Single-Trial Inhibition of Anterior Cingulate Disrupts Model-based Reinforcement Learning in a Two-step Decision Task.

Thomas Akam[1,2], Inês Rodrigues-Vaz[1,3], Xiangyu Zhang[4], Michael Pereira[1], Rodrigo Oliveira[1], Peter Dayan[4], Rui M. Costa[1,3].

Affiliations:

1. Champalimaud Neuroscience Program, Champalimaud Centre for the Unknown, Lisbon, Portugal

2. Department of Experimental Psychology, Oxford University, Oxford, UK

3. Department of Neuroscience, Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA.

4. RIKEN-MIT Center for Neural Circuit Genetics at the Picower Institute for Learning and Memory, Department of Biology and Department of Brain and Cognitive Sciences. Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

5. Gatsby Computational Neuroscience Unit, UCL, London, UK.

Abstract:

The anterior cingulate cortex (ACC) is implicated in learning the value of actions, and thus in allowing past outcomes to influence the current choice. However, it is not clear whether or how it contributes to the two major ways such learning is thought to happen: model-based mechanisms that learn action-state predictions and use these to infer action values; or model-free mechanisms which learn action values directly through reward prediction errors. Having confirmed, using a classical probabilistic reversal learning task, that optogenetic inhibition of ACC neurons on single trials indeed affected reinforcement learning, we examined the consequence of this manipulation in a novel two-step decision task designed to dissociate model-free and model-based learning mechanisms in mice. On the two-step task, silencing spared the influence of the trial outcome but reduced the influence of the experienced state transition. Analysis using reinforcement learning models indicated that ACC inhibition disrupted model-based RL mechanisms.

Introduction:

The anterior cingulate cortex (ACC) has long been implicated in reward guided decision making (Rushworth et al., 2004; Rushworth and Behrens, 2008). ACC neurons encode diverse decision variables (Cai and Padoa-Schioppa, 2012; Ito et al., 2003; Matsumoto et al., 2003; Sul et al., 2010), but ACC has been particularly associated with action reinforcement (Hadland et al., 2003; Kennerley et al., 2006; Rudebeck et al., 2008). However, instrumental learning is not a unitary phenomenon but rather is thought to be mediated by parallel control systems which use different computational principles to evaluate choices (Balleine and Dickinson, 1998; Daw et al., 2005; Dolan and Dayan, 2013). It has recently become a pressing problem to understand the neural underpinnings of these controllers and their interactions.

In familiar environments when executing well practiced actions, behaviour is apparently controlled by a habitual system thought to employ model-free reinforcement learning (RL) (Sutton and Barto, 1998). Model-free RL uses reward prediction errors to acquire or cache preferences between actions. However, when the environment or motivational state changes, model-free preferences can become out of date, and actions are instead determined by a goal-directed system believed to follow the precepts of model-based RL (Sutton and Barto, 1998). Model-based RL learns a predictive model of the consequences of actions, i.e. the states and rewards to which they typically immediately lead, and evaluates options by using the model to simulate or otherwise estimate their resulting long-run outcomes. Such a dual controller approach is beneficial because model-free and model-based RL possess complementary strengths, the former allowing quick and computationally cheap decision making at the cost of slower adaptation to changes in the environment, the latter flexible and efficient use of new information at the cost of computational effort and decision speed.

On specific anatomical and physiological grounds, we hypothesised that ACC is a component of the model-based control system. Firstly, the ACC provides a massive input to posterior dorsomedial striatum (Oh et al., 2014; Hintiryan et al., 2016), a region critical for model-based control as assessed through outcome-devaluation (Yin et al., 2005a, 2005b; Hilario et al., 2012). Secondly, decision related signals in ACC suggest that it plays a role in representing task contingencies beyond model-free cached values (Daw et al., 2011; Cai and Padoa-Schioppa, 2012; Karlsson et al., 2012; O'Reilly et al., 2013; Doll et al., 2015). We therefore sought to test the role of ACC in a reward guided decision task able to dissociate model-based and model-free mechanisms.

The classical approach to dissociating the systems in the laboratory involves outcome devaluation (Adams and Dickinson, 1981; Colwill and Rescorla, 1985). A subject is first trained to perform an action to receive a reward. The reward is then devalued, e.g. through pairing with illness, and the subject's subsequent tendency to perform the action is tested in extinction, i.e. without further rewards being delivered. If the action is mediated by a model-based prediction of the specific outcome to which it leads, devaluing that outcome will reduce the tendency to perform the action. If, on the other hand, the action is mediated by a cached model-free action value, devaluation will have no effect (Balleine and Dickinson, 1998; Daw et al., 2005). Learned actions often transition from being devaluation sensitive or goal-directed early in learning to being devaluation insensitive or habitual after extensive training (Dickinson et al., 1983; Dickinson, 1985). Lesion and inactivation studies using outcome devaluation indicate that goal-directed and habitual behaviours rely on partially separate cortical-basal ganglia circuits (Balleine et al., 2003; Killcross and Coutureau, 2003; Ostlund and Balleine, 2005; Yin et al., 2004, 2005b; Hilario et al., 2012; Gremel and Costa, 2013a, 2013b).

Unfortunately, outcome devaluation has limitations as a paradigm for decision neuroscience. Firstly, the critical devaluation test during which behavioural strategies are dissociated must be short because it is performed in extinction, limiting the number of choices or actions performed. Secondly, devaluation is a unidirectional single-shot manipulation of value. Neurophysiology thrives on behavioural paradigms that generate large decision datasets with parametric variation of decision variables. However, in workhorse tasks such as perceptual decision making or probabilistic reversal learning, the only uncertainty about the outcome of each decision is whether reward will be directly delivered. Thus, model-based prediction of future state and model-free prediction of future reward are ineluctably confounded.

Instead, at least for human subjects, novel tasks have recently been developed which aim to distinguish model-free and model-based reasoning in a stable manner over many trials. These tasks generally require subjects to take multiple steps through a decision tree to reach rewards, thus licensing the simulation-based search that is characteristic of the model-based controller (Daw et al., 2011; Simon and Daw, 2011; Huys et al., 2012). The most widely used is the so called two-step task (Daw et al., 2011), in which a choice between two actions leads probabilistically to one of two different states, in which further actions lead probabilistically to reward. Daw's two-step task has been used to assess the influence on behavioural strategy of behavioural (Otto et al., 2013, 2014) and neuronal manipulations (Wunderlich et al., 2012; Smittenaar et al., 2013), genetic factors (Doll et al., 2016), psychiatric illness (Sebold et al., 2014; Voon et al., 2015), and variants have also been used to examine more mechanistic aspects of interaction between the systems (Lee et al., 2014; Keramati et al., 2016; Doll et al., 2015). There is substantial interest from a number of groups in developing versions of the task for animal subjects to permit the use of more powerful neuroscience tools (Miller at al. Soc. Neurosci. Abstracts 2013, 855.13, Groman et al. Soc. Neurosci. Abstracts 2014, 558.19, Miranda et al. Soc. Neurosci. Abstracts 2014 756.09).

Here, we report our adaptation of the two-step task to study model-based and model-free learning in mice, and the use of our novel variant to probe the involvement of the anterior cingulate cortex (ACC), a region expected to be centrally involved. Based on an in depth computational analysis (Akam et al., 2015), we substantially modified the implementation and structure of the task, developing a new version in which both the reward probabilities in the leaf states of the decision tree and the action-state transition probabilities change over time. Here, detailed characterisation of subjects' behaviour indicated that, as in the human version, choices were guided by a mixture of model-based and model-free RL. However, we also observed a number of previously unexplored characteristics, including forgetting about actions that were not chosen, perseverative influences that spanned multiple trials, and representation of actions both in terms of the choices they represent and the motor output they require.

We found that optogenetic silencing of ACC neurons on individual trials reduced the influence of the experienced state transition on subsequent choice without affecting the influence of the trial outcome (rewarded or not). Analysis using RL models suggested this effect was due to reduced influence of model-based RL on ACC inhibition trials. For comparison purposes we performed the same ACC manipulation in a standard probabilistic reversal learning task, where it reduced the influence of the previous trial outcome on subsequent choice. These data are consistent with subjects using a combination of model-based and model-free RL in both tasks, but with the two-step task uniquely allowing a dissociation of their respective contributions to choice behaviour.

Results:

Single-trial inhibition of ACC impairs probabilistic reversal learning.

To confirm that ACC is involved in reward-guided decision making in mice, we first assessed whether optogenetic silencing of ACC neurons affected decision making in a standard probabilistic reversal learning task (Figure 1). Mice were trained to initiate each trial in a central nose-poke port which was flanked by left and right poke ports (Figure 1A). Trial initiation caused the left and right pokes to light up and subjects then chose between them for the chance of obtaining a water reward. Reward probabilities changed in blocks, with three block types; *left good* (left=0.75/right=0.25), *neutral* (0.5/0.5) and *right good* (0.25/0.75). Subject's choices tracked which option had higher reward probability (Figure 1B, C), choosing the correct option at the end of non-neutral blocks with probability 0.80 ± 0.04 (mean ± SD), and adapting to reversals in the reward probability with a time constant of 3.57 trials (exponential fit tau).

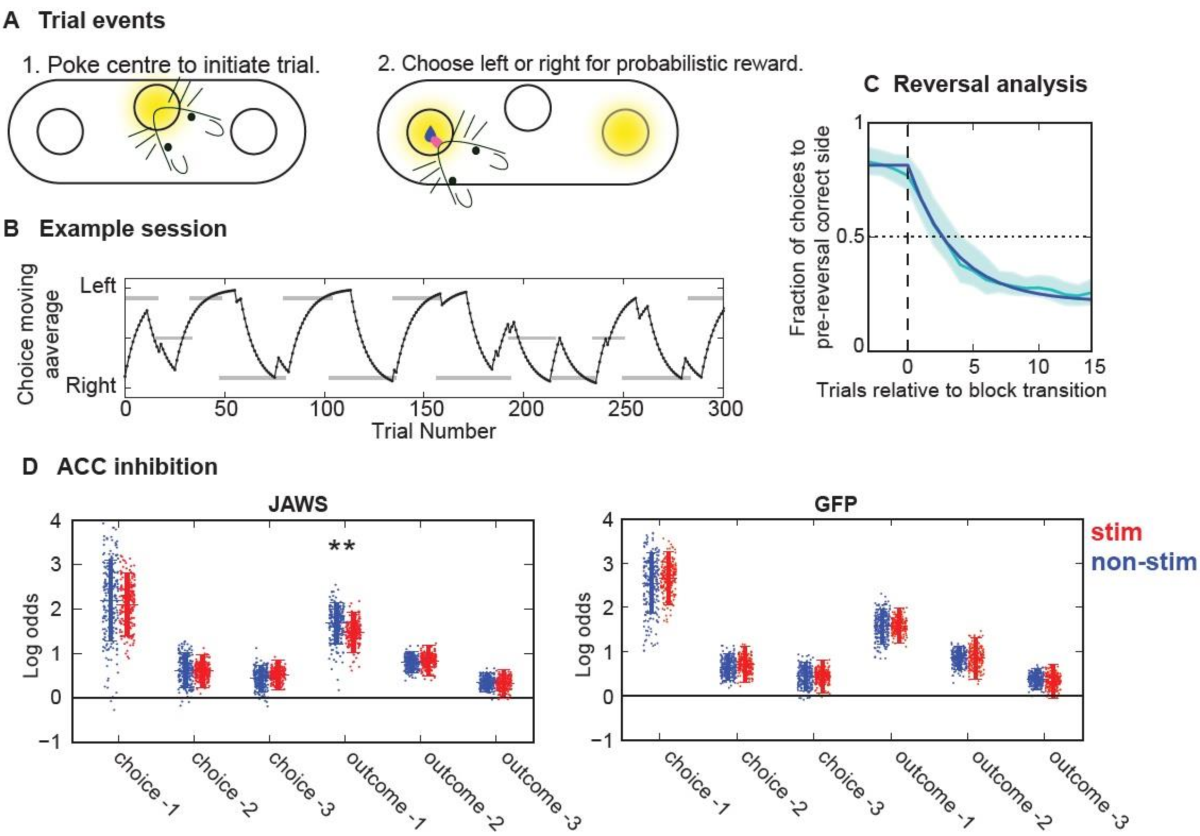


Figure 1. Optogenetic silencing of ACC in probabilistic reversal learning task. A) Diagram of apparatus and trial events. **B)** Example session, black line shows exponential moving average (tau = 8 trials) of choices, grey bars indicate reward probability blocks with y position of bar indicating whether left or right side has high reward probability or a neutral block. **C)** Choice probability trajectories around reversal in reward probabilities: Pale blue line – average trajectory, dark blue line – exponential fit, shaded area – cross-subject standard deviation. **D)** Logistic regression analysis showing predictor loadings for stimulated (red) and non-stimulated (blue) trials, for the ACC JAWS (left panel) and GFP controls (right panel). Bars indicate ±1 standard deviation of the population level distributions, dots indicate maximum a posteriori session fits. ** indicates significant difference (P<0.01) between stimulated and non-stimulated trials.

The following figure supplements are available for figure 1.

Figure supplement 1. **JAWS inhibition of ACC neurons.**

Figure supplement 2. **Average JAWS expression.**

We silenced the activity of ACC neurons on individual trials using the red-shifted halorhodopsin JAWS (Chuong et al., 2014). An AAV viral vector expressing JAWS-GFP under the CaMKII promotor was injected bilaterally into ACC of experimental animals (n=10 JAWS), while control animals (n=10) were injected with an AAV expressing GFP under the CaMKII promotor. Illumination was provided by a high power red LED chronically implanted above the cortical surface (Figure 1 - figure supplement 1). Electrophysiological recordings in animals implanted with micro-wire bundles (n=2) confirmed that red light (50mW, 630nM) from the implanted LEDs robustly inhibited ACC neurons (Figure 1- figure supplement 1). ACC neurons were inhibited using JAWS on a randomly selected 1/6 trials, with a minimum of two non-stimulated trials between each stimulated trial. Stimulation was delivered from when subjects poked in the side poke and received the trial outcome until the subsequent choice. The dataset comprised 12855 stimulated and 65186 non-stimulated trials for the JAWS animals and 11096 stimulated and 55913 non-stimulated trials for the controls.

We assessed the effect of ACC silencing using a logistic regression analysis with previous choices and outcomes as regressors. We separately analysed choices made during stimulation and on non-stimulated trials and used permutation tests to identify significant differences between the predictor loadings in the two conditions (Figure 1D). Previous choices predicted current choice with decreasing loading at increasing lag relative to the current trial. Obtaining reward further predicted repeating the rewarded choice, again with decreasing loading at increasing lag. ACC inhibition significantly reduced the influence of the most recent outcome (i.e., whether reward was received) on subsequent choice (permutation test $P = 0.004$ uncorrected, $P = 0.024$ Bonferroni corrected for 6 predictors), but did not affect the influence of either previous choices or earlier outcomes ($P > 0.18$ uncorrected). Light stimulation did not affect the influence of previous outcomes or choices on subsequent choice in the GFP controls ($P>0.38$ uncorrected) and the stimulation-by-group interaction was significant for the influence of the most recent outcome on choice ($P = 0.014$, permutation test).

These data indicate that transient ACC silencing disrupted reward-guided decision making in the probabilistic reversal learning task, however this task does not discriminate whether this was due to an effect on model-free mechanisms which learn action values directly, or model-based mechanisms which learn action-state transition probabilities and use these to guide choice. We therefore performed the same optogenetic manipulation in a multi-step decision task designed to dissociate the contribution of model-based and model-free reinforcement learning.

Development of a novel two-step task for mice

The task was based on that developed for humans by Daw et al. (2011) but both the physical format in which it was presented to subjects and the task structure were heavily adapted for use with mice. We first summarise changes to the task structure and their rationale before detailing the task implementation. As in the Daw two-step task, our version consisted of a choice between two 'first-step' actions which lead probabilistically to one of two 'second-step' states where reward could be obtained. Unlike the Daw task, in each second-step state there was a single action rather than a choice between two actions available, reducing the number of reward probabilities the subject must track from four to two (Figure 2 – figure supplement 1). In the original task, the stochasticity of the state transitions and reward probabilities caused both model-based and model-free control to obtain rewards at a rate negligibly different from random choice at the first-step (Akam et al., 2015; Kool et al., 2016). To promote task engagement, we increased the contrast between good and bad options by using a block-based reward probability distribution rather than the random walks used in the original, and by increasing the probability of common state transitions (see below) from 0.7 to 0.8. The final, and most significant, structural change was the introduction of reversals in the transition probabilities mapping the first-step actions to the second-step states. This step was taken to preclude subjects developing habitual strategies consisting of mappings from second-step states in which rewards had recently been obtained to specific actions at the first step (e.g. rewards in state X → chose action *x*, where action *x* is that which commonly leads to state X). Such strategies can, in principle, generate behaviour that looks very similar to model-based control despite not using a forward model which predicts the future state given chosen action (see Akam et al. (2015) for a detailed discussion).

We implemented the task using a set of four nose-poke ports: a low and a high poke in the centre, flanked by a left and a right poke (Figure 2A). Each trial started with the central pokes lighting up, mandating a choice. The resulting action led probabilistically to one of two states termed 'left-active' and 'right-active', in which respectively the left or right poke was illuminated. The subject then had to poke the illuminated side to gain a probabilistic water reward (Figure 2A,B). A 1 second inter-trial interval started from when the subject exited the side port at the end of the trial. The next trial then started with the illumination of the central pokes.

Both the transition probabilities linking the first-step actions to the second-step states, and the reward probabilities in each second-step state, changed in blocks (Figure 2C, D), such that each block was defined by the state of both the transition and reward probabilities. There were three possible states of the reward probabilities: *left good* (left=0.8/right=0.2), *neutral* (0.4/0.4) and *right good* (0.2/0.8). There were two possible states of the transition probabilities: *high→ right / low→left*, in which the high poke commonly (80% of trials) gave access to the right-active state and the low poke commonly gave access to the left-active state, and *high→ left / low→right* in which the high poke commonly gave access to the left-active, and the low poke commonly gave access to the right-active state. In either case, on 20% of trials, a rare transition occurred such that each first-step action gave access to the state commonly reached from the other first-step action. At block transitions, either the reward probabilities or the transition probabilities changed, except on transitions to neutral blocks, 50% of which were accompanied by a change in the transition probabilities (See Fig S3 for full block transition structure). Reversals in which first-step action (high or low) had higher reward probability, could therefore occur either due to the reward probabilities of the second-step states reversing, or due to the transition probabilities linking the first-step actions to the second-step states reversing. Block transitions were triggered based on a behavioural criterion (see methods) which resulted in block lengths of 63.6 ± 31.7 (mean ± SD) trials.

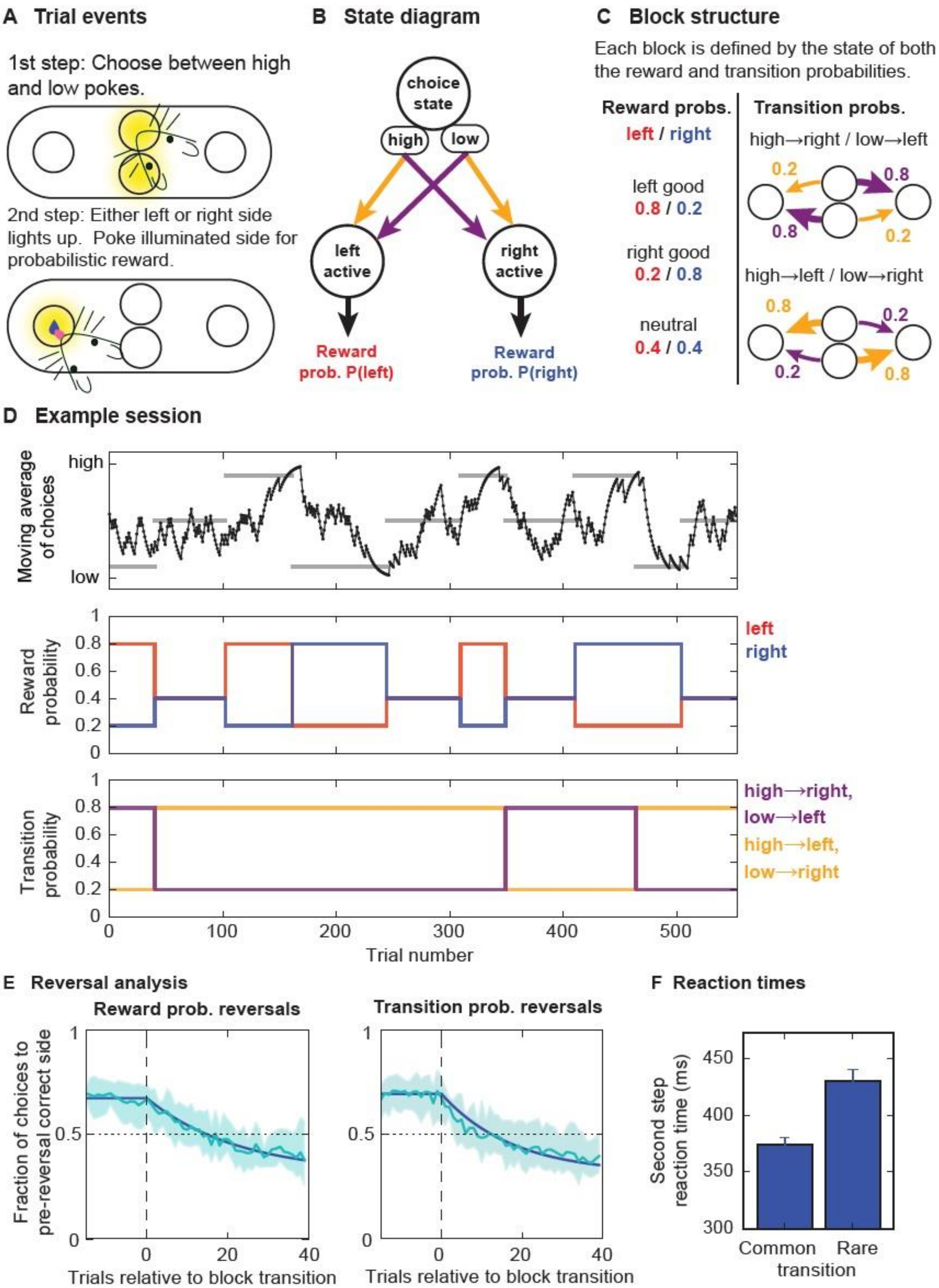


Figure 2. Two-step task. A) Diagram of apparatus and trial events. **B)** State diagram of task. **C)** Block structure, left side shows the three possible states of the reward probabilities, right side shows the two possible states of the transition probabilities. **D)** Example session: Top panel - Exponential moving average (tau = 8 trials) of choices. Horizontal grey bars show blocks, with correct choice (high, low or neutral) indicated by y position of bars. Middle panel – reward probabilities in left active (red) and right active (blue) states. Bottom

panel – Transition probabilities linking first-step actions (high, low pokes) to second step states (left/right active). **E)** Reversal analysis: Pale blue line – average trajectory, dark blue line – exponential fit, shaded area – cross-subject standard deviation. Left panel - reversals in reward probability, right panel – reversals in transition probabilities. **F)** Second step reaction times following common and rare transitions - i.e. the time between the first step choice and side poke entry. Error bars show cross-subject SEM.

The following figure supplements are available for figure 2.

Figure supplement 1. Comparison of original and new two-step task structures.

Figure supplement 2. Block transition probabilities.

Figure supplement 3. Body weight trajectory across training.

Subjects learned the task in 3 weeks with minimal shaping (see methods) and performed an average of 576 ± 174 (mean ± SD) trials per day thereafter. The baseline behavioural dataset consisted of sessions from day 22 of training onwards from 17 subjects, for a total of 400 sessions and 230237 trials. Subject's choices tracked which first-step action had higher reward probability (Figure 2D,E), choosing the correct option at the end of non-neutral blocks with probability 0.68 ± 0.03 (mean ± SD). Choice probabilities adapted faster (P = 0.009, bootstrap test) following block transitions in which the action-state transition probabilities reversed (exponential fit tau = 17.6 trials), compared with block transitions in which the reward probabilities in the two second-step states reversed (tau = 22.7 trials, Figure 2E). Reaction times at the second step, i.e. the latency from when the left or right side illuminated till the subject poked in the corresponding port, were faster following common than rare transitions ($P = 2.8 \times 10^{-8}$, paired t-test) (Figure 2F).

The choice probability trajectories around reversals show that subjects tracked which choice is best, but do not discriminate whether they used model-based or model-free RL. Both strategies are capable of tracking the best option, but do so in different ways: a model-based strategy learns estimates of the transition-probabilities linking the first-step actions to second-step states, and the reward probabilities in these states, and calculates the expected value of choosing each first-step action by combining these. By contrast, a model-free strategy directly learns action values for the first-step actions through the reward prediction errors that occur when the second-step is reached, and, via what is known as an eligibility trace, when the outcome (rewarded or not) is obtained after the second-step. As these different strategies learn different representations of the world, which are updated in different ways based on experienced events, it may be possible to dissociate them based on the fine structure of how events on each trial affect subsequent choices. We employ both of the two analysis approaches that are traditionally employed to do this: logistic regression showing how events on each trial affect subsequent choices, and direct fitting to the behavioural data of combined model-based and model-free reinforcement learning models. We detail these approaches below, and use them to unpick the effects of silencing the ACC.

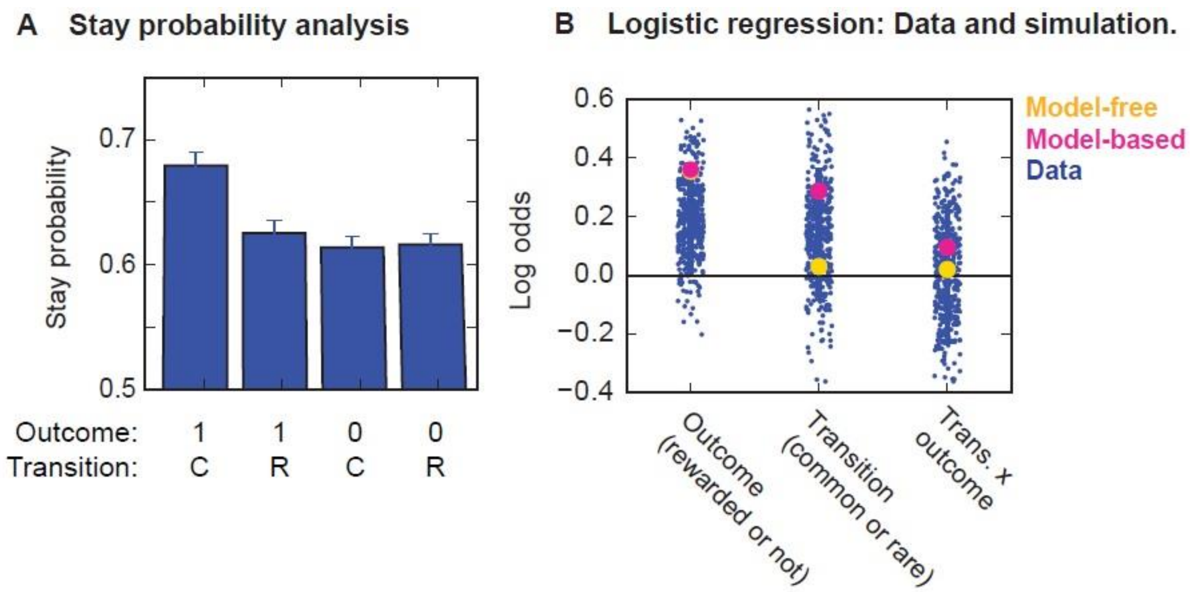


Figure 3. Stay probability and logistic regression analyses. A) Stay probability analysis. Fraction of trials the subject repeated the same choice following each combination of outcome (rewarded (1) or not (0)) and transition (common (C) or rare (R)). Error bars show cross-subject SEM. **B)** Logistic regression loadings for predictors; *outcome* (tendency to repeat choices following reward), *transition* (tendency to repeat choices following common transitions) and *transition-outcome interaction* (tendency to repeat choices following rewarded common transition trials and non-rewarded rare transition trials), comparing subject's data (blue) with simulated data from a model-free (yellow) and model-based (pink) agent fit to the subjects behaviour. For subjects data; blue bars indicate ±1 standard deviation of the population level distributions, blue dots indicate maximum a posteriori (MAP) session fits. The full set of predictor loadings is shown in figure supplement 1.

The following figure supplements are available for figure 3.

Figure supplement 1. Full logistic regression model fit.

Logistic regression analysis to disambiguate model-based versus model-free strategies

The simplest picture of behaviour is the raw so-called stay probabilities of repeating the first-step choice for the four possible combinations of transition and outcome (Figure 3A). Subjects were most likely to repeat choices following rewarded common transition trials, with a lower stay probability on rewarded rare-transition trials and non-rewarded trials. Logistic regression analyses of the relationship between choice and trial events test the nature of the interaction between transition and outcome, as this has historically been taken indicative of model-based reasoning. However, drawing such conclusions requires including various additional predictors in the model to capture strong, potentially confounding, effects. Some of these are conventional – for instance, accommodating perseveration or alternation between first-step choices and other direct biases of choice. However, we recently showed (Akam et al., 2015) the necessity of including an additional predictor which promotes repeating correct choices, as this avoids the effect of untoward correlations.

We therefore performed a logistic regression analysis which predicted stay probability as a function of trial events (outcome, transition and their interaction), with four additional regressors: the regressor discussed above which promoted repeating correct choices, a regressor which promoted repeating the previous choice, and two regressors capturing choice biases discussed below (Figure 3B, Figure 3 – figures supplement 1). Positive loading on the outcome predictor indicated that receiving reward was reinforcing (i.e. predicted staying) (P < 0.001, bootstrap confidence interval). Positive loading on the transition predictor indicated that experiencing common transitions was also reinforcing (P < 0.001). Loading on the transition-outcome interaction predictor was not significantly different from zero (P = 0.79). The absence of transition-outcome interaction has been used in the context of the traditional Daw two-step task (Daw 2011) to suggest that behaviour is model-free. However, we have previously shown (Akam et al. 2015) that this depends on the subjects not learning the transition probabilities from the transitions they experience. Such fixedness is reasonable for the traditional task, for which the probabilities are fixed and are known to be so by the human subjects. It is not for our task. Our analysis (Akam et al. 2015) suggests that when model-learning is included, loading in the logistic regression analysis is shifted off the interaction predictor and onto the outcome and transition predictors.

To understand more precisely the implications of this analysis, we simulated the behaviour of a model-based and a model-free RL agent, with the parameters of both fit to the behavioural data, and performed the logistic regression analysis on the data simulated from both models (Figure 2B). Data simulated from the model-free agent showed a large loading on the outcome regressor (i.e. rewards were reinforcing), but minimal loading on the transition and transition-outcome interaction regressors. By contrast, data simulated from the model-based agent showed a large loading on both outcome and transition predictors (i.e. both rewards and common transitions were reinforcing), and a small loading on the interaction predictor. The robust loading on the transition predictor observed in the experimental data in therefore consistent with subjects using model-based control as a component of their behavioural strategy.

In addition to the three predictors reflecting the influence of the previous trial's events, positive loading on the 'stay' predictor (Figure 3 – figure supplement 1, P < 0.001), indicated an overall tendency to repeat choices, consistent with the raw stay probabilities (Fig 3a). The 'correct' predictor also showed positive loading (P < 0.001) indicating that subjects were more likely to repeat choices to the correct, i.e. higher reward probability option irrespective of the experienced trial outcome. Subjects showed a small bias towards the high poke (P < 0.001) suggesting that the physical layout of the pokes made this action somewhat easier to execute. We included a second bias predictor which captured asymmetry in subject's bias dependent on the side they finish the previous trial on, i.e. a positive loading on this predictor promoted a bias towards the high poke if the previous trial ended on the left side, and towards the low poke if the previous trial ended on the right side. We term this a ‘rotational’ bias as positive loading promotes clockwise movement around the set of pokes (e.g. left→high, right→low), while negative loading promotes counter-clockwise movement. Though loading on this predictor was not on average different from zero (P = 0.092), it exhibited a substantial spread across the population of sessions such that a subset of sessions showed a strong rotational bias in either direction. Including this predictor substantially improved integrated Bayes Information Criterion (iBIC) scores for the regression model (Δ iBIC = 2639) indicating it captured a real feature of the data. Subjects may have developed this form of bias because it is the simplest fixed response pattern that was not penalised by the block transition rule: As block transitions were triggered based on a moving average of correct choices, developing an overall bias for the high or low poke resulted in the favoured poke spending most of the time as the bad option. Rotational bias may therefore be a default action which could be quickly executed when there was little evidence to suggest one option was better than the other.

Single-Trial Anterior Cingulate silencing in the two-step task impairs model based strategies

Parameters for optogenetic silencing in the two-step task were as closely as possible matched to those used in the probabilistic reversal learning task, with the same viral vector, injection sites and light stimulation. Again, optogenetic inhibition was delivered on a randomly selected 1/6 of trials, with a minimum of two non-stimulated trials between each stimulation trial. Inhibition was delivered from when the subject entered the side poke and received the trial outcome until the subsequent choice. The JAWS dataset comprised 11 animals with 12827 stimulated and 64523 non-stimulated trials, the GFP control dataset 12 animals, 11663 stimulated and 59408 non-stimulated trials.

We evaluated the effect of ACC inhibition on behaviour by performing the logistic regression analysis separately for choices which occurred during stimulation and on non-stimulated trials. As in the baseline dataset, both experimental and control animals showed positive loading on both the outcome and transition predictors on non-stimulated trials, indicating that both receiving reward and experiencing common transitions was reinforcing (Figure4 A,B). Optogenetic inhibition of ACC neurons reduced the influence of the previous state transition (common or rare) on subjects subsequent choice (P < 0.0002 uncorrected permutation test, P < 0.0006 Bonferroni corrected for multiple comparison of 3 predictors, stimulation by group interaction P = 0.029), but did not affect the influence of the previous reward (P = 0.94 uncorrected), or the transition-outcome interaction (P = 0.90 uncorrected).

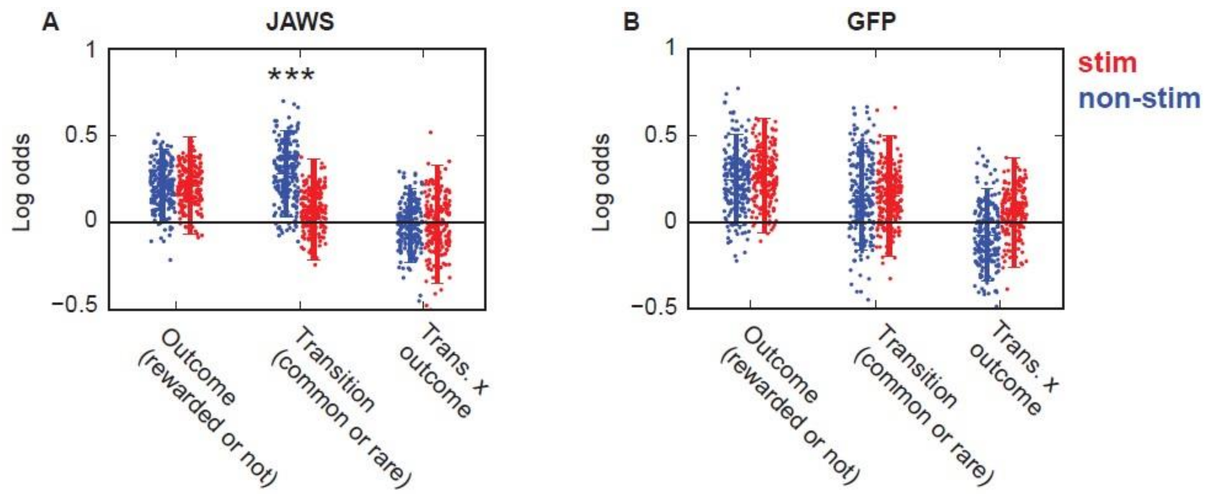


Figure 4. Optogenetic silencing of ACC in two-step task. A) Logistic regression analysis of ACC inhibition dataset showing loadings for the outcome, transition and transition-outcome interaction predictors for choices made on stimulated (red) and non-stimulated (blue) trials. **B)** As (a) but for GFP control animals. *** indicates significant difference (P<0.001) between stimulated and non-stimulated trials.

The following figure supplements are available for figure 4.

Figure supplement 1. ACC inhibition stay probabilities.

Figure supplement 2. ACC inhibition full logistic regression model fits.

Figure supplement 3. ACC inhibition reaction times.

This selective reduction in influence of the previous state transition while sparing the influence of the previous trial outcome is consistent with a shift from model-based towards model-free control as it is the transition predictor which most strongly differentiates behaviour generated by these two strategies (Figure 3B). Neither outcome, transition nor transition-outcome interaction predictors were affected by light stimulation in the GFP controls (Bonferroni corrected P > 0.2). In both experimental and control groups, light stimulation produced a small but significant bias towards the high poke, potentially reflecting an orienting response to the light (Bonferroni corrected P < 0.0015) (Figure 4 – figure supplement 1). Reaction times were not affected by light stimulation in either group (Paired t-test P > 0.36) (Figure 4 – figure supplement 2).

Reinforcement learning model analysis

To gain a sharper picture of the baseline behaviour and the effects of ACC silencing, we fitted and compared RL models to the respective datasets. Using our large baseline dataset, we performed an in-depth comparison of different RL models, as detailed in the supplementary material. Here, we summarise the principal findings. Our starting point was the RL agent used in the original Daw two-step task (Daw et al., 2011) in which behaviour is generated by a mixture of model-based and model-free strategies. Since the state transition probabilities change over time in our task, we modified the model to include ongoing learning about the transition probabilities.

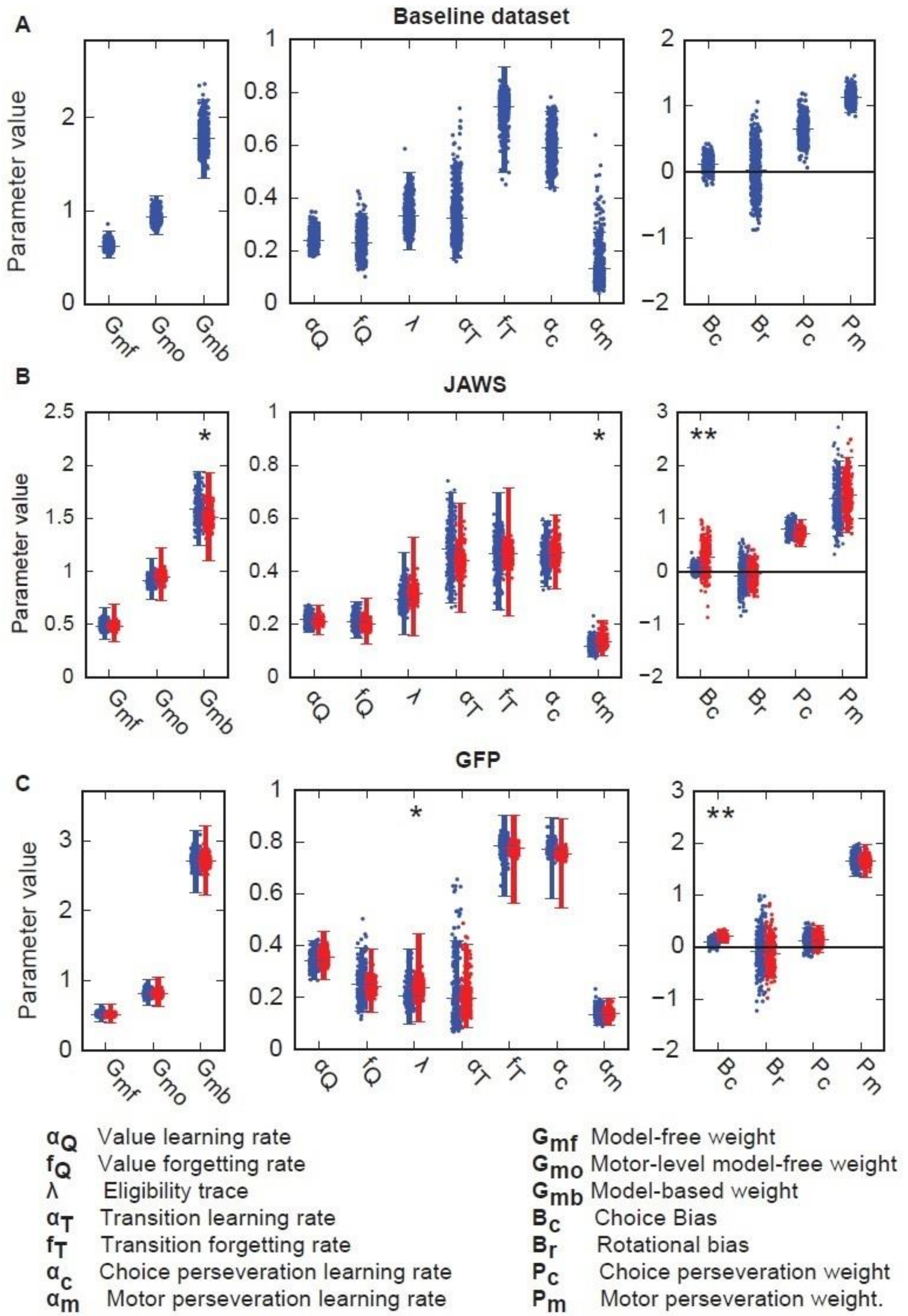


Figure 5. Reinforcement learning model fitting: A) Parameter values for best fitting RL model on baseline dataset. Bars indicate ±1 standard deviation of the population level distributions, dots indicate maximum a posteriori session fits. **B)** Reinforcement learning model fit to ACC inhibition dataset whose parameters take

separate values on stimulated (red) and non-stimulated (blue) trials. **C)** As (b) but for GFP control animals. * indicates significant difference (P<0.05) between stimulated and non-stimulated trials, ** indicates P < 0.01.

The following figure supplements are available for figure 5.

Figure supplement 1. Baseline dataset BIC score model comparison.

Figure supplement 2. Alternative RL model fits.

Figure supplement 3. Simulating effects of stimulation.

As with human behaviour on the Daw two-step task, the model (Figure 5A, Figure 5 - figure supplement 1) that best fit our baseline dataset used a mixture of model-based and model-free control. However, model comparison indicated the existence of a number of further structural features that have not previously been reported in models used for the Daw two-step task: forgetting about the values and state transitions for not-chosen actions, action perseveration effects spanning multiple trials, and representation of actions both at the level of the choice they represent (e.g. high poke) and the motor action they require (e.g. left→high movement). These are discussed in detail in the supplementary material. Taken together, the additional features produced a very substantial improvement in fit quality (Δ iBIC = 11018) over the model which lacked them (Figure 5 – figure supplements 1,2).

In seeking to use the model that fit the baseline dataset most parsimoniously to identify what aspect of learning or control was disrupted by ACC stimulation, we therefore had to understand their potential disrupting effects on telling apart model-based and model-free behaviour from data. As we also discuss in the supplementary material, this is a significant concern because either perseveration or model-free RL occurring at the level of motor actions rather than choices can generate loading on the transition predictor in the logistic regression (Figure 5 – figure supplement 3), breaking the simple pattern observed in figure 3B whereby only model-based RL gives substantial loading on the transition predictor.

We therefore sought to understand what aspect of learning or control was affected by the ACC inhibition by fitting a version of the RL model to the stimulation dataset in which parameters were free to take different values on stimulated and non-stimulated trials. In the JAWS animals (Figure 5B), the weighting parameter for the model-based system, which controls how strongly model-based action values influence choice, was significantly reduced on stimulation trials (P = 0.021, permutation test). This was not observed in control GFP animals (P = 0.348). We also found that the learning rate for motor-level perseveration was increased in stimulation trials (P = 0.01). The absolute size of the effects were not large, though this is likely influenced by the fitting procedure we used whereby we fit a version of the model in which parameters were constrained to take the same value on stimulated and unstimulated trials and then used this fit as the starting conditions for fitting the full model (see methods). Consistent with the logistic regression analyses, bias towards the high poke was significantly higher in both JAWS and GFP control animals on stimulation trials (P < 0.001), which likely reflects a bias caused by the light. The control animals also showed a significantly higher value for the eligibility trace parameter on stimulated trials (P = 0.027).

Taken in isolation this model fitting analysis would not be taken as robust support for an effect of ACC inhibition on model-based control because the effects would not survive multiple comparison correction for the large number of model parameters. However, we are not using this analysis to demonstrate the existence of an effect, but rather to test a hypothesis and probe the nature of the effect found in the regression analysis. Therefore, the lack of multiple comparison correction is appropriate here. We know that ACC inhibition affected some aspect of learning or control which causes experiencing a common transition to promote repeating the preceding choice (Figure 4A). Standard model-free RL does not predict any effect of transition type on choice while model-based RL does (Figure 3B), however we found that such an influence could also be generated by other factors, specifically perseveration or model-free RL occurring at the level of motor actions (Figure 5 – figure supplement 3). The RL analysis of the stimulation data supports the hypothesis that it is reduced influence of model-based RL on choice that explains the effect observed in the regression analysis as the weighting parameter for the model-based component was reduced on stimulation trials. The increased learning rate for motor-level perseveration should if anything increase loading on the transition predictor and hence could not explain the regression analysis effect. The probabilistic reversal learning task further argues against the effect of ACC inhibition being on outcome independent perseveration at the motor-level as in this task ACC inhibition reduced the influence of the most recent outcome.

Discussion:

We developed a novel two-step decision task for rodents that was designed to dissociate model-based and model-free RL. We used this task to probe the effect on reward guided behaviour of silencing ACC neurons, finding that optogenetic inhibition on individual trials reduced the influence of the experienced state transition, but not the trial outcome, on subsequent choice. Analysis using RL models suggested these effects were due to a disruption of model-based control.

The task was adapted from the two-step decision making task developed for human subjects by Daw and colleagues (Daw et al., 2011). The Daw two-step has been widely adopted because it offers the possibility of dissociating control strategies during ongoing learning and decision making, and generates large decision datasets well suited to behavioural modelling, manipulations and neurophysiology. However, in Akam et al. (2015) and here, we identified and addressed a significant challenge for the presently popular programme of developing versions of this task for animal subjects – that subjects may develop habitual mappings from where rewards are received to first step actions (referred to as extended state representations) which can generate behaviour that closely resembles model-based strategies. This is a particular concern in animal studies due to the different way subjects learn the task. Human subjects participating in the Daw two-step task are given detailed information about the structure of the task beforehand such that they start with a largely correct model, and then perform a limited number (~200) of trials. By contrast, animal subjects are typically extensively trained to reach the required performance level before recordings or manipulations are performed, giving ample opportunity to learn alternative strategies. In humans, extensive training renders apparently model-based behaviour resistant to a cognitive load manipulation (Economides et al., 2015) which normally disrupts model-based control (Otto et al., 2013), suggesting that it is possible to develop automatized strategies which closely resemble planning.

Motivated by this concern, we modified the task structure, introducing reversals into the transition probabilities mapping the first-step actions to the second-step states. This breaks the long term predictive relationship between where rewards are obtained and which first-step action has higher value, precluding a habit-like strategy that exploits this simple relationship, but not confounding a model-based strategy beyond requiring ongoing learning about the current state of the transition probabilities. The resulting task is quite complex compared with typical rodent decision tasks, and it is notable that mice are capable not just of learning it, but of doing so in a few weeks with minimal shaping. A further advantage of introducing reversals in the transition probabilities is that over the course of a session, the action-state transition probabilities, first-step action-values, and second-step state values are mutually decorrelated from each other. This should provide rich opportunity for future work identifying these decision variables in neural activity.

Our approach to developing a rodent two-step task contrasts with that taken by Miller et al. (Miller et al., 2016b) who retained the fixed transition probabilities of the original Daw task. Model-free use of extended state representations can produce a similar pattern of regression loadings to those observed by Miller et al., but interpreted by them in model-based terms. Indeed, the rats in the Miller et al. study showed little or no evidence of classical model-free behaviour leading to their conclusion that the behaviour is dominated by model-based planning. This might be surprising as even humans who have been explicitly told the correct structure of the Daw two-step task show an approximately even mix of model-based and model-free strategies.

Using our large baseline dataset, we performed a detailed characterisation of subject's behaviour on the new task, including an extensive process of RL model comparison. This indicated that subjects used a mixture of model-based and model-free RL, consistent with human subjects on the Daw two-step task. The model comparison also revealed a number of unexpected features of the behaviour; forgetting about value and state transition probabilities for not chosen actions, perseveration effects spanning multiple trials, and representation of actions both in terms of the choice they represent and the motor action they require. We are not aware of studies which have yet compared models including these elements on human two-step task data.

In retrospect, given the finding that representations at the motor-level influenced choice behaviour, the physical implementation of the task we used had a significant shortcoming: The action required to execute a given first step choice was different depending on the state reached at the second step on the previous trial. This caused unnecessary ambiguity in interpreting regression loadings in terms of control strategy and should be remedied in future work with this class of tasks by modifying the physical layout of the apparatus.

As a target for silencing, we chose the cingulate cortex between AP +1 and AP -0.5 (Figure 2 – figure supplement 2), which a recent cytoarchitectural study classifies as straddling the boundary between anterior-cingulate regions 24a and 24b and mid-cingulate regions 24a' and 24b' (Vogt and Paxinos, 2014). Although it has not hitherto been studied in the context of distinguishing actions and habits, there are anatomical, physiological and lesion-based reasons in rodents, monkeys and humans for considering this particular role for the structure. First, neurons in rat (Sul et al., 2010) and monkey (Ito et al., 2003; Matsumoto et al., 2003; Kennerley et al., 2011; Cai and Padoa-Schioppa, 2012) ACC carry information about chosen actions, reward, action values and prediction errors during decision making tasks. Where reward type (juice flavour) and size were varied independently (Cai and Padoa-Schioppa, 2012), a subset of ACC neurons encoded the chosen reward type rather than the reward value, consistent with a role in learning action-state relationships. In a probabilistic decision making task in which reward probabilities changed in blocks, neuronal representations in rat ACC underwent abrupt changes when subjects detected a possible block transition (Karlsson et al., 2012). This suggests that the ACC may represent the block structure of the task, a form of world model used to guide action selection, albeit one based on learning about latent states of the world (Gershman and Niv, 2010; Akam et al., 2015), rather than the forward action-state transition model of classical model-based RL.

Second, neuroimaging in the Daw two-step task has identified representation of model-based value in the BOLD signal in anterior- and mid-cingulate regions (Daw et al., 2011; Doll et al., 2015).

Likewise, neuroimaging in a saccade task in which subjects constructed and updated a model of the location of target appearance observed ACC activation when subjects updated an internal model of where saccade targets were likely to appear, (O'Reilly et al., 2013).

Third, ACC lesions in macaques produce deficits in tasks which require learning of action-outcome relationships (Hadland et al., 2003; Kennerley et al., 2006; Rudebeck et al., 2008), though the designs do not identify whether it is representation of the value or other dimensions of the outcome which were disrupted. Lesions of rodent ACC produce selective deficits in cost benefit decision making where subjects must weigh up effort against reward size (Walton et al., 2003; Rudebeck et al., 2006); however, again, the associative structures concerned are not clear.

Finally, the ACC provides a massive innervation to the posterior dorsomedial striatum (Oh et al., 2014; Hintiryan et al., 2016), a region necessary for learning and expression of goal directed action as assessed by outcome devaluation (Yin et al., 2005a, 2005b; Hilario et al., 2012).

We duly found that silencing ACC neurons on individual trials produced a selective change in how the previous trials events affected choice, reducing the influence of the previous state transition, while sparing the influence of reward. This appeared to be due reduced influence of model-based control on stimulated trials.

Recent discussion has focussed on whether ACC plays a direct role in decision making by calculating decision variables such as the expected value of possible courses of action, or a higher level function of deciding how much computational effort to expend on a decision (Kolling et al., 2016; Shenhav et al., 2016). Our results do not discriminate between these theories, because a shift in the balance between model-based and model-free control could occur either due to directly disrupting the model-based controller, or disrupting a higher-level system which arbitrated between their usage.

In sum, we suggest that our study offers a pioneering example of both the prospects and perils for the development of a new class of behavioural neuroscience investigations. We showed that it is possible to fashion sophisticated behavioural tasks that even mice can acquire quickly and effectively, thus affording all the benefits of modern genetic tools. However, in doing so, we showed the necessity for examining the behaviour in painstaking detail, lest one be misled by surface characteristics. We then provided suitably qualified support for the involvement of a key region of the brain in a cognitive trade-off of great contemporary interest. Our methods should offer rich opportunities for addressing this and other questions concerning the implementation and interaction of different neural control systems.

Methods:

Animals. All procedures were reviewed and performed in accordance with the Champalimaud Centre for the Unknown Ethics Committee guidelines. 59 male C57BL mice aged between 2 – 3 months at the start of experiments were used in the study. Mice were housed socially, except for 1 week in individual housing post-surgery where applicable. Animals were housed under a 12 hours light/dark cycle with experiments performed during the light cycle. 17 subjects were used in the two-step task baseline behaviour dataset. 14 subjects (8 JAWS, 6 GFP controls) were used for the two-step task ACC manipulation only. 14 subjects (8 JAWS, 6 GFP controls) were used for the probabilistic reversal learning task ACC manipulation only. 14 subjects (8 JAWS, 6 GFP controls) were first trained and tested on the two-step ACC manipulation, then retrained for a week on the probabilistic reversal learning task and tested on the ACC manipulation in this task. 7 JAWS-GFP animals were excluded from the study due to poor or mislocated JAWS expression. In the group that was tested on both tasks, 1 Jaws and 2 control animals were lost from the study before optogenetic manipulation on the probabilistic reversal learning task due to failure of the LED implants. The resulting group sizes for the optogenetic manipulation experiments were as reported in the results section.

Behaviour

Mice were placed on water restriction 48 hours before the first behavioural training session, and given 1 hour ad libitum access to water in their home cage 24 hours before the first training session. Mice received 1 training session per day of duration 1.5 – 2 hours, and were trained 6 days per week with 1 hour *ad libitum* water access in their home cage on their day off. During behavioural training mice had access to dry chow in the testing apparatus as we found this increased the number of trials performed and amount of water consumed. On days when mice were trained they typically received all their water in the task (typically 0.5-1.25ml), but additional water was provided as required to maintain a body weight >85% of their pre-restriction weight. Under this protocol, bodyweight typically dropped to ~90% of pre-restriction level in the first week of training, then gradually increased over weeks to reach a steady state of ~95-105% pre-restriction body weight (Figure 2 – figure supplement 3).

Behavioural experiments were performed in 14 custom made 12x12cm operant chambers using pyControl (http://pycontrol.readthedocs.io/en/latest/), a behavioural experiment control system built around the Micropython microcontroller. The pyControl task definition files are included in supplementary material. The apparatus, trial structure and block structure of the two-step task are described in the results section. Block transitions were triggered based on subject's behaviour, occurring 20 trials after an exponential moving average (tau = 8 trials) of subject's choices crossed a 75% correct threshold. The 20 trial delay between the threshold crossing and block transition allowed subjects performance at the end of blocks to be assessed without selection bias due to the block transition rule. In neutral blocks where there was no correct choice, block transitions occurred with 0.1 probability on each trial after the 40^{th}, giving a mean neutral block length of 50 trials. Subjects started each session with the reward and transition probabilities in the same state that the previous session finished on.

Subjects encountered the full trial structure from the first day of training. The only task parameters that were changed over the course of training were the reward and state transition probabilities and the reward sizes. These were changed to gradually increase task difficulty over days of training, with the typical trajectory of parameter changes as follows:

Session number	Reward size (ul)	Transition probabilities (common / rare)	Reward probabilities (good / bad side)
1	10	0.9 / 0.1	First 40 trials all rewarded, subsequently 0.9 / 0.1
2 - 4	10	0.9 / 0.1	0.9 / 0.1
5 - 6	6.5	0.9 / 0.1	0.9 / 0.1
7 - 8	4	0.9 / 0.1	0.9 / 0.1
9 - 12	4	0.8 / 0.2	0.9 / 0.1
13+	4	0.8 / 0.2	0.8 / 0.2

The trials structure and block structure of the probabilistic reversal learning task are described in the results section. Block transitions from non-neutral blocks were triggered 10 trials after an exponential moving average (tau = 8 trials) crossed a 75% correct threshold. Block transitions from neutral blocks occurred with probability 0.1 on each trial after the 15^{th} of the block to give an average neutral block length of 25 trials.

Optogenetic Inhibition

Experimental animals were injected bilaterally with *AAV5-CamKII-Jaws-KGC-GFP-ER2* (UNC vector core, titre: 5.9 x 10^{12}) using 16 injections each of 50nL (total 800nL) spread across 4 injection tracks (2 per hemisphere) at coordinates: AP: 0, 0.5, ML: ±0.4, DV: -1, -1.2, -1.4, -1.6mm relative to dura. Control animals were injected with *AAV5-CaMKII-GFP* (UNC vector core, titre: 2.9 x 10^{12}) at the same coordinates. Injections were performed at a rate of 4.6nL/5 seconds, using a Nanojet II (Drummond Scientific) with bevelled glass micropipettes of tip diameter 50-100um. A circular craniotomy of diameter 1.8mm was centred on AP: 0.25, ML: 0, and a high power red led (Cree XLamp XP-E2) was positioned above the craniotomy touching the dura. The LED was mounted on a custom designed insulated metal substrate PCB (Figure 1 – figure supplement 1A). The LEDs were powered using a custom designed constant current LED driver built around the AL8805 integrated circuit. Light stimulation (50mW, 630nM) was delivered on stimulation trials from when the subject entered the side poke until the subsequent choice, up to a maximum of 6 seconds. Stimulation was delivered on a randomly selected 17% of trials, with a minimum of 2 non-stimulated trials between each stimulation trial followed by a 0.25 probability of stimulation on each subsequent trial. At the end of behavioural experiments, animals were sacrificed and perfused with paraformaldehyde (4%). The brains were sectioned in 50um coronal slices and the location of viral expression was characterised with fluorescence microscopy (Figure 1 – figure supplement 2).

Two animals were injected unilaterally with the JAWS-GFP virus using the coordinates described above and implanted with the LED implant and a movable bundle of 16 tungsten micro-wires of 23µm diameter (Innovative-Neurophysiology) to record unit activity. After 4 weeks of recovery, recording sessions were performed at 24 hour intervals and the electrode bundle was advanced by 50 um after each session, covering a depth range of 300 – 1300um from dura over the course of recordings. During recording sessions mice were free to move inside a sound attenuating chamber. Light pulses (50mW power, 5 second duration) were delivered at random intervals with a mean inter-stimulus interval of 30 seconds. Neural activity was acquired using a Plexon recording system running Omniplex v. 1.11.3. The signals were digitally recorded at 40000 Hz and subsequently band-pass filtered between 200 Hz and 3000 Hz. Following filtering, spikes were detected using an amplitude threshold set at twice the standard deviation of the bandpass filtered signal. Initial sorting was performed automatically using Kilosort (Pachitariu et al., 2016). The results were refined via manual sorting based on waveform characteristics, PCA and inter-spike interval histogram. Clusters were classified as single units if well separated from noise and other units and the spike rate in the 2ms following each spike was less than 1% of the average spike rate.

Behavioural analysis: All analysis of behaviour was performed in Python, full analysis code and behavioural data is included in supplementary material.

Logistic regression model

The logistic regression model for the two-step task predicted the probability of choosing the high poke as a function events on the previous trial using the following set of predictors:

Variables used to define two-step task regression predictors	
C	+1 if previous choice to high poke, -1 if previous choice to low poke
O	+1 if previous trial rewarded, -1 if previous trial not rewarded
T	+1 if previous trial had common transition, -1 if previous trial had rare transition
R	+1 if previous choice to correct (higher reward probability) option, -1 if previous choice to incorrect (lower reward probability) option, 0 if neutral block

Predictors used in two-step task logistic regression	
Bias: high/low	1 for all trials. (Promotes choosing high poke)
Bias: clockwise /counter-clockwise	0.5 if previous trial ended on left side, -0.5 if right side. (Promotes choosing high following trials ending on left, low following trials ending on the right)
Stay	$0.5\ C$ (Promotes repeating previous Choice)
Correct	$0.5\ C\ R$ (Promotes repeating correct choices)
Outcome	$0.5\ C\ O$ (Promotes repeating rewarded choices)
Transition	$0.5\ C\ T$ (Promotes repeating choices following common transitions)
Transition outcome interaction	$0.5\ C\ T\ O$ (Promotes repeating choices following rewarded common transitions and non-rewarded rare transitions).

Note, regression predictors were scaled to take values of ±0.5 such that the loading are in units of log-odds. The two-step task logistic regression excluded the first 20 trials after each reversal in the transition probabilities as it is ambiguous which transitions are common and rare at this point. This resulted in ~9% of trials being excluded from the logistic regression analysis.

The logistic regression analysis for the probabilistic reversal learning task predicted the probability of choosing the left poke as a function of events on the previous 3 trials, using the following set of predictors:

Variables used to define probabilistic reversal learning task regression predictors	
C_{-t}	1 if left poke chosen on trial $-t$, -1 if right poke chosen.
O_{-t}	1 trial $-t$ rewarded, -1 if trial $-t$ not rewarded.
Predictors used in probabilistic reversal learning task logistic regression	
Bias	1 for all trials (Promotes choosing left poke)
$Choice_{-t}$	$0.5\ C_{-t}$ for $t \in \{1,2,3\}$ (Promotes repeating choices)
$Outcome_{-t}$	$0.5\ C_{-t}O_{-t}$ for $t \in \{1,2,3\}$ (Promotes repeating rewarded choices)

Reinforcement learning modelling:

The following variables and parameters were used in the RL models:

RL model variables	
R	Reward obtained on trial (0 or 1)
a_1	Action taken at first step (high or low poke)
a_2	Action taken at second step (left or right poke)
a'_1	Action not taken at first step (high or low poke)
a'_2	Action not taken at second step (left or right poke)
m_1	Motor-level action taken at first step (e.g. left→high)
m'_1	Motor-level action not taken at first step
s_1	First step state (choice state)
s_2	Second step state (left-active or right-active)
s'_2	State not reached at second step (left-active or right-active)

$Q_{mf}(s, a)$	Model-free action value for action *a* in state *s*
$Q_{mo}(s_1, m)$	Motor-level model-free action value for motor action *m* following in state s_1
$P(s\|a)$	Estimated transition probability of reaching state *s* after taking action *a*
$C(s_1, a)$	Choice perseveration variable
$M(s_1, m)$	Motor perseveration variable
$B(s_1, a_i)$	Choice bias variable
$R(s_1, m_i)$	Rotational bias variable.
RL model parameters	
α_Q	Value learning rate
f_Q	Value forgetting rate
λ	Eligibility trace parameter
α_T	Transition learning rate
f_T	Transition forgetting rate
α_c	Learning rate for choice perseveration
α_m	Learning rate for motor-level perseveration
G_{mf}	Model-free action value weight
G_{mo}	Motor-level model free action value weight
G_{mb}	Model-based action value weight
B_c	Choice bias (high/low)
B_r	Rotational bias (clockwise/counter-clockwise)
P_c	Choice perseveration strength

P_m	Motor-level perseveration strength

RL Model equations:

Model-free RL: The action value update used by the model-free RL component was:

$$Q_{mf}(s_1, a_1) \leftarrow (1 - \alpha_Q)Q_{mf}(s_1, a_1) + \alpha_Q \left(Q_{mf}(s_2, a_2) + \lambda \left(R - Q_{mf}(s_2, a_2)\right)\right)$$

$$Q_{mf}(s_2, a_2) \leftarrow (1 - \alpha_Q)Q_{mf}(s_2, a_2) + \alpha_Q R$$

In models that included value forgetting this value of not chosen actions was updated as:

$$Q_{mf}(s_1, a'_1) \leftarrow (1 - f_Q)Q_{mf}(s_1, a'_1)$$

$$Q_{mf}(s'_2, a'_2) \leftarrow (1 - f_Q)\, Q_{mf}(s'_2, a'_2)$$

Model-based RL: The model-based component updated its estimate of the state transition probabilities mapping first-step action to second-step state as:

$$P(s_2|a_1) \leftarrow (1 - \alpha_T)P(s_2|a_1) + \alpha_T$$

$$P(s'_2|a_1) \leftarrow (1 - \alpha_T)P(s'_2|a_1)$$

In models that included transition probability forgetting, the state transition probabilities for the not chosen action decayed towards a uniform distribution as:

$$P(s_2|a'_1) \leftarrow (1 - f_T)P(s_2|a'_1) + 0.5f_T$$

$$P(s'_2|a'_1) \leftarrow (1 - f_T)P(s'_2|a'_1) + 0.5f_T$$

At the start of each trial, model-based first step action values were calculated as:

$$Q_{mb}(s_1, a_i) = Q(s_1, a_i) = \sum_j P\left(s_j \middle| a_i\right)Q_{mf}(s_j, a_2)$$

Motor-level model-free RL: Agents which included motor-level model-free RL learned values for the first step actions represented as motor movements (e.g. left→high). The motor movement m_i for a given choice a_i (high or low) at the first step is dependent on the second-step state (left or right) the previous trial ended on. Motor-level model-free action values were updated as:

$$Q_{mo}(s_1, m_1) \leftarrow (1 - \alpha_Q)Q_{mo}(s_1, m_1) + \alpha_Q \left(Q_{mf}(s_2, a_2) + \lambda \left(R - Q_{mf}(s_2, a_2)\right)\right)$$

In models with motor-level model-free RL and value forgetting, all motor-level model-free values except that of the action taken decayed as:

$$Q_{mo}(s_1, m'_1) \leftarrow (1 - f_Q) Q_{mo}(s_1, m'_1)$$

Perseveration: Choice perseveration was modelled using variables $C(s_1, a)$ which reflected the previous choice history. In models using a single trial choice kernel these were updated as:

$$C(s_1, a_1) \leftarrow 0.5$$

$$C(s_1, a'_1) \leftarrow 0$$

In models which used an exponential choice kernel, $C(s_1, a)$ were updated as:

$$C(s_1, a_1) \leftarrow (1 - \alpha_c) C(s_1, a_1) + 0.5\, \alpha_c$$

$$C(s_1, a'_1) \leftarrow (1 - \alpha_c) C(s_1, a'_1)$$

In models which used motor-level perseveration this was modelled using variables $M(s_1, m)$ which reflected the previous history of motor actions at the first step. The motor-preservation variable for the motor action executed was updated as:

$$M(s_1, m_1) \leftarrow (1 - \alpha_m) M(s_1, a_1) + 0.5\, \alpha_m$$

The motor perseveration variables for all other motor actions was updated as:

$$M(s_1, a'_1) \leftarrow (1 - \alpha_m) M(s_1, a'_1)$$

Biases: A bias for the high/low poke was modelled with a bias variable B which took values:

$B(s_1, a_i)$ = 0.5 if a_i is high poke, -0.5 if a_i is low poke.

The rotational bias (see results section) was modelled with a variable $R(m_i)$ which took values:

$R(s_1, m_i) = 0.5$ if m_i is a clockwise movement (left→high or right→low)

$R(s_1, m_i) =$ -0.5 if m_i is a counter-clockwise movement (left→low or right→high)

Combined action values: Model-free, motor-level model-free and model-based action values were combined with perseveration and bias terms to give the net action values that drove choice behaviour.

$$Q_{net}(s_1, a_i) = G_{mf} Q_{mf}(s_1, a_i) + G_{mo} Q_{mo}(s_1, m_i) + G_{mb} Q_{mb}(s_1, a_i) + P_c\, C(s_1, a_i) + P_m\, M(s_1, m_i) + B_c B(s_1, a_i) + B_r R(s_1, m_i)$$

Where G_{mf}, G_{mo} and G_{mb} are weights controlling the influence of respectively the model-free, motor-level model-free and model-based action values, P_c & P_m control the strength of choice- and motor-level perseveration, and B_c & B_r control the strength of choice and rotational biases, m_i is that motor action which equates to choice a_i given the second step state reached on the previous trial.

Given the net action values for the two first step actions, choice probability was given by the softmax decision rule:

Probability of choosing action $a_i = \frac{e^{Q_{net}(s_1,a_i)}}{\sum_j e^{Q_{net}(s_1,a_j)}}$

Hierarchical modelling:

Both the logistic regression analyses and reinforcement learning model fitting used a Bayesian hierarchical modelling framework (Huys et al., 2011), in which parameter vectors $\boldsymbol{h}_i$ for individual sessions were assumed to be drawn from Gaussian distributions at the population level with means and variance $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. The population level prior distributions were set to their maximum likelihood estimate:

$$\boldsymbol{\theta}^{ML} = argmax_{\boldsymbol{\theta}}\{p(D|\boldsymbol{\theta}\}$$

$$= argmax_{\boldsymbol{\theta}}\{\prod_i^N \int d\,\boldsymbol{h}_i\, p(D_i|\boldsymbol{h}_i)p(\boldsymbol{h}_i|\boldsymbol{\theta})\}$$

Optimisation was performed using the Expectation-Maximisation algorithm with a Laplace approximation for the E-step at the k-th iteration given by:

$$p(\boldsymbol{h}_i^k|D_i) = N(\boldsymbol{m}_i^k, \boldsymbol{V}_i^k)$$

$$\boldsymbol{m}_i^k = argmax_{\boldsymbol{h}}\{p(D_i|\boldsymbol{h})p(\boldsymbol{h}|\boldsymbol{\theta}^{k-1})\}$$

Where $N(\boldsymbol{m}_i^k, \boldsymbol{V}_i^k)$ is a normal distribution with mean $\boldsymbol{m}_i^k$ given by the maximum a posteriori value of the session parameter vector $\boldsymbol{h}_i$given the population level means and variance $\boldsymbol{\theta}^{k-1}$, and the covariance $\boldsymbol{V}_i^k$given by the inverse Hessian of the likelihood around $\boldsymbol{m}_i^k$. For simplicity we assumed that the population level covariance $\boldsymbol{\Sigma}$ had zero off-diagonal terms. For the k-th M-step of the EM algorithm the population level prior distribution parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ are updated as:

$$\boldsymbol{\mu}^k = \frac{1}{N}\sum_{i=1}^N \boldsymbol{m}_i^k$$

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^N \left[\left(\boldsymbol{m}_i^k\right)^2 + \boldsymbol{V}_i^k\right] - \left(\boldsymbol{\mu}^k\right)^2$$

Parameters were transformed before inference to enforce constraints ($0 < \{G_{mf}, G_{mo}, G_{mb}\}$, $0 < \{\alpha_Q,\ f_Q, \lambda, \alpha_T, f_T, \alpha_c, \alpha_m\} < 1$).

To avoid local minima reinforcement learning models fits were repeated 16 times with the means of the population level prior distributions initialised to random values, the repeat with the best likelihood was then used.

Model comparison:

To compare the goodness of fit for models with different numbers of parameters we used the integrated Bayes Information Criterion (iBIC) score. The iBIC score is related to the model log likelihood $p(D|M)$ as:

$$\log p(D|M) = \int d\boldsymbol{\theta}\ p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}|M)$$

$$\approx -\frac{1}{2} iBIC = \log p(D|\,\boldsymbol{\theta}^{ML}) - \frac{1}{2}|M|\log|\mathrm{D}|$$

Where |M| is the number of fitted parameters of the prior, |D| is the number of data points (total choices made by all subjects) and iBIC is the integrated BIC score. The log data likelihood given maximum likelihood parameters for the prior $\log p(D|\,\boldsymbol{\theta}^{ML})$ is calculated by integrating out the individual session parameters:

$$\log p(D|\,\boldsymbol{\theta}^{ML}) = \sum_i^N log \int d\boldsymbol{h}\ p(D_i|\boldsymbol{h})p(\boldsymbol{h}|\,\boldsymbol{\theta}^{ML})$$

$$\approx \sum_i^N log \frac{1}{K}\sum_{j=1}^K p(D_i|\boldsymbol{h}^j)$$

Where the integral is approximated as the average over K samples drawn from the prior $p(\boldsymbol{h}|\boldsymbol{\theta}^{ML})$. Bootstrap 95% confidence intervals were estimated for the iBIC scores by resampling from the population of samples drawn from the prior.

Permutation testing:

Permutation testing was used to assess the significance of differences in model fits between stimulated and non-stimulated trials. For the logistic regression analyses, the regression model was fit separately to stimulated and non-stimulated trials to give two sets of population level parameters $\boldsymbol{\theta_s} = \{\boldsymbol{\mu_s}, \boldsymbol{\Sigma_s}\}$ and $\boldsymbol{\theta_n} = \{\boldsymbol{\mu_n}, \boldsymbol{\Sigma_n}\}$, where $\boldsymbol{\theta_s}$ are the parameters for the stimulated trials and $\boldsymbol{\theta_n}$ are the parameters for the non-stimulated trials. The distance between the population level means for the stimulated and non-stimulated conditions were calculated as:

$$\Delta_{true} = |\boldsymbol{\mu_s} - \boldsymbol{\mu_n}|$$

An ensemble of L permuted datasets was then created by shuffling the labels on trials such that trials were randomly assigned to the 'stimulated' and 'non-stimulated' conditions. The model was fit separately to the stimulated and non-stimulated trials for each permuted dataset and the distance between population level means in the stimulated and non-stimulated conditions was calculated for each permuted dataset i as:

$$\Delta_{perm}^{i} = |\boldsymbol{\mu_s^i} - \boldsymbol{\mu_n^i}|$$

The distribution of distances Δ_{perm} over the population of permuted datasets approximates the distribution of distances under the null hypothesis that stimulation does not affect the model parameters. The P-values for the observed distances Δ_{true} are then given by:

$$\boldsymbol{P} = \frac{1}{L}\sum_{i=1}^{L} \boldsymbol{x}^i$$

where $\boldsymbol{x}^i = 1$ for $\Delta_{perm}^{i} \geq \Delta_{true}$, $\boldsymbol{x}^i = 0$ for $\Delta_{perm}^{i} < \Delta_{true}$

In addition to testing for a significant main effect of the stimulation we tested for significant stimulation by group interaction. We first evaluated the true distance between the effect sizes for the two groups as:

$$\Delta_{true} = |(\boldsymbol{\mu_s^{JAWS}} - \boldsymbol{\mu_n^{JAWS}}) - (\boldsymbol{\mu_s^{GFP}} - \boldsymbol{\mu_n^{GFP}})|$$

The approximate distribution of this distance under the null hypothesis that there was no difference between the groups was evaluated by creating an ensemble of permuted datasets in which we randomly assigned subjects to the JAWS and GFP groups and the interaction P value was calculated as above.

For reinforcement learning models, the model cannot be fitted separately to stimulated and non-stimulated trials because of the serial dependence of decision variables from trial to trial. We therefore created RL models where all or a subset of the model parameters took separate values on stimulated and non-stimulated trials, such that if the base model had n parameters the resulting model had 2n parameters. To test for significant differences between parameters on stimulated and non-stimulated trials, the model was fit to give a set of population level parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, of which a subset $\boldsymbol{\mu_s}, \boldsymbol{\Sigma_s}$ were active on stimulation trials and their counterparts $\boldsymbol{\mu_n}, \boldsymbol{\Sigma_n}$ were active on non-stimulation trials. As before the distances between the stimulated and non-stimulated parameter values were calculated as $\Delta_{true} = |\boldsymbol{\mu_s} - \boldsymbol{\mu_n}|$ and permutation testing otherwise proceeded as described above for the regression models.

The following procedure was used to minimise problems with local minima when these high parameter count RL models were fitted to stimulation data. We first fitted a version of the model in which the parameters were the same for stimulated and non-stimulated trials. This fit was repeated 16 times with randomised initial values for the population level prior means. The fit with the best likelihood across repeats was used to initialise the population level prior distribution for the full model in which parameters were free to take different values on stimulated and non-stimulated trials, such the stim and non-stim parameters started the fitting procedure with the same values. For permutation testing the same initial fit was used for the true and permuted datasets. To ensure that permutation test results were not dependent on the specific initial fit found, the whole procedure was repeated 20 times and the mean P value across the 20 repeats was taken. Permutation tests were run on the Oxford Advanced Research Computing (ARC) facility.

Bootstrap test for reversal analysis:

The speed of behavioural adaptation to reversals in the transition and reward probabilities was evaluated by fitting exponentials to the average choice probability trajectories following each type of reversal (Figure 1E). To test whether adaptation following reversals in transition probabilities was significantly faster than that following reversals in reward probabilities, we constructed a bootstrap confidence interval for the difference $\Delta_\tau = \tau_R - \tau_T$, where τ_R and τ_T are respectively the exponential time constants following reversals in the reward and transition probabilities. The bootstrap confidence interval was evaluated by creating an ensemble of L resampled datasets by drawing subjects with replacement from the set of subjects that comprised the baseline dataset. The bootstrap P-value was then evaluated as:

$$\boldsymbol{P} = \frac{1}{L}\sum_{i=1}^{L} \boldsymbol{x}^i$$

where $\boldsymbol{x}^i = 1$ for $\Delta_\tau < 0$, $\boldsymbol{x}^i = 0$ for $\Delta_\tau \geq 0$.

Logistic regressions of simulated data:

To evaluate the logistic regression loadings expected for a model-based and model-free agent on the task (Figure 2B), we first fitted each agent type to our baseline behavioural dataset using the hierarchical framework outlined above. The agents used were a model-free agent with eligibility traces and value forgetting, and a model-based agent with value and transition probability forgetting. We then simulated data (4000 sessions each of 500 trials) from each agent, drawing parameters for each session from the fitted population level distributions for that agent. We performed the logistic regression on the simulated data, again using the hierarchical framework as for the logistic regression analysis of experimental data.

Simulating effects of single trial inhibition

In Figure 5 – figure supplement 3 we simulated the effects of lesioning on 'stimulation' trials individual components of that RL model found to give the best fit to the baseline dataset. This was done by setting the weighting parameter for the relevant component to zero on stimulation trials, removing its influence on choice on that trial. The components lesioned and their respective weighting parameters were; choice-level model-free RL (G_{mf}), motor-level model-free RL (G_{mo}), model-based RL (G_{mb}), motor-level perseveration (P_m). For each lesion simulation, a simulated dataset (4000 sessions each of 500 trials) was generated using parameters for each session drawn from the population level distribution of the model fit to the baseline dataset. The logistic regression analysis of the simulated data was performed as on the experimental data by fitting the regression model separately to choices made on stimulated and non-stimulated trials.

Acknowledgements:

The authors thank Zach Mainen, Joe Patton, Mark Walton, Tim Behrens, Nathaniel Daw, Kevin Miller and Bruno Miranda for discussions about the work. The authors acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility (http://dx.doi.org/10.5281/zenodo.22558).

Competing interests:

The authors have no competing interests to report.

References:

Adams, C.D., and Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. Q. J. Exp. Psychol. Sect. B *33*, 109–121.

Akaishi, R., Umeda, K., Nagase, A., and Sakai, K. (2014). Autonomous Mechanism of Internal Choice Estimate Underlies Decision Inertia. Neuron *81*, 195–206.

Akam, T., Costa, R., and Dayan, P. (2015). Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. PLoS Comput Biol *11*, e1004648.

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

Balleine, B.W., Killcross, A.S., and Dickinson, A. (2003). The Effect of Lesions of the Basolateral Amygdala on Instrumental Conditioning. J. Neurosci. *23*, 666–675.

Cai, X., and Padoa-Schioppa, C. (2012). Neuronal encoding of subjective value in dorsal and ventral anterior cingulate cortex. J. Neurosci. *32*, 3791–3808.

Chuong, A.S., Miri, M.L., Busskamp, V., Matthews, G.A.C., Acker, L.C., Sørensen, A.T., Young, A., Klapoetke, N.C., Henninger, M.A., Kodandaramaiah, S.B., et al. (2014). Noninvasive optical inhibition with a red-shifted microbial rhodopsin. Nat. Neurosci. *17*, 1123–1129.

Colwill, R.M., and Rescorla, R.A. (1985). Postconditioning devaluation of a reinforcer affects instrumental responding. J. Exp. Psychol. Anim. Behav. Process. *11*, 120–132.

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

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., and Dolan, R.J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron *69*, 1204–1215.

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. Philos. Trans. R. Soc. B Biol. Sci. *308*, 67–78.

Dickinson, A., Nicholas, D.J., and Adams, C.D. (1983). The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. Q. J. Exp. Psychol. *35*, 35–51.

Dolan, R.J., and Dayan, P. (2013). Goals and Habits in the Brain. Neuron *80*, 312–325.

Doll, B.B., Duncan, K.D., Simon, D.A., Shohamy, D., and Daw, N.D. (2015). Model-based choices involve prospective neural activity. Nat. Neurosci. *18*, 767–772.

Doll, B.B., Bath, K.G., Daw, N.D., and Frank, M.J. (2016). Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning. J. Neurosci. *36*, 1211–1222.

Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M., and Dolan, R.J. (2015). Model-Based Reasoning in Humans Becomes Automatic with Training. PLoS Comput Biol *11*, e1004463.

Gershman, S.J., and Niv, Y. (2010). Learning latent structure: carving nature at its joints. Curr. Opin. Neurobiol. *20*, 251–256.

Gold, J.I., Law, C.-T., Connolly, P., and Bennur, S. (2008). The Relative Influences of Priors and Sensory Evidence on an Oculomotor Decision Variable During Perceptual Learning. J. Neurophysiol. *100*, 2653–2668.

Gremel, C.M., and Costa, R.M. (2013a). Premotor cortex is critical for goal-directed actions. Front. Comput. Neurosci. *7*.

Gremel, C.M., and Costa, R.M. (2013b). Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. Nat. Commun. *4*.

Hadland, K.A., Rushworth, M.F.S., Gaffan, D., and Passingham, R.E. (2003). The Anterior Cingulate and Reward-Guided Selection of Actions. J. Neurophysiol. *89*, 1161–1164.

Hilario, M., Holloway, T., Jin, X., and Costa, R.M. (2012). Different dorsal striatum circuits mediate action discrimination and action generalization. Eur. J. Neurosci. *35*, 1105–1114.

Hintiryan, H., Foster, N.N., Bowman, I., Bay, M., Song, M.Y., Gou, L., Yamashita, S., Bienkowski, M.S., Zingg, B., Zhu, M., et al. (2016). The mouse cortico-striatal projectome. Nat. Neurosci.

Huys, Q.J.M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R.J., and Dayan, P. (2011). Disentangling the Roles of Approach, Activation and Valence in Instrumental and Pavlovian Responding. PLoS Comput Biol *7*, e1002028.

Huys, Q.J.M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., and Roiser, J.P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput. Biol. *8*, e1002410.

Ito, M., and Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J. Neurosci. *29*, 9861–9874.

Ito, M., and Doya, K. (2015). Distinct Neural Representation in the Dorsolateral, Dorsomedial, and Ventral Parts of the Striatum during Fixed- and Free-Choice Tasks. J. Neurosci. *35*, 3499–3514.

Ito, S., Stuphorn, V., Brown, J.W., and Schall, J.D. (2003). Performance Monitoring by the Anterior Cingulate Cortex During Saccade Countermanding. Science *302*, 120–122.

Karlsson, M.P., Tervo, D.G., and Karpova, A.Y. (2012). Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. Science *338*, 135–139.

Kennerley, S.W., Walton, M.E., Behrens, T.E.J., Buckley, M.J., and Rushworth, M.F.S. (2006). Optimal decision making and the anterior cingulate cortex. Nat Neurosci *9*, 940–947.

Kennerley, S.W., Behrens, T.E., and Wallis, J.D. (2011). Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. Nat. Neurosci. *14*, 1581–1589.

Killcross, S., and Coutureau, E. (2003). Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats. Cereb. Cortex *13*, 400–408.

Kolling, N., Wittmann, M.K., Behrens, T.E.J., Boorman, E.D., Mars, R.B., and Rushworth, M.F.S. (2016). Value, search, persistence and model updating in anterior cingulate cortex. Nat. Neurosci. *19*, 1280–1285.

Kool, W., Cushman, F.A., and Gershman, S.J. (2016). When Does Model-Based Control Pay Off? PLOS Comput Biol *12*, e1005090.

Matsumoto, K., Suzuki, W., and Tanaka, K. (2003). Neuronal correlates of goal-based motor selection in the prefrontal cortex. Science *301*, 229–232.

Miller, K., Shenhav, A., and Ludvig, E. (2016a). Habits without Values. bioRxiv 067603.

Miller, K.J., Botvinick, M.M., and Brody, C.D. (2016b). Dorsal hippocampus plays a causal role in model-based planning. bioRxiv 096594.

Oh, S.W., Harris, J.A., Ng, L., Winslow, B., Cain, N., Mihalas, S., Wang, Q., Lau, C., Kuan, L., Henry, A.M., et al. (2014). A mesoscale connectome of the mouse brain. Nature *508*, 207–214.

O'Reilly, J.X., Schüffelgen, U., Cuell, S.F., Behrens, T.E., Mars, R.B., and Rushworth, M.F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. Proc. Natl. Acad. Sci. *110*, E3660–E3669.

Ostlund, S.B., and Balleine, B.W. (2005). Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. J. Neurosci. *25*, 7763.

Otto, A.R., Gershman, S.J., Markman, A.B., and Daw, N.D. (2013). The Curse of Planning Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive. Psychol. Sci. *24*, 751–761.

Otto, A.R., Skatova, A., Madlon-Kay, S., and Daw, N.D. (2014). Cognitive control predicts use of model-based reinforcement learning. J. Cogn. Neurosci.

Pachitariu, M., Steinmetz, N., Kadir, S., Carandini, M., and Harris, K.D. (2016). Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels. bioRxiv 061481.

Paxinos, G., and Franklin, K.B. (2007). The mouse brain in stereotaxic coordinates -3rd Edition (Academic Press).

Rudebeck, P.H., Walton, M.E., Smyth, A.N., Bannerman, D.M., and Rushworth, M.F.. (2006). Separate neural pathways process different decision costs. Nat. Neurosci. *9*, 1161–1168.

Rudebeck, P.H., Behrens, T.E., Kennerley, S.W., Baxter, M.G., Buckley, M.J., Walton, M.E., and Rushworth, M.F.S. (2008). Frontal Cortex Subregions Play Distinct Roles in Choices between Actions and Stimuli. J. Neurosci. *28*, 13775–13785.

Rushworth, M.F.S., and Behrens, T.E.J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. Nat. Neurosci. *11*, 389–397.

Rushworth, M.F.S., Walton, M.E., Kennerley, S.W., and Bannerman, D.M. (2004). Action sets and decisions in the medial frontal cortex. Trends Cogn. Sci. *8*, 410–417.

Sebold, M., Deserno, L., Nebe, S., Schad, D.J., Garbusow, M., Hägele, C., Keller, J., Jünger, E., Kathmann, N., Smolka, M., et al. (2014). Model-Based and Model-Free Decisions in Alcohol Dependence. Neuropsychobiology *70*, 122–131.

Shenhav, A., Cohen, J.D., and Botvinick, M.M. (2016). Dorsal anterior cingulate cortex and the value of control. Nat. Neurosci. *19*, 1286–1291.

Simon, D.A., and Daw, N.D. (2011). Neural Correlates of Forward Planning in a Spatial Decision Task in Humans. J. Neurosci. *31*, 5526–5539.

Smittenaar, P., FitzGerald, T.H.B., Romei, V., Wright, N.D., and Dolan, R.J. (2013). Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. Neuron.

Sul, J.H., Kim, H., Huh, N., Lee, D., and Jung, M.W. (2010). Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. Neuron *66*, 449–460.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement learning: An introduction (The MIT press).

Thorndike, E.L. (1911). Animal intelligence: Experimental studies.

Vogt, B.A., and Paxinos, G. (2014). Cytoarchitecture of mouse and rat cingulate cortex with human homologies. Brain Struct. Funct. *219*, 185–192.

Voon, V., Derbyshire, K., Rück, C., Irvine, M.A., Worbe, Y., Enander, J., Schreiber, L.R.N., Gillan, C., Fineberg, N.A., Sahakian, B.J., et al. (2015). Disorders of compulsivity: a common bias towards learning habits. Mol. Psychiatry *20*, 345–352.

Walton, M.E., Bannerman, D.M., Alterescu, K., and Rushworth, M.F.. (2003). Functional specialization within medial frontal cortex of the anterior cingulate for evaluating effort-related decisions. J. Neurosci. *23*, 6475.

Wunderlich, K., Smittenaar, P., and Dolan, R.J. (2012). Dopamine Enhances Model-Based over Model-Free Choice Behavior. Neuron *75*, 418–424.

Yin, H.H., Knowlton, B.J., and Balleine, B.W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur. J. Neurosci. *19*, 181–189.

Yin, H.H., Knowlton, B.J., and Balleine, B.W. (2005a). Blockade of NMDA receptors in the dorsomedial striatum prevents action–outcome learning in instrumental conditioning. Eur. J. Neurosci. *22*, 505–512.

Yin, H.H., Ostlund, S.B., Knowlton, B.J., and Balleine, B.W. (2005b). The role of the dorsomedial striatum in instrumental conditioning. Eur. J. Neurosci. *22*, 513–523.

Figure supplements:

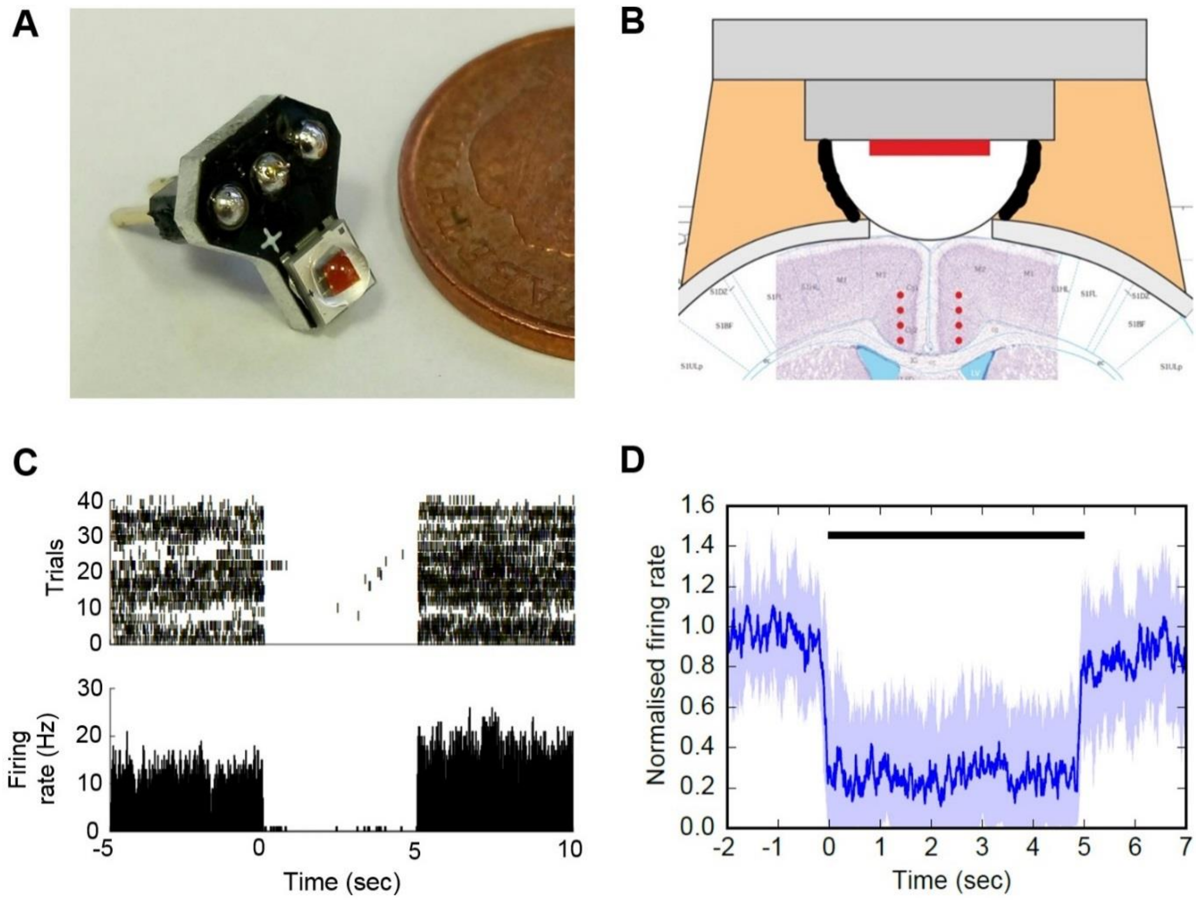


Figure 1 - figure supplement 1. JAWS inhibition of ACC neurons. A) LED implant. B) Implantation diagram, red dots indicate location of virus injections. **C)** Inhibition of example cell, top panel – spike raster, bottom panel average firing rate. **D)** Normalised firing rate for significantly inhibited cells (Kruskal-Wallis P < 0.05, 67/249 cells), dark blue line – median, shaded area 25 – 75 percentile.

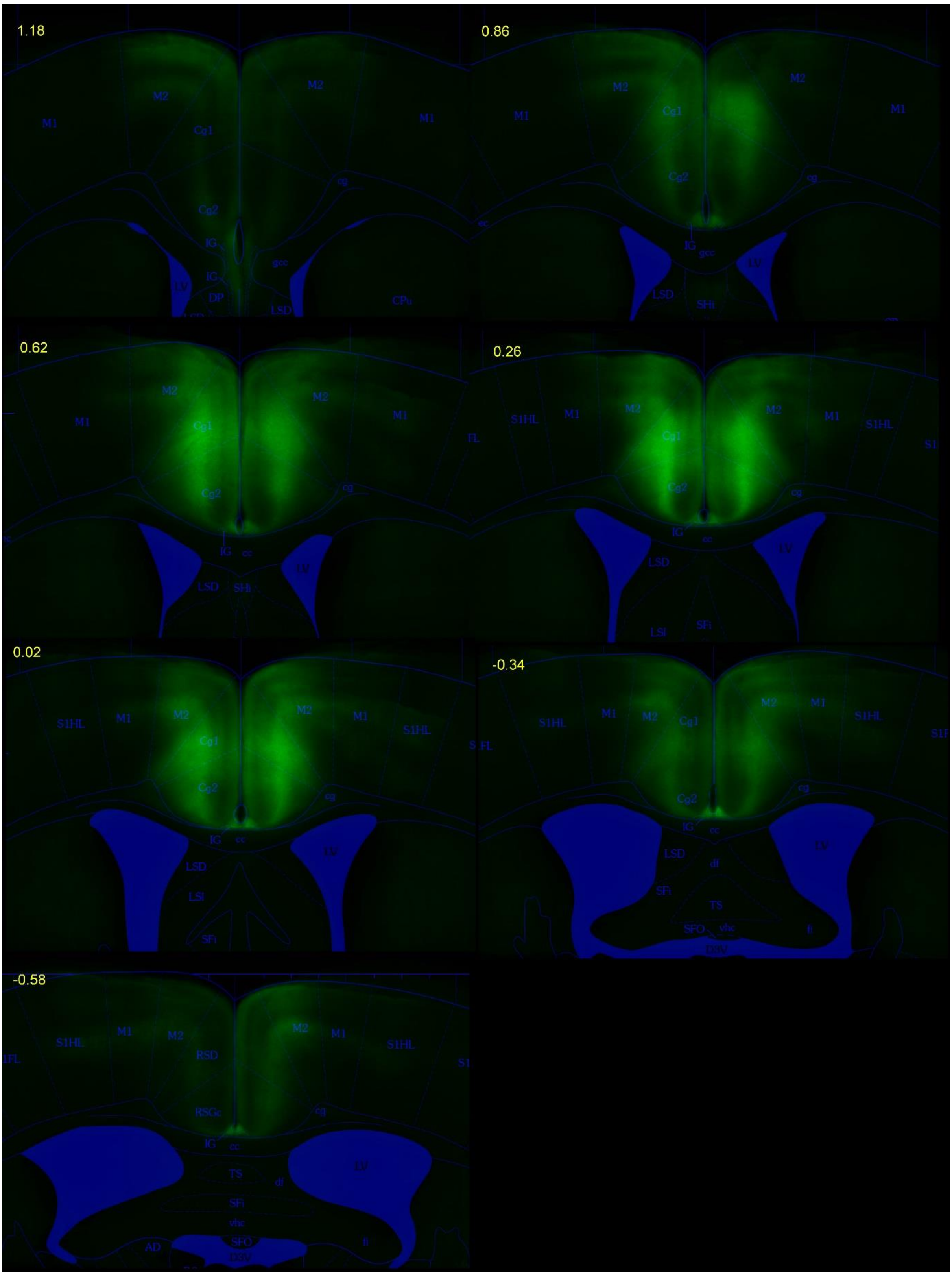
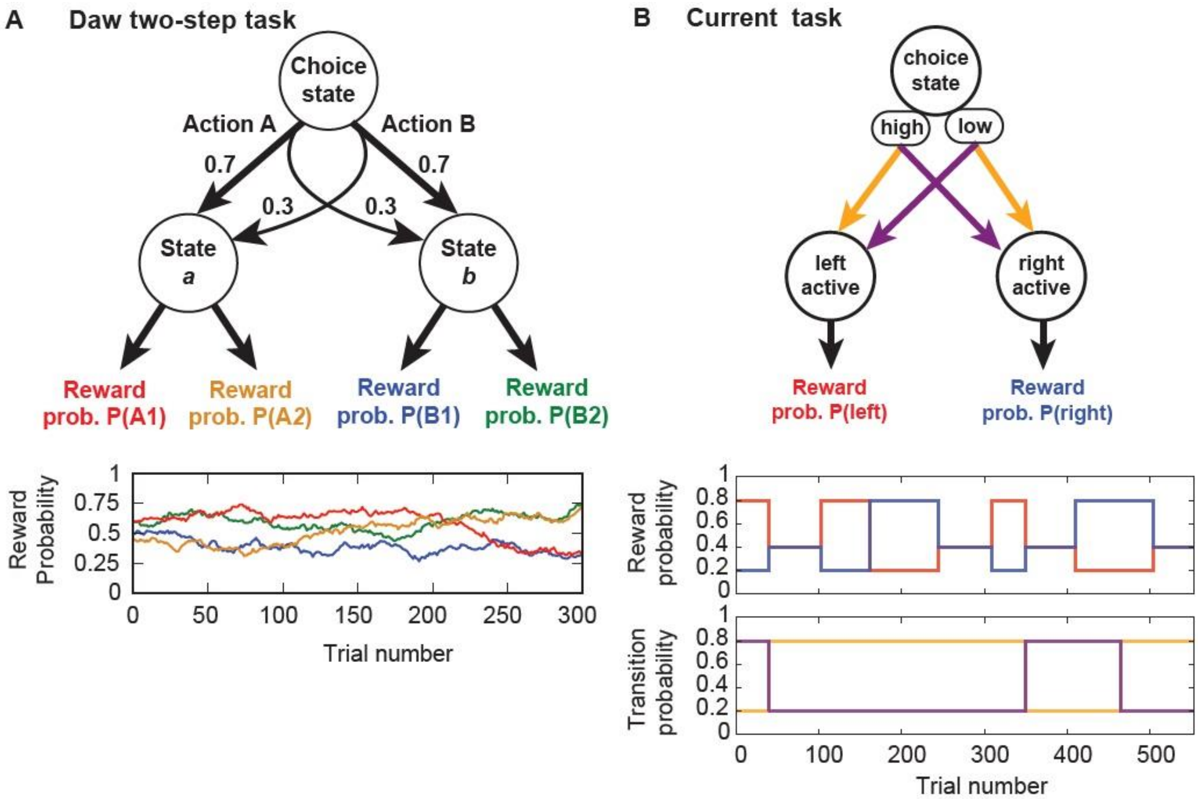
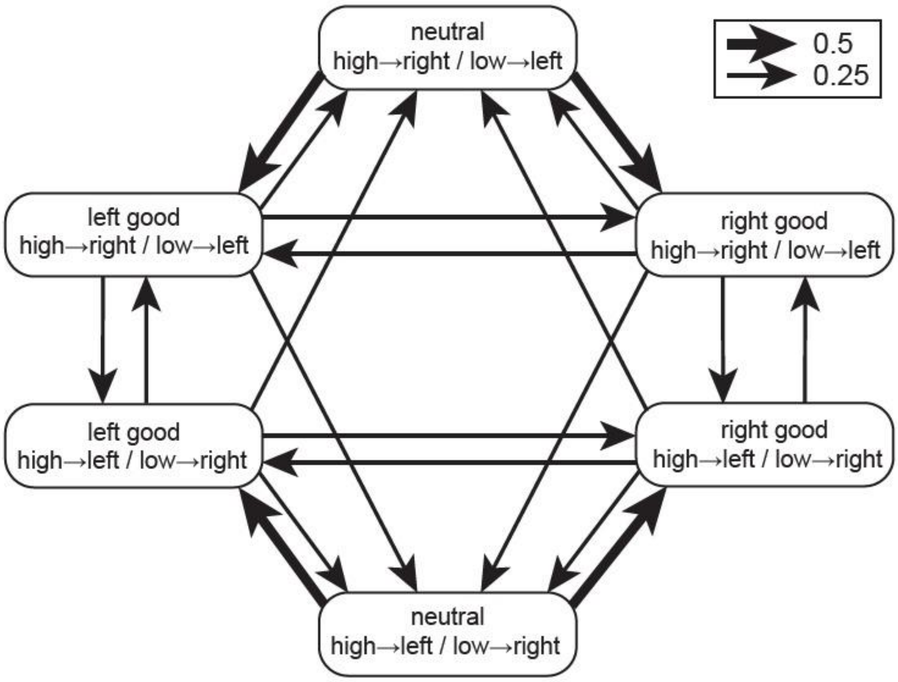


Figure 1 – figure supplement 2. Average JAWS expression. Average JAWS-GFP fluorescence for all JAWS-GFP animals included in the study aligned onto reference atlas (Paxinos and Franklin, 2007). Numbers indicate anterior-posterior position relative to bregma (mm).



[Figure 2](#) - **figure supplement 1. Comparison of original and new two-step task structures. A)** State diagram of the original Daw two step task with example reward probability trajectories. **B)** State diagram of the two-step task used in the current study with example reward probability and transition probability trajectories.



[Figure 2](#) - **figure supplement 2. Block transition probabilities.** Diagram of block transition probabilities for the two-step task used in the current study.

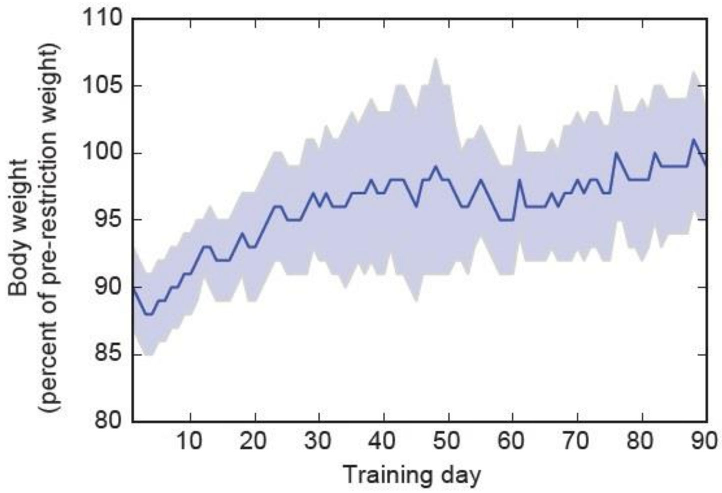


Figure 2 - figure supplement 3. Body weight trajectory across training: Mean (blue line) and standard-deviation (shaded area) of subject's body weight trajectory across days of training.

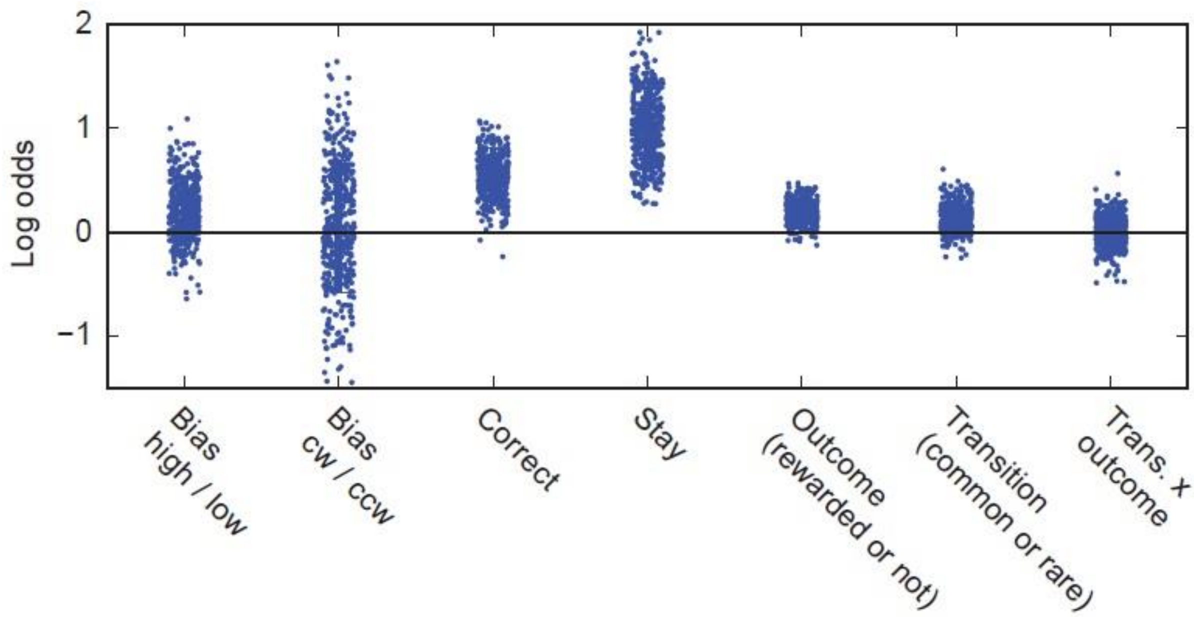


Figure 3 - figure supplement 1. Full logistic regression model fit. Fit of the logistic regression model to the baseline dataset showing loadings for all 7 parameters. Bars indicate ±1 standard deviation of the population level distributions, dots indicate maximum a posteriori session fits. Predictors: *Bias high/low* – tendency to choose the high poke, *bias clockwise /counter-clockwise* – tendency to choose high following left and low following right, *Correct* – tendency to choose the correct option, i.e. that option which commonly leads to state with higher reward probability, *Stay* – tendency to repeat choices irrespective of subsequent trial events, *Outcome* – tendency to repeat choices following reward, *Transition* – tendency to repeat choices following common transitions, *Transition-outcome interaction* – tendency to repeat choices following rewarded common transition trials and non-rewarded rare transition trials.

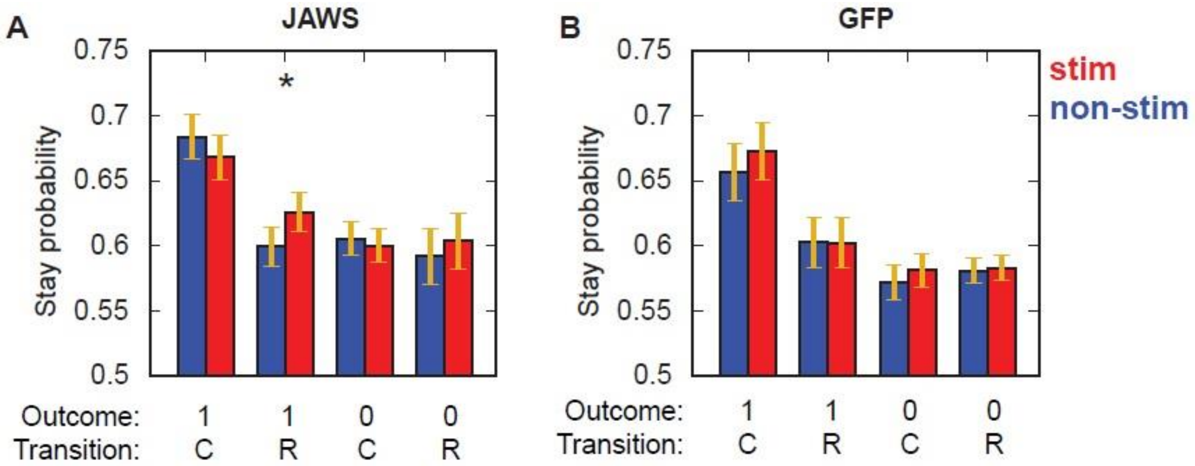


Figure 4 - figure supplement 1. ACC inhibition stay probabilities Stay probability analysis for JAWS (**A**) and GFP control (**B**) animals showing fraction of trials the subject repeated the same choice following each combination of outcome (rewarded (1) or not (0)) and transition (common (C) or rare (R)). Stay probabilities were evaluated separately for trials with (red) and without (blue) light stimulation delivered from the trial outcome to the subsequent choice. Error bars show cross-subject SEM. * indicates paired t-test P value < 0.05.

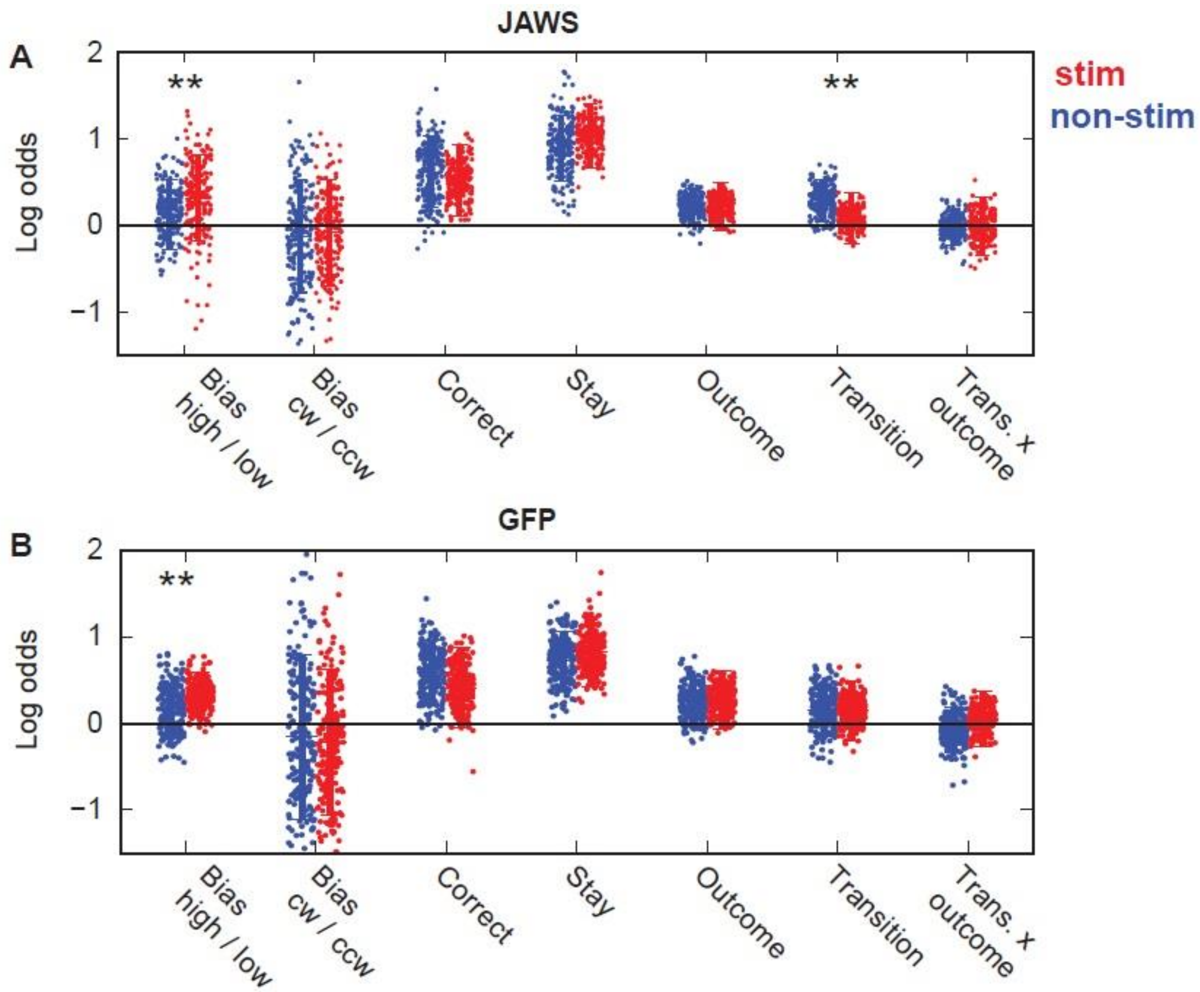
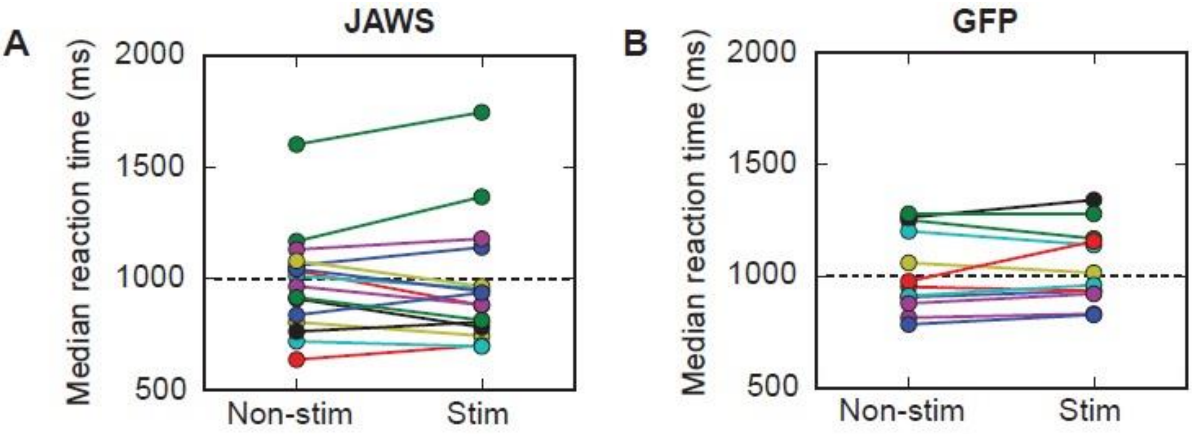
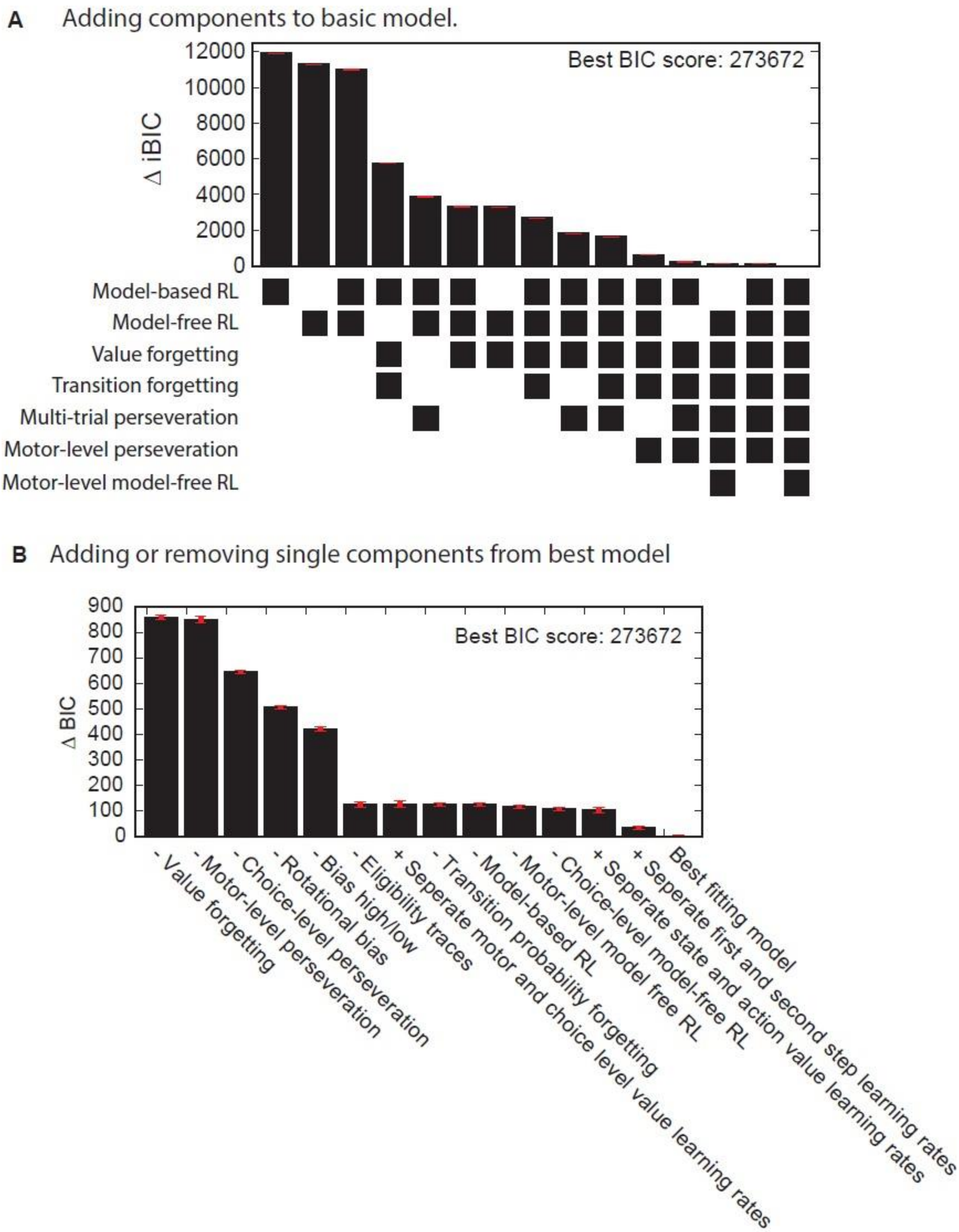


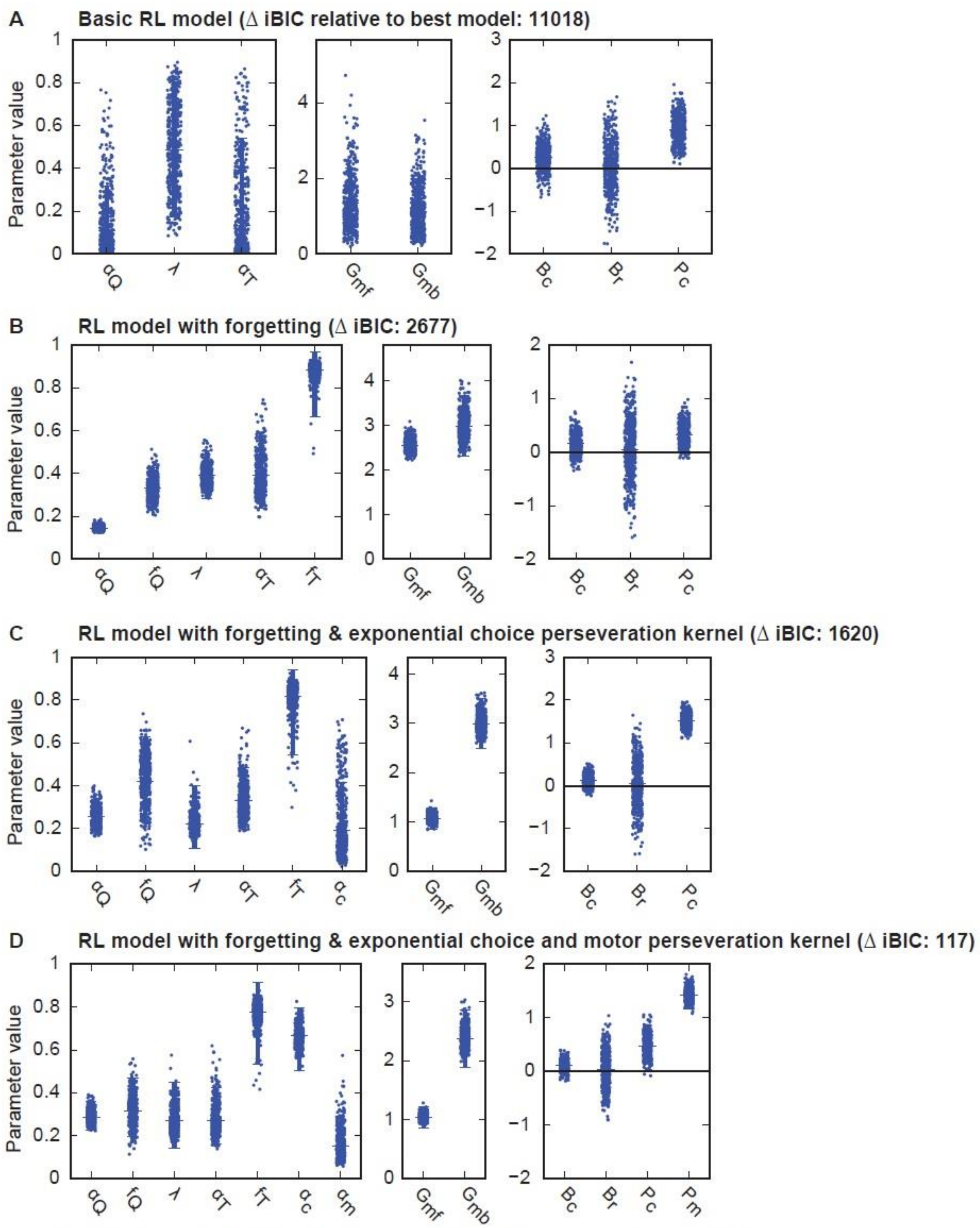
Figure 4 - figure supplement 2. ACC inhibition full logistic regression model fits. Fit of the logistic regression model to the JAWS ACC inhibition (**A**) and GFP controls (**B**) showing loadings for all 7 parameters. Bars indicate ±1 standard deviation of the population level distributions, dots indicate maximum a posteriori session fits.



1018
1019 **Figure 4 - figure supplement 3. ACC inhibition reaction times.** Reaction times for first-step choice on
1020 stimulated and non-stimulated trials. Reaction time is measured from the start of the ITI when the subject
1021 exits the side poke at the end of the previous trial, until the next high or low poke. The dashed line indicates
1022 the end of the ITI at which point the high and low pokes become active.



1024 **Figure 5 - figure supplement 1. Baseline dataset BIC score model comparison. A)** iBIC score comparison for
1025 set of RL models on baseline behavioural dataset. The set of models was constructed as described in
1026 supplementary results by iteratively adding features to the RL model. The grid below the plot indicates which
1027 features were included in each model. **B)** iBIC score comparison on the baseline dataset for set of RL models
1028 created by adding or removing a single feature at a time from the best fitting model. The text below each bar
1029 indicates what feature has been added or removed. Error-bars indicate the bootstrap 95% confidence interval
1030 on the BIC score.



α_Q: Value learning rate, f_Q: Value forgetting rate, λ: Eligibility trace, α_T: Transition learning rate, f_T: Transition forgetting rate, α_c: Choice perseveration learning rate, α_m: Motor perseveration, learning rate, G_{mf}: Model-free weight, G_{mb}: Model-based weight, B_c: Choice Bias, B_r: Rotational bias, P_c: Choice perseveration weight, P_m: Motor perseveration weight.

Figure 5 **- figure supplement 2. Alternative RL model fits.** Fit of Reinforcement learning models of different levels of complexity. Model complexity increases from **A** to **D** as features are added to the basic RL model. For each fit, bars indicate ±1 standard deviation of the population level distributions, dots indicate maximum a posteriori session fits. For each model the difference in iBIC score between this model and the best fitting model is reported.

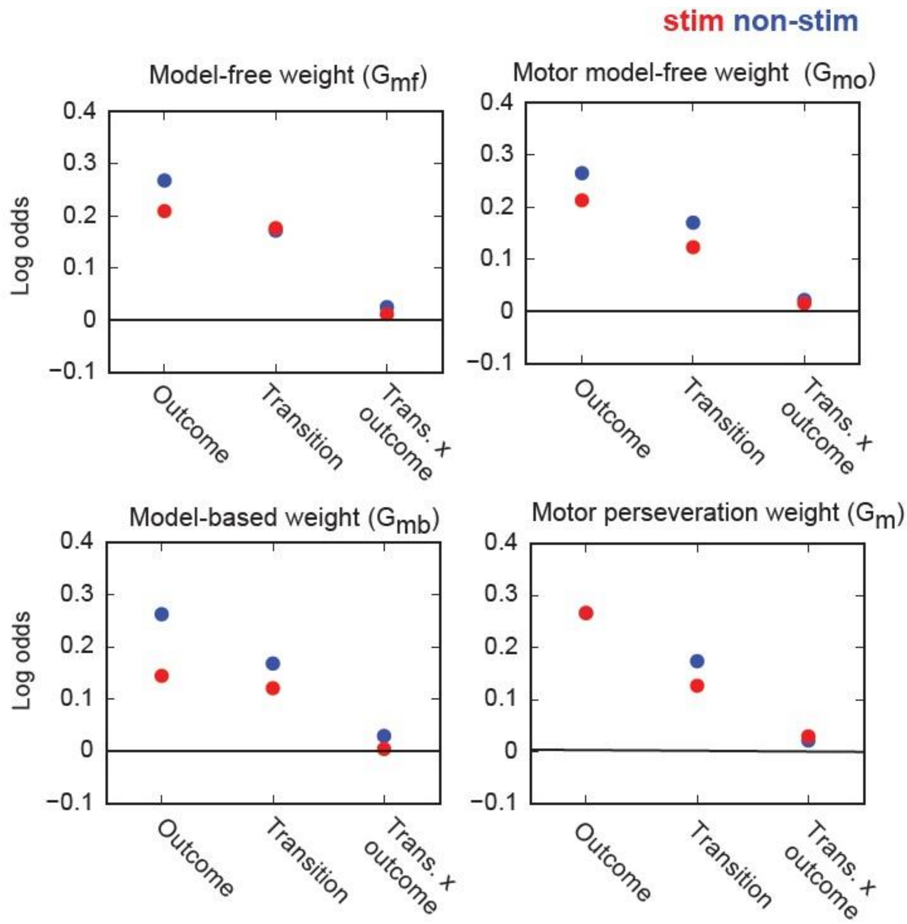


Figure 5 - figure supplement 3. Simulating effects of stimulation: Simulation of the effects of lesioning different components of the best fitting RL model on stimulation trials. Model lesioning was implemented by setting individual parameters to zero on stimulation trials. Panels show logistic regression loadings for stimulated and non-stimulated trials. For each panel the title indicates which model-parameter was set to zero on stimulation trials.

Supplementary Material:

Model comparison:

The starting point for our model comparison process was the RL agent used in the original Daw two-step task (Daw et al., 2011). As the action-state transition probabilities in our task were not fixed, we modified the model-based component of the agent to update its estimate of the transition probabilities for the chosen action on each trial using an error driven learning rule. As in the original Daw agent we included a perseveration parameter which promoted repeating the previous choice. Based on the evidence for response biases from the logistic regression, we additionally included in the RL agent two parameters capturing a bias towards the high/low poke and the rotational bias described in the results section. We compared the goodness of fit of a pure model-free agent, a pure model-based agent, and an agent which used a mixture of both strategies. The mixture agent provided a better fit to the data than either the pure model-free (Δ iBIC = 264, Figure3B) or pure model-based agent (Δ iBIC = 888), and the mixture model fit suggested an approximately equal contribution of model-based and model-free control (Figure 5 – figure supplement 2A). As the task is novel and hence we do not know what features may be present in the behaviour, we performed an exploratory process of model comparison to better understand whether the RL model was providing a good description of the behaviour. This identified a number of additional features which greatly improved fit quality when added to the model.

RL models typically assume that action values of options that are not chosen remain unchanged. However, it has been reported that model-fits in some rodent decision making tasks are substantially improved by including forgetting about the value of not chosen actions, typically implemented as action value decay towards zero (Ito and Doya, 2009, 2015). Including such action value forgetting in the mixture agent produced a dramatic improvement in iBIC score for our data (Δ iBIC = 7698). Including forgetting about action-state transition probabilities, implemented as a decay of transition probabilities for the not chosen action towards a uniform distribution, further improved the goodness of fit (Δ iBIC = 643). The mixture agent including value and transition probability forgetting again showed approximately equal weighting of the model-based and model-free action values in controlling behaviour (Figure 5 – figure supplement 2B). When forgetting was included for each agent the mixture agent provided a better fit to the data than either a pure model-free (Δ iBIC = 612) or pure model-based (Δ iBIC = 3066) agent.

Forgetting decreases the value of not chosen relative to chosen options, and therefore promotes perseveration of choice. It is therefore possible that if subjects are in fact strongly perseverative, this could be mistakenly identified as forgetting in the RL fit. Though the model included a perseveration parameter for repeating the previous choice, several studies have reported perseveration effects spanning multiple trials, even in tasks where decisions optimally should be treated as independent (Gold et al., 2008; Akaishi et al., 2014). We therefore tested whether goodness of fit was improved by an exponential choice kernel through which prior choices directly influenced the current choice with exponentially decreasing weight at increasing lag (Figure 5 – figure supplement 2C). This is equivalent to the decision inertia model of Akaishi et al. (2014) in which choice is influenced by a variable they term the choice estimate CE, an average of previous choices updated following each decision using the error driven learning rule $CE_{n+1} = CE_n + \alpha\,(C_n - CE_n)$, where C_n is the choice on trial n and α is a learning rate. The addition of this exponential choice kernel dramatically improved fit quality when added to the mixture agent without forgetting (Δ iBIC = 7133). However even with the exponential choice kernel included, value forgetting substantially improved goodness of fit (Δ iBIC = 2071), and transition probability forgetting further increased goodness of fit (Δ iBIC = 194). These results indicate that forgetting about values and transitions for not chosen options is a genuine feature of the behaviour and not an artefact due to a tendency to perseverate. They further indicate that subjects do in fact show a strong tendency to perseverate over multiple trials, which is not captured even by forgetting RL models, presumably because it is independent of the recent reinforcement history. Forgetting may be a heuristic used in dynamic environments where evidence becomes less reliable with the passage of time due to state of the world changing. Alternatively, forgetting may occur due to limitations of the learning systems involved, perhaps due to differences between the rapidly changing reward statistics in the task and those typical of natural environments.

The choice kernel assumes that perseveration occurs at the level of the decision between the high and low pokes, however it is also possible that the perseverative tendency is at the lower level of motor actions. In the current task, a given choice (high or low) entails a different motor action depending on which side (left or right) the previous trial ended on. We therefore considered a model with perseveration at the motor level such that the choice on a given trial only increased the probability of repeating that same motor action in future, e.g. a choice taken by moving from the left to high poke only increased the probability of choosing high in future following trials which ended on the left side (Figure 5 – figure supplement 2D). Motor perseveration was modelled by maintaining separate moving averages of choices following trials that ended on the left and right, updated using the error driven learning rule described above, which each influenced choices following trials ending on their respective sides. Replacing the exponential choice kernel with this motor perseveration substantially improved fit quality (Δ iBIC = 1004). However, including perseveration both at the level of choice, (high vs low, independent of motor action), and at the motor level, further improved fit quality (Δ iBIC = 499), indicating that subjects exhibit perseverative tendencies at both the choice and motor level that are not predicted by the RL component of the model. These data support the existence of mechanisms which reinforce selected behaviours in a reward-independent fashion, i.e. simply choosing to execute a behaviour increases the chance that behaviour will be executed in future. This is consistent with previous reports from perceptual (Gold et al., 2008; Akaishi et al., 2014) and reward-guided decision making tasks (Miller et al., 2016a), and we think is a parsimonious explanation for our results. Such perseveration is somewhat puzzling from a normative perspective but may be a signature of a mechanism for automatizing behaviour by reinforcing chosen actions. Thorndike proposed such a 'law of exercise' (1911) and the idea has recently been revisited by Miller et al. (2016a) who suggest that habit formation occurs through outcome-independent reinforcement of chosen actions. This framework views habit formation as a supervised learning process in which behaviour generated by value sensitive systems, i.e. model-free and model-based RL, is used to train value-independent learning systems. Such a mechanism could account for the perseveration observed in our data assuming it operated both on actions represented at the level of the choice they represent and the level of motor actions. An alternative mechanism which could give rise to perseveration would be subjects sampling an option multiple times between choices, which may be adaptive if the decision process is costly in time or effort. However, this explanation does not account for the observation in our data that perseveration occurred at the level both of choices and of motor actions, with different timescales for each (see respective learning rates, Figure 5).

Evidence that perseveration occurred both at the level of choice and motor action raises the question of whether reward driven learning also occurs at both levels of representation. This might be expected from the architecture of parallel cortical-basal ganglia loops, with circuits linking somatosensory and motor cortices to dorsolateral striatum learning values over low level motor representations, and circuits linking higher level cortical regions to medial and ventral striatum learning values over more abstract state and action representations. We therefore tested an agent in which model-free action values were learned in parallel for actions represented both in terms of choice (high/low) and motor action (e.g. left→high). This improved goodness of fit (Δ iBIC = 117) and the resulting model fit indicated that motor-level model-free values had a somewhat stronger influence on behaviour than the choice level model-free values (Figure 3a). With the perseveration kernels and motor level representations included in each model, the mixture agent again provided a better fit to the data than either a pure model-free (Δ iBIC = 127) or pure model-based (Δ iBIC = 227) agent. We tested a number of other modifications to the model including separate learning rates at the first and second step, but did not find further improvements in fit quality (Figure 5 – figure supplement 1A). Finally, as adding features to the model may make other features which previously improved the fit unnecessary, we tested whether removing any individual component from the model improved fit quality but again did not find further improvements (Figure 5 – figure supplement 1B).

Lesioning Full RL model

The simulations presented in Figure 3b indicated that data simulated from a model-based RL agent showed loading on the transition and outcome predictors while data simulated from a model-free RL agent showed loading only on outcome. This suggests that reduced influence of model-based and increased influence of model-free RL could produce the observed effect of ACC inhibition. However, the full RL model arrived at in the model comparison process included additional features not included in those simulations which may complicate the relationship between behavioural strategy and regression loadings. Specifically, we were concerned that perseveration or model-free RL for actions represented at the motor level (i.e. as a movement from left to high poke, rather than as a choice of the high poke irrespective of where the movement started) could produce loading on the transition predictor. This is because the state transition determines which second-step state the subject ends up in, and hence which motor action they must take to make a given choice on the next trial. We therefore performed a set of simulations where we set the influence on choice of different components of the model to zero on stimulation trials, which we term lesioning a model component (Figure 5 – figure supplement 3). This confirmed that consistent with Figure 2B, lesioning the choice-level model-free system selectively reduced loading on the outcome predictor, while lesioning the model-based system reduced loading on outcome and transition, and to a lesser extent on the interaction predictor. However, lesioning the motor-level model free system (which learned model-free action values for individual motor actions such as left→high), also reduced loading on the outcome and transition predictors, while lesioning motor-level perseveration reduced loading only on the transition predictor. These simulations suggest that the reinforcing effect of experiencing a common transition is mediated in part by the use of model-based RL but also in part by perseveration and model-free RL occurring at the level of motor actions.