Flexibility to contingency changes distinguishes habitual and goal-directed strategies in humans

Julie Lee[1,#a*], Mehdi Keramati[1,#b]

[1]Gatsby Computational Neuroscience Unit, University College London, London, UK

[#a]Current Address: Institute of Ophthalmology, University College London, London, UK

[#b]Current Address: Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, London, UK

* Corresponding author

E-mail: julie.lee.15@ucl.ac.uk (JL)

Abstract

Decision-making in the real world presents the challenge of requiring flexible yet prompt behavior, a balance that has been characterized in terms of a trade-off between a slower, prospective goal-directed model-based (MB) strategy and a fast, retrospective habitual model-free (MF) strategy. Theory predicts that flexibility to changes in both reward values and transition contingencies can determine the relative influence of the two systems in reinforcement learning, but few studies have manipulated the latter. Therefore, we developed a novel two-level contingency change task in which transition contingencies between states change every few trials; MB and MF control predict different responses following these contingency changes, allowing their relative influence to be inferred. Additionally, we manipulated the rate of contingency changes in order to determine whether contingency change volatility would play a role in shifting subjects between a MB or MF strategy. We found that human subjects employed a hybrid MB/MF strategy on the task, corroborating the parallel contribution of MB and MF systems in reinforcement learning. Further, subjects did not remain at one level of MB/MF behaviour but rather displayed a shift towards more MB behavior over the first two blocks that was not attributable to the rate of contingency changes but was rather a more general effect of block order. The extent to which each subject used MB control was also related to reward earned, with a correlation between MB weight and reward rate. We demonstrate that flexibility to contingency changes can distinguish MB and MF strategies, with human subjects utilising a hybrid strategy that shifts towards more MB behavior over blocks, consequently corresponding to a higher payoff.

Introduction

To make optimal decisions, humans must learn to associate the choices they make with the outcomes that arise from them. Classical learning theories suggest that this problem is addressed by habitual or goal-directed strategies for reinforcement learning [1, 2]. These strategies differ in that habitual behavior seeks simply to reinforce responses based on environmental cues, whereas goal-directed behavior considers action-outcome relationships – that is, contingencies – in the environment. Habitual and goal-directed strategies have been implemented in model-based (MB) and model-free (MF) reinforcement learning algorithms, respectively. Both algorithms make decisions by estimating action values and choosing the actions that maximize reward in the long term [3, 4]. The MF system achieves this retrospectively, caching past rewards using a reward prediction error signal [5] whereas the MB system achieves this prospectively, planning using a learned internal model of the state transitions and rewards in the environment.

Recent studies have emphasized that MB and MF systems work in parallel rather than in isolation [4, 6-8]. Early studies discerned MB and MF contributions using manipulations of reward values, such as in reward devaluation paradigms, but did not seek to quantify their relative contributions [1]. A recent study [6] addressed this by developing the hallmark "two-step" task in which, using reward value changes, each trial was informative of the MB/MF tradeoff, thereby permitting model-fitting analyses to quantify their relative influence in decision-making. Human subjects showed a hybrid MB/MF strategy in the task, a result that has been widely replicated under different manipulations [9, 10] and extended to the non-human animal literature (Groman et al. Soc. Neurosci. Abstracts 2014, 558.19, Miranda et al. Soc. Neurosci. Abstracts 2014 756.09, Akam et al. Cosyne Abstracts 2015, II-15; Hasz & Redish, Soc. Neurosci. Abstracts 2016 638.08; [11]).

Theory predicts that flexibility to transition contingency changes can – like flexibility to reward value changes – determine the relative influence of MB and MF strategies [4, 12]. The advantage of manipulating transitions, rather than reward values, is apparent when contrasting the model-based system to a successor representation (SR) [13]. The successor representation caches transitions in a model-free fashion, but learns reward values in a model-based fashion; thus, changes to reward values cannot distinguish MB and SR representations. In contrast, transition changes ensure that consequent choices only can be explained by an MB system. Two studies have examined the flexibility of MB and MF systems to global contingency changes [14, 15]. However, quantification of the MB/MF tradeoff was limited as these studies manipulated contingency and tested flexibility to the change of contingency in separate phases; at these timescales, it becomes difficult to exclude the effect of adaptation on MB/MF weights. Therefore, we developed a novel two-level contingency change task containing multiple, frequent and interleaved transition contingency changes that elicit different consequent actions by the MB and MF systems. Our design, like the two-step task [6] and its variants, therefore permits model-fitting analyses to robustly determine the relative influence of the MB/MF systems. The contingency change task is structured such that actions following frequent contingency changes are distinctly attributed to either a MB or MF strategy; this then permits quantification of the degree to which each system is in control.

On top of a hybrid MB/MF strategy, subjects may not remain at one level of MB/MF control but instead shift their relative weight in accordance with environmental factors. In general, animals show habit formation with time, a robust effect reported since early reward devaluation studies [16] in which extensive training stamped in habits, resulting in insensitivity to reward devaluation; in contrast, limited training retained goal-directed behavior. Sensitivity to contingency degradation (the omission of a previously-learned contingency between actions and outcomes) also decreases with overtraining, likewise reflecting a trend towards habitization with time [17]. In the original two-step task, the MB/MF trade-off was designed to be stable [6], but will shift under manipulations such as limited time [8] or cognitive load [18]. However, habits are not guaranteed to form with time; even after extended training, rats can show residual responding following outcome devaluation, indicating that they retained goal-directed behavior despite overtraining [19]. In another study using the two-step task [20], the level of MB/MF control in fact increased in favour of more MB control (i.e. towards less habitual behavior) over three days of training. However, general shifts in MB/MF control should be disentangled from the effects of environmental volatility, which are known to affect the MB/MF balance [21]. Thus, in this study, we examined whether the MB/MF relationship is affected by environmental stability, or whether it shifts more generally over time.

We found that human subjects indeed showed a hybrid strategy in reacting to contingency changes in our task, with an increased influence of MB control over the first two blocks. However, relative MB/MF control did not significantly differ across rates of contingency changes; thus, the increase in MB control may be a more global effect of "anti-habitization" over time. The increased reliance on the MB system was associated with a higher proportion of highly rewarded actions and consequently a higher reward rate, indicating that as subjects proceeded through the session, they became more proficient at exploiting their learned internal model of the task structure to maximize their reward.

Results

Subjects (N=16) performed a two-level contingency change task which consisted of 600 trials (Fig 1). Each trial began at either the first level (S0) with 50% probability, or the second level with 50% probability – 25% for each of the two states at this level (S1 or S2). If a trial started at the first level, a two-alternative choice was possible between two abstract stimuli. Each first-level action deterministically always led to the same second-level state, i.e. A1 to S1 and A2 to S2. Critically however, transitions from the second-level states to the terminal states flipped between two contingencies every 3-14 trials. Each of the two terminal states was then associated with either high or low reward, with the exact reward values drifting across trials (see Methods for details). Thus, flexibility to contingency changes was essential for maximizing reward.

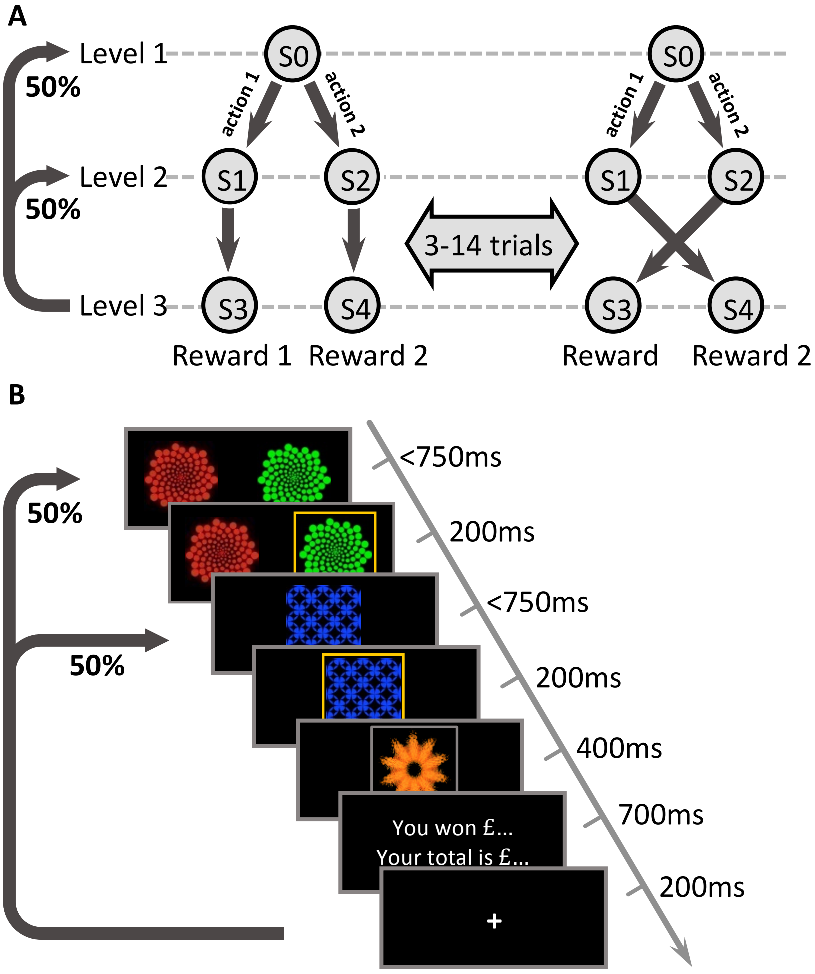


Fig 1. Schematic of the experimental design. (A) Each trial started from either the first-level state (S0), with 50% probability, or one of the two second-level states (S1 or S2), each with 25% probability. While two choices were available at S0, only a forced choice was available at the second-level states. The transition structure from the second-level states to terminal states repeatedly flipped after a random number of trials (every 3- 14), in an unsignalled fashion. One of the two terminal states (S3 or S4) was associated with a high reward outcome and the other with a low reward outcome. (B) Timeline of the task for one example trial.

If a contingency change occurred, subjects always experienced the new transition structure regardless of whether they started at the first or second level, as contingency could only change between second-level and terminal states. Therefore, provided that an action was possible at the next trial (i.e. that the next trial started at the first level) the MB system would plan using the updated causal structure and thus would take the action that led under the new transition contingencies to the high reward terminal state. However, if a contingency change trial started from the second level, the MF system would not choose the optimal action on the next trial, as neither the received reward nor the new contingency would update the cached values of first-level actions, simply because no first-level action was experienced on those trials. As a result, the relative contribution of MB and MF systems can be measured by the degree of behavioral flexibility on first-level trials following contingency change trials starting from the second level.

To examine the effect of environmental volatility on the contribution of the two systems, the frequency of contingency changes was varied – from 3-6 trials for 200 trials, to 7-10 trials for the next 200 trials, and then 11-14 trials for the final 200 trials. The order of fast and medium contingency changes was counterbalanced across two subject groups (n=8 each). Every 40 trials, assignment of the high and low reward states also flipped to prevent formation of habits over an extended state representation, which could masquerade MF as MB behavior [22].

Simulated choices on the task were implemented according to MB and MF reinforcement learning algorithms (see Methods for details). For each system, we measured a "stay probability" index which followed the logic of contingency change trials described above. This index differs from classic stay probabilities [6] as trials starting from the second level do not have any choices to "stay". Instead, stay probability in our task was defined as the probability of choosing the first-level action that results in the same second-level state as the previous trial. Since first-level to second-level contingencies were fixed, this modified measure provided stay probabilities on any trial, regardless of whether it started at the first or second level. Stay probability was measured for four different conditions: whether the reward received in the previous trial was "high" or "low", and whether the transition experienced in the previous trial, relative to the trial before that, was "changed" or remained "fixed". In all cases, analyses were restricted to trials starting from the first level, following a contingency change trial starting at the second level, since only these could distinguish MB and MF strategies.

Across these conditions, MB and MF systems showed different stay probability patterns. The MF system, having no experience of the action that led to the new contingency, was more likely to stay on the action leading to the high reward state, and shift on the action leading to the low reward state, under "fixed" than "changed" conditions ($p < 0.01$), indicating it was not flexible to changes in contingencies (Fig 2A). However, the MB system could immediately adapt with the correct next action, staying on the action if it would lead to the high-reward state but shifting if it would lead to the low-reward state, with a main effect of reward ($p < 0.01$) regardless of contingency condition (Fig 2B). As expected, for contingency changes from the first level, MB and MF systems did not differ in stay probability patterns, as the MF system was able to update its action values accordingly, given that it directly experienced the action leading to the new contingency (S1 Fig). In addition to pure MF and pure MB strategies, we simulated a hybrid model that linearly weights MB and MF action values according to a parameter *wMB*. The stay probability pattern produced by this hybrid system reflected a mixture of the effects observed for the pure MF and MB stay probabilities – that is, showing a main effect of reward ($p < 0.01$), but also an interaction between reward and contingency ($p < 0.01$) (Fig 2C).

Fig 2. Stay probability patterns predicted by simulating model-based (A), model-free (B), and hybrid (C) reinforcement learning algorithms. Stay probability measures the probability of choosing the first-level action that results in the same second-level state as the previous trial. This index was measured when the reward received in the previous trial was "high" or "low", and when the transition experienced in the previous trial (relative to the trial before that) had its contingency "changed" or remained "fixed". Stay probabilities are plotted for trials following a change trial that started at the second level, as these distinguish model-based and model-free strategies. * $p < 0.05$, ** $p < 0.01$

Subjects showed hallmarks of both MB and MF strategies in reacting to contingency changes (Fig 3A), showing a main effect of reward, $F(1,60) = 24.65$, $p < 0.01$, as well as a reward/contingency interaction, $F(1,60) = 13.60$, $p < 0.01$. Therefore, subjects did not solely use a MB or MF strategy when reacting to contingency changes, but rather displayed a hybrid MB/MF strategy.

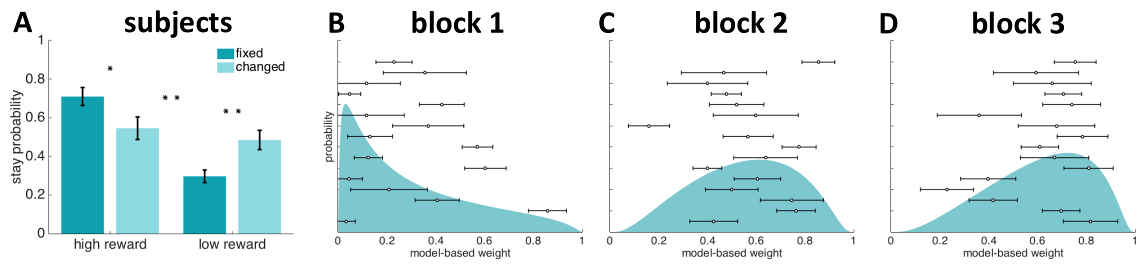


Fig 3. Experimental results. (A) Stay probability pattern from human subjects (N=16) showed significant effects of both model-based (p<0.01) and model-free (p<0.01) strategies. * p < 0.05, ** p < 0.01 (B) Probability density function over the model-based weight parameter, estimated in three different blocks of the first, middle and last 200 trials (out of 600 trials). Overlaid are the individual subjects' model-based weight parameter estimates for each block type. Error bars represent standard deviation.

To characterize the effect of contingency changes over multiple consecutive trials, lagged logistic regression was performed (S2 Fig). This analysis computes the influence of reward and contingency conditions (predictors) from past trials (lags) on choice probabilities [22, 23]. MB and MF systems differed in the extent to which past predictors influenced the current choice, with the MB system showing more flexibility to recent changes – and less influence of past predictors – than the MF system; this was evidenced by a smoother predictive weight over lags for the MF system than the MB system (S2 Fig). As expected, the pattern of the predictive weights for subjects' choices and the simulated hybrid model reflected a mixture of the MF and MB systems' patterns.

While stay probabilities excluded a purely MB or purely MF strategy, this measure could not quantify the degree to which subjects used the hybrid strategy; therefore, we used a hierarchical Bayesian method to fit candidate models of behavior to the subjects' data, to determine which model best explained subjects' choices and to obtain parameter estimates for the MB/MF weighting used by the subjects. The models tested included a pure MB model, a pure MF model, a hybrid model with one constant weight *wMB* across the session, a hybrid model with three separate *wMB* weights for the three experimental blocks (which differed in terms of frequency of contingency changes: fast, medium, or slow), and a hybrid model with three separate *wMB* weights for each range of contingency changes rates. The last two models served to test whether the relative contribution of the two systems depended on volatility of transition structure, or instead more generally trial order. Model-fitting was confirmed to be able to recover true parameter values, as median estimated parameter values from model-fitting (see Methods for details) were well-correlated to known parameter values from simulations, $r \geq 0.99$, $p < 0.01$.

Model-fitting results supported the existence of a hybrid MB/MF strategy in our task. Candidate models were compared using two criteria – integrated Bayesian Information Criterion which controls for number of parameters (iBIC) [24] and exceedance probabilities [25] (S2 Table). The hybrid model with three *wMB* weights over blocks outperformed the other candidate models on both criteria, with the lowest iBIC and a probability of 89.4% that it was the most common of the four models across subjects. Thus, from here we only discuss the results of best-fit model, the three-block hybrid model.

The median fitted *wMB* weights in the three-block hybrid increased across the three blocks (Fig 3B-D), indicating some extent of "anti-habitization" rather than habit formation. The increase of *wMB* from block 1 to block 2, but not the increase from block 2 to block 3, was significant according to permutation tests, $p < 0.01$. Stay probability analyses were not conducted on the three separate blocks, as slower contingency changes meant that the later blocks had fewer samples of contingency changes for comparison. The increase in *wMB* across blocks was not attributable to differences in quality of fit from the model-fitting procedure, as the log-likelihood of parameter estimates did not differ significantly across blocks, $F(2,45) = 1.42$, $p > 0.05$. Strength of correlations between fitted and simulated *wMB* weights were also similar across blocks (block 1: $r = 0.99$, block 2: $r = 1.00$, block 3: $r = 0.99$; $p < 0.01$ for all blocks). Therefore, the significant increase in *wMB* from the first to second block was not caused by differences in quality of model fit.

To confirm that the increase in model-based weight was not due to differences in the rate of contingency changes, we further analysed the fitted weights from the three-frequency hybrid model, which had a different *wMB* assigned to each range of contingency change rates, i.e. fast (every 3-6 trials), medium (every 7-10 trials) and slow (every 11-14 trials) contingency change blocks. The estimated *wMB* weights (S3 Fig) were not significantly different between fast vs. medium, or medium vs. slow frequency of contingency change blocks in permutation tests, $p > 0.05$. Thus, the increase in *wMB* in our study seemed to be an effect of block order rather than environmental volatility from differences in contingency change rates. In summary, subjects became more model-based across the first two blocks but did not differ in MB influence between different rates of contingency changes; therefore, it seems that block order, but not contingency change volatility, affects *wMB* in our task.

As subjects became more model-based, high reward choices and consequently reward rate also increased. Choice probabilities for the high reward action differed over blocks, $F(2,45) = 5.77$, $p < 0.01$, with post-hoc tests finding a significant increase between the first and third blocks ($p < 0.01$) and the second and third blocks ($p < 0.05$). Additionally, there was a significant difference in reward rate across blocks, $F(2,45) = 3.83$, $p < 0.05$, increasing between the first and third blocks ($p < 0.05$). Mean reaction time and number of missed trials due to timeout did not significantly change across blocks, $p < 0.05$; therefore, the increase in high reward choices over blocks was not necessarily because subjects were worse at the task to begin with. Two analyses were performed to rule out the possibility of practice effects driving the association between reward rate and model-based weight. Within each block, there was a significant correlation of each subject's median *wMB* and reward rate (block 1: $r = 0.66$, $p < 0.01$, block 2: $r = 0.65$, $p < 0.01$, block 3: $r = 0.56$, $p < 0.05$), indicating that on an individual subject basis, the extent of MB control was related to reward earned. Since these analyses were conducted within blocks, the association with reward rate could not be accounted for by block order. Additionally, the hybrid model was simulated using a range of MB weights (0, 0.2, 0.4, 0.6, 0.8 and 1) using the one-weight hybrid model for simplicity. There was a significant effect of MB weight on reward rate, $F(5,90) = 8.5$, $p < 0.01$. In all, these findings suggest that MB influence in this task truly corresponded to a better "payoff" in terms of reward gained.

Discussion

We developed a novel two-level contingency change task in which flexibility to frequently-changing transition contingencies between states could determine whether subjects were using a model-based or a model-free strategy. Subjects showed a hybrid strategy when reacting to contingency changes, corroborating recent evidence of the parallel contribution of MB and MF systems in reward-guided decision-making. Importantly, this finding confirmed that changes to transition contingencies can elicit a balance of MB and MF behavior akin to changes to reward values. Model-fitting analyses indicated that a hybrid model with three MB weights best explained subjects' choices, with relative MB control increasing over blocks. The rate of contingency changes did not significantly shift the MB/MF balance; rather, MB control increased over the first two blocks of trials. This increase in MB control was concurrent with an increased proportion of high reward choices and consequently increased reward rate; individually, each subject's MB was also correlated with reward gained in the same block.

In all, these results illustrated that not only do subjects use a mixed MB/MF strategy, but within this hybrid strategy, the trade-off shifts towards "anti-habitization" across the first two blocks. This agrees with a previous study [20] that used the two-step task over three days, reporting that their subjects' MB weight increased across days. One distinction between our findings is that in [20], subjects started relatively model-based (i.e. median $wMB > 0.5$)

whereas in our case, subjects began relatively model-free (i.e. median *wMB* < 0.5). This difference in starting MB weight simply may be due to individual differences, which is evident even within our subject pool. Alternatively, differences could be accounted for by the relatively short reaction time limit in our task compared to theirs (750ms in ours vs. 2000ms). A shorter reaction time limit is known to provide a depth-of-planning pressure and favor more MF control [8]. Hence, our subjects may have started more model-free and only become more model-based once they mastered prospective planning of the task structure. This is supported by the lack of significant changes in reaction time across blocks, suggesting that subjects may have used the full extent of their time and eventually learned more efficient planning under time pressure, therefore showing increased MB influence over blocks.

These findings of an increase in MB control over blocks, however, goes against another study [26] using a similar task to the two-step task, that found an exponential decay in MB weight over the experimental session, or habit formation. This difference in findings is likely because they used a fixed rather than drifting amount of reward; in stationary environments such as these, habit formation can occur from overtraining, manifesting in an increase in MF rather than MB behavior [21]. Thus, these results point to the importance of maintaining a changing environment, as subjects can otherwise adapt to the change and become habitized.

Manipulations of the rate of contingency changes did not seem to affect MB/MF control. While it has been shown that environmental volatility can influence MB/MF levels in the context of reward value changes [21], in our case, the kind and range of contingency change volatility did not elicit a significant difference in relative MB/MF control. Further work is certainly needed to definitively rule out the possibility that environmental volatility in the form of the rates of contingency changes does not affect MB weight, but in the present study, we find that subjects did not change their use of MB control with contingency change volatility, but rather increased MB influence more generally with block order.

In conclusion, in a two-level contingency change task, subjects showed a hybrid MB/MF strategy, emphasizing their parallel contribution in reacting to changes in transition contingencies. The inclusion of multiple, frequent changes allowed us to perform model-fitting; by doing so, we found an increase in MB control over the first two blocks, a result not detectable in model-agnostic analyses alone. Our results build on the literature of a hybrid MB/MF strategy in reacting to changes in reward values, demonstrating a mixture of strategies in reacting to multiple, frequent contingency changes that has yet been unexplored. This novel paradigm therefore provides another avenue for exploring the relationship between MB and MF control for future studies in neuropsychiatric disorders that may differentially implicate this balance between changes in transition contingencies and changes in reward values.

Methods

Subjects

Sixteen subjects (nine males, mean age 24 years) took part. The study was approved by the University College London Research Ethics Committee (Project ID 3450/002). All subjects provided written informed consent.

Experimental procedure

Subjects performed 600 trials of three blocks (200 each) which differed in frequency of contingency changes: fast (every 3-6 trials), medium (contingency change every 7-10 trials) or slow (every 11-14 trials). Each subject was assigned to one of two groups (n=8 each), which differed by the order of presentation of fast and medium contingency change blocks, i.e. half of the subjects had fast, medium, then slow contingency changes, and the other half started with medium, fast, then slow frequency of contingency changes.

To ensure subjects understood the task structure, they were first trained with practice stimuli (35 trials) then trained on novel test stimuli without reward (55 trials) before starting the experimental session. Subjects were informed that contingency changes would occur, but did not know the frequency of changes nor that those rates would vary across the session.

At the first level, subjects had a two-alternative forced choice between two actions (pressing 'S' for the action available on the left side of the screen, 'L' for the right) with the presentation of stimuli randomized for the left/right side of the screen. To ensure that subjects recognized second-level states, they had to press 'D' if they encountered one of these states, and 'K' for the other. Both responses had a time limit of 750ms, following which the trial would end with no reward. Missed trials were not repeated.

Payoff at the high-reward terminal state varied with a drift rate of 0.2 and offset of 0.15 to the bound of £1, with payoff at the low-reward terminal state being £1 minus the reward of the high-reward terminal state. Subjects received a fixed proportion of their total reward gained, with payoff bounded between £5 and £25. To make the task adequately difficult and prevent formation of complex state-space representations [22], high- and low-reward assignments switched every 40 trials between the two terminal states. This change was designed never to co-occur with contingency changes.

Model

Both model-free and model-based algorithms seek to estimate the values of state-action pairs in order to choose the actions which can maximize expected future rewards. The state space was modelled as having a first-level state s_0 with two actions a_1 and a_2, two possible second-level states s_1 and s_2, and two possible terminal states s_3 and s_4. There was only one action available on second-level and terminal states, as the subject did not have any choices at these levels.

The model-free algorithm updates values of state-action pairs using temporal difference Q-learning [3, 27]. The reward r_t is used to compute a reward prediction error δ_t which updates action values for that state s and action a at time t, $Q_{MF}(s_t, a_t)$. At the first level r_t is set to be 0 as there is no reward at this level.

$$\delta_t = r_t + \max_{a'}[Q_{MF}(s_{t+1}, a')] - Q_{MF}(s_t, a_t)$$

$$Q_{MF}(s_t, a_t) = Q_{MF}(s_t, a_t) + \alpha_{MF}\lambda\delta_t$$

The reward prediction error updates existing action values according to a learning rate α_{MF} and modified by the eligibility parameter λ. Eligibility governs how much credit past actions were given for outcomes, with $\lambda = 0$ corresponding to a pure TD algorithm whereby first-stage actions are updated only by the second-level action values, which in turn is updated by terminal state rewards. In contrast, $\lambda = 1$ means the algorithm updates first-level actions only using the final reward from the terminal state reached on that trial.

The model-based algorithm learns both transition probabilities P_T and reward probabilities R_T. The transition probabilities track the transition contingencies P_T between states s and subsequent states s'. Upon encountering a contingency change, the model-based system always updated its knowledge of both transitions.

$$P_T\left(s_1 \overset{a}{\rightarrow} s_3\right) = \begin{cases} 1, & \text{if } s' = s_3, s = s_1 \\ 0, & \text{otherwise} \end{cases}$$

$$P_T\left(s_2 \overset{a}{\rightarrow} s_4\right) = \begin{cases} 1, & \text{if } s' = s_4, s = s_2 \\ 0, & \text{otherwise} \end{cases}$$

$$P_T\left(s_1 \overset{a}{\rightarrow} s_4\right) = 1 - P_T\left(s_1 \overset{a}{\rightarrow} s_3\right)$$

$$P_T\left(s_2 \overset{a}{\rightarrow} s_3\right) = 1 - P_T\left(s_2 \overset{a}{\rightarrow} s_4\right)$$

The reward probabilities R_T use the reward r_t to update its subjective reward R for that state s and action a at time t.

$$R(s_t, a_t) = R(s_t, a_t) + \alpha_{MB}(R(s_t, a_t) - r_t)$$

These learned transition and reward functions are then used to update the action values for the model-based system, Q_{MB}.

$$Q_{MB}(s_t, a_t) = P_T\left(s \overset{a}{\rightarrow} s_3\right) \cdot R(s_3, a) + P_T\left(s \overset{a}{\rightarrow} s_4\right) \cdot R(s_4, a)$$

Other parameters from the simulated models included learning rates for model-based and model-free systems, α_{MB} and α_{MF}, and a stay bias which temporarily increased the action value for the previously-selected action regardless of outcome, to quantify a perseveration bias. These additional parameters improved fit even when controlling for model complexity (S3 Table).

For both systems, values for the non-selected action were updated as well, assuming that subjects knew that the reward for the selected action and reward for the non-selected action were negatively related, according to proposals of fictive reward [28]. Action values were updated for both visited and non-visited states, with the action values of non-visited states corresponding to $1 - Q(s_t, a_t)$ of the visited states. The inclusion of fictive reward updates resulted in a better fit to the subjects' choices (S3 Table).

The hybrid model weighted MB and MF action values according to a parameter wMB, with $wMB = 1$ indicating fully MB control:

$$Q_{hybrid}(s_t, a_t) = wMB \cdot Q_{MB}(s_t, a_t) + (1 - wMB) \cdot Q_{MF}(s_t, a_t)$$

Action selection was then determined for all models according to a "softmax" rule which computes action probabilities as proportional to the exponential of the action values.

$$p(a_t = a_1 | s_t) = \frac{\exp\big(\beta \cdot Q(s_t, a_1)\big)}{\exp\big(\beta \cdot Q(s_t, a_1)\big) + \exp\big(\beta \cdot Q(s_t, a_2)\big)}$$

The inverse temperature β determined the extent to which action selection was stochastic or deterministic from action values, quantifying an exploration/exploitation trade-off.

Simulations

To best replicate the subjects' data of 600 trials for 16 subjects, each simulation was run for 16 initializations of 600 trials each. All reported simulations used fitted parameters from the three-block hybrid model for the learning rates α_{MF} and α_{MB}, inverse temperature β, eligibility trace λ and stay bias (S1 Table). *wMB* values were 1 for pure MB and 0 for pure MF models.

Model-fitting

Subjects' data were fit to the models using mixed effects hierarchical model fitting. Estimation-maximisation was used which iteratively generates group-level distributions over individual subject parameter estimates, choosing the parameters that maximizes the likelihood of the data given those estimates. Parameters were estimated by minimizing the negative log-likelihood of parameter estimates using *fminunc* in Matlab (MathWorks).

To ensure the efficacy of *wMB* parameter estimation for the candidate model, each block *wMB* was simulated for 11 different parameter values: 0, 0.1, 0.2, … 1. These resulted in a total of 33 parameter settings for *wMB1, wMB2, wMB3*, with 16 iterations per setting. All other parameters in the simulations were set constant as the median parameter estimates taken from the hybrid three-block model from model-fitting on the subjects' data. The same model-fitting procedure was performed on the simulated data and estimated parameter values were extracted.

The integrated Bayesian information criterion (iBIC) [24] was used to compare the fits of candidate models to the data, with lower scores indicating better fit; this criterion penalizes more complex models. Finally, Bayesian model selection [25] was used to examine the prevalence of each model in the participant population. This quantifies an exceedance probability, the probability that each model is the most common in the subject pool.

Permutation tests

Permutation tests were run to evaluate the probability that *wMB* could differ across blocks by chance. Subjects' blocks were randomly permuted such that each "block" contained a mixture of true first, second and third blocks. Model-fitting was run on each permutation to extract parameter estimates of *wMB* for each new "block". The probabilities $p(wMB_{block\ 2} > wMB_{block\ 1})$, and $p(wMB_{block\ 3} > wMB_{block\ 2})$ were then evaluated for each permutation. The occurrences of the random permutations which had a smaller $p(wMB_{block\ 2} > wMB_{block\ 1})$, and $p(wMB_{block\ 3} > wMB_{block\ 2})$ than the true permutation were then tallied.

Likewise, to evaluate the effect of frequency of contingency changes, permutation tests were run to compare *wMB* for fast, medium and slow contingency change blocks. Each subject was randomly assigned to one of the two groups (which differed in the order of fast and medium contingency change blocks) then *wMB* of each frequency block was computed for each permutation. Both the aforementioned one-tailed permutation test and a two-tailed Hellinger distance permutation test were used.

Acknowledgements

Thanks to Peter Dayan for supervision and comments, and Thomas Akam for comments on the manuscript. JL is supported by a Wellcome Trust doctoral fellowship. MK is supported by the Gatsby Charitable Foundation.

References

1. Adams CD, Dickinson A. Instrumental responding following reinforcer devaluation. The Quarterly journal of experimental psychology. 1981;33(2):109-21.

2. Dickinson A, Balleine B. Motivational control of goal-directed action. Animal Learning & Behavior. 1994;22(1):1-18.

3. Sutton RS, Barto AG. Reinforcement learning: An introduction: MIT press Cambridge; 1998.

4. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience. 2005;8:1704-11. doi: 10.1038/nn1560.

5. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997;275(5306):1593-9.

6. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. Neuron. 2011;69:1204-15. doi: 10.1016/j.neuron.2011.02.027.

7. Keramati M, Dezfouli A, Piray P. Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS computational biology. 2011;7:e1002055. doi: 10.1371/journal.pcbi.1002055. PubMed PMID: 21637741.

8. Keramati M, Smittenaar P, Dolan RJ, Dayan P. Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. PNAS. 2016:1-16. doi: 10.1073/pnas.1609094113. PubMed PMID: 27791110.

9. Otto AR, Skatova A, Madlon-Kay S, Daw ND. Cognitive control predicts use of model-based reinforcement learning. Journal of cognitive neuroscience. 2014.

10. Wunderlich K, Smittenaar P, Dolan RJ. Dopamine Enhances Model-Based over Model-Free Choice Behavior. Neuron. 2012;75:418-24. doi: 10.1016/j.neuron.2012.03.042. PubMed PMID: 22884326.

11. Miller KJ, Botvinick MM, Brody CD. Dorsal hippocampus plays a causal role in model-based planning. bioRxiv. 2016. doi: 10.1101/096594.

12. Balleine BW, O'Doherty JP. Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. Neuropsychopharmacology. 2010;35:48-69. doi: 10.1038/npp.2009.131.

13. Dayan P. Improving generalization for temporal difference learning: The successor representation. Neural Computation. 1993;5(4):613-24.

14. Momennejad I, Russek EM, Cheong JH, Botvinick MM, Daw N, Gershman SJ. The successor representation in human reinforcement learning. bioRxiv. 2016:1-27. doi: 10.1101/083824.

15. Gershman SJ, Markman AB, Otto AR. Retrospective revaluation in sequential decision making: A tale of two systems. Journal of Experimental Psychology: General. 2014;143(1):182.

16. Adams CD. Variations in the sensitivity of instrumental responding to reinforcer devaluation. The Quarterly Journal of Experimental Psychology. 1982;34(2):77-98.

17. Dickinson A. Omission learning after instrumental pretraining. The Quarterly Journal of Experimental Psychology: Section B. 1998;51(3):271-86.

18. Otto AR, Gershman SJ, Markman AB, Daw ND. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. Psychological science. 2013;24(5):751-61.

19. Colwill RM, Rescorla RA. Instrumental responding remains sensitive to reinforcer devaluation after extensive training. Journal of Experimental Psychology: Animal Behavior Processes. 1985;11(4):520.

20. Economides M, Kurth-Nelson Z, Lübbert A, Guitart-Masip M, Dolan RJ. Model-Based Reasoning in Humans Becomes Automatic with Training. PLoS computational biology. 2015:1-19. doi: 10.1371/journal.pcbi.1004463.

21. Simon DA, Daw ND. Environmental statistics and the trade-off between model-based and TD learning in humans. Advances in Neural Information Processing Systems (NIPS). 2011:1-9.

22. Akam T, Costa R, Dayan P. Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-step Task. PLoS computational biology. 2015:021428. doi: 10.1101/021428.

23. Miller KJ, Brody CD, Botvinick MM. Identifying Model-Based and Model-Free Patterns in Behavior on Multi-Step Tasks. bioRxiv. 2016:096339.

24. Huys QJ, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP. Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput Biol. 2012;8(3):e1002410.

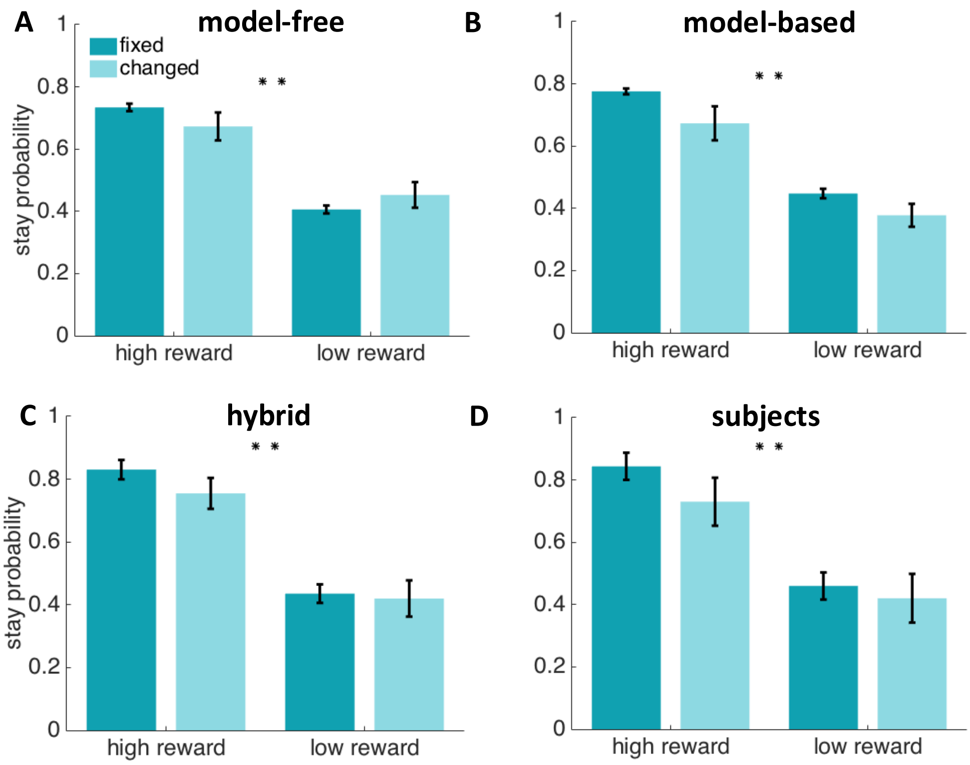
25. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. NeuroImage. 2009;46:1004-17. doi: 10.1016/j.neuroimage.2009.03.025.

26. Gläscher J, Daw Nathaniel D, Dayan P, O'Doherty JP. States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron. 2010;66:585-95. doi: 10.1016/j.neuron.2010.04.016.

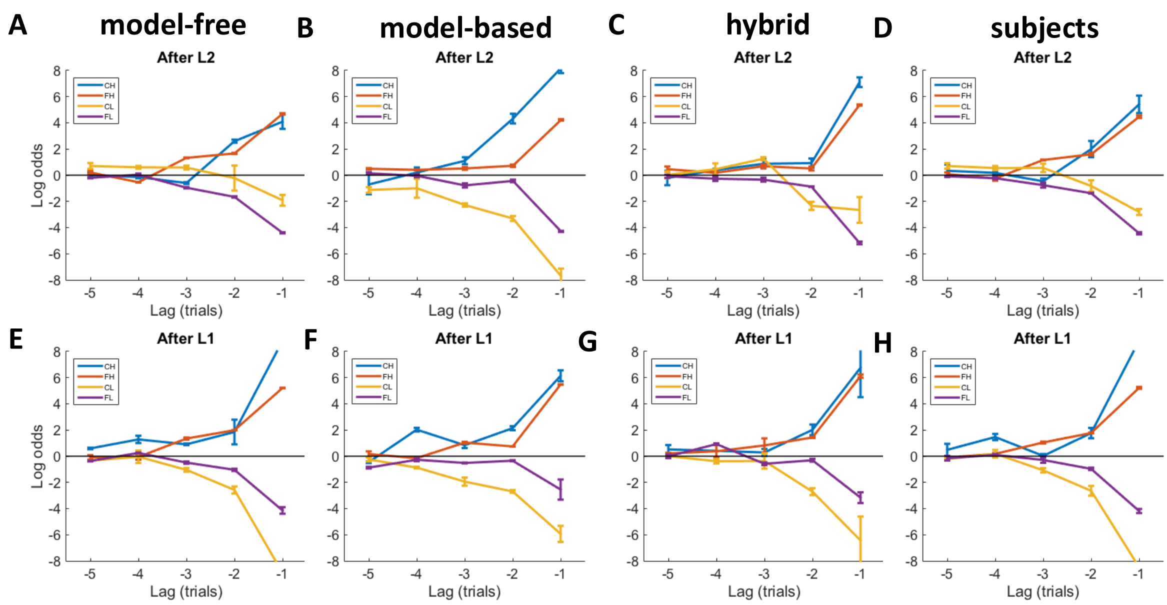
27. Watkins CJ, Dayan P. Q-learning. Machine learning. 1992;8(3-4):279-92.

28. Lohrenz T, McCabe K, Camerer CF, Montague PR. Neural signature of fictive learning signals in a sequential investment task. Proceedings of the National Academy of Sciences. 2007;104(22):9493-8.

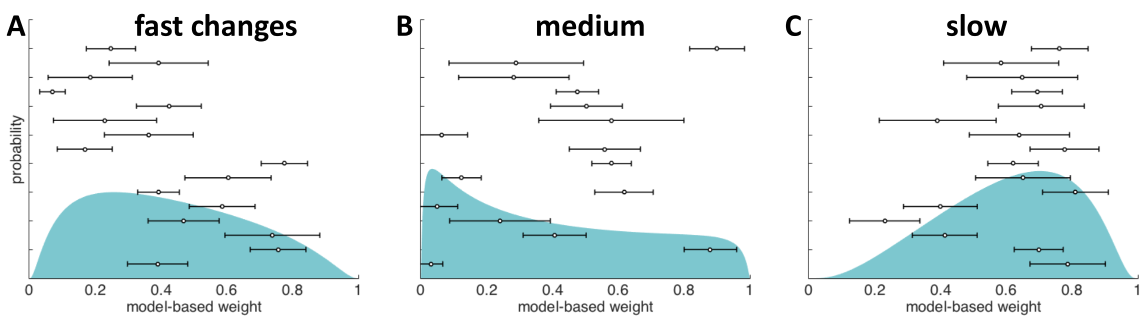
Supporting Information



S1 Fig. Stay probability patterns after first level contingency changes predicted by simulating model-based (A), model-free (B), and hybrid (C) reinforcement learning algorithms, along with experimental results (D). Stay-probability measures the probability of choosing the first-level action that results in the same second-level state as the previous trial, following a trial that started at the first level. For each system, this index was measured under four different conditions: when the reward received in the previous trial was "high" or "low", and when the transition experienced in the previous trial (relative to the trial before that) "changed" or remained "fixed". * $p < 0.05$, ** $p < 0.01$



S2 Fig. Predictive weights from lagged logistic regression after second level (A-D) or first level (E-H) contingency changes predicted by simulating model-based (A, E), model-free (B, F), and hybrid (C, G) reinforcement learning algorithms, along with experimental results (D, H). Lagged logistic regression measures the influence of different predictors over several trials in the past, in this case, five trials. For each system, this index was measured under four different conditions: when the reward received in the previous trial was high (H) or low (L), and when the transition experienced in the previous trial (relative to the trial before that) "changed" (C) or remained "fixed" (F).



S3 Fig. Model-based weights for different frequencies of contingency changes. Probability density function over the model-based weight parameters estimated from model-fitting, for the blocks of fast (every 3-6 trials), medium (every 7-10 trials) and slow (every

11-14 trials) frequency of contingency changes. Overlaid are the individual subjects' parameter estimates for each block type. Error bars represent standard deviation.

S1 Table. Median Plus Quartile Group-level Parameter Estimates. Best-fitting parameter estimates over the subjects from model-fitting.

	β	Stay bias	α_{MB}	α_{MF}	λ	wMB (block 1)	wMB (block 2)	wMB (block 3)
1st quartile	1.88	0.04	0.55	0.03	0.30	0.10	0.40	0.48
Median	2.99	0.10	0.70	0.30	0.47	0.23	0.57	0.63
3rd quartile	4.73	0.22	0.81	0.85	0.65	0.46	0.71	0.76

S2 Table. Model Comparison of Candidate Models. Integrated Bayesian Information Criterion (iBIC) and negative log-likelihood of all candidate models from model-fitting. The models tested were: pure model-free ("MF"), pure model-based ("MB"), hybrid MB/MF ("hybrid"), hybrid MB/MF with different weights fitted for each of the three 200-trial blocks ("three-block hybrid"), and a hybrid model with different weights fitted for each frequency of contingency changes ("three-frequency hybrid"). The winning model was the three-block hybrid, highlighted in gray, according to iBIC and Bayesian model selection [25].

Model	**Model-free**	**Model-based**	**Hybrid**	**Three-block hybrid**	**Three-frequency hybrid**
Parameters	4	3	6	8	8
iBIC	9051	8091	7938	7687	7711
Negative Log Likelihood	4489	4018	3914	3770	3782

S3 Table. Model Comparison of Additional Parameters. Integrated Bayesian Information Criterion (iBIC) and negative log-likelihood of the winning three-block hybrid model with different weights fitted for each of the three 200-trial blocks and the same model without stay bias, with $\lambda = 1$, with $\lambda = 0$, with only one learning rate for both MF and MB systems, and without updating fictive reward. The full model fit better to the data than the same model without each of the aforementioned parameters, even when controlling for model complexity in the iBIC.

Model	**Full model**	**No stay bias**	$\boldsymbol{\lambda = 1}$	$\boldsymbol{\lambda = 0}$	**One learning rate**	**No fictive reward**
Parameters	8	7	7	7	7	8
iBIC	7687	8047	7702	7754	7774	8024
Negative Log Likelihood	3770	3959	3787	3813	3823	3939