

Supplementary Methods

Reinforcement Learning Models

Q-Learning. These algorithms learn what actions to take when in a given state by learning a value of the reward that is expected after taking that action. The simplest form, as depicted in the *Methods* section in the paper, updates the value of the action via a simple Rescorla-Wagner (RW) rule (Rescorla and Wagner, 1972). On a trial t in which action a is selected, the value of action a is updated via a prediction error δ :

$$V_a(t+1) = V_a(t) + \eta\delta(t), \quad (\text{s1})$$

where η is the learning rate. The prediction error $\delta(t)$ is calculated by comparing the actual reward received $r(t)$ after choosing action a with the expected reward for that action:

$$\delta(t) = r(t) - V_a(t). \quad (\text{s2})$$

Thus, action values reflect the immediate subsequent reward that is expected after taking that action. This can be extended so as to learn the cumulative rewards expected in the future, as a consequence of taking a given action. In general, an exponentially discounted measure of future expected reward is used, in which more weight is given to rewards expected in the near future in comparison to rewards expected in the far future:

$$E(R) = \sum_{t=1}^{\infty} \gamma^t E(r_t) \quad , 0 < \gamma < 1 \quad (\text{s3})$$

where γ is the discount factor that indicates how near and far future rewards are weighed. Q-learning can be extended to learn future discounted expected rewards via Temporal Difference Learning (TD), in which the prediction error used in the value update is:

$$\delta(t) = r(t) + \gamma V_a(t+1) - V_a(t). \quad (\text{s4})$$

The effective reward obtained in the next time step due to action \mathbf{a} , is a sum of the immediate reward $\mathbf{r}(\mathbf{t})$ and the discounted expected reward due to action \mathbf{a}' in the next time step. When using Q-learning with temporal difference prediction errors, Q(td), we fitted a variable number of intermediate states in between the time of choice ($t=1$) and the time of reward outcome ($t=T$) (as used in O'Doherty et al., 2003b). Both Q-learning algorithms, Q(rw) and Q(td), then stochastically choose the action with most value, as explained in *Methods* (equation 3).

Actor-Critic. Instead of learning action values directly as in Q-learning, an alternative approach is to separate value learning and action selection into two stages (Sutton and Barto, 1981). The first stage or ‘Critic’ involves policy evaluation, in which the expected future rewards that follow from being in a particular state (corresponding to the average of the expected rewards available for all selected actions in that state) is learnt. The second stage or ‘Actor’, uses the prediction error signal ($\delta(t)$) derived from the critic to modify the probability of choosing a given action so as to increase the average expected reward of the whole policy. For this, action values (m_a) are learnt using the prediction error signal and probabilities of taking a given action (eq. s5a), and the probability of choosing action a is computed through a softmax (eq. s5b).

$$m_a(t+1) = m_a(t) + \eta[\delta_{ab} - P_b]\delta(t) \quad (\text{s5a})$$

$$P_b = \frac{e^{\beta m_b}}{\sum_i e^{\beta m_i}} \quad (\text{s5b})$$

The AC(rw) algorithm tested in this paper uses a Rescorla-Wagner rule (eq. s2) to calculate the prediction error $\delta(t)$, and the AC(td) algorithm uses Temporal Differences (eq. s4) with intermediate state steps between the time of choice and the time of reward outcome to calculate the prediction error $\delta(t)$.

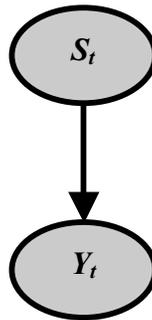
Advantage Learning. Advantage Learning is an extension of the actor-critic (Baird, 1993; see also Dayan and Balleine, 2002). Although advantage learning differs in a number of respects from the actor-critic, the main difference in terms of the behavioral fitting of the two models in the present case is that in advantage learning the value of the initial state (time of choice, $t=1$) is not directly estimated but instead is set according to:

$$V(t = 1) = P_a V_a(t = 2) + P_b V_b(t = 2) \quad (s6)$$

Bayesian Hidden State Markov Models

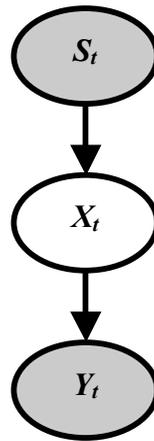
The key to Bayesian inference is an effective rule, *Bayes' rule*, which allows one to *infer* the probability of a hidden cause from observable variables.

Imagine a medical situation. A patient visits you repeatedly. Upon visit t , you prescribe her medication S_t . The result is a set of symptoms Y_t .



In simple reinforcement learning, you learn the relationship between the medication S_t and subsequent symptoms Y_t . When this patient returns for visit $t+1$, you use the direct relationship between both variables which you learnt from all the past pairs of observables (S_t, Y_t) , to determine the choice of medication S_{t+1} . Reinforcement learning is flexible enough to allow for changes in this direct relationship through learning. Effectively, reinforcement learning discounts observable pairs from earlier visits more than from more recent visits.

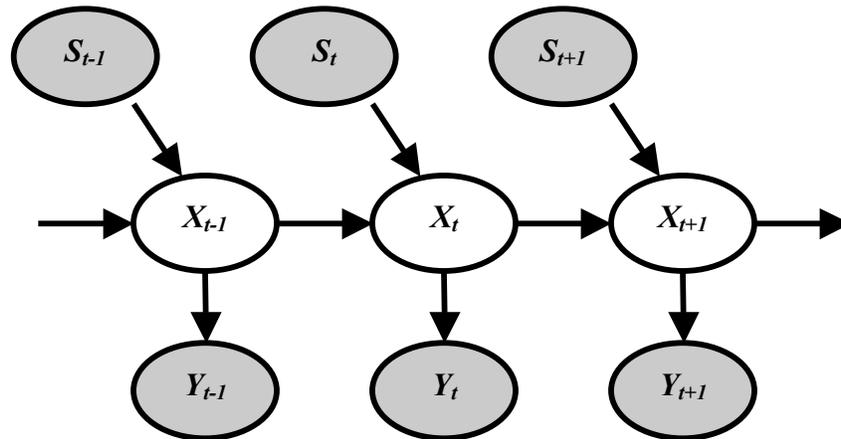
Reinforcement learning is only concerned with observables. Bayesian inference, in contrast, allows for the possibility that there is a hidden cause X_t that determines the effectiveness of the medication, i.e., the relationship between the medication S_t and the symptoms Y_t . Bayes' rule allows one to infer what X_t is given the past observable pairs and knowledge on how the hidden causes X_t generate the observable symptoms Y_t .



To make the example more concrete, one can imagine that X_t takes on 4 values: $X_t=0$ corresponds to no illness; $X_t=1$ corresponds to a viral infection; $X_t=2$ denotes bacterial infection; $X_t=3$ stands for allergic reaction. Y_t is a measure of nasal symptoms related to colds, flu and allergic reactions. And S_t stands for medication ($S_t=0$: none; $S_t=1$: anti-viral; $S_t=2$: antibiotic; $S_t=3$: antihistamine).

Bayesian inference allows one to model explicitly how the “cause” X_t changes over time – including as a result of the choice variable S_t (medication, in the case of the example).

It is often assumed that such changes are *Markov*, which means that the change in X from visit t to $t+1$ is influenced only by the immediate past (X_t).



The interpretation of the various components of this *Bayesian hidden state Markov model* in the paper is as follows: S_t denotes the choice of the subject ($S_t=0$ means “stay”; $S_t=1$ corresponds to “switch”); X_t denotes the correctness of the choice ($X_t=0$ means “incorrect choice”; $X_t=1$ means “correct choice”); Y_t is the observed reward (Y_t is a continuous monetary value, rewards being positive and punishments negative).

Further information on Bayes’ rule, Bayesian inference and Bayesian hidden state Markov models can be found by D.MacKay, [Information Theory, Inference and Learning Algorithms](#); Z.Ghahramani, [Bayesian Machine Learning](#), M.Jordan, [Graphical Models](#), and a [Bayesian methods reading list](#) by T.Griffiths.

References

Baird LC (1993) Advantage updating. Dayton, OH: Wright Patterson Air Force Base.

Dayan P, Balleine BW (2002) Reward, motivation, and reinforcement learning. *Neuron* 36:285-298.

Sutton RS, Barto AG (1998) Reinforcement learning. Cambridge, MA: MIT.

Supplementary Table 1: Mean parameters across subjects for the behavioral models.

Q-learning (Rescorla-Wagner)		
Learning Rate η	α	$\log_{10}\beta$
0.84 ± 0.03	0.0 ± 0.02	0.6 ± 0.2

State-based decision model				
Correct mean μ_C	Incorrect mean μ_I	Transition prob. δ	α	$\log_{10}\beta$
11.3 ± 1.0 cents	-7.5 ± 1.3 cents	0.24 ± 0.03	0.55 ± 0.02	1.4 ± 0.2

Supplementary Table 2**a) Prior correct activity**

brain region	laterality	x	y	Z	Z-score
Ventral medial PFC	L/R	6	57	-6	5.33
Posterior amygdala / anterior hippocampus	R	33	-15	-18	4.68
Dorsomedial PFC (frontopolar gyrus)	L	-9	66	21	4.47
Dorsomedial PFC (frontopolar gyrus)	R	6	66	24	4.46
Medial OFC / Subgenual cingulate cortex	L/R	0	39	-6	4.17
Medial OFC	L/R	0	33	-24	4.04
Posterior cingulate cortex	L/R	-9	-57	24	3.98

b) Posterior correct – Prior correct activity

brain region	laterality	x	y	z	Z-score
Dorsomedial PFC	L	-6	54	24	3.54
Ventral striatum	L	-24	3	-9	4.64
Ventral striatum	R	18	3	-15	4.48

c) Prior incorrect activity

brain region	laterality	x	y	z	Z-score
Dorsolateral PFC	R	39	36	33	4.30
Anterior insula/frontal operculum	R	48	15	9	3.96
Anterior cingulate cortex	R	6	21	45	3.37

d) Switch-Stay activity

brain region	laterality	x	y	z	Z-score
Anterior cingulate cortex	L/R	-3	24	30	4.54
Anterior insula	R	51	21	3	4.23
Anterior insula	L	-39	18	-12	4.26