no actions were performed by either the observee or the participant) across 16 trials (4 with each beaker) to identify the devalued beaker. Again, participants were told that they would not lose or gain any of the displayed points during this phase.

Extinction test. Finally, having correctly identified the devalued beaker, participants were given the opportunity to regulate the system themselves for personal monetary reward. During this test phase, all text messages, indicating gains or losses, system balance checks, and regulation charges, were covered up to prevent additional learning (i.e., simulating extinction). Participants were instructed that, despite these gray strips, they should assume that all was exactly as they had learned before; that is, they would still lose points whenever the system was not balanced, there was still a cost for regulating the system, and the previously identified irrelevant beaker was still irrelevant for system balance. Importantly, given the small charge for regulating the system, refilling the now irrelevant beaker actually resulted in a net loss. The test phase consisted of 11 trials with each beaker, including the devalued one, for a total of 44 trials.

Imaging protocol

Previous imaging studies on experientially acquired instrumental actions have provided evidence for a dissociation between human goal-directed and habitual performance at the level of the striatum, with anterior caudate contributing to goal-directed performance (Tanaka et al., 2008) and the posterior caudate/putamen (Valentin et al., 2007; Tricomi et al., 2009) contributing to habitual control. In addition, the inferior parietal lobule (IPL) and ventromedial prefrontal cortex have both been implicated in human goal-directed performance: most notably, respectively, in action-outcome contingency learning (Liljeholm et al., 2011) and outcome devaluation (Valentin et al., 2007). We predicted that these areas would also be differentially recruited across conditions during observational learning. Since the graphics of response sequences ostensibly reflected the actions of another individual performing the task, we further hypothesized that effects would emerge in areas previously found to be active during action observation, including the premotor cortex, primary motor cortex (M1), and inferior and superior parietal lobules (Caspers et al., 2010).

Acquisition and preprocessing. We used a 3 tesla scanner (MAGNETOM Trio; Siemens) to acquire structural T1-weighted images and T2*-weighted echoplanar images (repetition time, 2.65 s; echo time, 30 ms; flip angle, 90°; 45 transverse slices; matrix, 64 × 64; field of view, 192 mm; thickness, 3 mm; slice gap, 0 mm) with BOLD contrast. To recover signal loss from dropout in the medial orbitofrontal cortex (O'Doherty et al., 2002), each horizontal section was acquired at 30° to the anterior commissure—posterior commissure axis.

Image processing and analyses were performed using SPM5 (http://www.fil.ion.ucl.ac.uk/spm). The first four volumes of images were discarded to avoid T1 equilibrium effects. Remaining volumes were corrected for differences in slice acquisition, realigned to the first volume, spatially normalized to the Montreal Neurological Institute (MNI) echoplanar imaging template, and spatially smoothed with a Gaussian kernel (8 mm, full-width at half-maximum). We used high-pass filter with cutoff = 128 s.

Imaging analysis. We specified a separate linear model for each subject, with 32 regressors, one for each instrumental action, in each of four blocks of observational instrumental learning for each of the two conditions (i.e., C-R and R-O). Two regressors accounting for the matching trials in the C-R and R-O condition and six regressors accounting for the residual effects of head motion were also included. For instrumental regressors, we modeled the period from the onset of the abstract cue to the final press in the response sequence performed by the observee (Fig. 1*A*). For the matching trials, we modeled the period between the onset of the cue and the onset of the matching screen (Fig. 1*B*). All regressors were convolved with a canonical hemodynamic response function. Group-level statistics were generated by entering contrast estimates for each condition into between-subjects analyses assessing the interactions [R-O > C-R (instrumental > matching)] and [C-R > R-O (instrumental > matching)].

Small volume corrections (svc) were performed on three a priori regions of interest using a 10 mm sphere. We used coordinates identified in previous studies of goal-directed [anterior caudate: −15/15, 9, 15) (Liljeholm et al., 2011)] and habitual [tail of caudate: ±27, −36, 12 (Valentin et al., 2007); posterior putamen: ±33, −44, 0 (Tricomi et al., 2009] learning. Unless otherwise indicated, all other effects were reported at $p < 0.05$, using cluster size thresholding (cst) to adjust for multiple comparisons (Forman et al., 1995). AlphaSim, a Monte Carlo simulation (AFNI) was used to determine cluster size and significance. An individual voxel probability threshold of $p = 0.005$ indicated that using a minimum cluster size of 122 MNI transformed voxels resulted in an overall significance of $p < 0.05$. For display purposes, statistical maps in all figures are shown at an uncorrected threshold of $p < 0.005$.

To eliminate nonindependence bias for plots of parameter estimates, a leave-one-subject-out (LOSO) (Esterman et al., 2010) approach was used, in which 19 GLMs were run with one subject left out in each, and with each GLM defining the voxel cluster for the omitted subject. Spheres (10 mm) centered on the LOSO peaks (identified within ROIs for small volume corrections) were then used to extract mean beta weights for each condition; these were averaged across subjects to plot overall effect sizes.

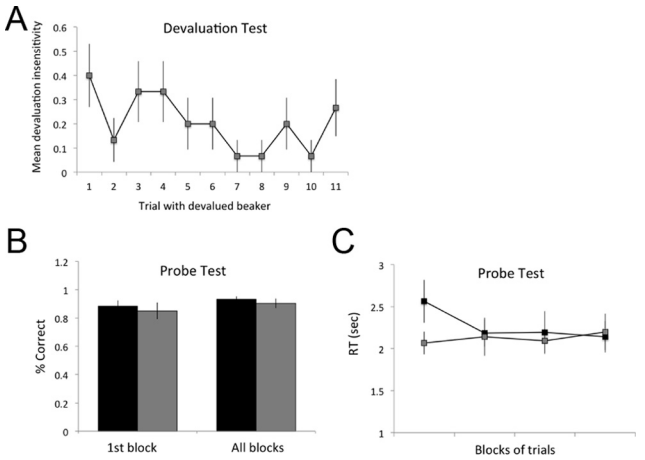

Figure 3. Behavioral results from the devaluation and probe tests, shown for the C-R (gray) and R-O (black) conditions. ***A***, Mean responding, across subjects, in the C-R condition, on each trial with the devalued beaker. ***B***, Mean accuracy on the probe test on the first trial with each action (first block) and across all blocks. ***C***, Mean response times (from cue onset to last element in response sequence) in each block of the probe test for the C-R and R-O conditions. Error bars are SEM.

Results

Behavioral results

Three participants, all of whom were included in our statistical analyses, requested and received additional observational learning trials after completing the probe test before proceeding to the devaluation phase; one of these participants requested additional trials in both the C-R and R-O conditions, while the remaining two only did so in the C-R condition. No scanning was performed during the additional observational learning trials.

There were no differences between conditions in participant's ability to identify the devalued beaker at the end of the devaluation procedure; all participants successfully identified the devalued beaker in both conditions. Performance in the final extinction test indicated that our novel paradigm did indeed produce devaluation insensitive performance under observational conditions: having learned through observation how to regulate the system, and having correctly identified the devalued beaker, participants responded on trials with the devalued beaker at a significantly higher rate in the C-R condition (mean rate = 0.21; SEM = 0.06) than in the R-O condition (mean rate = 0; $p < 0.005$). Indeed, in the R-O condition, not a single participant refilled the irrelevant beaker, suggesting that the disadvantage of doing so given the response cost was apparent. Mean responding on each trial with the devalued beaker in the C-R condition is shown in Figure 3*A*.

Although there was no significant interaction between training conditions and counterbalancing order in the devaluation test ($p = 0.07$, $F_{(1,13)} = 3.76$), there was a clear trend: the degree of devaluation insensitivity in the C-R condition was greater when this was the first condition (mean = 0.31, SEM = 0.1, $n = 8$) than when it followed the R-O condition (mean = 0.11, SEM = 0.04, $n = 7$), suggesting that previous exposure to the instrumental task or to the devaluation procedure improved devaluation sensitivity. Of course, a similar effect may have been present for the R-O condition had it not been for the apparent floor effect. Since, across participants, there was not a single response for the devalued beaker in the R-O condition when it was presented first, there was no room for any improvement when this condition was presented second. More generally, we note

that this floor effect might obscure a potentially even greater difference in devaluation sensitivity between our two training conditions. Regardless, the difference between conditions in devaluation sensitivity was equally significant across the two orders of presentation; both $ps = 0.02$.

The difference between conditions was reliable on the very first trial with the devalued beaker (C-R condition mean = 0.4; SEM = 0.13; $p < 0.01$), suggesting that it cannot be attributed to differential learning occurring during the test phase. Nor can the difference in outcome devaluation sensitivity between conditions be attributed to differences in difficulty levels. In the probe test (data shown in Fig. 3*B*,*C*) conducted immediately after the observational learning phase, mean accuracy did not differ significantly between the C-R and R-O conditions on the first trial of performing each action ($p = 0.55$) nor across all probe trials ($p = 0.29$). Probe test response times (RTs), measured from the onset of the cue to the last element of the response sequence differed only in the first block of four probe trials (1 with each action, block randomized), such that RTs were slightly longer in the R-O than in the C-R condition in this block, $p < 0.05$. When collapsing across all probe trials, there was no difference between conditions on this measure ($p = 0.49$). There was no influence of counterbalancing order on the probe test measures in either condition (all $ps > 0.2$). Finally, there was no significant correlation between the degree of insensitivity to devaluation in the extinction test and the level of accuracy during the probe test ($r = -0.03$, $p = 0.89$).

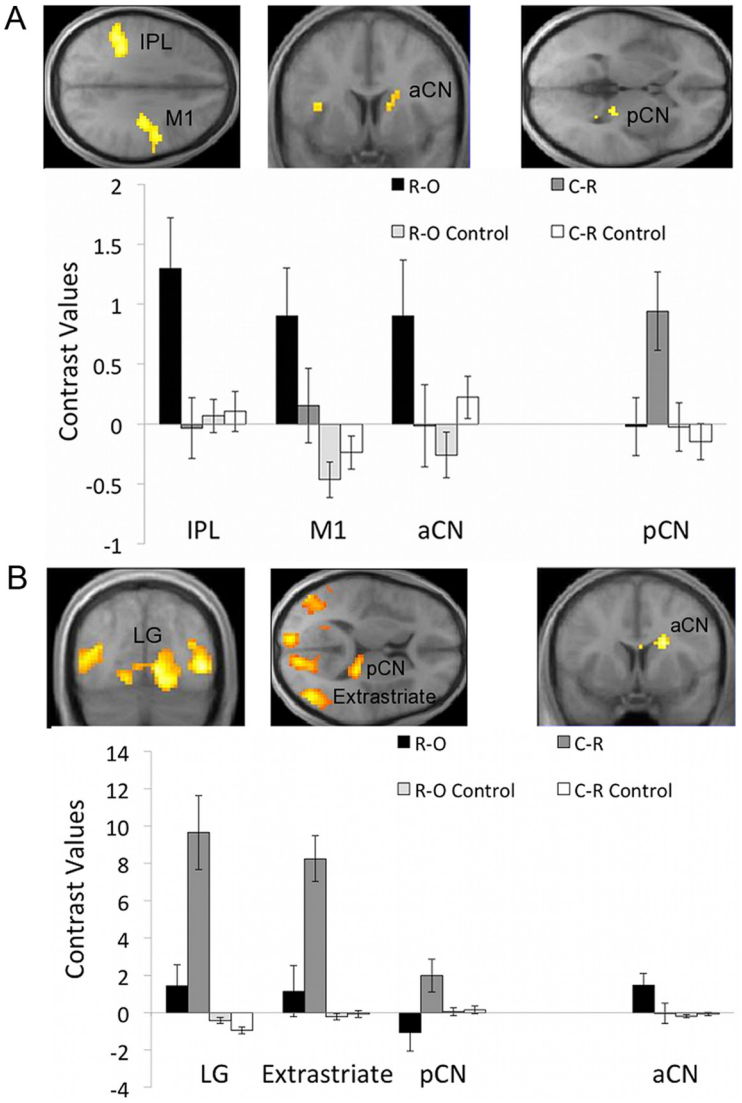

Figure 4. fMRI results for interaction contrasts assessing differences between the R-O and C-R conditions across the instrumental task and the control (matching) task. ***A***, Maps of the *t* statistics for tests of neural activity during the first block (25%) of observational learning trials, showing effects for the [R-O > C-R] contrast in M1, the IPL, and the aCN, and for the [C-R > R-O] contrast in the pCN. ***B***, Map of the *t* statistics for tests assessed across all trials of observational learning, with the [C-R > R-O] contrast revealing effects in the LG, the extrastriate cortex, and the pCN, and with the [R-O > C-R] contrast yielding effects in the aCN. Bar graphs show contrast values at LOSO coordinates; error bars are SEM.

Imaging results

Experimental versus control conditions

A hallmark feature of goal-directed performance is that it typically dominates during early learning, with what is commonly referred to as undertraining (Dickinson et al., 1995; Balleine and Dickinson, 1998). In addition, several imaging studies on instrumental reward learning have reported training-dependent decreases in neural activity in the inferior parietal lobule, medial frontal gyrus, and caudate nucleus (Delgado et al., 2005; Koch et al., 2008), areas commonly associated with goal-directed instrumental performance (Tanaka et al., 2008; Liljeholm et al., 2011). To accommodate these potential temporal dynamics, we begin our imaging analyses with assessing differences between conditions during the earliest block of observational learning, specifically the first quartile of trials (Fig. 4*A*). For the test of activity that was greater in the R-O than the C-R condition, we found effects in the supramarginal gyrus of the left IPL ($x, y, z = -48, -42, 33$; cst), the M1 ($x, y, z = 36, -15, 36$; cst), and in the dorsal anterior caudate (aCN; $x, y, z = 15, 6, 18$; svc). The test for activity that was greater in the C-R than the R-O condition revealed significant effects only in the right posterior caudate (pCN; $x, y, z = 24, -30, 6$; svc). In contrast, when neural effects were assessed across all training blocks (Fig. 4*B*), the [R-O > C-R] test only yielded effects in the aCN ($x, y, z = 18, 9, 24$; svc), while the reverse [C-R > R-O] test revealed extensive effects in the extrastriate cortex ($x, y, z = 45, -72, 9$; cst) and lingual gyrus (LG; $x, y, z = 12, -69, 0$; cst), as well as in the dorsomedial frontal cortex (DMFC; $x, y, z = 9, 49, 32$; cst) and, again, in the pCN ($x, y, z = 21, -30, 9$; svc).

As the difference between corresponding matching control conditions was subtracted from each of the contrasts reported above, it is unlikely that the effects reflected differences in visual attention. Instead, we attribute these results to the differential recruitment of areas involved in representing the goals of observed actions and those that support the mapping of observed actions to eliciting stimuli. As can be seen in Figure 4, contrast values for the two control conditions did not differ at all, or differed in a direction opposite to that observed for experimental conditions. When looking only at the difference between match-

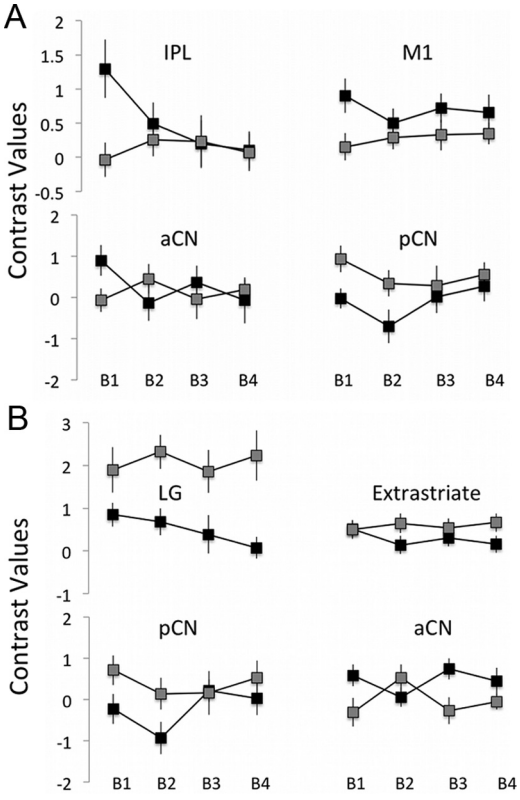

Figure 5. Contrast values in each of the four blocks of observational training (B1–B4), in the R-O (black) and C-R (gray) conditions, estimated at LOSO coordinates based on the statistical maps shown in Figure 4, *A* and *B*, for ***A*** and ***B***, respectively. Error bars are SEM.

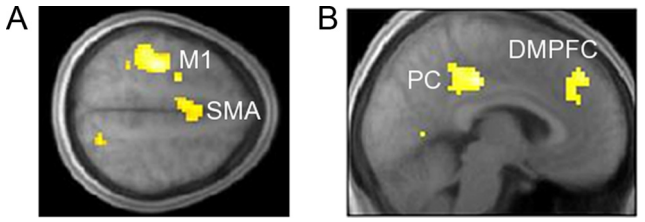

Figure 6. Neural activations for tests of changes across blocks of observational learning trials. ***A***, Map of the statistics for the conjunction test of linearly decreasing activity across training blocks in both conditions, showing effects in the supplementary motor area (SMA), left M1, and left postcentral gyrus. ***B***, A test for decreasing activity across blocks in the R-O condition and increasing activity across blocks in the C-R condition revealed effects in the DMFC and posterior cingulate (PC).

Table 1. Peak coordinates for the disjunction analysis of decreasing activity across training blocks in the R-O but not in the C-R condition

Region	MNI coordinates			
	x	*y*	*z*	*Z*
Supplementary motor area	−9	−0	51	4.43
Anterior cingulate cortex	9	30	24	4.26
DMFC	−3	24	42	3.82
Right inferior frontal gyrus	48	12	21	3.88
Left M1	−39	−18	60	4.62
Right somatosensory cortex	48	−24	45	3.56
Right IPL	30	−45	45	4.74
Left IPL	−45	−33	27	3.88
Left transverse temporal gyrus	−33	−33	12	4.01
Cuneus	15	−90	6	3.52
Right ventral lateral nucleus (thalamus)	12	−9	12	3.81
Left anterior nucleus (thalamus)	−9	−6	15	3.82
Right aCN	12	12	9	3.20

ing control conditions, activity in the cuneus was greater for the R-O (i.e., beaker) match than for the C-R (i.e., cue) match (x, y, $z = 9, -81, 21$; cst). No effects were found for the reversed contrast (i.e., C-R match > R-O match) at our criteria of significance, although weak effects emerged at an uncorrected threshold of $p < 0.001$ in the dorsolateral prefrontal cortex (x, y, $z = 39, 18, 21$; $Z = 3.21$).

Training-dependent changes in neural activity

The source of the asymmetries between tests of early learning versus tests including all training trials is apparent in contrast values estimated for each training block in the two conditions (Fig. 5), which reveal decreasing activity across blocks of trials in the R-O but not the C-R condition (presumably attenuating the effects in IPL, M1, and aCN, while enhancing of those in the LG and extrastriate cortex). To further explore training-dependent changes in neural activity, we specified a factorial model with training block as a factor, and added linear weights to the blocks. A conjunction test, assessing neural activity that decreased linearly across training blocks in both the R-O and C-R condition, revealed effects in the supplementary motor area (SMA) and the left precentral and postcentral gyrus (Fig. 6*A*). We then performed a disjunction test for activity that decreased in the R-O, but not the C-R, condition and found significant effects throughout the frontoparietal network, including the left superior and right inferior parietal lobules, anterior cingulate, and DMFC, as well as bilateral thalamus extending into dorsal aCN (Table 1).

No effects were found at our criteria of significance for the reversed disjunction of decreasing activity in the C-R but not the R-O condition, nor for a test of increasing activity in the C-R but not the R-O condition. Notably, under normal training conditions, habits have been shown to control performance only with extensive training (Tricomi et al., 2009), suggesting that one might perhaps expect to see an increase in neural activity across training blocks in the C-R condition. However, our task was designed specifically to encourage dependence on stimulus–response associations from the onset of learning in this condition (see Introduction and Materials and Methods, above); it is not surprising therefore, that activity in areas responding preferentially to the C-R condition appears to have remained relatively stable throughout observational learning. Nevertheless, we did find significant effects for a test of activity that increased across blocks in the C-R condition while also decreasing across blocks in the R-O condition in the posterior cingulate and DMPFC (Fig. 6*B*). We conjecture that this result, reflecting opposite changes in training-dependent neural activity across conditions, may be related to the dynamic competition between goal-directed and habitual learning systems.

Neural correlates of devaluation performance

To relate the neuroimaging data to our behavioral effects, we tested whether a difference in neural activity between the C-R and R-O conditions during observational learning correlated with the degree of subsequent devaluation insensitivity. This was indeed the case. On a participant level, those with stronger activation of the dorsal premotor cortex (dPMC; x, y, $z = 30, 3, 57$; cst) and superior parietal lobule (SPL; x, y, $z = 21, -54, 60$; cst] in the C-R, relative to the R-O, condition responded on a greater proportion of devalued trials in the subsequent test (Fig. 7). In contrast, tests correlating differences in neural activity during observational learning with differences in accuracy or RT on the

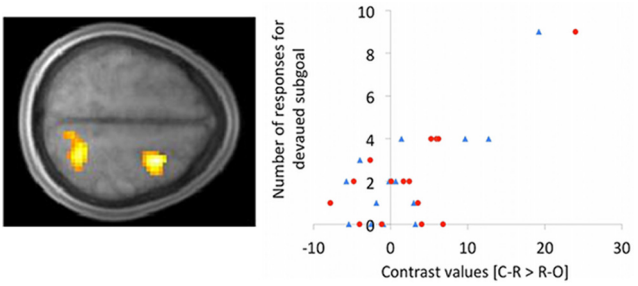

Figure 7. Correlation between the difference in devaluation insensitive performance (i.e., responding to fill up the devalued beaker) across the C-R and R-O conditions and the BOLD response to the [C-R > R-O] contrast during observational learning. Effects were found in the SPL and dorsal premotor (dPM) cortex (left). Scatter plot (right) shows devaluation insensitive performance in the C-R condition as a function of [C-R > R-O] contrast values in the SPL (blue triangles) and dPMC (red circles), estimated at LOSO coordinates.

probe test, or with probe test accuracy averaged across conditions, did not yield any effects at our criteria for significance, perhaps because accuracy was generally very high, with very small differences between conditions and very low variability across individuals.

Discussion

In this study, we explored whether goal-directed and habitual behavioral control strategies, frequently studied in the experiential domain, might also govern vicarious learning of instrumental actions. Specifically, we sought to determine whether outcome insensitivity of actions could be established through observation, and to elucidate the neural substrates mediating this process. Using a novel observational learning paradigm we found that, in subsequent performance, participants were more likely to respond for a devalued outcome when each observed action had been uniquely signaled by one of several discriminative cues than when each action had obtained a unique outcome. At the neural level, we found that activity in premotor and parietal areas during observational learning correlated with the degree of subsequent insensitivity to outcome devaluation. In addition, distinct areas of the action observation network (AON) and of the striatum discriminated between our two observational learning conditions.

Importantly, while our two training conditions differed significantly in the degree to which they supported sensitivity to outcome devaluation, this difference was relative rather than absolute, such that a fairly high level of devaluation sensitivity was observed even in the C-R condition. However, current models of action selection generally assume that goal-directed and habitual systems interact to control performance (Daw et al., 2005), so that it is only under extreme conditions that an action would be entirely insensitive to devaluation. Consequently, we interpret the relative insensitivity to outcome devaluation in the C-R condition as reflecting an increased, albeit not complete, dependence on habitual control.

It is important, however, to consider alternative sources of the observed difference in devaluation sensitivity. As noted previously, although the stimulus materials actually presented on the screen were identical across our two conditions, the features to which participants were required to pay attention to successfully regulate the system—the abstract cues versus the set of beakers—differed in nature as well as complexity. With respect to the imaging data, our matching control conditions should account for any differences in neural activity due to differences in relevant stimulus properties. However, we cannot rule out the possibility that the behavioral devaluation effect is attributable to such differences. For example, it is possible that the complex visual features making up the antecedent cues are more likely to generate reflexive, outcome-insensitive response strategies than the relatively simple display of the set of beakers does. It should be noted, however, that the fluid in all of the beakers were constantly fluctuating, so that detection of a below-threshold drop in a particular beaker was not trivial. Another possibility is that the spatial nature of the beaker stimuli facilitated outcome-sensitive responding; further research is needed to evaluate the contribution of perceptual stimulus properties to the arbitration between outcome-sensitive and -insensitive instrumental action selection.

We have suggested that, whereas the R-O condition involves learning about the relationship between an action and a subgoal, a much simpler, stimulus–response association is learned in the C-R condition. It is important to note, however, that the filling up of emptied beakers was an explicitly stated subgoal in both conditions (see Materials and Methods, above); that is, in both conditions, participants were instructed that an emptied beaker caused system imbalance and that their task was to keep the system balanced by filling up any emptied beakers using the different actions. This general emphasis on the beakers might have prompted participants in the C-R condition to attempt to generate a complex, hierarchical, conditional rule, entailing representations of both cues and beakers (i.e., if beaker 1 is empty and cue 1 is present then action 1 will fill up beaker 1) rather than a simple stimulus–response rule. In contrast, in the R-O condition, participants could focus on the direct link between actions and outcomes, ignoring all other aspects of the stimulus environment. Previous work has shown that direct links and higher-level hierarchical action representations recruit dissociable areas of premotor and prefrontal cortex, indicating a rostrocaudal gradient (Badre et al., 2010). However, to our knowledge, there is no theoretical or empirical basis for predicting differences between the two types of decision strategies in sensitivity to outcome devaluation. Nonetheless, a clear direction for future work is to assess correspondence between the direct versus hierarchical distinction and between goal-directed and habitual instrumental performance.

Alternative explanations for our behavioral effect notwithstanding, the neural data seems to indicate the use of distinct strategies during the observational phase, consistent with the observed difference during subsequent performance. The bulk of evidence for a neural distinction between strategies of experiential learning comes from rodent lesion studies demonstrating the respective involvement of the dorsomedial and dorsolateral striatum in goal-directed and habitual control (Yin et al., 2005a,b). More recently, human neuroimaging studies have implicated the human dorsomedial striatum (i.e., aCN) in goal-directed learning (Tanaka et al., 2008; Liljeholm et al., 2011), while the pCN and posterior putamen have been associated with behavioral insensitivity to outcome value, indicative of habits (Valentin et al., 2007; Tricomi et al., 2009). Our results suggest that these striatal dissociations, which appear to be relatively preserved across species (Balleine and O'Doherty, 2010), also underlie analogous strategies of observational learning. This finding is consistent with a recent study showing reward prediction errors in the aCN when human participants observed a confederate perform instrumental actions to obtain juice reward, as well as when the participant performed the actions and obtained the rewards themselves (Cooper et al., 2012).

We found effects in several areas previously implicated in ac-

tion observation and execution, including the IPL, LG, extrastriate cortex, and M1 (Hari et al., 1998; Astafiev et al., 2004; Järveläinen et al., 2004; Williams et al., 2006). The dissociation demonstrated here between the IPL and M1 on the one hand, and the LG and extrastriate cortex on the other, based on whether observed actions obtain distinct goals or are signaled by distinct cues, suggests that a functional separation of action–outcome and stimulus–response learning exists in the AON. This finding can be related to previous studies aimed at separating object information from action kinematics. For example, Järveläinen et al. (2004) reported that M1 activity discriminated between videos in which chopsticks were used to transfer items from one plate to another, based on whether the items were actually touched and moved or the act was simply pantomimed. Using a similar comparison of pantomimed and object-oriented observed actions, Buccino et al. (2001) found that, whereas both types of actions activated the premotor cortex, object-oriented actions selectively increased activity in the IPL. Here, we relate such results, which suggest that the IPL and M1 encode the physical consequences of actions to strategies governing action selection.

It should be noted, however, that the present study adopts a very different experimental approach to that which has featured in fMRI studies of action observation to date. Whereas the typical study on action observation measures neural responses during the observation of physical limb movements with or without a visually depicted object target, in the present study we test for the neural underpinning of learning associations between different components of a decision task (i.e., between discriminative stimuli, actions, and outcomes) in the absence of overt depictions of physical motor performance. Our approach yields unique insights into the nature of the associative processes being implemented within parts of the action observation network. For example, while Buccino et al. (2001) found activity in response to object-related, but not pantomimed, observed actions in the superior parietal lobule, in the current study we found this region to be correlated with the degree of devaluation insensitive action replication, suggesting that at the level of associative encoding, this region is in fact involved in stimulus–response and not goal-directed processing. Likewise, contrasting conditions in which to-be-imitated finger movements reached toward a location that was either marked or unmarked, and that was either parallel or diagonal (i.e., contralateral) to the initial finger position, Koski et al. (2002) found that activity in the dPMC was selective for the marked and contralateral movements. They interpreted these results as evidence for goal-related encoding by dPMC. In contrast, as with the SPL, the current results suggest that this area may contribute to habit formation, a finding that is more in line with its previously demonstrated role in conditional action selection (Grafton et al., 1998).

A possible explanation for the discrepancies between our findings and the action observation studies discussed in the previous paragraph is the fact that, although goal-directed evaluation likely entails object representations, an object-oriented action is not necessarily goal-directed. Specifically, it is difficult to determine whether, in the above studies, object features were processed by participants as action outcomes or as habit-eliciting discriminative stimuli. Indeed, grasp selection in actual interactions with everyday objects has been shown to depend on both habitual and goal-directed systems (Herbort and Butz, 2011), and evidence from animal conditioning studies suggests that the sensory features of action outcomes may trigger actions through habitual, stimulus–response associations (Ostlund and Balleine, 2007). In the current study, the use of a devaluation procedure overcomes this issue by providing a direct test of the nature of the associations underlying performance in a given task condition.

We found decreasing activity across blocks of observational learning throughout the frontoparietal network, including the IPL and dorsomedial frontal cortex, in the R-O, but not the C-R, condition. These effects may reflect a gradual disengagement of executive control processes due to increased automaticity (Poldrack et al., 2005). However, training-dependent decreases in neural activity specific to the R-O condition could also be due to a cumulative suppression of areas supporting habit formation. More generally, we note that, although the distinction between goal-directed and habitual learning is likely related to that made between declarative and procedural memory (Poldrack et al., 2001), a strong resistance to dual task interference [the behavioral paradigm commonly used to distinguish multiple memory systems (Foerde et al., 2006)] does not necessarily imply a decreased ability to suppress responding for devalued goals (for related findings on automaticity and response inhibition, see Cohen and Poldrack, 2008). Future work is needed to determine the relationship between instrumental control systems and multiple memory systems.

The distinction between habitual and goal-directed instrumental performance is similar to that made by theories of social learning, between imitation—merely copying observed actions—and emulation—learning about the causal relationships between objects (Horner and Whiten, 2005). Although behavioral evidence for such a dichotomy in observational learning has been found in human children and adults, as well as in a range of other animals (Tennie et al., 2006; Miller et al., 2009; McGuigan et al., 2011), very little is known about the distinct associative structures that respectively support emulative versus imitative strategies. According to a recent proposal (Seymour et al., 2009), imitative and habitual actions are associatively the same, in the sense that both are detached from the probability and value of their consequences and simply mapped to the stimuli making up the environment in which they occur. Despite this potential representational similarity, the relationship between the two learning strategies has not been empirically assessed. Specifically, previous tests of imitative versus emulative learning have not used the assays necessary to determine whether subsequent instrumental performance is in fact habitual. Our use of a devaluation procedure allowed a direct test of, and yielded positive evidence for, the idea that experiential and observational learning strategies depend on similar associative structures.

In the experiential domain, it has been suggested that goal-directed action selection can be accounted for in terms of a specific type of computational process termed model-based reinforcement learning (RL). According to this theory, goal-directed learners use an internal model of the environment to generate decisions based on state transition probabilities and outcome utilities. In contrast, habitual performance has been explained as model-free RL, in which actions are selected on the basis of cached values that contain no information about the identity or current utility of contingent outcome states (Doya et al., 2002; Daw et al., 2005). The results of the present study suggest that this framework might be extendable to observational learning, leading to a more general formal theory of the bases for behavioral control. The most striking implication of this extension is that behavioral automaticity might come about through the mere observation of other individuals. The current findings provide an initial step toward characterizing the computational and neural bases of such vicarious transmission of habits.

References

Astafiev SV, Stanley CM, Shulman GL, Corbetta M (2004) Extrastriate body area in human occipital cortex responds to the performance of motor actions. Nat Neurosci 7:542–548.

Balleine BW, Dickinson A (1998) Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology 37:407–419.

Balleine BW, O'Doherty JP (2010) Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology 35:48–69.

Badre D, Kayser A, D'Esposito M (2010) Frontal cortex and the discovery of abstract action rules. Neuron 66:315–326.

Buccino G, Binkofski F, Fink GR, Fadiga L, Fogassi L, Gallese V, Seitz RJ, Zilles K, Rizzolatti G, Freund HJ (2001) Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. Eur J Neurosci 13:400–404.

Caspers S, Zilles K, Laird AR, Eickhoff SB (2010) ALE meta-analysis of action observation and imitation in the human brain. Neuroimage 50:1148–1167.

Cohen JR, Poldrack RA. (2008) Automaticity in motor sequence learning does not impair response inhibition. Psychon Bull Rev 15:108–115.

Cooper JC, Dunne S, Furey T, O'Doherty JP (2012) Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. J Cogn Neurosci 24:106–118.

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

Delgado MR, Miller MM, Inati S, Phelps EA (2005) An fMRI study of reward-related probability learning. Neuroimage 24:862–873.

Dickinson A, Balleine BW, Watt A, Gonzales F, Boakes RA (1995) Overtraining and the motivational control of instrumental action. Anim Learn Behav 22:197–206.

Doya K, Samejima K, Katagiri K, Kawato M (2002) Multiple model-based reinforcement learning. Neural Comput 14:1347–1369.

Esterman M, Tamber-Rosenau BJ, Chiu YC, Yantis S (2010) Avoiding nonindependence in fMRI data analysis: leave one subject out. Neuroimage 50:572–576.

Foerde K, Knowlton BJ, Poldrack RA (2006) Modulation of competing memory systems by distraction. Proc Natl Acad Sci U S A 103:11778–11783.

Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC (1995) Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn Reson Med 33:636–647.

Grafton ST, Fagg AH, Arbib MA (1998) Dorsal premotor cortex and conditional movement selection: a PET functional mapping study. J Neurophysiol 79:1092–1097.

Hari R, Forss N, Avikainen S, Kirveskari E, Salenius S, Rizzolatti G (1998) Activation of human primary motor cortex during action observation: a neuromagnetic study. Proc Natl Acad Sci U S A 95:15061–15065.

Herbort O, Butz MV (2011) Habitual and goal-directed factors in (everyday) object handling. Exp Brain Res 213:371–382.

Horner V, Whiten A (2005) Causal knowledge and imitation/emulation switching in chimpanzees (*Pan troglodytes*) and children (*Homo sapiens*). Anim Cogn 8:164–181.

Järveläinen J, Schürmann M, Hari R (2004) Activation of the human primary motor cortex during observation of tool use. Neuroimage 23:187–192.

Koch K, Schachtzabel C, Wagner G, Reichenbach JR, Sauer H, Schlösser R (2008) The neural correlates of reward-related trial-and-error learning: an fMRI study with a probabilistic learning task. Learn Mem 15:728–732.

Koski L, Wohlschläger A, Bekkering H, Woods RP, Dubeau MC, Mazziotta JC, Iacoboni M (2002) Modulation of motor and premotor activity during imitation of target-directed actions. Cereb Cortex 12:847–855.

Liljeholm M, Tricomi E, O'Doherty JP, Balleine BW (2011) Neural correlates of instrumental contingency learning: differential effects of action–reward conjunction and disjunction. J Neurosci 31:2474–2480.

McGuigan N, Makinson J, Whiten A (2011) From over-imitation to supercopying: adults imitate causally irrelevant aspects of tool use with higher fidelity than young children. Br J Psychol 102:1–18.

Miller HC, Rayburn-Reeves R, Zentall TR (2009) Imitation and emulation by dogs using a bidirectional control procedure. Behav Processes 80:109–114.

O'Doherty JP, Deichmann R, Critchley HD, Dolan RJ (2002) Neural responses during anticipation of a primary taste reward. Neuron 33:815–826.

Ostlund SB, Balleine BW (2007) Selective reinstatement of instrumental performance depends on the discriminative stimulus properties of the mediating outcome. Learn Behav 35:43–52.

Poldrack RA, Clark J, Paré-Blagoev EJ, Shohamy D, Creso Moyano J, Myers C, Gluck MA (2001) Interactive memory systems in the human brain. Nature 414:546–550.

Poldrack RA, Sabb FW, Foerde K, Tom SM, Asarnow RF, Bookheimer SY, Knowlton BJ (2005) The neural correlates of motor skill automaticity. J Neurosci 25:5356–5364.

Seymour B, Yoshida W, Dolan R (2009) Altruistic learning. Front Behav Neurosci 3:23.

Tanaka SC, Balleine BW, O'Doherty JP (2008) Calculating consequences: brain systems that encode the causal effects of actions. J Neurosci 28:6750–6755.

Tennie C, Call J, Tomasello M (2006) Push or pull: imitation vs. emulation in great apes and human children. Ethology 112:1159–1169.

Tricomi E, Balleine BW, O'Doherty JP (2009) A specific role for posterior dorsolateral striatum in human habit learning. Eur J Neurosci 29:2225–2232.

Valentin VV, Dickinson A, O'Doherty JP (2007) Determining the neural substrates of goal-directed learning in the human brain. J Neurosci 27:4019–4026.

Williams JH, Waiter GD, Gilchrist A, Perrett DI, Murray AD, Whiten A (2006) Neural mechanisms of imitation and 'mirror neuron' functioning in autistic spectrum disorder. Neuropsychologia 44:610–621.

Yin HH, Knowlton BJ, Balleine BW (2005a) Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. Eur J Neurosci 22:505–512.

Yin HH, Ostlund SB, Knowlton BJ, Balleine BW (2005b) The role of the dorsomedial striatum in instrumental conditioning. Eur J Neurosci 22:513–523.