The Neural Bases of Action-Outcome Learning in Humans

Richard W. Morris, Amir Dezfouli, Kristi R. Griffiths, Mike E. Le Pelley, and Bernard W. Balleine

1Centre for Translational Data Science, University of Sydney, Sydney, NSW 2006, Australia, 2Data61, Commonwealth Scientific and Industrial Research Organisation, Sydney, NSW 2015, Australia, 3Brain Dynamics Centre, Westmead Institute for Medical Research, University of Sydney, Sydney, NSW 2145, Australia, and 4School of Psychology, University of New South Wales Sydney, Sydney, NSW 2052, Australia

From an associative perspective the acquisition of new goal-directed actions requires the encoding of specific action-outcome (AO) associations and, therefore, sensitivity to the validity of an action as a predictor of a specific outcome relative to other events. Although competitive architectures have been proposed within associative learning theory to achieve this kind of identity-based selection, whether and how these architectures are implemented by the brain is still a matter of conjecture. To investigate this issue, we trained human participants to encode various AO associations while undergoing functional neuroimaging (fMRI). We then degraded one AO contingency by increasing the probability of the outcome in the absence of its associated action while keeping other AO contingencies intact. We found that this treatment selectively reduced performance of the degraded action. Furthermore, when a signal predicted the unpaired outcome, performance of the action was restored, suggesting that the degradation effect reflects competition between the action and the context for prediction of the specific outcome. We used a Kalman filter to model the contribution of different causal variables to AO learning and found that activity in the medial prefrontal cortex (mPFC) and the dorsal anterior cingulate cortex (dACC) tracked changes in the association of the action and context, respectively, with regard to the specific outcome. Furthermore, we found the mPFC participated in a network with the striatum and posterior parietal cortex to segregate the influence of the various competing predictors to establish specific AO associations.

Key words: caudate nucleus; goal-directed action; Kalman filter; medial prefrontal cortex; posterior parietal cortex; prediction error

Significance Statement
Humans and other animals learn the consequences of their actions, allowing them to control their environment in a goal-directed manner. Nevertheless, it is unknown how we parse environmental causes from the effects of our own actions to establish these specific action-outcome (AO) relationships. Here, we show that the brain learns the causal structure of the environment by segregating the unique influence of actions from other causes in the medial prefrontal and anterior cingulate cortices and, through a network of structures, including the caudate nucleus and posterior parietal cortex, establishes the distinct causal relationships from which specific AO associations are formed.

Introduction
There is now considerable evidence from various species that animals encode the consequences of their behavioral responses and use this information to select, evaluate and initiate future goal-directed actions (Balleine and Dickinson, 1998; Tanaka et al., 2008; Balleine and O’Doherty, 2010; Alexander and Brown, 2011; Doll et al., 2012; Balleine, 2019). At a neural level such action-outcome (AO) learning has been found to depend on a highly conserved circuit involving regions of medial prefrontal cortex (mPFC) and their projections to the dorsal striatum; the dorsomedial striatum in rodents or caudate nucleus in humans (Balleine and O’Doherty, 2010; O’Doherty et al., 2017; for review, see Balleine, 2019). Nevertheless, although it is clear that goal-directed action depends on AO learning, the rules that govern this learning have yet to be established (Perez and Dickinson, 2020) and so the computations undertaken by the corticostratial circuit remain unresolved.

One recent account suggests AO learning is derived indirectly from a predictive model of the environment that allows the actor to search possible future states and select the action likely to achieve the highest reward value based on that search. This idea has been formalized within computational models (Alexander and Brown, 2011, 2019; Doll et al., 2012; O’Doherty et al., 2017)
and manipulations of action values yield data consistent with this general suggestion (Tanaka et al., 2008; Liljeholm et al., 2011). However, at least in rats, changes in response selection have been reported following manipulations of the AO association that leave action values unchanged (Colwill and Rescorla, 1986; Dickinson and Mulatero, 1989; Balleine and Dickinson, 1998). For example, when two actions earn distinct but equally rewarding outcomes, presenting one of the outcomes unpaired with the actions without altering the probability that the actions will earn the paired outcomes is sufficient to reduce selection of the action associated with the unpaired outcome while leaving the performance of the other action intact.

These findings suggest that AO learning is driven by the causal status of an action with respect to its particular outcome, i.e., the strength of the AO association reflects the evidence that A causes O relative to other potential causes. In situations where an outcome is delivered unpaired, a likely alternative cause is the experimental context and, indeed, it has been established that, in rats, overshadowing the context by signaling unpaired outcomes using another stimulus, such as a light, restores the predictive status of the action (Dickinson and Charnock, 1985; Colwill and Rescorla, 1986). It is unknown, however, if the same effect is observed in humans nor is it known how actions and the context compete in the brain to alter the AO association. Prior studies in humans have manipulated a single AO relationship and found that BOLD responses in ventromedial PFC and caudate track increments (Tanaka et al., 2008), whereas anterior cingulate and inferior frontal gyrus track decrements in the AO association (Liljeholm et al., 2011), but to date, no study has manipulated the AO contingency in humans unconfounded by changes in action value (Shanks and Dickinson, 1991; Griffiths et al., 2014; O’Callaghan et al., 2019).

Here, we sought to understand the computations within the circuitry mediating AO learning in humans by manipulating specific AO associations while keeping action values constant. We adapted the behavioral designs used in rodent studies to humans, degrading the causal relationship between specific AO associations and then reversing the effect of degradation by signaling the outcomes delivered unpaired with the actions. To decode the neural data, we developed a Kalman filter (Dayan et al., 2000; Krusche, 2008; Gershman, 2015) which learned the specific AO relationships by attributing prediction-errors to different causal variables adjusted by their covariance with specific outcomes. We found that learning about the causal effects of specific actions reflects competition between various predictors encoded as a probability distribution over AO contingencies; that activity in mPFC and dorsal anterior cingulate cortex (dACC) tracked these changes and that prefrontal connections with caudate and parietal cortex integrated these associations to form specific AO associations.

Materials and Methods

As detailed below, we conducted two fMRI studies using a free-operant task in which human participants could perform different actions to earn distinct food outcomes in a self-paced manner. In these experiments we trained hungry participants on distinct actions for specific food outcomes and then sought to degrade the causal relationship between action and outcome without changing action values. To achieve this, we used a variation of the procedure reported by Hammond (1980): outcomes continued to be delivered contingent on the action at a fixed probability, with one outcome delivered unpaired with performance at the same probability. Thus, whereas delivery of the unpaired outcome diminishes the reward value of both actions equally (since reward can now be obtained without taking either action), depending on the identity of the unpaired outcome, this treatment will selectively degrade the causal status of actions paired with that outcome relative to other actions (Dickinson and Mulatero, 1989; Balleine and Dickinson, 1998).

Participants

In experiment 1, 31 right-handed English-speaking volunteers, aged between 19 and 51 (mean age 30.5) were scanned. One participant was removed because of excessive head movement (>2 mm), thus n = 30 (18 females). We conducted a fMRI-based power analysis (Mumford and Nichols, 2008) of the mPFC effects in experiment 1 to determine a sufficient sample size to detect the same effects in a new sample. We selected the smallest effect in the mPFC reported in experiment 1: AAO response in BA9 (Z = 4.71). The power analysis was conducted on the BA9 functional region of interest (fROI), with a p-value threshold of 0.05 for a one-sided hypothesis test. The analysis revealed that N = 20 would provide >95% power. On the basis of this analysis, scans from 23 right-handed, English-speaking volunteers, aged between 17 and 32 (mean age 26) were considered for experiment 2. Three participants were removed because of excessive head movement (>2 mm), thus n = 20 (11 females). All participants were free of food allergies, neurologic or psychiatric disease, and psychotropic drugs, and reported strong liking of the snack foods we provided as reward. Informed consent to participate was obtained and the study was approved by the Human Research Ethics Committee at the University of Sydney (HREC no. 12812). Participants were reimbursed 45 Australian dollars in shopping vouchers, in addition to the snack foods they earned during the test session.

AO contingency degradation task

In each experiment, participants were instructed not to eat 3 h before the appointment. Pretesting involved obtaining preference ratings on a seven-point scale for each of three snacks (M&Ms, BBQ flavored crackers, chocolate cookies), from which the two most similarly preferred snacks were selected for the experiment.

Experiment 1 involved learning two concurrent AO associations (Fig. 1A). Participants were instructed they could liberate snack foods from a vending machine by tilting it to the left or right by pressing either a left or right button, and that sometimes the vending machine would also release a snack for free. They were also instructed to find the best action for releasing snacks. Outcomes were indicated by the presentation of a visual stimulus depicting the snack for 1-s duration (a particular snack food, e.g., M&M or BBQ cracker), during which further outcomes could not be earned. The relationships between actions and outcomes were constant across blocks for each participant (e.g., left = M&M and right = BBQ crackers for all blocks). Each block lasted 120-s, and the software controlling the task (PsychoPy2 v1.8; Peirce, 2007, 2009) divided each block into 120 1-s intervals to determine the outcome rate. Participants were unaware of the 1-s intervals, and they responded freely using the index finger on their right hand to press the left or right button on a Lumina MRI-compatible response pad (LU-400, Cedrus Corporation). An action (tilt left or tilt right) earned a particular outcome with a constant probability p = 0.2 if that action had occurred in the preceding 1-s interval. If both actions occurred in the preceding interval, then only the most recent action was considered for reward (Hammond, 1980). One of the two outcomes was sometimes delivered unpaired (i.e., in the absence of a response), thus degrading the specific AO relationship involving that outcome. Specifically, an unpaired outcome was delivered with p = 0.2 in each second in which neither action had been made. This schedule ensured two important features: (1) that there was no serendipitous contingency between an action and an unpaired outcome, which would result in a higher reward contingency for the contingent action (Dickinson and Mulatero, 1989); and (2) the paired outcome appeared at a varying interval up to 1 s after a successful action, which is sufficient to introduce ambiguity into the perceived AO contingency (Shanks and Dickinson, 1991). Participants completed six blocks; the outcome (BBQ cracker or M&M) that was delivered unpaired was counterbalanced across blocks (ARBAAB; for an example of the event order, see Fig. 1C). At the end of each block, participants rated how
causal each action was with respect to each outcome on a Likert scale from 1 (not at all) to 7 (very causal) in response to a reward specific question, i.e., “Did tilting the machine cause an M&M to drop?” In experiment 1, a follow-up test was conducted after the scan. The test setting, duration and programmed AO contingencies in the follow-up test were the same as in the scanner, with the addition of a 1-s yellow light cue displayed on the front of the virtual vending machine as a signal immediately before the delivery of each unpaired outcome (for an example of the event order, see Fig. 1B). At the end of all testing, participants received all snacks that had been delivered onscreen during test.

Experiment 2 involved learning a single AO contingency. The session was arranged in 12 blocks of 60-s duration, and in each block the participant responded freely for a single snack-food reward (counterbalanced between BBQ crackers and M&Ms). As before, in each block the probability of the outcome given that the action had been performed was \( P(O|A) = 0.2 \) in every second a response was made. The probability of an unpaired outcome in every second when no response was made, i.e., \( P(O|\neg A) \), varied between 0, 0.1 and 0.2 across blocks in a counterbalanced order. Consequently, the AO contingency, derived from the difference in these possibilities [i.e., \( \Delta P = P(O|A) - P(O|\neg A) \)], varied from 0.2, 0.1 to 0, respectively. In addition, experiment 2 varied whether the unpaired outcome was the same snack or a different snack to that earned by the button press in each block in an ABBA order. In half the blocks, therefore, the paired and unpaired outcomes were different whereas in the remaining blocks the unpaired outcomes were the same as the paired outcome. Providing unpaired outcomes that were different to the paired outcome is, with respect to the competing context predictions, equivalent to signaling the unpaired outcomes, and so experiment 2 replicates the design of experiment 1 and extends it by conducting the signaled test session during the scan.

Behavior data analysis

The behavioral data consisted of the rate of responding during each block and the causal ratings obtained at the end each block. Experiment 1 tested for differences between contingent (Con) and degraded (Deg) actions in the proportion of total responses, as well as mean causal ratings. In each case, a Shapiro–Wilk test confirmed the data did not violate the assumption of normality and differences were assessed by paired \( t \) tests (two-tailed). Experiment 2 tested the main effect of the outcome condition (whether the unpaired outcome was the Same or Different to that earned by pressing the button) and its linear interaction with the contingency condition (\( \Delta P = 0.2, 0.1, \) and 0.0), using a 2 \( \times \) 3 repeated measures ANOVA (two-tailed). Mauchly’s test was used to detect violations of sphericity, in which case the Greenhouse–Geisser correction was applied.

Model-based analysis

The learning models described below used a relatively coarse temporal resolution (1 s) to match the 1-s trial structure of outcome delivery just described. Thus, the occurrence of each event was converted to a binary result (0, 1) in each 1-s bin. Each outcome was linked to the preceding action (or unlinked) according to the delay in seconds (D) and an arbitrary and idiosyncratic temporal threshold (k), which was a free parameter fit to each individual.

Q-learning model

We modeled action-value learning using a Q-learning algorithm (Watkins, 1989) with linear function approximation. Input features are denoted by binary vector \( \mathbf{a} \) of size 3 by 1, in which each entry represents the presence of the context (X) and each action (A1 and A2). For example, \( \mathbf{a} = [1, 0, 1]^{T} \) implies that the context was present (the first entry) and action A2 was taken (the last entry: operator \( ^{T} \) denotes matrix transpose). We then calculated the reward that was expected for the current input vector using a linear function \( Q(\mathbf{a}) = \mathbf{w}^{T}\mathbf{a} \), which represents the value of input features \( \mathbf{a} \) modulated by the weight attached to each feature, represented by the \( 3 \times 1 \) vector \( \mathbf{w} \). Based on this, after each action the corresponding weights were updated as follows:

\[
\Delta \mathbf{w} = \alpha (r - \mathbf{w}^{T}\mathbf{a}) \mathbf{a},
\]

where \( \alpha \) is the learning rate, and \( r \) is a scalar representing the value of the reward that was actually delivered, with \( r = 0 \) if no reward occurred, and \( r = 1 \) if either outcome O1 or O2 (e.g., M&Ms or BBQ crackers) was delivered. Consequently, Equation 1 implies that the weights corresponding to each present input feature (context and taken action) were updated using a prediction-error (\( \delta \)) rule. For action selection, the value of action A1 was calculated as:

\[
Q(A1) = \mathbf{w}^{T} [1, 1, 0] \text{ and } Q(A2) = \mathbf{w}^{T} [1, 0, 1].
\]

See below for how these values translated to choice probabilities. Note that, in this model, the delivery of unpaired outcomes has a symmetrical impact on both actions (on the action delivering the same outcome as the unpaired outcome, and the action delivering a different outcome from the unpaired outcome).

Kalman algorithm

The aim of the Kalman algorithm was to learn the unique causal weight of each action over and above the context. Unlike Q-learning, in this model the associations between actions and outcomes are learned separately (whereas the Q-learning model learns a single value prediction for each action). In addition, the associations were updated based on the uncertainty of predictions (whereas in the Q-learning model weights are assigned using a fixed learning rate, \( \alpha \)). The Kalman algorithm builds a probabilistic representation of the causal weights \( \mathbf{w} \) of each input (actions and context cues) predicting each outcome \( \mathbf{o} \), representing causal beliefs. The causal beliefs are represented by a multivariate normal density \( N(\mathbf{w} | \mu, \Sigma) \) with a prior Gaussian distribution, and after observing each outcome the causal weights are updated by changes in the mean and variance of each distribution (see below). The matrix of means \( \mu \) (of size 3 by 2) summarizes the beliefs in the causal relationship of each input feature \( \mathbf{X}, \mathbf{A1}, \mathbf{A2} \) with each of the two outcomes \( \mathbf{O1}, \mathbf{O2} \), while
the covariance matrix $C$ (of size $3 \times 3$) captures the uncertainty around the predictions for each input feature. When the variance is large, there is a correspondingly large uncertainty regarding the true causal strength. The updating equations for $\mu$ and $C$ have the following form:

$$\Delta \mu = (\nu + a^T Ca)^{-1} (\alpha - \mu T a) Ca,$$

$$\Delta C = -(\nu + a^T Ca)^{-1} C a a^T C,$$

where $\nu$ is a free parameter capturing outcome variance to accommodate unmodelled differences in learning rate (similar to $a$ in Eq. 1). For each second, the execution of an action and the constant context are represented as a binary input vector $(a)$ and an outcome vector $(\alpha)$ represents the delivery of the outcomes (O1 and O2), e.g., $a = [0, 1]$ denotes delivery of the O2. The first term $(\nu + a^T Ca)^{-1}$ represents the total certainty (inverse sum of outcome uncertainty and belief uncertainty) and it governs the learning rate. $\mu a$ is a vector representing the learned causal weights on the basis of the current inputs $(a)$. The difference between the observed outcome and learned causal weights $(\alpha - \mu a)$ is a vector of outcome-specific prediction-errors, each of which represents a summed error-term for O1 and O2. The rightmost term $\text{Ca}$ is the product of the covariance matrix and input vector, and it allows for changes to the mean belief about actions otherwise correlated with the context but absent on the present trial (i.e., “counterfactual” learning). In this manner, the Kalman filter is able to distinguish the unique influence of actions from the context during unpaired outcomes. Importantly, the covariation between each action and the context is tracked in the off-diagonal elements of $C$, which allowed us to test a unique prediction of this model. Thus, $\Delta \mu$ is a matrix which contains changes to mean beliefs about the association between each action and each outcome ($\Delta A1O1$, $\Delta A1O2$, $\Delta A2O1$, $\Delta A2O2$) as well as the context/background and outcomes ($\Delta XO1$, $\Delta XO2$), and these changes depend on the prediction-error as well as the covariance matrix $C$. Note that in this model delivery of unpaired outcomes will have an asymmetrical effect on the predictions for each action. For the action delivering the same outcome as the unpaired outcome the associations will decrease, but for the other action (delivering a different outcome to the unpaired outcome) the associations will remain intact. We define the value of each action in this model as the maximum causal belief that it maintains with all the outcomes, $Q_{a} = \arg \max_{a} \mu_{a} \alpha$.

Null model

For comparison, we also described a null model without any temporal dynamics but ideal asymptotic performance. The null model assumed that the Q-value of each action was proportional to the final contingency obtained in each block, so $Q(A1) = P(O|A1) - P(O|\neg A1)$ and $Q(A2) = P(O|A2) - P(O|\neg A2)$. $\mathbf{Policy}$

The probability of selecting each action was based on Q-values and in accordance with the softmax rule:

$$\pi_{A} = e^{\theta Q_{A,i}} / (e^{\theta Q_{A,1}} + e^{\theta Q_{A,2}}),$$

in which $\tau$ is the inverse temperature parameter and determines the rate of exploration.

Model selection

To decide between the competing learning models, we compared the approximate log evidence of each learning model against the null model in a random effects analysis using each subject’s BIC score as a summary statistic of model adequacy. The results of both a classical paired $t$ test and a Variational Bayes comparison against the null model are reported (Stephan et al., 2009; Rigoux et al., 2014). We also report a likelihood ratio test (LRT) against a chance model (i.e., where $\tau = 0$, as well as the pseudo-$R^2$ for each model for a comprehensive comparison.

Simulations

Simulations with a generative model of the Kalman filter were run to confirm the Kalman filter generated behavioral data consistent with experiment 1. The generative model was able to produce two actions ($A1$, $A2$) as well as a third action (waiting) to mimic the option of waiting in the free-response experiment. Each action was produced stochastically over 120 time points to mimic the 120-s block in experiment 1, according to a Dirichlet distribution over the learned action values of the model. Two outcomes (O1, O2) were generated with equal conditional probabilities by each action, $P(O1|A1) = 0.2$ and $P(O2|A2) = 0.2$, and one outcome (O1) was generated in the absence of both actions $P(O1|\neg A1, \neg A2) = 0.2$. In this manner we arranged $AP Con = 0.2$ and $AP Deg = 0$ and ensured that the contingent action would not result in more outcomes, which mimics the experimental contingencies used in experiment 1. For $N=30$ simulations, the model generated behavior using the same parameters as the maximum likelihood estimates of the Kalman algorithm fit over the whole-group in experiment 1 ($\nu = 8.6, \tau = 11.74$; Daw, 2011). We also performed $N=1000$ simulations over a range of parameter values ($\nu = 0.50, \tau = 0.50$) to examine how causal learning was generated over a plausible parameter space.

Imaging

MRI data were acquired on a 3-Tesla GE Discovery using a 32-channel head coil. A T1-weighted high-resolution structural scan was acquired for each subject for screening and registration with a 1-mm3 voxel resolution (TR: 7200 ms, TE: 2.7 ms, 176 sagittal slices, 1 mm thick, no gap, 256 × 256 × 256 matrix). For BOLD acquisition, we acquired echoplanar imaging (EPI) volumes comprising 52 axial slices in an ascending, interleaved acquisition order (TR: 2910 ms, TE: 20 ms, FA: 90°, FOV: 240 mm, matrix: 128 × 128, acceleration factor: 2, slice gap: 0.2 mm) with a voxel resolution of 1.88 × 1.88 × 2.0 mm. Slices were angled 15° from AC-PC to reduce signal loss in the OFC. In experiment 1, 343 EPIs were acquired while in experiment 2, 260 EPIs were acquired.

Data were analyzed using SPM8 (www.fil.ion.ucl.ac.uk/spm). Preprocessing and statistical analysis were conducted separately for each experiment. The first four images were automatically discarded to allow for T1 equilibrium effects, then images were slice-time corrected to the middle slice and realigned with the first volume. The structural image was co-registered to the mean functional image, segmented and warped to MNI space. The warp parameters were then used to normalize the resampled functional images (2 mm3). Images were then smoothed with a Gaussian kernel of 8-mm full-width half maximum to improve sensitivity for group analysis.

Model-based fMRI analysis

For each first-level General Linear Model (GLM) analysis, we constructed vectors for each action causal beliefs ($\Delta A0$) and background causal beliefs ($\Delta XO$) from the matrix of $\delta$ values (i.e., changes to beliefs $\Delta \mu_i$), generated with the parameters provided by the group maximum likelihood estimation (MLE; Daw, 2011). For $\Delta A0$, the $\delta$ values were derived for the current action contingency (i.e., $A1$-O1 or $A2$-O2), whereas for $\Delta XO$, the $\delta$ values were derived for all occurrences of each outcome (including paired and unpaired). To test for brain activity tracking the unique changes in each vector, we entered $\Delta AO$ and $\Delta XO$ as parametric modulators of a stick function that included both response and reward times in an event-related design. While these update signals will fluctuate independently, there will be some collinearity when the covariance is zero. Collinearity is a problem when trying to determine unique effects associated with each regressor. However, the variance inflation factor can be used to indicate if a collinearity problem is present. The variance inflation factor was 1.23, which is within the bounds of a conservative threshold <5 (Mumford et al., 2015). Nevertheless, to remove any residual collinearity between these regressors, each regressor was entered as the second modulator to ensure it was adjusted for the prior regressor using the default orthogonalizing routine in Statistical Parameter Mapping (SPM)-8. Each GLM also contained rating periods and six movement parameters; $\beta$ were estimated with a 128-s high-pass filter and AR1 correction for autocorrelation. The resulting $\beta$ images were included in a group-level random effects analysis in SPM one-sample $t$
Dynamic causal modeling (DCM) analysis
A DCM analysis (Daunizeau et al., 2011) was designed to further interrogate the role of the caudate-PFC interactions implied by the group results of the GLM analysis from experiment 1. A BA9 volume of interest (VOI) was defined by the significant cluster from the analysis of ΔAO in the group results. This significant group cluster was used to construct a binary mask and this mask was then used to define the VOI and extract results of the GLM analysis from experiment 1. A BA9 volume of interest (VOI) was defined by the significant cluster from the analysis of ΔAO in the group results. This significant group cluster was used to construct a binary mask and this mask was then used to define the VOI and extract the first eigenvector for all individuals, adjusted for the binary mask and this mask was then used to define the VOI and extract the significant cluster from the analysis of ΔXO in the group results. The significant cluster from the analysis of ΔXO was used to construct a VOI.

Psychophysiological interaction (PPI) analysis
A whole-brain PPI analysis was designed to determine whether cortico-striatal interactions varied with causal learning during experiment 2. The psychological term was the block condition: whether or not the unpaired outcomes were distinguishable within that block (Same or Different). The physiological term was the timeseries from the anterior caudate in each participant (n = 20) using a group rROI mask from the GLM analysis of the covariance timeseries. We constructed the interaction term in SPM8 (per defaults) and included all three terms in a subject-level GLM. Finally, we tested for regions of interaction at the group-level across the whole-brain, corrected for multiple comparisons FDR-q ≪ 0.05.

Data availability
Data are available on request. Unthresholded statistical maps are available for viewing and downloading at http://neurovault.org/collections/VXWZKTW/. Experimental programs and MATLAB code to generate simulations can be downloaded from https://github.com/datarichard/Algorithmic-Neuroanatomy.

Results
AO degradation reveals people learn the unique effects of their actions
Figure 2A shows that, although performance of the two actions was initially indifferent, introduction of the unpaired outcome generated a clear preference for the contingent action (Con) over time as people were exposed to the difference between each AO contingency. Figure 2B, left panel, shows that, overall, the mean number of contingent actions (Con) was significantly greater than the degraded actions (Deg), paired \( t \) test \( t_{(29)} = 4.15, p = 0.0002 \). Causal ratings collected at the end of each 2-min block confirmed people judged the Con action as more causal than the Deg action (Fig. 2C, left panel), paired \( t \) test \( t_{(29)} = 3.94, p = 0.0004 \). Note there is some apparent heterogeneity in the actions and ratings among participants. However, this heterogeneity is explained when we examine the correlation between actions and ratings (Fig. 2D). There was a strong correlation between, in z score units, (1) the difference in performance of the contingent and degraded actions and (2) the difference in participants’ causal judgments of the contingent and degraded actions, \( r_{(20)} = 0.73, p = 0.00002 \). That is, participants who tended to act indiscriminately also tended to rate the actions as more similarly causal, whereas participants who acted in a goal-directed manner (more Con actions than Deg actions) also detected the difference between the causal effect of each action in their ratings.

Signaling the unpaired outcomes restores the performance of the degraded action
The selective impact of unpaired outcomes on actions and causal judgments reveals human participants were sensitive to the unique effect of their actions, even when the probability of observing an outcome among actions was equal. The unpaired outcomes made it more difficult for participants to distinguish the outcome their actions caused, resulting in a reduction in the perceived causal efficacy of that action relative to the action which delivered an outcome that differed from the unpaired outcome. We conducted a follow-up test after the fMRI scan, under the same contingencies, with the addition that each unpaired outcome was now signaled by a yellow cue-light (Fig. 1B). The signal reduced the causal uncertainty generated by the unpaired outcomes and allowed participants to distinguish the unique causal effect of their own actions. The results found the addition of the signal restored responding (Fig. 2B, Signaled), paired \( t \) test \( t_{(29)} = 0.75, p = 0.46 \), as well as causal judgments (Fig. 2C, Signaled), paired \( t \) test \( t_{(29)} = 0.88, p = 0.39 \), of the degraded action (Deg) to the same level as the contingent action (Con). The restoration of both action performance and causal judgments by the signal indicates the perceived background rate of a specific reward, i.e., whether it is predicted by the context or some other event, plays a key role in learning the unique causal strength of individual actions.

Action-selection reflects changes in the experienced AO correlation
To check our experimental control over each contingency in this free-response task, we confirmed that the unpaired outcomes selectively degraded the experienced contingency of the degraded action: post hoc analysis revealed the mean experienced contingencies [i.e., \( P(O|A) - P(O|\neg A) \)] for the Con and Deg actions were 0.18 and 0.07, respectively, paired \( t \) test \( t_{(29)} = 12.06, p = 0.8E-12 \), whereas ignoring unpaired outcomes, the probability of a paired outcome for the Deg action was \( P(O|A) = 0.17 \), which did not differ significantly from the Con action \( P(O|A) = 0.18 \); paired \( t \) test \( t_{(29)} = 0.90, p = 0.37 \), confirming that experienced differences in the probability of contingent reward were not responsible for the results (Extended Data Fig. 2-1). Furthermore, there was no experienced correlation between reward and the number of actions between individuals: Figure 2E, blue panel, shows the correlation between the number of Con actions and the total number of outcomes, which was close to zero across participants, \( r_{(28)} = 0.05, p = 0.80 \). Conversely there was no significant correlation, and certainly no negative correlation, between Deg actions and total outcomes (Fig. 2E, red panel), \( r_{(28)} = 0.11, p = 0.58 \). Furthermore, the distribution of delays between each outcome and the preceding action did not differ substantially or significantly within a 10-s interval for Con and Deg actions (Fig. 2F), Kolmogorov–Smirnov \( D_{28} = 0.09, p = 0.09 \), confirming that the immediate temporal contiguity with reward was indistinguishable and so was unlikely to have differentially influenced action-selection. Finally, pretest preference ratings of the snack food outcomes confirmed both rewards were similarly liked. The mean [95% confidence interval (CI)] ratings on a seven-point Likert scale were 5.8 CI[5.5, 6.2] and 6.3 CI[5.9, 6.6], for BBQ crackers and M&Ms, respectively, with substantial overlap between the 95%Cs. Thus, action-selection did not reflect any significant post hoc or experienced differences in
reward contingency or contiguity; these analyses confirm that we were able to alter the causal status of the action without affecting action values. It is clear, therefore, that models that rely on differences in action value to explain differences in performance cannot explain these results. In summary, these results make clear that the degradation procedure did not affect the relative rate of reward, i.e., the probability that an action will produce an outcome, nor did it affect the delay between making an action and the outcome being delivered and that neither of these factors contributed to the degradation effect.

**Action selection was best explained by the Kalman algorithm**

Data from experiment 1 were used to calculate the posterior probability of the Kalman algorithm as well as a prediction-error model and a null model. The null model used the asymptotic AO contingencies of each block as action values; thus, it was not a learning model but a static model with no temporal dynamics. Comparison with the null model determined whether each learning model explained how choices depend on the unique trajectories of trial-by-trial feedback. Table 1 shows the aggregate negative log likelihoods, model summary statistics and planned comparisons of each model relative to the null model. After fitting each participants’ data by maximum-likelihood estimation, the results of an LRT indicated both learning models predicted significantly more behavior than chance. However, the Kalman filter had the highest pseudo-$R^2$, indicating it explained more choices than the other models, with the planned random-effects analyses further revealing that the Kalman filter was significantly superior to the null model according to the classical $t$ test ($p = 0.047$), as well as the variational Bayes test (posterior expected model proportion $<r_k> = 0.72$ and the protected exceedance probability $\phi = 92.34\%$; for details, see Stephan et al., 2009). Hence the Kalman filter was both more likely and explained the acquisition of causal learning over trials better than an ideal asymptotic model, while the Q-learning model did neither. A (post hoc) direct comparison between the Kalman filter and the Q-learning model also confirmed the Kalman filter was superior ($t_{29} = 2.68, p = 0.01; <r_k> = 0.78$, protected $\phi = 99.11\%$). Thus, the majority of choices and the total evidence favors this simple Kalman algorithm (Table 1).
We also confirmed via simulation that the Kalman filter was able to generate choices and exhibit a preference for the more causal action over the degraded action (Fig. 3A), as well as recover the group parameter fit from experiment 1. The mean (±SEM) of the recovered parameter values were similar to the whole-group parameters (v = 8.48 ± 1.19, τ = 11.76 ± 0.74; for individual results, see Extended Data Fig. 3-1), confirming that parameter recovery was accurate and reliable. Moreover, these parameter values generated learned action values (Fig. 3A) and behavioral data consistent with that observed in experiment 1 (Fig. 3B). Furthermore, these results were not dependent on the exploration or learning parameters chosen (Fig. 3C,D).

The mPFC distinguishes causal actions from the context
We evaluated whether the brain attributed prediction-errors to different causal variables so as to uniquely predict their specific outcomes by regressing the model-derived learning estimates against the image data collected in experiment 1. Separate estimates for the signed changes in beliefs (i.e., learning) about actions and context (ΔAO and ΔXO; Materials and Methods) were included as parametric modulators of a stick (δ) function of response and outcome times. We included outcomes in the δ function to include times when the action was present as well as absent (the context was assumed to always be present). Whole-brain analysis revealed that learning about actions and the context occurred in distinct regions of the mPFC (Fig. 4A,B). As we have reported previously (Morris et al., 2014), action learning (ΔAO) appeared in a medial region of the superior frontal gyrus (BA9, global peak MNI co-ordinates: −15 +47 + 40, Z = 4.71, F(1,29) = 37.12, FWE = 0.031). At the same time, learning-related changes to the context estimates (ΔXO) appeared in the dACC (BA32, global peak MNI: −9 +41 + 22, Z = 5.19, F(1,29) = 50.20, FWE = 0.004), as well as smaller changes in the left caudate (MNI: −15 +14 +7, Z = 4.88, F(1,29) = 41.80, FWE = 0.017), and cuneus (MNI: −3

Table 1. Model evidence and comparison scores

<table>
<thead>
<tr>
<th>Free parameters</th>
<th>Negative log likelihood</th>
<th>LRT χ²</th>
<th>LRT p</th>
<th>Pseudo-R²</th>
<th>t&lt;sub&gt;150&lt;/sub&gt; (p)</th>
<th>&lt;t₁₅₀&gt; (protected φ₁)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kalman filter v, k, and τ</td>
<td>14,056</td>
<td>χ² = 7790</td>
<td>&lt;1E−30</td>
<td>0.24</td>
<td>2.07 (0.047)</td>
<td>0.72 (92.34%)</td>
</tr>
<tr>
<td>Q-learning α, k, and τ</td>
<td>16,343</td>
<td>χ² = 5124</td>
<td>&lt;1E−30</td>
<td>0.13</td>
<td>0.85 (0.403)</td>
<td>0.26 (5.75%)</td>
</tr>
<tr>
<td>Null model (H₀) τ</td>
<td>15,992</td>
<td>χ² = 5826</td>
<td>&lt;1E−30</td>
<td>0.13</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For each model, the model evidence (aggregate negative log likelihood scores), model significant differences from chance (LRT), model fit (pseudo-R²), and the results of the planned random-effect analyses comparing subject-specific BIC scores relative to the informed but static model (H₀), for classical (t test) and variation Bayes (posterior expectations of model probabilities −<t₁₅₀> and protected exceedance probabilities −<φ₁>), v is outcome variance, α is learning rate, k is the AO delay temporal threshold, and τ is inverse temperature (exploitation/exploitation).

Figure 3. Experiment 1 simulation results. Mean (±SEM) learned action values (µ; A) and action choice probabilities (B) from N = 30 simulations of the Kalman filter, using the group parameters fit from the data of experiment 1. Importantly, this result confirmed the Kalman algorithm was able to learn and distinguish the causal action from the degraded action, with the fit parameter values from experiment 1 (as well as within the time duration of each 120-s block). Learned action value of the contingency action (µ<sub>Con</sub>) and the corresponding probability (B) that µ<sub>Con</sub> is greater than µ<sub>Deg</sub> as a function of two free parameters, exploration (τ) and learning rate (v). Darker blue indicates better causal learning, and the prevalence of causal learning throughout the parameter space confirmed the fitted model results were not substantially dependent on the learning rate parameters chosen since the differences appeared across a wide range of parameters. See also Extended Data Figure 3-1, which shows the recovered parameters from each N = 30 simulation.

The posterior parietal cortex tracks the covariance between causes
The results so far indicate causal actions are distinguished from the context in the mPFC. A unique feature of the Kalman filter is that the covariance term in this model distinguishes the influence of candidate causes by updating ΔAO and ΔXO together when the covariance is positive but updates them in opposite directions when the covariance is negative (Fig. 4C). We tested whether any...
brain regions tracked the covariance between actions and context by entering the covariance values as a parametric modulator. A whole-brain analysis revealed bilateral activity in posterior parietal cortex (BA40; red voxels FWE cluster p = 0.007). At the same time, model updates to the context (ΔX0) occurred in the dACC (blue voxels FWE cluster, p < 0.001). 

Figure 4. A corticostriatal network mediates causal learning. A, Model-derived learning variables were tracked in the mPFC: model updates to actions (ΔA0) occurred in the mPFC (BA9; violet voxels FWE cluster p = 0.007). At the same time, model updates to the context (ΔX0) occurred in the dACC (blue voxels FWE cluster, p < 0.001). B, A cut-away representation showing the spatial relationship of the corticostriatal network, including model covariance in the right posterior parietal cortex (BA40; red voxels FWE cluster p = 0.007). C, The Kalman algorithm adjusts causal beliefs over time according to the prediction-error adjusted by the covariance between other possible causes. Under a degradation schedule, causal beliefs, i.e., the attribution of causal strength, to the action and to the context initially increase together as actions co-occur with outcomes. However, as unpaired outcomes become more prevalent, the causal strength attributed to the context diverges from the action. Changes in the context and action occur in the same direction when covariance between the action and context is positive, but when the covariance is negative then the attribution of causal strength moves in opposite directions. An illustration of the action and outcome events over time in this example is shown in the lower part of this panel. See also Extended Data Figure 3-1 for data showing parameter recovery, action values, exploration, and learning rate parameters from simulations of Kalman filter. D, ROI analysis in the striatum: red voxels (image threshold p < 0.5 svc) in the anterior caudate tracked the covariance, whereas green voxels (image threshold p < 0.5 svc) in the caudate body tracked summed prediction-errors. Overlap is indicated in yellow. E, Sagittal view of ROI results. F, Diagram of the DCM assuming the caudate integrates information from separate regions in the mPFC (dACC and BA9), modulated by the covariance between potential causes. G, An alternate DCM showing the caudate segregating information in the mPFC, modulated by the covariance between potential causes. H, The probability the data supports the segregate model (SEG) is more likely than the integrate model (INT). I, The posterior probability of each model (INT vs SEG) generating the observed data.

Prediction-error and covariance converge in caudate

Some evidence suggests that the ventral striatum tracks or receives reward prediction-errors (O’Doherty et al., 2004; Pessiglione et al., 2006; Tobler et al., 2006), while dorsal striatal regions track action values (O’Doherty et al., 2004). In the present example of AO learning, the prediction-error represents the deviations between the observed outcome and the summed total causal expectancy (based on both the action performed and the context). We tested whether the striatum tracks this summed error term, by including it as a parametric modulator in an anatomic ROI analysis of the striatum. Figure 4D,E shows BOLD responses in a posterior region of the caudate body (green) tracked the summed errors (ROI peak MNI: +15 11 4, Z = 4.17, t29 = 4.94, FWE = 0.002 svc), whereas activity in the anterior caudate (red) was associated with the covariance between actions and context (ROI peak MNI: 15 23 7, Z = 3.42, t29 = 3.84, FWE = 0.029 svc). These regions were more medial and dorsal than those implicated in reward prediction-error signals but similar to regions previously implicated in instrumental learning (Balleine and O’Doherty, 2010). Thus, the caudate appears to receive sufficient information to segregate the influence of different events and so play a role in selectively distinguishing causal actions from context effects.

The caudate segregates the effect of the action from the context

To further determine the caudate’s role in distinguishing action control, we performed a DCM analysis (Daunizeau et al., 2011). We tested two possibilities shown in Figure 4F,G. In model 1 the caudate is a site of convergence of the updated values from the PFC to enable action-selection. In model 2 the caudate segregates the prediction-error to update the estimates of the action and context separately. Bayesian model selection revealed the relative log-evidence for model 2 was 85.31, which corresponds to strong evidence in favor of segregation. A random-effects analysis revealed a large majority of participants were significantly more consistent with the segregation than the integration model; the exceedance probability was 99% (Fig. 4H), and it was more likely to be true for any random subject; the posterior probability of the segregated model was 83.03 (Fig. 4I).

Cortex and caudate interact to distinguish causal actions by their covariance

Experiment 2 replicated key behavioral and fMRI results in an independent sample of naive participants, using a design that allowed us to assess the effect of distinguishing the identity of the unpaired outcomes on the corticostriatal network we identified above. The experiment used a single AO contingency and varied whether or not the unpaired outcomes were distinguishable
from the paired outcomes in a block-by-block fashion so as to test the interaction between causality and caudate activity. For half the blocks we used the same outcome for both unpaired and paired outcomes (i.e., as in experiment 1), whereas in the other half the paired outcomes were different from the unpaired outcomes. Using distinct outcomes in half the blocks allowed the participants to discern the causal effect of their actions (i.e., equivalent to the effect of signaling the unpaired outcomes, as in the follow-up to experiment 1). As shown in Figure 5, responses (Fig. 5A) and causal ratings (Fig. 5B) were higher when the unpaired outcomes were different from the paired outcomes; a 2 (Same vs Different unpaired outcomes) by 3 (contingency: 0.2, 0.1, or 0; see Materials and Methods) ANOVA of responses and causal ratings confirmed the main effect of unpaired outcomes was significant in each case ($F_{(1,19)} = 12.57, p = 0.02$ and $F_{(1,19)} = 33.82, p = 0.6E-4$, respectively). As expected, the effect of distinct unpaired outcomes on response rate and causal ratings also interacted with contingency (interaction $F_{(1,5,30)} = 7.24, p = 0.005$ and $F_{(2,30)} = 10.07, p = 0.002$, respectively), and in a manner that was neither explicable by the rate of earned or of freely delivered outcomes (Extended Data Fig. 2-1).

As before, we fitted the Kalman algorithm to the data using maximum likelihood estimation. The optimal model predicted significantly more choices than chance, the mean group average likelihood per trial was 57% (95%CI: 53–60). A fROI analysis using masks generated from the significant results in the mPFC and caudate of experiment 1 confirmed that learning about the context ($\Delta XO$) occurred in the same dorsal ACC region (Fig. 5C, blue), ROI peak MNI: /C0/313613, $Z = 3.94, t_{(19)} = 5.06, FWE = 0.02$. Meanwhile learning values for the action ($\Delta AO$) occurred in the same region of BA9 (Fig. 5C, violet), ROI peak MNI: /C0/2147146, $Z = 3.04, t_{(19)} = 3.56, svc p = 0.001$. Covariance between the action and context was tracked in the caudate (Fig. 5C, right panel), ROI peak MNI: $-2+47+46, Z = 3.04, t_{(19)} = 3.56, svc p = 0.001$. Covariance between the action and context was tracked in the caudate (Fig. 5C, right panel), ROI peak MNI: $-12+20+1, Z = 3.91, t_{(19)} = 5.01, FWE = 0.002$. We used a whole-brain PPI analysis to determine whether any cortical regions interacted with the caudate when the unpaired outcomes were the same as the paired outcomes. A single region in the right parietal junction interacted with the caudate when unpaired outcomes were the same (vs different), shown in Figure 5D, global peak MNI: $+54-58+30, Z = 4.68, F_{(1,19)} = 24.90, FWE = 0.01$. This region overlapped with the right posterior parietal cortex (BA40).
identified in experiment 1 (Fig. 5E, red). Together with the results from experiment 1, this experiment provides converging evidence implicating the posterior parietal cortex in tracking the covariance between different potential causes and suggests that it interacts with the caudate when the unique effects of our actions need to be distinguished from other potential causes.

Discussion

We sought to establish the learning rules that govern AO learning in instrumental conditioning and their neural bases. We found that the mPFC participates in a circuit that detects and segregates the unique causal effects of actions from the context. More importantly, this segregation was generated by a prediction-error, described by a Kalman filter, that uses a summed prediction-error term along with the covariance between potential causes to distinguish the unique effect of actions. Furthermore, the caudate appears to be a key point of integration of the covariance term and prediction-error and segregates the summed prediction-error into separate update values for each causal belief. Thus, this model represents a simple, iterative Bayesian model of diction-error into separate update values for each causal belief.

More importantly, this segregation was generated by a prediction-error, described by a Kalman filter, that uses a summed prediction-error term along with the covariance between potential causes to distinguish the unique effect of actions. Furthermore, the caudate appears to be a key point of integration of the covariance term and prediction-error and segregates the summed prediction-error into separate update values for each causal belief. Thus, this model represents a simple, iterative Bayesian model of diction-error into separate update values for each causal belief.

Thus, this model represents a simple, iterative Bayesian model of diction-error into separate update values for each causal belief. The role of this process, process-level, account of AO learning that can be instantiated in the neural code (Guest and Love, 2017; Gershman and Uchida, 2019).

Many results have emphasized the critical role of the mPFC in AO learning, however the exact nature of this role has remained unclear. There is a wealth of evidence that, in the rat, area 32, the prelimbic region of the mPFC, is critical for the acquisition of goal-directed actions based on the effect of manipulations of the prelimbic cortex on sensitivity to contingency degradation and outcome devaluation (Balleine and Dickinson, 1998; Corbit and Balleine, 2003; Dias-Ferreira et al., 2009; Naneix et al., 2009; for review, see Balleine, 2019). Generally, the same has proven true in studies of goal-directed action in humans (Balleine and O’Doherty, 2010). One major difference between animal and human studies, however, is the acute nature of manipulations in the latter, where training and testing occur within a single session (vs multiple sessions over days or weeks in animal studies). This is likely to be important. mPFC involvement is limited to early in training in rodents, something that has been suggested to reflect involvement of a working memory process, and although longer term involvement of mPFC has not been assessed, it is likely the same in humans (Euston et al., 2012; Reber et al., 2017). The role of this process is likely variable; accounts have emphasized the role of mPFC in generating both positive and negative prediction-errors during learning (Alexander and Brown, 2011, 2019). The relationship to negative prediction-errors is particularly germane to the current case, as they may be implicated in distinguishing the effect of the context relative to the action during contingency degradation. Importantly, these changes were rapid and reflected changes in outcome-specific predictions, something that is anticipated on the causal account assessed here (there relies on a vector of outcome-specific prediction-errors) but not on reinforcement learning accounts that derive a single (scalar) reward value and so predict degradation will affect all actions associated with reward in the degradation context. An implication of our results is, therefore, that the mPFC (near the dACC) tracks outcome-specific changes in value, which extends our understanding of its computational role beyond scalar reinforcement learning. We observed that changes in dACC activity were consistent with changes to the AO contingency when the covariance between the action and context was positive; particularly early in learning when the unique effect of the action was yet to be fully distinguished from the effect of any potential background causes. However, once the negative covariance between action and background (context) causes had been acquired, the dACC tracked changes to the background in the opposite direction to the changes to the AO contingency, consistent with a competitive learning process.

Our results from both experiments also distinguished a separate region of the mPFC, near the medial surface of BA9, the activity of which represented updates to the unique effect of the action (ΔAO; Fig. 5C). The PPI analysis showed (Fig. 5D,E) that the caudate interacts with the parietal cortex when the unpaired outcome was the same as, or indistinguishable from, the paired outcome. The selective interaction between parietal cortex and caudate arises only when there is no observable information to distinguish control. When unpaired outcomes are indistinguishable, the covariance between our actions and the context is the only information that can be used to distinguish them. The right posterior parietal junction identified in the PPI was the same cortical region identified in experiment 1 as tracking the covariance term (along with the caudate in the subcortex). It also overlaps with the cortical region previously implicated in learning the transition matrix during model-based reinforcement learning (Gläscher et al., 2010) consistent with the claim, across studies and laboratories, that this region represents the covariance structure of the environment. Although PPI does not indicate the direction of influence, these results are consistent with a network extending from the parietal cortex to caudate to mPFC that tracks the covariance between actions and other events and then segregates the error-term to learn about and distinguish between the influence of different causes.

We found the competitive allocation of causal beliefs to actions relative to the context is a form of error-related selective learning closely related to cue-competition models in associative learning (Rescorla and Wagner, 1972). An important difference is the covariance matrix of the Kalman filter, in which the off-diagonal terms track the covariation between events. In our model, the covariance allowed the learner to distinguish or segregate the effects of an action from the context in the absence of that action, as shown in Figure 5. This allowed the model to reason counterfactually about what would have happened if an action had not occurred. In this manner, the covariance is analogous to heuristically motivated formalizations of within-event learning (e.g., negative α), which allows learning about absent events in more modern versions of cue-competition models (Van Hamme and Wasserman, 1994; Dickinson and Burke, 1996).

Unlike some other computational models of causal learning (e.g., causal power models; Cheng, 1997) the causal estimates learned by our Kalman filter converge on ΔP, a normative measure of causal strength (Allan, 1980; Shanks, 1993). Other researchers have argued that ΔP does not provide the best approximation of human causal inference because changing the base-rate probability of an outcome while holding ΔP constant modulates causal judgements (e.g., “the base-rate illusion”). However, the base-rate illusion is considerably weaker in free-response, instrumental learning where trials are not explicitly segmented (Wasserman et al., 1983, 1993; Vallee-Tourangeau and Murphy, 1999; Vallee-Tourangeau et al., 2005). Furthermore, when learning about causal effects, active intervention...
is a more reliable guide to causal relations than observation, largely
because actions constitute one basic way to control for possible
alternative causes (Steyvers et al., 2003; Pearl, 2009; Holyoak
and Cheng, 2011; Cheng and Buehner, 2012). Humans, and
potentially rats (Laurent and Balleine, 2015), are able to reason
counterfactually about what should be expected to happen if an
intervention is or is not made, and rapid fluctuations in striatal
dopamine (Kishida et al., 2016) in the context of our Bayesian
model could reflect these counterfactual action values. For these
reasons, AO learning may not suffer the same biases as forms of
causal learning that are based on passive observation.

Another recent view of goal-directed instrumental action sug-
gests that it is the experienced correlation between the rate at
which an action is performed and the rate of outcome delivery
that determines the strength of the AO association (Perez and
Dickinson, 2020). On this view, by selectively increasing the
rate of outcome delivery, the addition of unpaired outcomes
during contingency degradation reduces the experienced
 correlation between the rate of that outcome and its associ-
ated action and so will reduce the strength of the AO associ-
ation and the performance of that specific action. Thus, like
the current account, the correlational account predicts that
contingency degradation reduces the strength of the AO associa-
tion (cf. Crimmins et al., 2021). Nevertheless, as,
on the correlational account, AO learning is not a competi-
tive process, it is not immediately clear why signaling the
unpaired outcome should restore the positive correlation. It
is possible that the signal encourages the actor to ignore the
unpaired outcomes or to treat them as if occurring in a dif-
ferent context but, without further investigation, how this
account differs from that we advance here and whether it
could be implemented using the same form of Kalman algo-
rum remains to be determined.

Nevertheless, whatever the fate of alternative accounts, learn-
ing about the causal effects of our actions, as required for goal-
directed learning and as investigated here, appears to reflect fea-
tures of traditional associative models such as competition for
predictive value, as well as more recent conventions such as envi-
ronmental structure (covariance). In our hands, these features
were combined in a highly simplified, iterative Kalman filter that
learned a probability distribution over AO contingencies to pro-
vide a novel account of AO learning. In our results there was im-
pressive agreement across experiments and replications that
distinct regions of a corticostriatal network distinguished the
unique causal effect of actions from other causes, most notably
those of the context.

References

Alexander WH, Brown JW (2011) Medial prefrontal cortex as an action-out-


Allan LG (1980) A note on measurement of contingency between two binary

Balleine BW (2019) The meaning of behavior: discriminating reflex and voli-

and incentive learning and their cortical substrates. Neuropsychopharmacology
57:407–419.

trol: corticostriatal determinants of goal-directed and habitual action.
Neuropsychopharmacology 35:48–69.

Psychol Rev 104:367.


