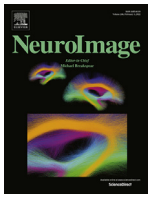
NeuroImage 246 (2022) 118780

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage



Brain signals of a Surprise-Actor-Critic model: Evidence for multiple learning modules in human decision making

Vasiliki Liakoni[a,1,*], Marco P. Lehmann[a,1], Alireza Modirshanechi[a], Johanni Brea[a], Antoine Lutti[b], Wulfram Gerstner[a,2], Kerstin Preuschoff[c,2]

[a] *École Polytechnique Fédérale de Lausanne (EPFL), School of Computer and Communication Sciences and School of Life Sciences, Lausanne, Switzerland*
[b] *Laboratoire de recherche en neuroimagerie (LREN), Department of Clinical Neurosciences, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland*
[c] *Geneva Finance Research Institute & Interfaculty Center for Affective Sciences, University of Geneva, Geneva, Switzerland*

ARTICLE INFO

Keywords:
Reinforcement learning
Surprise
Human learning
Sequential decision making
Behavior
fMRI

ABSTRACT

Learning how to reach a reward over long series of actions is a remarkable capability of humans, and potentially guided by multiple parallel learning modules. Current brain imaging of learning modules is limited by (i) simple experimental paradigms, (ii) entanglement of brain signals of different learning modules, and (iii) a limited number of computational models considered as candidates for explaining behavior. Here, we address these three limitations and (i) introduce a complex sequential decision making task with surprising events that allows us to (ii) dissociate correlates of reward prediction errors from those of surprise in functional magnetic resonance imaging (fMRI); and (iii) we test behavior against a large repertoire of model-free, model-based, and hybrid reinforcement learning algorithms, including a novel surprise-modulated actor-critic algorithm. Surprise, derived from an approximate Bayesian approach for learning the world-model, is extracted in our algorithm from a state prediction error. Surprise is then used to modulate the learning rate of a model-free actor, which itself learns via the reward prediction error from model-free value estimation by the critic. We find that action choices are well explained by pure model-free policy gradient, but reaction times and neural data are not. We identify signatures of both model-free and surprise-based learning signals in blood oxygen level dependent (BOLD) responses, supporting the existence of multiple parallel learning modules in the brain. Our results extend previous fMRI findings to a multi-step setting and emphasize the role of policy gradient and surprise signalling in human learning.

1. Introduction

When visiting a new city we may by chance find a new nice restaurant. Suppose we enjoyed our meal there and, two weeks later, decide to return. There are then two possibilities; it could be that, during our walk in the neighbourhood of the restaurant, we were able to build a mental map of the environment. In this case, even if at our later visit the main road is blocked due to construction, we would be able to plan a different path and return to the restaurant. On the other hand, it may be that we were not able to build a mental map, but two weeks later we encounter, by chance, an intersection where turning right feels preferable, even if we do not immediately remember that this turn leads to the great restaurant.

These two types of experiences closely correspond to two classes of reinforcement learning (RL) algorithms (Daw et al., 2005; Sutton and Barto, 1998), viz., those that use a model of the environment ("world-model") in the form of explicit knowledge about the consequences of actions in different states (*"if I turn left here I reach that intersection and if I turn right there I reach the restaurant"*), and those that do not use an explicit model but instead learn the preferences of actions in different states (*"turning right here is preferable"*). However, it is often neglected that each of the two classes contains different types of algorithms that make a fundamentally different use of similar building blocks.

The class of algorithms without a world-model (model-free algorithms) includes policy gradient methods and Temporal-Difference (TD) methods. In *policy gradient algorithms* (Peters, 2010; Schulman et al., 2015; Sutton and Barto, 1998; Williams, 1992), rewards directly trigger changes of the parameters of the policy that the agent uses to choose actions. In multi-step decision tasks, the parameter updates of policy gradient methods also depend on eligibility traces (Fig. 1A) that keep a decaying memory of previous state-action pairs (Gerstner et al., 2018; Lehmann et al., 2019; Peters, 2010; Schulman et al., 2015; Sutton and Barto, 1998; Williams, 1992).

* Corresponding author.
E-mail address: vasiliki.liakoni@epfl.ch (V. Liakoni).
[1] These authors contributed equally.
[2] These authors contributed equally.

https://doi.org/10.1016/j.neuroimage.2021.118780.
Received 26 March 2021; Received in revised form 3 August 2021; Accepted 4 December 2021
Available online 5 December 2021.

In contrast, *model-free TD methods* determine the policy of an agent only indirectly. First, "goodness" values are assigned either to the states (V-values) or to state-action pairs (Q-values), where the values give an estimate of the expected reward that can be collected starting from a given state or state-action pair. In model-free TD algorithms (Sutton and Barto, 1998), these values are learned via the reward prediction error (RPE) which quantifies the discrepancy between the value that has been expected and the one that is perceived. The actual policy that guides action choices is then derived from the Q-values or V-values. Similarly to policy gradient methods, data-efficient versions of TD algorithms use eligibility traces (Fig. 1A).

All methods in the class of algorithms with a world-model (*model-based algorithms*) build a memory of experienced state-action-state triplets in order to estimate transition probabilities from the present state to a next state when choosing an action. In other words, in model-based RL, agents learn a model of the environment, i.e. how states are connected and where rewards are located. Given the world-model, agents can flexibly update values through mental simulation, at the expense of higher computational costs. Learning the world-model is mediated by a state prediction error (SPE) or by surprise (Fig. 1A), which express the discrepancy between the expected and the experienced state in the world (Faraji et al., 2018; Gläscher et al., 2010; Liakoni et al., 2021; Yu and Dayan, 2005). In standard model-based RL, the transition matrix of the world-model can either be used to update (in the background) the estimate of Q-values or V-values (Sutton and Barto, 1998), or to directly plan ahead, via "mentally" sampling from the transition matrix in order to predict possible future trajectories. Furthermore, as we will show in this paper, the transition matrix could also be used for the extraction of surprise signals.

While classic work in neuroscience compared model-based with model-free TD learning, it soon became apparent that a binary segregation between model-free and model-based is oversimplifying (Collins and Cockburn, 2020; Daw, 2015; 2018; Langdon et al., 2018), suggesting the presence of parallel learning systems in the brain (Geerts et al., 2020). Importantly, these modules could be always active and learning but take over control only in the appropriate conditions (Daw et al., 2005; Lee et al., 2014). The hypothesis of parallel learning systems in the brain is supported by various experimental findings. First, humans and animals exhibit behaviors consistent with both or a mixture of the two types of learning in different circumstances (Daw et al., 2011; 2005; Geerts et al., 2020; Xu et al., 2021). Second, in the brain, the neural substrates of the two strategies are often shared (Doll et al., 2012; Gremel and Costa, 2013; Langdon et al., 2018; Tanaka et al., 2015). For example, while dopaminergic neurons have been traditionally thought to convey a model-free RPE, they are also sensitive to sensory prediction errors (Howard and Kahnt, 2018; Takahashi et al., 2017), indicating an involvement in model-based learning. Similarly, the striatum, traditionally thought to be involved in model-free learning, has been found to also perform model-based computations (Daw et al., 2011).

Given the large number of human behavioral RL studies highlighting different RL algorithms (Anggraini et al., 2018; Cushman and Morris, 2015; Daw et al., 2011; Deserno et al., 2015; Dezfouli et al., 2014; Doll et al., 2015a; 2015b; Economides et al., 2015; Fermin et al., 2016; Gershman et al., 2014; Gläscher et al., 2010; Huys et al., 2012; Kroemer et al., 2019; Lee et al., 2014; Otto et al., 2013a; 2013b; Simon and Daw, 2011; Wimmer and Shohamy, 2012; Wunderlich et al., 2012a; 2012b), we ask here whether we can find evidence supporting all three classes of algorithms we identified above. In particular, we ask whether we can find signatures of three parallel learning systems in human behavior linked to three different brain signals: (i) the RPE at non-goal states used in model-free TD learning; (ii) surprise (derived, for example, from the SPE) used in updating a world-model; and (iii) the reward at the goal state (dissociated from the RPE at non-goal states) as well as policy preferences of policy gradient learning. Importantly, if we assume that the three learning systems always run in parallel even if they are not used in a specific task, we should be able to identify brain signals for latent learning systems, i.e. ones that do not currently make a significant contribution to the observed behavior.

Classical experimental research trying to address the above questions has mostly employed two-step decision tasks (Daw et al., 2011; Gläscher et al., 2010) or variations thereof (Cushman and Morris, 2015; Deserno et al., 2015; Dezfouli et al., 2014; Doll et al., 2015a; 2015b; Economides et al., 2015; Kroemer et al., 2019; Otto et al., 2013a; 2013b; Wunderlich et al., 2012b). However, two-step tasks may implicitly favour model-based approaches since the mental load for planning and updating is small (Silva and Hare, 2020), and their simple structure may not allow for experimental manipulations necessary for dissociating different entangled learning signals (Daw, 2018; Fouragnan et al., 2018; Pernet, 2014). Therefore, in this paper, we focus on a multi-step decision task that corresponds more closely to the more complex tasks of the real world (Tartaglia et al., 2017) and gives us sufficient degrees of freedom for experimental manipulations. Our hypothesis has been that in such a more complex task, characteristic behavioral and physiological features of model-based and model-free RL algorithms could become transparent. However, scaling up the task complexity alone in brain imaging experiments causes additional challenges for dissociating different entangled learning signals. To facilitate dissociation, we introduce a class of surprise trials that should help to distinguish brain signals of model-free TD and model-based approaches.

The space of potential RL algorithms is large (Daw, 2018; Silva and Hare, 2020; Sutton and Barto, 1998), since modules of different learning systems can be combined in multiple interaction patterns (Fig. 1A). In this work, we test how well more than ten different representative combinations of RL modules can explain human behavior in a novel complex multi-step task. Since policy gradient methods have not yet received the equivalent amount of attention in human experimental studies (Coddington and Dudman, 2019; Li and Daw, 2011; O'Doherty et al., 2004), we also include policy gradient approaches into our model comparison. In particular, we introduce a novel RL algorithm that we call "Surprise Actor-critic". Our algorithm combines a model-free actor that optimizes the parameters of its policy via policy gradient, with a model-free critic that optimizes values via TD learning, and with a model-based world-model that optimizes transition estimates via a surprise measure. Importantly, the world-model is not used for planning or for the estimation of model-based Q-values, but only to detect surprising events and potentially influence the learning rate of the actor. The influence of surprise on the learning rate is motivated by the design of our experiment and derived from a normative approach within a Bayesian framework for outlier detection. Our results extend previous fMRI findings to a multi-step scenario, support the existence of multiple parallel learning modules in the brain and indicate surprise and policy gradient as important building blocks of human learning.

2. Materials and methods

2.1. Experimental paradigm

2.1.1. Task layout

We designed a multi-step decision making task (Tartaglia et al., 2017) situated in a state-space with 7 circularly arranged fractal images (states), with 2 possible actions at each state (apart from the goal which is a terminal state) (Fig. 1B). At the beginning of each episode, participants were shown an initial state, randomly chosen among two possible initial states. As soon as a participant chose an action, there was a random time interval of 2 to 7 seconds, after which a different image was shown (Fig. 1C). Participants continued to choose actions until they reached the goal, which completes an episode (Fig. 1B and Fig. 1C). They were instructed that their task was to reach the goal in the smallest number of actions. The image of the goal state was visually distinguishable from all other states and known to participants beforehand. State transitions were deterministic, with occasional "surprise tri-

als", explained further below (subsection 2.1.2). In what follows, we use the terms "state transition" and "trial" interchangeably.

From each state, the two actions led to new states at different distances from the goal. The "correct action" is defined as the one that brings the participant closer (in terms of number of actions) to the goal state. Since two actions (leading to two different target states s') can be taken in each state s, we characterize a state s by the "distance from goal" of its two possible target states s'. For example, from the state $s3$ of Fig. 1B the correct action (green arrow) leads to the goal (0 actions from goal), whereas the other action (red arrow) leads to state $s6$ located 3 actions away from the goal. We therefore denote state $s3$ as "0-or-3". States that have equal minimum distance from goal may differ in the resulting distance from goal when the "wrong" action is chosen, or, in other words, in their "difference in correctness" between the two available actions. For example, for two states "0-or-3" and "0-or-1" choosing the correct action brings participants to the goal, but choosing the wrong one is more detrimental for the "0-or-3" state.

The state images and their locations on the screen were randomized across participants (for example, state $s1$ was not physically next to $s2$, but could be positioned at any location). The assignment of actions to left or right button presses were also randomized. We defined two different underlying transition matrices (one of them is depicted in Fig. 1B) also in a randomized way across participants. The distance of each state from goal and the actions' "correctness" for both graphs are provided in the Appendix (Table A.1). The task was implemented in Matlab using the Psychophysics Toolbox (Brainard, 1997).

2.1.2. Surprise trials

During the experiment we ran in the background the model-free (MF) SARSA-λ algorithm (Sutton and Barto, 1998) and the model-based (MB) Forward Learner (Gläscher et al., 2010) with the participant's choices as inputs. This gave us an online estimate of the values and the transition probabilities computed by the two algorithms, respectively. The SARSA-λ uses the RPE to update approximate Q-values $Q(s,a)$, i.e. the expected sum of future discounted rewards starting from state s and action a. The Forward Learner updates the transition probabilities via the SPE and uses them to directly compute the $Q(s,a)$ values via the Bellman equation. More details on the algorithms are provided in subsection 2.4 and in the Appendix subsection A.8. For the online RPE and SPE computations, the choice of the parameters (e.g. learning rates α) was based on pilot experiments performed prior to this study.

After participants had encountered the goal four times (i.e. completed 4 episodes of the task), we introduced "surprise trials". On a surprise trial, participants transited to a state s'' other than the one they had learned to expect as the outcome of action a from state s (Fig. 1D). We employed two types of surprise trials: (i) purely random transitions, and (ii) transitions that met a threshold criterion on V-values. For the latter, if a participant expected to transit from s to s', the other state s'' was chosen such that s' and s'' had similar V-values according to the MF SARSA-λ, i.e. $\left|V(s') - V(s'')\right| \leq \Delta V$ where $V(s) = \max_a Q(s,a)$ and ΔV was a small threshold. This manipulation does not affect the MF system, since the experienced RPE stays the same. For learned transitions, in particular, the RPE will take low values. At the same time, the experienced MB SPE will be high, since the learned transition has been violated (Fig. 1D).

In more detail, after each action taken by a participant, from a state s to (an expected state) s', we checked whether the following conditions were fulfilled: (i) the transition from s to s' has been learned according to what has been previously experienced by the Forward Learner running in the background, (ii) there is a state s'' that meets the aforementioned V-value threshold criterion, according to the SARSA-λ running in the background, and (iii) more than 3 trials have occurred since the last surprise trial. If these conditions were fulfilled, the participant transited to s'', and this constituted a surprise trial/surprising transition. Moreover, if during 8 consecutive trials no surprise trial had occurred, i.e. the above conditions were not fulfilled, a randomly chosen unexpected transition was enforced in order to ensure some variability – excluding the goal state and the current state from possible landing states.

Without unexpected transitions, the SPE would quickly decrease to low or zero values in a deterministic graph after experience, and unexpected transitions without the V-value criterion explained above would lead to a high SPE as well as a high RPE, resulting in a high correlation between these two signals. Our novel experimental manipulation with online monitoring, however, enables us to increase the variability of the SPE signal and to decorrelate the RPE from the SPE at the same time. In order to prevent participants from resetting their estimations and starting learning from scratch after an unexpected transition, i.e. prevent them changing their policy, we informed participants that the task graph does not change during the experiment and the surprise trials are outliers. We note that our task design was based on previous literature (Daw et al., 2011; Doll et al., 2015a; Economides et al., 2015; Gläscher et al., 2010; Lee et al., 2014; Otto et al., 2013a) and thus on the assumption that SARSA and Forward Learner are likely the correct underlying algorithms (see Appendix subsection A.5 for more details and for the case that this assumption is violated).

2.2. Participants

Twenty-three healthy adults (average age 23.9 years old, right-handed, 10 female) were recruited to participate in our experiment. The choice of the number of participants was based on earlier human RL fMRI studies (Daw et al., 2011; Gläscher et al., 2010). All participants provided written informed consent, and the experiment was conducted in accordance with the ethics commission of the Canton de Vaud, Switzerland. Participants performed the task in a 3T Siemens Prisma MRI Scanner at the Laboratoire de recherche en neuroimagerie (LREN) at the Centre hospitalier universitaire vaudois (CHUV). Prior to the experiment, participants were informed about the number of states and possible actions, as well as the image indicating the goal state. They were instructed that their task is to reach the goal state in the least number of steps and got familiar with the task outside and inside the scanner during short sessions of two episodes each, with different images and transitions than the ones used during the experiment. Furthermore, participants were beforehand informed on the existence of surprise trials and on the fact that the underlying transition matrix does not change. Participants performed the task in a block of 20 minutes and were compensated with a fixed monetary amount for their participation, plus a small extra performance-based amount.

We excluded two participants (one male and one female) from both our behavioral and fMRI analysis; one participant performed less than half of the average number of episodes than the rest of the participants and likely did not understand the task, and another participant was falling asleep and his brain images exhibited a high degree of movement artifacts. The remaining 21 participants performed on average 54 full episodes (std: 5.24), consisting of on average 188 trials (std: 8.33, minimum: 174, maximum: 201). Approximately 17% of the trials (std: 1.5 %) were surprise trials.

We emphasize that the nature of "surprise" is that it must be against reliable expectations. Therefore, the fraction of surprise trials has been kept relatively low. If we increased the fraction of surprise trials, the model itself would be better described by a stochastic model with high probabilities for several transitions and would likely be perceived by participants as such. Our design was intended to suggest to participants that the task is *deterministic* with a few outliers, and it was communicated as such to them.

2.3. fMRI data acquisition and preprocessing

We acquired functional data of the 21 participants on a 3T Siemens Prisma MRI Scanner, using a custom-made T2*-weighted 2D echo planar imaging (EPI) sequence (Lutti et al., 2013) (442 volumes, 34 slices/volume, slice thickness of 2.5 mm, 20% interslice gap, repetition

time 2720 ms, flip angle 90°, matrix size 64x64, field of view 192 mm^2). We used parallel imaging with an acceleration factor of 2 along the phase-encoding direction and images were reconstructed using GRAPPA (Griswold et al., 2002). To minimize signal dropouts in the EPI images and achieve optimal BOLD sensitivity in all brain regions, we acquired three echo images following each radio-frequency excitation (echo times = 17.4 ms, 35.2 ms, 53 ms) (Poser et al., 2006). Slices were tilted by -20° off the line connecting the anterior-posterior commissure. Brain coverage included the orbitofrontal cortex and subcortical structures and excluded some posterior-superior frontal and parietal regions. We used B0-field maps, obtained from double-echo FLASH acquisitions (64 slices; matrix size 64 × 64, spatial resolution 3 mm; short echo time 10 ms, long echo time 12.46 ms; repetition time 1020 ms), to correct the EPI images for distortions along the phase-encode direction (Hutton et al., 2002).

For the preprocessing of the fMRI images, we used the SPM12 software[3] For each image volume of the EPI time-series, we first combined the three echo images using a simple summation (Poser et al., 2006). The resulting combined images, with maximal BOLD sensitivity in all brain areas, were used for all subsequent analysis. We performed slice-timing correction and corrected the images based on the obtained B0-field maps. Next, the images were realigned to the first functional image (rigid-body realignment), spatially normalized to standard Montreal Neurological Institute (MNI) coordinates (via co-registration of the structural T1 images to the mean functional image and tissue segmentation), and smoothed with a Gaussian kernel of 8 mm (Full width at half maximum - FWHM). We used the resulting images for the statistical analysis.

2.4. RL Algorithms for behavioral modelling

We considered several classes of possible strategies that a participant may follow to accomplish the task: purely MF strategies, purely MB, MF strategies with surprise modulation, and hybrid strategies, defined as combinations of MF and MB strategies (Fig. 1A).

In this section, we present a brief overview of our candidate algorithms. Our focus will be more on the algorithms that ranked best according to our statistical model comparison (subsection 3.2). More details for all algorithms can be found in the Appendix (subsection A.8).

2.4.1. Model-free strategies

A MF strategy with RPE-mediated updates does not take into account the information about the existence of surprise trials and includes the values of surprising transitions together with non-surprising ones. By construction of the task, for most of the surprise trials, this does not affect substantially the value estimation, since the landing state of a surprise trial is chosen so that it has a value similar to the expected state (apart from the purely random transitions).

In the family of purely MF algorithms we considered the TD algorithm SARSA-λ (Sutton and Barto, 1998), the policy gradient REINFORCE (Williams, 1992), the policy gradient REINFORCE with a baseline (Williams, 1992), and the Actor-critic (Sutton and Barto, 1998). All MF algorithms we considered use exponentially decaying memory (eligibility) traces to backpropagate updates to preceding actions (Fig. 1A, box A).

SARSA-λ estimates the $Q(s,a)$ values with RPE-mediated updates and with eligibility traces (Fig. 1A, boxes A, B, and D). The values are then used to compute the policy $\pi(s,a)$, i.e. the probability of taking an action a at a state s, via a softmax function (see subsection 2.4.5).

The REINFORCE algorithm and the REINFORCE with a baseline consider a parametrized policy, i.e. proportional to $\exp(p(s,a)/\tau)$ (see Eq. 9), where the parameters $p(s,a)$ are also called "policy preferences" (Fig. 1A, box E). They then optimize the policy preferences $p(s,a)$ of all the preceding within-episode decisions directly with gradient ascent using the discounted cumulative reward, called return, potentially corrected by a baseline (Fig. 1A, boxes A, C, and E).

The Actor-critic involves both value learning and policy parameter optimization (Fig. 1A, boxes A, B, C, D, and E). Values are used for optimizing the policy preferences, and the policy preferences are used for action selection. The V-values are estimated by the critic via the RPE (Fig. 1A, boxes B and D), defined as

$$\mathrm{RPE}_t = r_{t+1} + \gamma\, V(s_{t+1}) - V(s_t), \tag{1}$$

where r_{t+1} is the immediate reward received upon landing on the next s_{t+1}, and $\gamma \in [0,1]$ is the discount factor that controls the importance of distant rewards. The RPE is used to update the V-values incrementally, in a way similar to SARSA-λ, and is also fed into the actor to modify the policy preferences $p(s,a)$ of all pairs of states and actions (Fig. 1A, arrow from box B to E – see Appendix subsection A.8 for more details)

2.4.2. Model-based strategies

A MB strategy attempts to estimate a model of the world (Fig. 1A, box F), summarized by the transition matrix $T(s_t,a_t,s_{t+1})$, i.e. the probability of transiting from a state s_t to s_{t+1} when selecting a_t, and the reward function $\bar{R}(s_t,a_t,s_{t+1})$, expressing the expected immediate reward due to the transition from s_t to s_{t+1} when selecting a_t. It then uses its estimated world-model to compute the values of states and actions (Fig. 1A, boxes F and H), via the (Bellman) equation $Q(s_t,a_t) = \sum_{s'} \hat{T}(s_t,a_t,s')\big(\hat{\bar{R}}(s_t,a_t,s') + \gamma V(s')\big)$, where $\hat{T}$ and $\hat{\bar{R}}$ are the estimations of the true T and $\bar{R}$ by the MB learner. These values are then used to compute action selection probabilities via a softmax policy, similar to the MF SARSA-λ (Sutton and Barto, 1998).

Analogous to Gläscher et al. (2010), we assumed that the reward function was known to participants through the instructions and familiarization with the task, i.e. that only one of the images was rewarding. However, we considered two ways of estimating the transition matrix of the task, a sub-optimal and a Bayes-optimal way.

First, we used a traditional approach for learning the transition matrix through a delta-rule based on the SPE (Fig. 1A, box G), called Forward Learner (FWD) (Daw et al., 2011; Gläscher et al., 2010). The $\mathrm{SPE}_{\mathrm{FW}}$ of the FWD learner is defined as

$$\mathrm{SPE}_{\mathrm{FW}}^{(t)} = 1 - \hat{T}_{\mathrm{FW}}(s_t,a_t,s_{t+1}), \tag{2}$$

where $\hat{T}_{\mathrm{FW}}(s_t,a_t,s_{t+1})$ is the transition matrix estimated by the FWD. The SPE is a measure of surprise or state discrepancy. In the FWD, the SPE is used to incorporate new information to the old estimate of the transition matrix with a constant learning rate. In this case, surprising trials are essentially treated as stochasticity in the environment.

Going beyond common methods for transition matrix estimation in human studies, we, secondly, adopted a normative approach and used a Bayesian framework for estimating the transition matrix by assuming prior knowledge on the structure of the task (generative model), formed by the instructions we gave to participants. More precisely, we developed an approximate Bayesian model learning algorithm – a particle filter (Doucet, Godsill, Andrieu, 2000; Gordon, Salmond, Smith, 1993; Särkkä, 2013)– that exploits the fact that participants were informed that there would be occasional unexpected transitions, while the underlying structure of the task would remain unchanged.

The derived update rule of our algorithm involves a recently developed measure of surprise (Liakoni et al., 2021), the "Bayes Factor Surprise" $\mathbf{S}_{\mathrm{BF}}$ (Fig. 1A, box G), defined for our task as

$$\mathbf{S}_{\mathrm{BF}}^{(t)} = \frac{\mathrm{Const.}}{\hat{T}_{\mathrm{BF}}(s_t,a_t,s_{t+1})}, \tag{3}$$

where $\hat{T}_{\mathrm{BF}}(s_t,a_t,s_{t+1})$ is the Bayesian estimated probability of transiting to s_{t+1} from the current (s,a) pair at time t (See Appendix subsection A.6).

Our algorithm essentially implements outlier detection; high values of surprise in this task signal a surprising transition that should be ignored, since the underlying graph connectivity does not change. Hence

[3] https://www.fil.ion.ucl.ac.uk/spm/software/spm12/.

in our task the Bayesian update of the transition probability from s_t and a_t to any other state s' has the form of

$$\hat{T}_{\mathrm{BF}}(s_t, a_t, s') \leftarrow (1 - \gamma^{(t)}_{\mathbf{S}_{\mathrm{BF}}})\hat{T}^{\mathrm{integration}}_{\mathrm{BF}}(s_t, a_t, s') + \gamma^{(t)}_{\mathbf{S}_{\mathrm{BF}}}\hat{T}_{\mathrm{BF}}(s_t, a_t, s'), \tag{4}$$

where $\gamma^{(t)}_{\mathbf{S}_{\mathrm{BF}}} \in (0, 1)$ is a learning (adaptation) rate that approaches 1 if surprise $\mathbf{S}_{\mathrm{BF}}$ is large, and $\hat{T}^{\mathrm{integration}}_{\mathrm{BF}}$ is the transition matrix after the Bayesian incorporation of the observed transition. In contrast to the FWD, we have a trade-off between (perfectly) integrating the observed transition to the estimated transition matrix $\hat{T}^{\mathrm{integration}}_{\mathrm{BF}}$ and maintaining the current estimation of the transition matrix $\hat{T}_{\mathrm{BF}}$ (i.e. ignoring the transition). This trade-off is controlled by a surprise-dependent learning rate $\gamma_{\mathbf{S}_{\mathrm{BF}}}$. Derivations and more details are provided in the Appendix (subsection A.6 and subsection A.7). The exact online computation of Eq. 4 is computationally expensive and grows intractable in time. We therefore approximate it with a particle filter (subsection A.6).

For our task and for the *same* estimate of the transition matrix $\hat{T}_{\mathrm{BF}}$ (see Appendix subsection A.11 for more details), we have the following non-linear relationship between $\mathrm{SPE}_{\mathrm{BF}}$ and $\mathbf{S}_{\mathrm{BF}}$

$$\mathbf{S}^{(t)}_{\mathrm{BF}} = \frac{\mathrm{Const.}}{1 - \mathrm{SPE}^{(t)}_{\mathrm{BF}}}. \tag{5}$$

This means that an agent (or the brain), that learns via approximate Bayesian inference, could potentially estimate first an $\mathrm{SPE}_{\mathrm{BF}}$ based on the experienced transition and then convert it to $\mathbf{S}_{\mathrm{BF}}$ downstream, or vice versa. There is no a-priori reason to prefer one measure of surprise over the other – in this task and for a Bayesian learner – nor a way to distinguish them. We exploit this relationship of Eq. 5 later in the statistical analysis of the fMRI data, and we report our results for both measures of surprise, $\mathrm{SPE}_{\mathrm{BF}}$ and $\mathbf{S}_{\mathrm{BF}}$, computed by the same algorithm. Importantly, in both cases the transition matrix $\hat{T}_{\mathrm{BF}}$ is calculated by our outlier detection algorithm driven by $\mathbf{S}_{\mathrm{BF}}$.

Either of the two ways to estimate the world-model (delta-rule with constant learning rate and approximate Bayesian inference) can be coupled to a RL procedure that estimates the Q-values through value iteration (e.g. the FWD (Daw et al., 2011; Gläscher et al., 2010)) or an approximation thereof (e.g. Prioritized Sweeping (Moore and Atkeson, 1993; Seijen and Sutton, 2013)), leading to a total of four MB algorithms. For our behavioral fitting analysis, we considered two of these: the FWD that estimates the transition probabilities via $\mathrm{SPE}_{\mathrm{FW}}$ (same as Gläscher et al. (2010)), and a particle filter (approximate Bayesian algorithm) for the model estimation combined with Prioritized Sweeping (PS with PF) that estimates the transition probabilities via $\mathrm{SPE}_{\mathrm{BF}}$ or $\mathbf{S}_{\mathrm{BF}}$.

2.4.3. Surprise-modulated model-free strategies

As a possible way that the MF and MB systems interact and are combined in the brain, we considered strategies where surprise signals ($\mathbf{S}_{\mathrm{BF}}$ or $\mathrm{SPE}_{\mathrm{BF}}$) derived from the world-model provide additional information to the MF system (Xu et al., 2021). For example, a high value of $\mathbf{S}_{\mathrm{BF}}$ is likely to correspond to the detection of a surprise trial and may dampen the update of the MF values (in SARSA) or the policy preferences (in REINFORCE) of a particular state-action pair (Fig. 1A, arrow from box G to box D or E). Surprise in this case is not used for planning and for the computation of MB values, but only as a modulatory signal to MF updates. We introduce the Surprise REINFORCE, the Surprise Actor-critic and the Surprise SARSA-λ as MF algorithms with a learning rate modulated by surprise, where surprise is derived from a world-model learning module that performs outlier detection.

More specifically, in the Surprise Actor-critic (Fig. 1E), the critic computes the TD V-values via the RPE (Eq. 1), similar to the standard Actor-critic algorithm, i.e.

$$\begin{aligned} &V(s) \leftarrow V(s) + \alpha_C\, \mathrm{RPE}_t\, e^C_t(s) \quad \forall\, s \in \mathcal{S} \\ &e^C_t(s) = \begin{cases} 1, & \text{if } s_t = s \\ \gamma\lambda_C e^C_{t-1}(s), & \text{otherwise,} \end{cases} \end{aligned} \tag{6}$$

where α_C is the learning rate of the critic, $e^C_t(s)$ are exponentially decaying eligibility traces of the critic, and λ_C the eligibility trace decay factor of the critic.

The RPE is then fed to the actor module to influence the update of the policy preferences (Fig. 1E). Contrary to the standard Actor-critic, a world-model is learned via a particle filter (see Appendix subsection A.6), and a surprise signal ($\mathbf{S}_{\mathrm{BF}}$) derived from it modulates the actor's learning rate (Fig. 1E)

$$\begin{aligned} &p(s, a) \leftarrow p(s, a) + \alpha^{(t)}_{\mathbf{S}_{\mathrm{BF}}}\, \mathrm{RPE}_t\, e^A_t(s, a), \quad \forall\, s \in \mathcal{S}, a \in \mathcal{A} \\ &e^A_t(s, a) = \begin{cases} 1 - \pi(s, a), & \text{if } s_t = s, a_t = a \\ \gamma\lambda_A e^A_{t-1}(s, a), & \text{otherwise,} \end{cases} \end{aligned} \tag{7}$$

where $e^A_t(s, a)$ are exponentially decaying eligibility traces of the actor, with eligibility trace decay factor λ_A, and $\alpha^{(t)}_{\mathbf{S}_{\mathrm{BF}}}$ is the time-dependent surprise-modulated learning rate of the actor

$$\alpha^{(t)}_{\mathbf{S}_{\mathrm{BF}}} = \alpha_A \cdot \left(\frac{1}{1 + m\mathbf{S}_{\mathrm{BF}}}\right), \tag{8}$$

dependent on the surprise $\mathbf{S}_{\mathrm{BF}}$ of Eq. 3 (and hence indirectly on the $\mathrm{SPE}_{\mathrm{BF}}$, according to Eq. 5), where m is a saturation parameter $m = \frac{p_j}{1-p_j}$, with $p_j \in (0, 1)$ the probability for a surprise trial to take place (hazard rate), and α_A is a constant learning rate of the actor (same as in the standard Actor-critic).

2.4.4. Hybrid strategies

Finally, we considered hybrid strategies that involve a weighted average of MF and MB computations (Fig. 1A), similar to Daw et al. (2011); Gläscher et al. (2010); Lee et al. (2014). In this category we included the Hybrid Learner-0 (Gläscher et al., 2010) which is a mixture of FWD and SARSA-0, and the Hybrid Learner-λ (Daw et al., 2011) which is a mixture of FWD and SARSA-λ. Moreover, we introduce the following hybrid algorithms: (1) Hybrid-λ-PS-PF that combines MF SARSA-λ and a MB particle filter (PF) with prioritized sweeping (PS) as a form of value updating; (2) the Hybrid Actor-critic that is a mixture of the Actor-critic and the FWD of Gläscher et al. (2010) (i.e. value iteration with a $\mathrm{SPE}_{\mathrm{FW}}$-mediated delta-rule that integrates new observations with a constant learning rate); and (3) the Hybrid Actor-critic PF that is a mixture of the Actor-critic and a FWD with surprise-modulated outlier detection implemented by a particle filter (i.e. value iteration with $\mathbf{S}_{\mathrm{BF}}$-mediated approximate Bayesian inference). Since we did not observe a clear benefit of using PS over value iteration, and for consistency with previous human studies that employed value iteration (Daw et al., 2011; Gläscher et al., 2010), we did not implement hybrid versions of Actor-critic coupled with PS.

Moreover, given our model fitting results (subsection 3.2), we did not further explore variants of MB or other hybrid MB and MF strategies, like successor representation methods (Dayan, 1993). Finally, we also included a random walk with a bias term as a null model.

2.4.5. The common form of policy for all algorithms

For all algorithms we used a softmax action selection policy

$$\pi(s, a) = e^{f(s,a)/\tau} / \sum_b e^{f(s,b)/\tau}, \tag{9}$$

with temperature parameter τ, where $f(s, a)$ corresponds to $Q(s, a)$ for value-based algorithms (e.g. SARSA-λ and FWD) and to $p(s, a)$ for policy-based algorithms (e.g. Actor-critic and REINFORCE).

2.5. Behavioral analysis

2.5.1. Behavioral and performance measures

We analyzed participants' behavior in terms of: average path length to the goal across episodes, percentage of correct actions across state visits and reaction time. For the analysis of the reaction time, we excluded the transitions with reaction times larger than 8 sec, considering them as outliers. However, even with the inclusions of these outliers our results remained significant, with similar p-values.

2.5.2. Parameter fitting and model selection

We fitted the above algorithms to the behavioral data (i.e. actions) of 21 participants, using the Metropolis-Hasting Markov Chain Monte Carlo (MCMC) method (Hastings, 1970), similar to Lehmann et al. (2019) (see subsection A.4 for more details). At the end of each MCMC run, we registered the parameter values that maximize the log-likelihood LL of the data, and we repeated several runs of the MCMC procedure.

In order to perform model comparison, we approximated each model's log-evidence using cross-validation. This method is similar to approaches used in statistics and economics (Berger and Pericchi, 1996; Fong and Holmes, 2020; Rust and Schmittlein, 1985; Wang and Pericchi, 2020). While the models' log-evidence can be approximated Akaikes Information Criterion (AIC) and Bayesian information criterion (BIC), cross-validation is often considered a more robust method for model comparison (Ito and Doya, 2011).

We performed 3-fold cross-validation, and obtained the sum of the out-of-sample LL (i.e. on the test set) of the 3 folds. The MCMC procedure includes some randomness, due to random starting points and random moves in the parameter space. In order to deal with this source of noise and to make more informed conclusions about model selection, we repeated the 3-fold cross-validation procedure 5 times, for each algorithm. We computed, finally, the mean sum of out-of-sample LL and its standard error across the 5 rounds. We report this quantity as an approximation of the log-evidence. The penalty for high complexity comes naturally through cross-validation, and the algorithm with the highest log-evidence is the winning model. For completeness, we also report the maximum LL when fitting the whole dataset and the corresponding BIC in the Appendix subsection A.10. Finally, we performed a random effects analysis (Rigoux et al., 2014; Stephan et al., 2009) on the model log-evidence of each subject, as calculated through the cross-validation procedure.

2.6. Model recovery and posterior predictive checks

For the Surprise Actor-critic (which we later focus on as our winning model of model comparison), we performed (i) posterior predictive checks as well as (ii) a model recovery and (iii) a parameter recovery analysis (Nassar and Frank, 2016; Wilson and Collins, 2019). Our goal was to investigate whether (i) Surprise Actor-critic's choices mirror those of the real participants, (ii) we can identify it as the winning model, and (iii) recover its parameter values given its generated data.

To do so, we simulated 21 agents employing the Surprise Actor-critic strategy (using the same set of parameter we later use in the analysis of the fMRI data), with different random seeds, and repeated this three times – resulting in 3 sets of 21 simulated participants. We then computed the behavioural performance measures (average path length to the goal across episodes and percentage of correct actions across state visits) and performed the model comparison and parameter fitting procedures, exactly as we did for the real participants.

2.7. Model comparison at the neural level

Similar in spirit to earlier work (Mack et al., 2013; Turner et al., 2017; 2013), we used fMRI data in combination with the results of our model comparison in order to further distinguish between computational models. In contrast to earlier work (Turner et al., 2017; 2013), we proceeded sequentially. After fitting the algorithms to behavior (see subsection 2.5.2), we built a general linear model (GLM) (Friston et al., 1994; Penny et al., 2011) for each algorithm that was in the group of winners in our model comparison, using as regressors its main relevant signals (e.g. RPE) – more details on the GLMs and the statistical analysis of the fMRI are provided in the next subsection.

We next computed, for each algorithm and for each voxel in the (whole) brain, the adjusted coefficient of determination, *adjusted* R^2 (Razavi et al., 2003; Soch and Allefeld, 2018), using the MACS toolbox[4] (Soch and Allefeld, 2018). The adjusted R^2 quantifies how well the linear model approximates the data, corrected for the number of regressors included in the model (Soch and Allefeld, 2018). We then obtained the median adjusted R^2 across the whole brain for each algorithm as a more robust indicator of overall performance. We, finally, compared the median adjusted R^2 between algorithms, considering a Bonferroni corrected statistical threshold to control for multiple comparisons.

We also investigated whether the neural data can further constrain the parameters of the Surprise Actor-critic by repeating the procedure described above using parameter sets obtained from independent optimization runs (see subsection A.12).

2.8. fMRI data statistical analysis

After fitting the algorithms to the behavioral data, we computed, for each participant, their trial-by-trial learning signals (e.g. RPE and $\mathrm{SPE}_{\mathrm{BF}}$ or $\mathbf{S}_{\mathrm{BF}}$) and used them as regressors against the fMRI data in a GLM (Friston et al., 1994; Penny et al., 2011). For this step, we used the population parameters, i.e. the ones fitted to the behavioral data of all participants together, as it is usually done and recommended in the analysis of fMRI data (Daw, 2011; Daw et al., 2011). First, in what we will refer to as "GLM_4" we included four regressors in the model: (i) one regressor for the intervals during which a state was on the screen (on-off boxcar events), (ii) the RPE, (iii) the $\mathrm{SPE}_{\mathrm{BF}}$ or $\mathbf{S}_{\mathrm{BF}}$, and (iv) one regressor for participants' actions (zero-duration events). We emphasize that $\mathrm{SPE}_{\mathrm{BF}}$ and $\mathbf{S}_{\mathrm{BF}}$ are a nonlinear function of each other (cf. Eq. 5) and represent the same learning module. We convolved all regressors with the canonical hemodynamic response function (HRF) (Worsley and Friston, 1995). For the case of $\mathbf{S}_{\mathrm{BF}}$ we first z-scored it within each participant, to ensure comparable signals and regressor coefficients across subjects (Rouault et al., 2019). The $\mathrm{SPE}_{\mathrm{BF}}$ (or $\mathbf{S}_{\mathrm{BF}}$) and RPE regressors were placed at the time of the states (as their parametric modulators). We entered all regressors without orthogonalization with respect to each other and let them compete for variance, in order to obtain and test for only what is uniquely explained by each signal (Anggraini et al., 2018; Gläscher et al., 2010; Mumford et al., 2015). To control for remaining motion artifacts, we included in the model the six rigid-body realignment motion parameters.

The estimated regression coefficients for each of the regressors of each participant were then taken to random effects group level analysis (one-sample t-test). For the statistical analysis at the group level, we performed nonparametric permutation testing (10,000 permutations) and controlled for multiple comparisons over the whole brain via the maximum statistic (Nichols and Holmes, 2002). We used cluster-wise correction with a cluster-defining threshold (CDT) of $p = 10^{-4}$ and a family wise error (FWE)-corrected threshold of $p = .05$. Note that this CDT is one order of magnitude lower (and hence stricter) than the one which is considered legitimate and usually used (Eklund et al., 2016). We also performed cluster-wise correction with the most commonly used CDT of $p = 10^{-3}$ and a FWE-corrected threshold of $p = .05$ and report it in the Appendix. We did not focus on a-priori selected brain regions of interest and plotted all statistical brain maps at the same threshold that we used at the inference step (i.e. corrected), without masking out any regions.

As we will see in the Results section, the RPE of the algorithms that explain behavior best is, with the fitted parameter values, highly correlated with the occurrence of the goal state. The goal state has a reward $r = 1$. However, the fitted learning rate of the critic was very small, so that the RPE derived from it was much smaller than 1 for all non-goal states but stayed, for several episodes, close to 1 at the goal. In other words, RPE and reward r are highly correlated, for these fitted parameter values, so that we cannot easily distinguish brain activity correlating with r from that correlating with RPE. However, since at all other states,

[4] https://github.com/JoramSoch/MACS

apart from the goal, the reward is always zero, reward and RPE are not correlated at non-goal states. To separate r from RPE at non-goal states, we thus implemented a second linear model ("GLM_7") where we explicitly included the goal as a separate regressor. We also included two more regressors, to control for all relevant learning signals in the same model. Thus, in GLM_7 we included seven regressors in the model: (i) the states (on-off boxcar events), (ii) the goal state, (iii) the RPE, (iv) the SPE_{BF} (or $\mathbf{S}_{BF}$), (v) the estimated V-values of the landing state, (vi) the participants' actions (zero-duration events), and (vii) the relative policy preferences $p(s, a_{\text{chosen}}) - p(s, a_{\text{not chosen}})$ at the current action (as its parametric modulator). All the analysis steps, inference, and corrections were the same as for GLM_4.

For the first-level statistical analysis of the fMRI images we used the SPM12 software and for the second-level statistical analysis the non-parametric SnPM13.1.05 software[5] The resulting statistical brain maps were plotted using the plotting utilities of the Nilearn software[6] (Abraham et al., 2014) and projected on the MNI152 atlas ICBM 2009c.

Finally, for the comparisons of different RL algorithms at a neural level (subsection 2.7) we performed the same analysis steps as described above. For each algorithm we included as regressors all their pertinent computed quantities: (a) REINFORCE: states, goal, actions, relative policy preferences; (b) Surprise REINFORCE: states, goal, actions, SPE_{BF}, relative policy preferences; (c) Actor-critic: states, goal, actions, RPE, V-values, relative policy preferences; (d) Hybrid Actor-critic: states, goal, actions, RPE, SPE_{FWD}, V-values, relative policy preference; (e) Surprise Actor-critic: states, goal, actions, RPE, SPE_{BF}, V-values, relative policy preference (i.e. GLM_7).

3. Results

3.1. Behavioral results

Participants were able to learn the task during a block of 20 minutes and reached the goal in 3.5 actions on average (minimum possible is 2) already at the 4th episode (Fig. 2A). The mean number of actions that participants took in the first episode, i.e. before they found the goal, is 8 $\pm$ 1.4 (mean $\pm$ standard error). Under a random search policy, the mean number of actions to goal is 6 (for all graphs and starting states; see Appendix subsection A.3) which is not significantly different from the observed one (one sample t-test, $p = .17$). This suggests that in the first episode human participants do not yet have a specific policy to move in this environment.

As an aside we note that in the classical two-stage decision tasks, a terminal state is always reached after two actions even under a random search policy. In that sense, our task structure is significantly more complex than traditional tasks used in fMRI studies.

We introduced surprise trials from the 5th episode onwards, which results in an increased average number of steps that participants took to reach the goal (Fig. 2A). After this point, the episode length gradually decreases again, indicating that participants were able to learn how to act, even in the presence of surprise trials.

The reaction time of participants is significantly higher after a surprising transition than a non-surprising transition (Fig. 2B and C, paired t-test, p=0.02 excluding the non-surprising trials of the first 4 episodes for more fair comparison, and p=0.04 including the first 4 episodes), indicating that participants did notice the unexpected transitions and that they may have developed a mental map of the task. Longer reaction times, which often serve as a behavioral signature of surprise (Huettel et al., 2002; Meyniel et al., 2016; Vassena et al., 2020), presumably reflect the cognitive processes involved in choosing a new action after landing in an unexpected state.

The task structure allows different paths to the goal state and therefore not every participant visited the same states in each episode. For a given state s, we therefore analyzed the percentage of correct action choices (i.e. those leading towards the goal) as a function of the number n of visits of that state (Fig. 2D). We find that participants' speed of learning at each state depends, qualitatively, on (i) the distance of the state from the goal (see Methods) and (ii) the "correctness difference" of the available actions at that state (Fig. 2D).

More specifically, participants reached higher performance levels much earlier for the state that is "0-or-3" actions away from the goal, compared to the "0-or-1" and "1-or-2" states (Fig. 2D, first three panels, respectively). In the state "0-or-3" (Fig. 2D, 1st panel), the 80% performance level (vertical dashed blue line) is reached after only 5 visits, whereas for the state "0-or-1" (Fig. 2D, 2nd panel) it is reached after approximately 14 visits. This is likely due to the fact that, even though the correct action is still only one step away from the goal, the "wrong" action has less negative effects. At the state with distance index "1-or-2" (Fig. 2D, 3rd panel), reaching a performance level of 80% takes 27 visits. On the other hand, for the state "2-or2" (Fig. 2D 4th panel), where any action brings the participant to a state 2 actions away from the goal, we do not observe a clear preference between the two.

In order to get an estimate of confidence intervals of the performance, computed as the number of visits until reaching a 80% level, we performed bootstrapping (Efron and Hastie, 2016). We re-sampled the 21 participants with replacement 200 times, and for each re-sampled selection of participants, we computed the percentage of correct actions in time averaged across all participants (as we did in Fig. 2D) and obtained the number of visits that the 80% performance level was reached. We then computed the mean and the standard deviation across these 200 values (Fig. 2E) (see Appendix subsection A.2 for more details). The height of each bar in Fig. 2E indicates the (bootstrapped) mean number of visits until reaching 80% performance for different states. Collectively, this tendency to choose the correct action when in closer proximity to the goal can be interpreted as a sign that information on reward decays as the distance to the goal increases.

To summarize, participants learn within a few episodes to efficiently move towards the goal (Fig. 2A) and acquire the correct action more rapidly in states close to the goal than in those several action steps away from goal (Fig. 2D and E). Since these findings are qualitatively consistent with a large class of RL algorithms, we now turn to a quantitative comparison of behavioral data with selected algorithms.

3.2. Actor-critic algorithms explain behavior best

The algorithms that are the most likely models of behavior are the Actor-critic, the Surprise REINFORCE, the REINFORCE baseline, the Hybrid Actor-critic, the REINFORCE, the Surprise Actor-critic, and the Hybrid Actor-critic PF, as indicated by the comparison of the negative log-evidence of each model (Fig. 3B). Differences in log-evidence larger than 3 are usually considered significant, and larger than 10 strongly significant (Efron and Hastie, 2016; Held and Ott, 2018; Neath and Cavanaugh, 2012). The Actor-critic is weakly, but not significantly, better than the rest, with a difference (Δ) in log-evidence of 2.29 ± 0.49, 2.46 ± 0.4, and 3.28 ± 0.65 compared to the Surprise REINFORCE, the Hybrid Actor-critic, and the Surprise Actor-critic, respectively (Fig. 3B). Even though the mean difference between Actor-critic and Surprise Actor-critic is just above 3, for some optimization runs this was not the case, as can be inferred by the error bars of the Surprise Actor-critic, which prevents us from declaring them significantly different. The Surprise REINFORCE, the REINFORCE baseline, the Hybrid Actor-critic, and the REINFORCE are essentially indistinguishable from each other ($\Delta < 1$). Furthermore, pooling the log-evidence of each of the best models and each subject to random effects analysis (Rigoux et al., 2014; Stephan et al., 2009) leads to a Bayesian omnibus risk (probability of no differences across model) of 0.98, and the protected exceedance probability (probability that each model is more likely than the rest) of the best algorithm (Actor-

[5] http://nisox.org/Software/SnPM13/.
[6] https://nilearn.github.io/index.html.

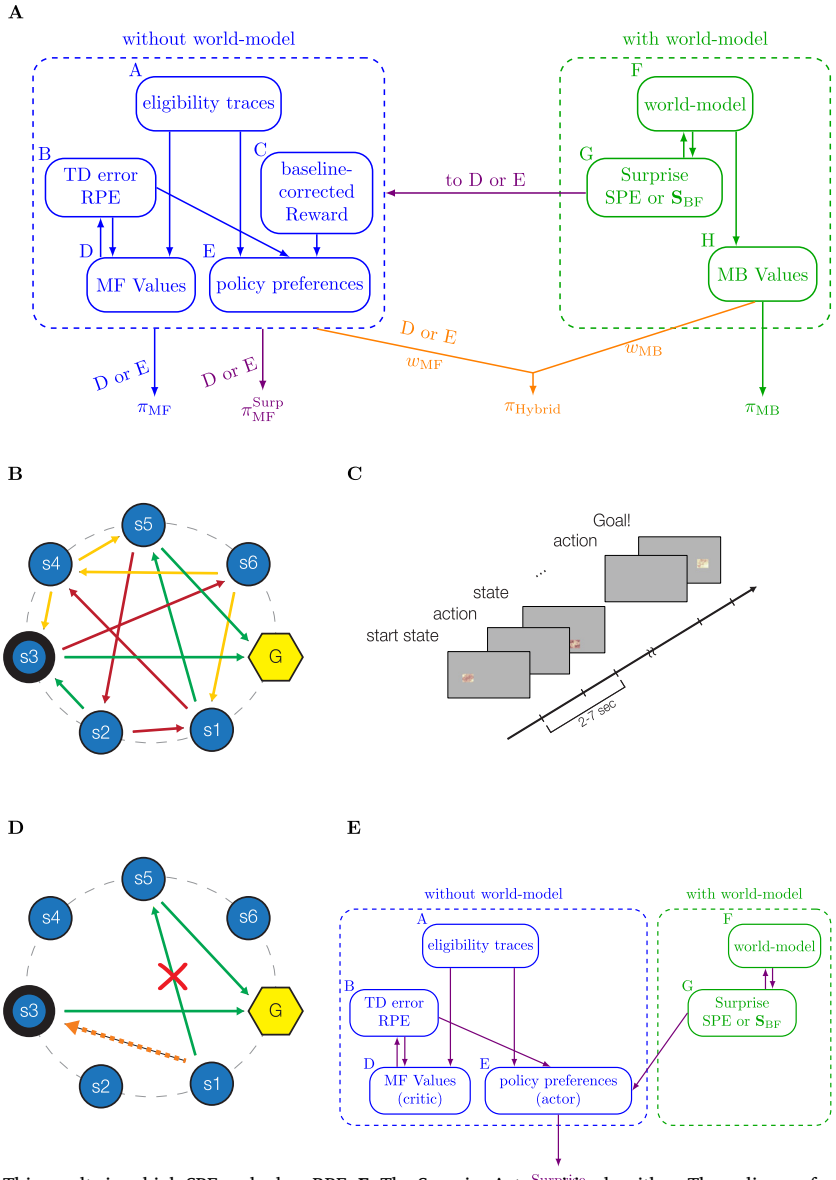


Fig. 1. Multi-step learning task designed to decorrelate learning signals. A. Schematic illustration of the families of algorithms considered. A classic distinction of RL algorithms is based on whether a world-model is (green) or is not (blue) acquired during learning. Powerful algorithms without a world-model implement eligibility traces (box A) as a decaying memory of previous state-action pairs. Learning is driven either by the RPE (box B) used to update V- or Q-values of MF TD learning (box D) or by the potentially baseline-corrected reward (box C) used to update the policy parameters (policy preferences) of policy-gradient RL (box E). In all Actor-critic models of this paper (but not in the REINFORCE models), there is information flow from box B to E since the RPE is used for the policy estimation. MF values (box D) or policy preferences (box E) are then used to compute the MF policy π_{MF}. Algorithms with a world-model (box F) implement estimates of the transitions matrix as a memory of previous state-action-state triplets. Learning can be driven by signals of surprise, such as the SPE or the $\mathbf{S}_{BF}$ (box G). The transition matrix can be used to estimate MB values (box H) and corresponding MB policies π_{MB}. Hybrid algorithms (orange) combine MB values with MF values or policy preferences (with weights π_{MB} and π_{MF}, respectively) to estimate their policy π_{Hybrid}. Finally, in surprise-modulated MF algorithms (magenta), the world-model is not used for the calculation of MB values; surprise signals derived from it modulate the rate of learning MF values (box D) or policy preferences (box E). These surprise-modulated estimates give then rise to the surprise-modulated MF policy π^{Surp}_{MF}. External reward is a potential input to all modules. **B**. The hidden graph of the task. There are seven states and two possible actions at each state. The goal state G is highlighted in yellow. Arrows mark the possible deterministic transitions between states. A "correct" action (green arrows) brings participants closer to the goal and a "wrong" one (red arrows) brings them to a state away from the goal. For example, in $s3$ the correct action leads directly to the goal (0 actions from G), whereas the wrong action leads to state $s6$ which is 3 actions away from the goal (e.g. $s4 \rightarrow s5 \rightarrow$ G). We denote this state as "0-or-3". For some states the two actions are equally "correct" (yellow arrows). **C**. Schematic of the timeline of a full episode. Participants view one state (specific fractal at a specific location) on the screen, choose one action that moves them to the next state, and continue until they eventually reach the goal. **D**. Example of a surprise trial. The expected transition for the action chosen by the participant is $s1 \rightarrow s5$, but the next image (and corresponding location) is the one of $s3$, which has a similar MF V-value to the one of $s5$. This results in a high SPE and a low RPE. **E**. The Surprise Actor-critic algorithm. The policy preferences (actor, box E) are updated using the RPE (box B), computed via the critic (box D), and a surprise signal (box G), derived from the world-model (box F).

critic) was not different than the other algorithms (Fig. 3C, 0.145 for Actor-critic). We interpret these results as evidence for the family of algorithms with policy learning and Actor-critic architecture as the most likely model of behavior, and as not strong evidence for selecting one among them compared to the rest in this family (see Appendix subsection A.9 and subsection A.10 for more analysis and interpretation). We additionally note that in the Actor-critic the resulting fitted learning rate of the critic was smaller than 0.0001, making it therefore very similar to the REINFORCE algorithm (see Appendix subsection A.12 for the fitted parameters of the best algorithms, as well as Table A.2 for the exact values of the log-evidence for each algorithm).

The purely MB algorithms are significantly less appropriate for explaining behavior in this task. Among the algorithms that combine MF and MB modules, the ones that achieve the highest log-evidence are the Hybrid Actor-critic and the Surprise Actor-critic. Overall, our results suggest that behavior is more consistent with model-free learning. However, the Surprise Actor-critic and the Hybrid Actor-critic do employ a model learning procedure, each with a different approach. While the first one learns to ignore the surprise trials, the second one integrates them in its estimate of the transition probabilities. Nevertheless, in terms of behavioral fitting, they are indistinguishable from each other and from a purely model-free Actor-critic.

We emphasize that our main goal in designing our experimental paradigm was to separate MF and MB learning signals in the brain in a multi-step decision task. Surprise trials were designed in a way to bring participants to another state with a value close to what they had expected; this particular feature enables us to dissociate MB learning signals from MF learning signals. A direct consequence of this feature, however, is that the MF learning system may not be affected by surprise trials most of the time. Therefore, it is not unexpected that different MF strategies with or without surprise modulation give rise to the same behavior or have similar power in explaining participants' action-sequences. We will, hence, use the neural data in the next section to further distinguish the winning algorithms.

In summary, the comparison of a large variety of RL algorithms indicates that MF policy gradient algorithms such as REINFORCE or Actor-critic are more suitable to explain the behavioral data than pure TD algorithms such as SARSA-λ or pure MB algorithms, and, thus, favor an estimation of policy parameters rather than Q-values. However, since algorithms that combine MB with MF approaches such as the Surprise

REINFORCE, Surprise Actor-critic, or the Hybrid Actor-critic are not significantly worse than pure MF approaches, we cannot exclude that human participants build a model of the world or assign values to states.

3.3. Model recovery and posterior predictive results

In a model recovery procedure we found for simulated behavioral data generated by the Surprise Actor-critic a group of winning models; the Surprise Actor-critic is not the single best algorithm (Fig. A.8 and Fig. A.9). As described in the previous section, this is because the Surprise Actor-critic and simpler models, such as the pure Actor-critic, can give rise to similar behaviour and our task was not designed to distinguish among them behaviorally. Importantly, the winners of the model comparison on the real data are also the winners in the model recovery approach; the policy gradient/Actor-critic family outperforms the other types of learning, same as in the real data (Fig. A.8).

The Surprise Actor-critic reproduces closely the behavioural signatures of the real participants. The average path length decreases over time and increases with the introduction of surprise trials, and learning occurs faster in states closer to the goal and with higher "correctness difference" (see Fig. 4 for one set of simulated participants and Fig. A.7 for the other two sets).

Also, a parameter recovery procedure showed similar variability as in the real participants' data (see Fig. A.6 and Fig. A.7). Some parameter of the simulated Surprise Actor-critic are recovered consistently (such as the learning rate of the critic), whereas other parameters show high variability. The reasons for this variability are either that one parameter becomes irrelevant due to another well-restricted parameter (for example the low learning rate of the critic renders the eligibility trace decay factor λ_C of the critic irrelevant), or that certain parameters are not well reflected in behavior in our experiment (such as the p_j). For more details see subsection A.12 and subsection A.16 of the Appendix. In the next subsections, we ensure that this variability in the parameter estimation does not affect our results of the fMRI data analysis (see Fig. 5 and subsection A.19).

In summary, our recovery analysis and our posterior predictive checks further support our conclusions that the algorithms in the policy gradient/Actor-critic family are equally likely to have produced the behavior of the real participants and that human participants potentially build a model of the world.

3.4. Actor-critic with a world-model explains neural data best

In our model fitting of the behavioral data we found that the policy gradient/Actor-critic algorithms explain behavior best and are indistinguishable in performance. We next leveraged our fMRI data in order to distinguish between these winning algorithms. The Surprise Actor-critic algorithm provides a significantly better fit for the fMRI data, compared to the REINFORCE, Surprise REINFORCE and the Actor-Critic (Fig. 5A – Wilcoxon signed rank test $p < .001$, i.e. passing a Bonferonni corrected threshold of 0.0125 for the 4 comparisons performed). The performance of the Surprise Actor-critic and of the Hybrid Actor-critic were statistically indistinguishable. For these two winning models, we performed a detailed statistical analysis of the fMRI data (see Methods subsection 2.8 and next section for results).

As a control, we also investigated whether we can select any of the different fitted parameter sets of the Surprise Actor-critic obtained across independent optimization runs (see subsection A.12) based on the fMRI data. We did not, however, observe any significant differences in the goodness of fit across these different parameters (Fig. 5B), consistent with findings in Wilson and Niv (2015). We, thus, repeated our fMRI data analysis (GLM_7) for all parameter sets and showed that the resulting statistical maps are overall robust against possible parameter misspecification (see Appendix subsection A.19).

3.5. Imaging results

The combination of our behavioral fitting, recovery analyses and neural model comparison support the possibility that participants learn a world-model (Fig. 3, Fig. 5, Fig. 4). Moreover, we have evidence in other aspects of behavior that surprise trials were detected by participants, indicating the existence of a MB learning component (Fig. 2A-C). We, thus, hypothesized that a potentially "latent" surprise signal would be present in the brain, performing "latent learning" (Gläscher et al., 2010; Tolman, 1948). Correlated activity with the SPE or $\mathbf{S}_{\mathrm{BF}}$ would indicate that the relevant information is available in the brain, even if it is not currently used in guiding behavior (Daw et al., 2005; Lee et al., 2014; O'Doherty et al., 2020). In order to study the correlation of brain activity with a potentially latent surprise signal, we used the linear model GLM_4 (see Methods). In order to distinguish between contributions of the RPE at non-goal states ($\mathrm{RPE}_{\mathrm{non-goal}}$) from those at the goal r, and to test the effect of other relevant quantities we used the linear model GLM_7 (see Methods). The $\mathrm{RPE}_{\mathrm{non-goal}}$ can, similarly, be interpreted as a potentially latent learning signal for use within the Actor-critic architecture whereas the r at goal states can be interpreted as a potentially latent learning signal in REINFORCE algorithms.

We tested the two winning algorithms of our combined model comparison, that include all three learning components, i.e. surprise, $\mathrm{RPE}_{\mathrm{non-goal}}$ and r, in our analysis of the brain data. The first one is the Hybrid Actor-critic which derives its policy by combining the policy preferences of the actor (Fig. 1A, box E) and the MB Q-values (Fig. 1A, box H), calculated via the $\mathrm{SPE}_{\mathrm{FW}}$ by the FWD (Fig. 1A, box G), i.e by integration of new observations with a constant learning rate. The second one is the Surprise Actor-critic which uses all modules of Fig. 1E with a world-model calculated via the $\mathbf{S}_{\mathrm{BF}}$ (or $\mathrm{SPE}_{\mathrm{BF}}$, cf. Eq. 5) by an approximate Bayesian approach (particle filter) implementing outlier detection (see Methods for more details). Interestingly, the results for both algorithms were very similar, indicating the calculation of the relevant learning signals independent of how they are used for action selection. We therefore focus on the Surprise Actor-critic, in this section and include the corresponding results for the Hybrid Actor-critic in the Appendix (subsection A.20).

3.5.1. Neural signatures of model-based surprise signals

For the MB $\mathrm{SPE}_{\mathrm{BF}}$, we found with the GLM_4 approach correlated activity in the supplementary motor area (SMA) (cluster FWE-corrected $p = .001$, cluster size $k = 104$, peak voxel: $6, 26, 47$; $T = 5.64$), the insula (right $p = .0022$, $k = 72$, peak: $36, 20, -1$; $T = 6.33$, left $p = .0262$, $k = 10$, peak: $-36, 20, -1$; $T = 6.33$), the middle frontal gyrus (right $p = .0009$, $k = 114$, peak: $51, 44, -4$; $T = 6.26$, left $p = .0027$, $k = 60$, peak: $-42, 11, 35$; $T = 6.26$), the angular gyrus, the supramarginal gyrus, the superior parietal lobule (right $p = .0003$, $k = 203$, peak: $57, -40, 41$; $T = 6.91$, left $p = .0050$, $k = 39$, peak: $-39, -46, 38$; $T = 5.67$) and the superior frontal gyrus (right $p = .0017$, $k = 79$, peak: $18, 23, 56$; $T = 5.33$) (Fig. 6A, CDT $p = 10^{-4}$, FWE-corrected at $p = .05$, whole brain, 21 subjects, random effects analysis). While some of these regions, namely the regions located around the intraparietal sulcus (angular gyrus, supramarginal gyrus and superior parietal lobule) and regions in prefrontal cortex were also reported in the two-step task of Gläscher et al. (2010), we found prefrontal activation also in more middle and superior locations, as well as in SMA and insula. All aforementioned regions exhibit overlap with components of the "salience network" (Seeley et al., 2007), previously associated with the detection of salient or novel stimuli and with error monitoring in order to guide actions. Interestingly, we do not find correlates of the $\mathrm{SPE}_{\mathrm{BF}}$ in subcortical structures (e.g. striatum). A full list of exact coordinates and p-values of all locations showing significant correlation are provided in subsection A.21.

In the GLM_7 approach, we found that inclusion of the goal regressor led to some of the regions in Fig. 6A to be sufficiently explained by the $\mathrm{RPE}_{\mathrm{non-goal}}$. The regions that are – in the GLM_7 approach – robustly and uniquely explained by the $\mathrm{SPE}_{\mathrm{BF}}$ are the right supramarginal gyrus and

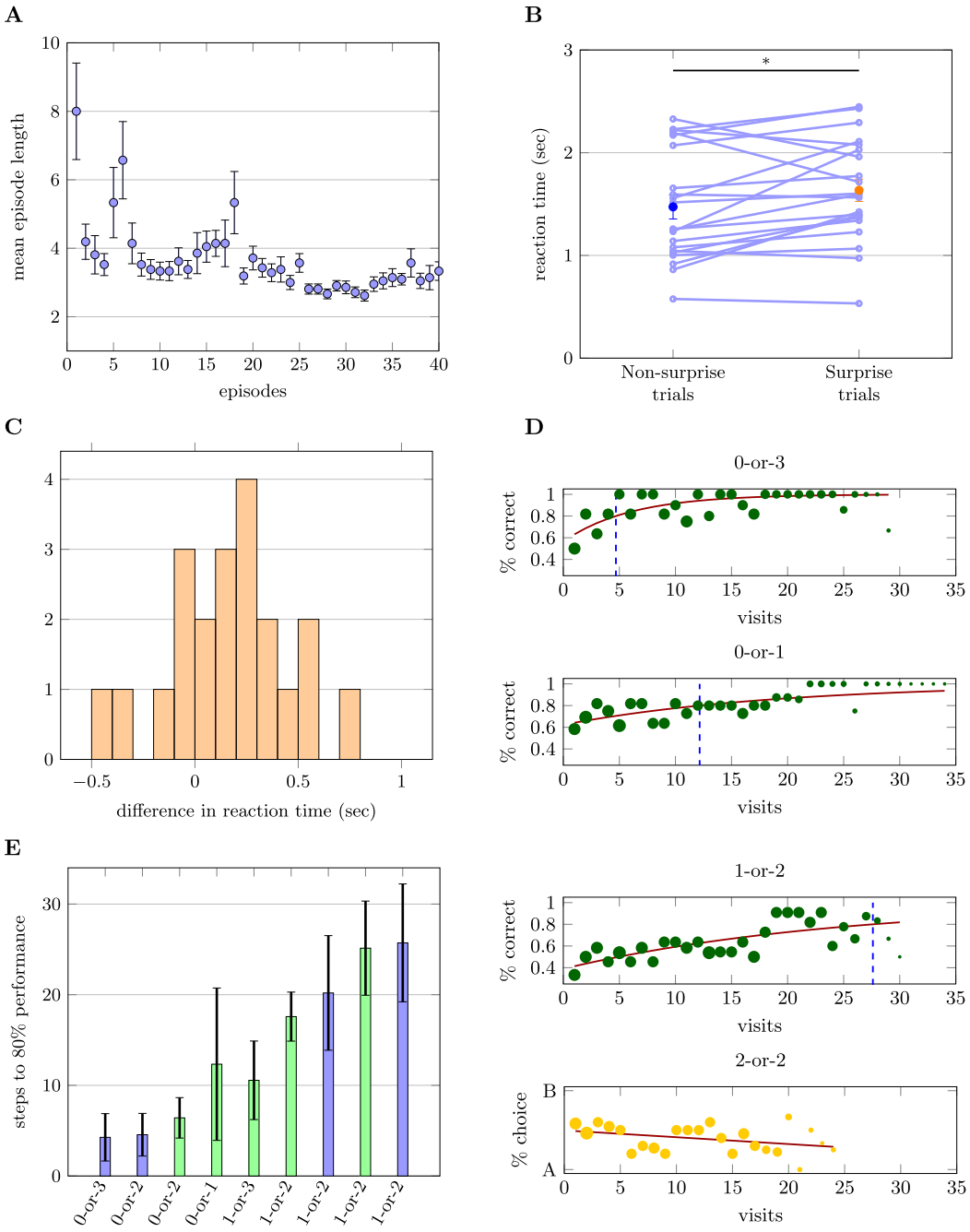


Fig. 2. Behavioural results. A. Mean length of each episode. The circles represent the number of actions per episode (from start to goal state) averaged across participants. The error bars mark the standard error of the mean. Already at the 2nd episode, participants reach the goal within about 4 actions (minimum possible is 2). From the 5th episode onwards, we introduce surprise trials and the average episode length increases at this point. **B**. For each participant, the reaction time is averaged across all non-surprise trials and all surprise trials (open connected circles). The mean of these 21 values is depicted as blue and orange filled circles, for non-surprise and surprise trials respectively (the error bars mark the standard error of the mean). The reaction time on surprise trials is significantly higher than on non-surprise trials (paired t-test, p=0.02). **C**. Histogram of the difference between mean reaction time for surprise versus non-surprise trials for all participants. The distribution takes mostly positive values, with a maximum difference of 0.8 sec. **D**. Percentage of selecting the "correct" action at states whose distance from goal is "0-or-3", "0-or-1", "1-or-2", and "2-or-2" actions, respectively, as a function of the number n of state visits. The vertical position of a filled green circle indicates the fraction of participants that selected the "correct" action, while the circle size represents the number of participants that visited this state n times. Only a few participants (small circles) have visited a state more than 20 times. The average learning curve (red line) is obtained by fitting a weighted (by number of participants) function approaching exponentially towards a baseline ($1 - a \cdot e^{-bn}$). The vertical dashed blue line indicates the time when the red learning curve reaches the 80% performance level. These graphs provide qualitative evidence that participants learn to choose the "correct" action faster for states that are closer to the goal and for which the "wrong" action has more negative consequences. **E**. Performance summary plot across all states. The height of each bar corresponds to the (bootstrapped) mean time that performance reaches 80% for the states in **D** and all other states (except the ones where both actions are equidistant from goal). The error bars correspond to the standard deviation, calculated via bootstrapping, which is an estimate of the standard error of the mean performance of participants. Different colors signify the two different task graphs employed randomly across participants (see Table A.1).

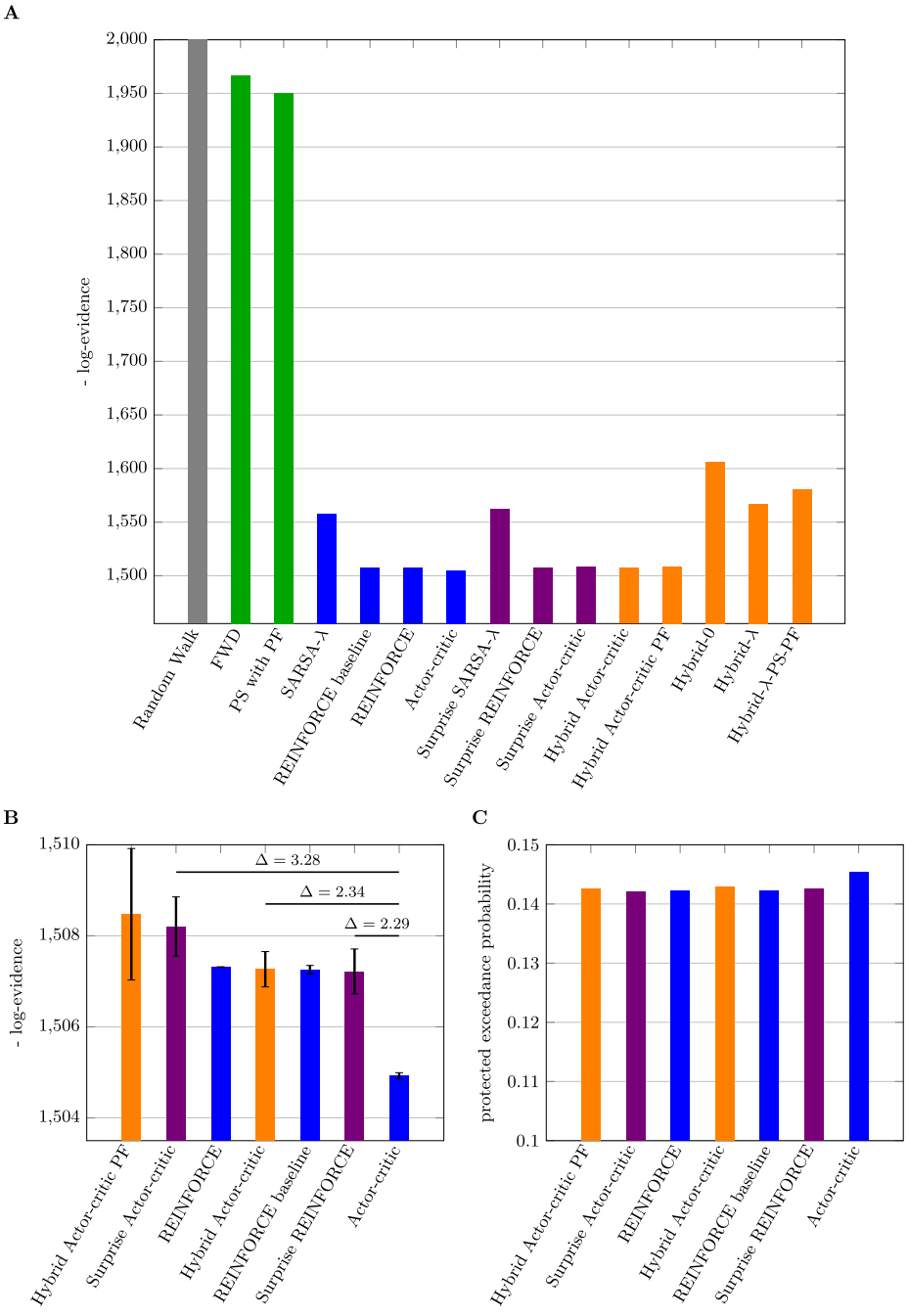


Fig. 3. Algorithm fit to behaviour. A. Negative model log-evidence for all algorithms. Smaller values indicate better performance. The color coding of the algorithms is the same as in Fig. 1A. The most likely models for behavior are the Actor-critic, the Surprise REINFORCE, the REINFORCE baseline, the Hybrid Actor-critic, the REINFORCE, the Surprise Actor-critic and the Hybrid Actor-critic PF. **B**. Negative model log-evidence for only the best algorithms, sorted with increasing performance. The error bars (not marked in A because they were not visible) indicate the standard error of the mean across 5 runs of a 3-fold cross-validation procedure. The log-evidence differences Δ between the Actor-critic and some of the rest closest winning algorithms are noted on the graph. Δ larger than 3 is usually considered significant, and larger than 10 strongly significant (Efron and Hastie, 2016; Held and Ott, 2018; Neath and Cavanaugh, 2012). We thus consider the Actor-critic only weakly better than the other winning algorithms. The Surprise REINFORCE, the REINFORCE baseline, the Hybrid Actor-critic and the REINFORCE are essentially indistinguishable from each other (Δ < 1). **C**. Protected exceedance probabilities after subjecting the model log-evidence of the best algorithms to a random effects analysis (Rigoux et al., 2014; Stephan et al., 2009). None of the algorithms is more likely than the rest (Bayesian omnibus risk = 0.98 (Rigoux et al., 2014)). *Abbreviations:* PS: Prioritized Sweeping, PF: Particle Filtering.

the right angular gyrus ($p = .0005$, $k = 114$, peak: $54, -37, 41$; $T = 6.84$), the right insula ($p = .0054$, $k = 29$, peak: $36, 20, -1$; $T = 6.49$), and the right middle frontal gyrus ($p = .0053$, $k = 30$, peak: $39, 59, 5$; $T = 5.75$, and $p = .0126$, $k = 15$, peak: $36, 14, 53$; $T = 5.03$) (Fig. 6A, CDT $p = 10^{-4}$, FWE-corrected at $p = .05$). The SMA correlates still significantly with the SPE_{BF}, but at the slightly higher CDT of $p = 10^{-3}$ (see Fig. A.12A, CDT $p = 10^{-3}$, FWE-corrected at $p = .05$).

Implementing GLM_4 and GLM_7 with $\mathbf{S}_{\text{BF}}$, instead of the SPE_{BF}, led to significant correlations of a smaller spatial extent in a subset of the regions we found for the SPE_{BF}, namely in the right middle frontal gyrus ($p = .02$, $k = 15$, peak: $42, 62, 5$; $T = 5.45$) and the right insula ($p = .0054$, $k = 42$, peak: $36, 20, 2$; $T = 6.73$) (GLM_4: Fig. 6C, GLM_7: Fig. 7C). As we saw in subsection 2.8, there is a deterministic non-linear relationship between the SPE_{BF} and the $\mathbf{S}_{\text{BF}}$ and no a-priori reason to choose one or the other for correlation with brain activity (see Eq. 5 and Supplementary Fig. A.5 for more details). At least in a linear model, however, the SPE_{BF} seems to lead to more regions of significant correlation with brain activity. We speculate that the SPE is a crucial signal for the brain, conveying a "state mismatch" (see Discussion).

The statistical maps of the SPE_{BF} using different fitted parameters for the Surprise Actor-critic are similar and thus robust against possible parameter misspecification (see Appendix Fig. A.14). For some parameter sets we observed bilateral activation in the supramarginal gyrus, the angular gyrus and the middle frontal gyrus (similar to what we found in GLM_4), but overall the activation on the right hemisphere is more robust and present in all repetitions of the analysis.

3.5.2. *Neural signatures of model-free signals*

First, with the GLM_4 approach, we found significant correlation of the RPE in the inferior frontal and orbitofrontal gyrus (right: $p = .0003$, $k = 201$, peak: $48, 38, -1$; $T = 7.48$, left: $p = .0047$, $k = 20$, peak:

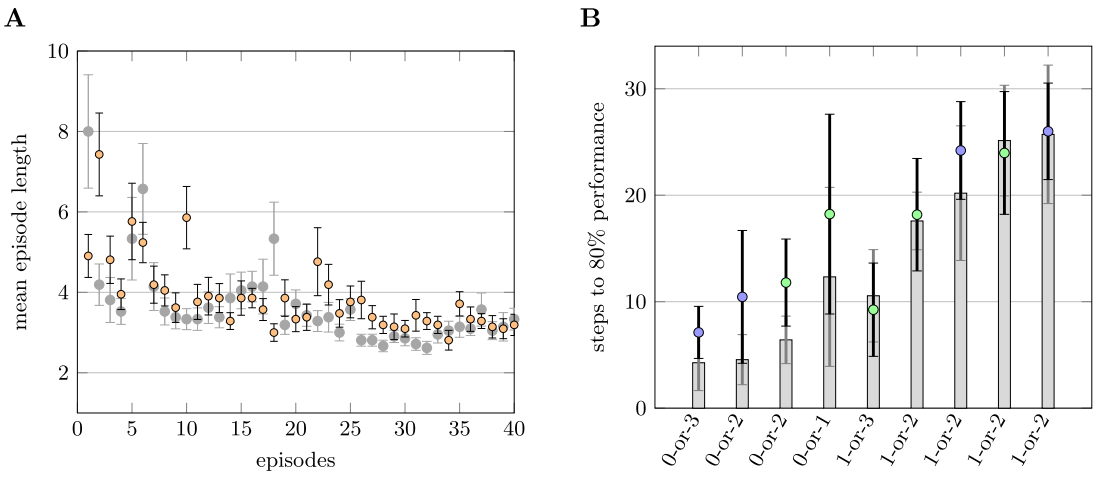


Fig. 4. Posterior predictive checks. A. Mean length of each episode for 21 simulated Surprise Actor-critic participants (orange circles) and for the 21 real participants (grey circles – same data as in Fig. 2A but re-plotted here for convenience). As in Fig. 2, the circles represent the number of actions per episode averaged across simulated participants and the error bars mark the standard error of the mean. The Surprise Actor-critic's choices mirror closely those of the real participants. **B.** Performance summary plot across all states, for 21 simulated Surprise Actor-critic simulated participants (colored circles) and for the 21 real participants (grey bars – same data as in Fig. 2E). See caption of Fig. 2 for more details. The performance across states of the simulated participants matches the one of the real participants.

$-24, 38, -7$; $T = 5.63$), the ventromedial prefrontal cortex (vmPFC) ($p = .0132$, cluster size $k = 8$, peak voxel: $6, 44, -13$; $T = 5.04$), the bilateral putamen and pallidum (right: $p = .0034$, $k = 26$, peak: $27, -13, -1$; $T = 5.61$, left: $p = .0205$, $k = 5$, peak: $-27, -7, -7$; $T = 5.47$), the inferior occipital gyrus (right: $p = .0013$, $k = 58$, peak: $54, -70, 2$; $T = 6.16$, left: $p = .0153$, $k = 7$, peak: $-48, -85, -7$; $T = 5.10$), the anterior cingulate gyrus ($p = .0261$, $k = 4$, peak: $-3, 29, -13$; $T = 5.07$), as well as in the temporal gyrus and the fusiform area (not shown, right $p = .0003$, $k = 268$, peak: $33, -82, -28$; $T = 8.77$) (Fig. 6B, CDT $p = 10^{-4}$, FWE-corrected at $p = .05$, whole brain, 21 subjects, random effects analysis). For a full list of locations with significant activity see subsection A.21. As mentioned earlier, the fitted learning rate of the critic was very low, meaning that there was a very small update of the critic's V-values on a trial-per-trial basis. Thus, the RPE takes most of the time, in non-goal states, very small values compared to those at the goal state. Hence, the neural correlates we found largely include regions that have been associated with reward delivery and values, e.g. the vmPFC and orbitofrontal cortex (OFC) (Behrens et al., 2008; Chase et al., 2015; Hare et al., 2008; Stalnaker et al., 2018; Wunderlich et al., 2012a), and to a lesser degree subcortical regions usually reported in the literature for RPE (e.g. striatum).

In the GLM_7 approach, where we include the goal regressor in order to dissociate the $\text{RPE}_{\text{non-goal}}$ component, we found significant correlation in the left putamen and pallidum ($p = .0034$, $k = 53$, peak: $-21, 8, 11$; $T = 5.37$), the superior parietal lobule ($p = .0056$, $k = 37$, peak: $-21, -61, 44$; $T = 6.11$), the SMA ($p = .0097$, $k = 26$, peak: $-6, 8, 53$; $T = 5.49$), and the right precuneus ($p = .0108$, $k = 24$, peak: $24, -55, 41$; $T = 5.34$) (Fig. 7B, CDT $p = 10^{-4}$, FWE-corrected at $p = .05$, whole brain, 21 subjects, random effects analysis). The putamen and pallidum were bilaterally active in the most commonly used CDT $p = 10^{-3}$ (Fig. A.13B). The striatal activity we found is slightly more dorsal than ventral, presumably because the RPE in our algorithm is also used to update the policy parameters; the dorsal striatum is implicated in action execution and stimulus-response learning (Balleine, 2005; Miller and Venditto, 2020), and has been hypothesized to implement the actor in an Actor-critic architecture (Joel et al., 2002; Takahashi et al., 2008). Due to the high correlation between the RPE and the reward in GLM_7, there are no areas uniquely correlated with the reward, and the reward-related regions we found in GLM_4 (Fig. 6B) are not attributed to any of the regressors in GLM_7. As a control, we also performed a version of GLM_7, where we orthogonalized (Mumford et al., 2015) the RPE with respect to the reward. The brain regions correlating with the reward in this case are, as expected, similar to the ones found for the RPE in GLM_4 (Fig. 6B), with activation of an even smaller extent in the putamen and pallidum in this case. The statistical maps of the RPE are, as for the SPE, similar when using different fitted parameters for the Surprise Actor-critic (see Appendix Fig. A.15). For some parameter sets we found additional correlated activity in the posterior cingulate gyrus.

Additionally, we found correlates of the relative policy preferences $p(s, a_{\text{chosen}}) - p(s, a_{\text{not chosen}})$ in the middle cingulate gyrus ($p = .0004$, $k = 270$, left peak: $-6, -4, 44$; $T = 7.07$, right peak: $9, -7, 44$; $T = 6.80$), in the putamen (left: $p = .0015$, $k = 91$, peak: $-30, -4, 5$; $T = 6.65$; right: $p = .0025$, $k = 71$, peak: $27, 5, 8$; $T = 6.66$), the fusiform gyrus and cerebellum exterior (left: $p = .0022$, $k = 76$, peak: $-36, -49, -22$; $T = 6.40$, right: $p = .004$, $k = 47$, peak: $27, -46, -22$; $T = 5.36$), and the middle temporal gyrus (left: $p = .0070$, $k = 29$, peak: $-54, -7, -28$; $T = 6.12$) (Fig. 7E) providing further support for the policy learning part of our algorithm, and consistent with the idea of the dorsal striatum implementing an actor (Joel et al., 2002; Takahashi et al., 2008).

In summary, our analysis of the fMRI data suggests that learning signals for three learning modules, i.e. signals of surprise for learning the world-model, as well as RPE, reward and policy preferences for both TD value estimation and policy updating, are all available in the brain.

4. Discussion

We have introduced a novel multi-step decision making task that allows the disentanglement of MF and MB learning signals in human BOLD responses. In our analysis, we considered various existing and novel RL algorithms. In particular, we developed a normative surprise-based particle filtering algorithm for model learning in our experiment, showed that it automatically leads to an outlier detection approach, and combined it with a MF Actor-critic to explain human behavior and brain activity.

We have found that human behavior is best explained by the Actor-critic and policy gradient framework. Contributions from the MB learning system are not readily detectable in terms of model fitting based solely on behavior, but we did find representations of MB learning signals in neural responses and in different aspects of behavior. We found signatures of RPE in the striatum, the SMA and in parietal regions whereas signals of surprise were correlated with activity in the middle frontal gyrus, the insula and the intraparietal sulcus. Our results confirm and extend previous fMRI results to a multi-step scenario and support the existence of parallel learning modules in the brain, impor-

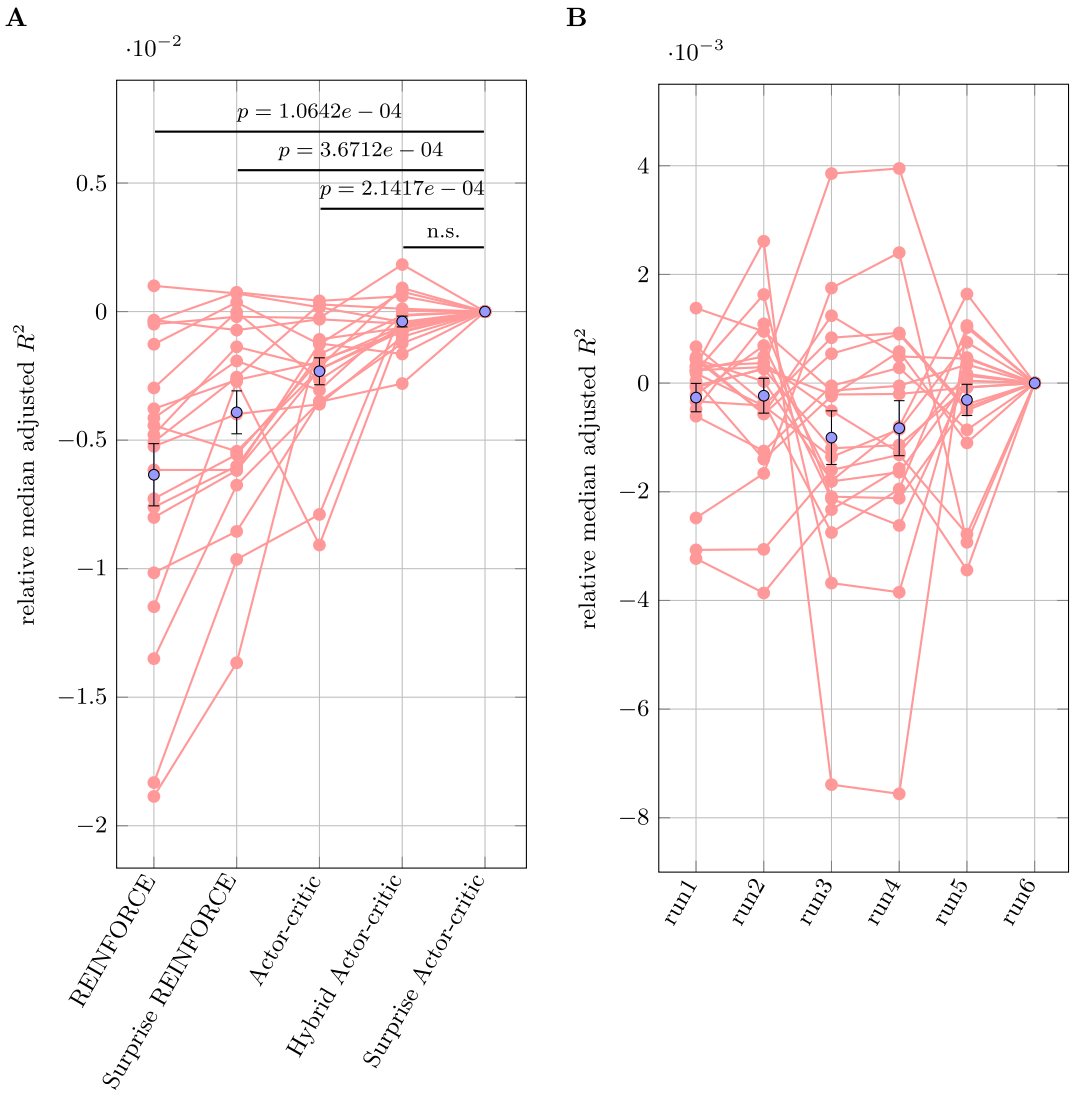


Fig. 5. Neural model comparison. A. Difference in median adjusted R^2 across the whole brain for the winning computational models and the Surprise Actor-critic. Each red line corresponds to a participant and is centered with respect to the Surprise Actor-critic. The median adjusted R^2 of the Surprise Actor-critic is significantly larger from the one of the REINFORCE, Surprise REINFORCE and the Actor-Critic (Wilcoxon signed rank test $p < .001$, i.e. passing a Bonferonni corrected threshold of 0.0125 for the 4 comparisons performed). The performance of the Surprise Actor-critic and of the Hybrid Actor-critic were not significantly different. **B**. Median adjusted R^2 across the whole brain for the Surprise Actor-critic with different sets of parameters (of independent optimization runs), centered with respect to run6 (corresponding to the randomly chosen parameter set used in the analysis of the fMRI data). The goodness of fit is not significantly different across the different parameter sets.

tantly policy learning and surprise signaling. Finally, our work adds to the collection of learning tasks towards gaining a better understanding of the various aspects of human learning. Thus, in face of replicability concerns often raised in neuroscience and psychology, our work helps confirm previous findings on brain structures involved in MF and MB learning and helps increase confidence that previous findings are likely not specific to task designs.

4.1. A multi-step decision making task with surprising transitions

Our experiment allows the detection of MF and MB brain signatures in a multi-step scenario, whereas the majority of human learning studies employ two-stage tasks, with few exceptions (see Simon and Daw (2011) for a task without signal de-correlation and Balaguer et al. (2016) for a study with focus on hierarchical planning). As a starting point for the task design, we considered the algorithms SARSA-λ (Sutton and Barto, 1998), Forward Learner and their hybrid combination, which have been shown to explain human behavior in numerous studies (Daw et al., 2011; Doll et al., 2015a; Economides et al., 2015; Gläscher et al., 2010; Lee et al., 2014; Otto et al., 2013a). However, the idea we follow for the decorrelation of MF and MB prediction errors could also be applied if we replaced SARSA−λ by Q−λ or other model-free value-based algorithms (but see subsection A.5 in the Appendix for a situation that can reduce the efficiency). We emphasize that our task design does not aim to distinguish algorithms at the level of behavior, but – given the view that humans implement multiple learning modules – to de-correlate prediction errors at the level of brain signals. Moreover, our task design does not seek to dissociate different possible MB signals from each other (i.e. SPE from other surprise signals), but MF signals from model learning ones.

We have shown that Bayesian inference on the generative model of our task leads to a surprise signal that inhibits learning, rather than accelerating it (see subsection A.6 for details and derivations). The concept of surprise having different effect on learning depending on the statistical context has been previously proposed and developed for tasks involving tracking of targets in Gaussian settings (d'Acremont and Bossaerts,

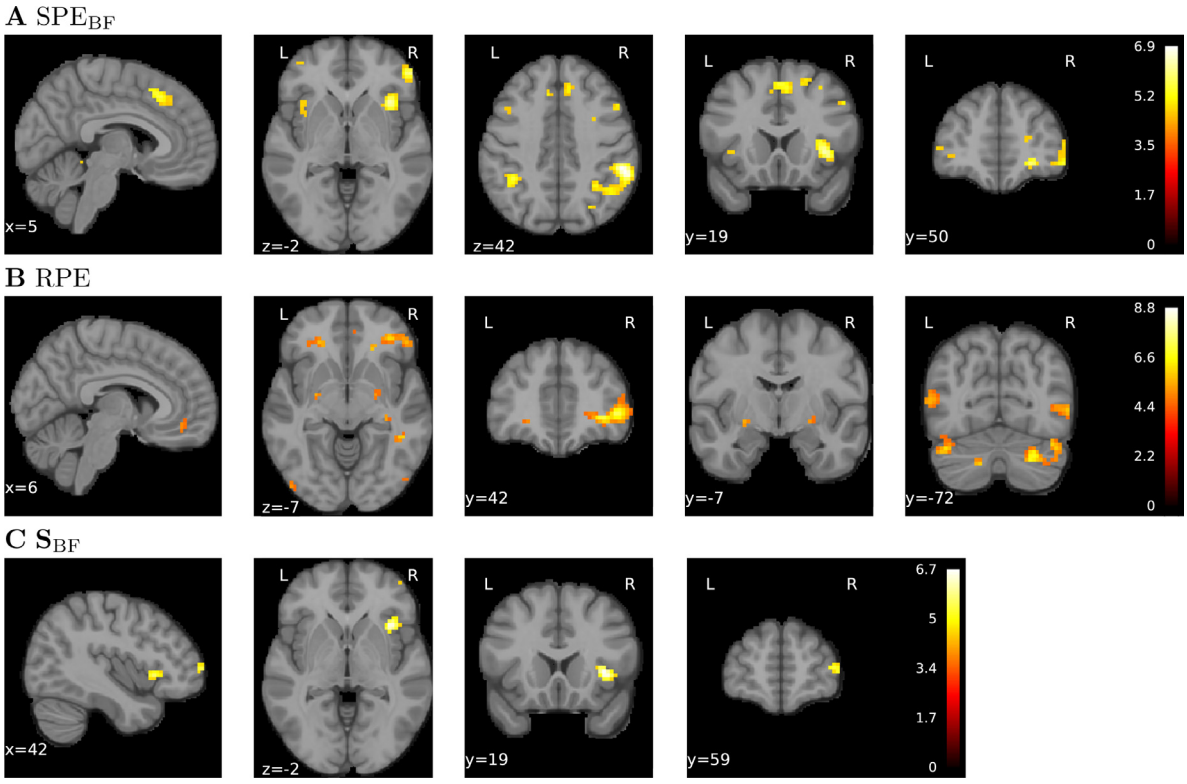


Fig. 6. Neural correlates of learning signals of the Surprise Actor-critic – GLM$_4$. T-statistic maps (21 subjects, random effects whole brain analysis, cluster-wise correction with a cluster-defining threshold (CDT) of $p = 10^{-4}$ and a FWE-corrected threshold of $p = .05$, nonparametric permutation test with maximum statistic approach) of **A**. SPE$_{BF}$. We find significant correlation in SMA, insula, middle frontal gyrus, angular gyrus, supramarginal gyrus and in the superior frontal gyrus. **B**. RPE. We find significant correlation in the inferior frontal and orbitofrontal gyrus, the striatum (putamen and pallidum), the vmPFC, and the inferior occipital gyrus. **C**. **S**$_{BF}$. We find significant correlation in the right insula and the right middle frontal gyrus. For the statistical maps with the more commonly used CDT of $p = 10^{-3}$ see Fig. A.12.

2016; Nassar et al., 2019). Here, we started from a general generative model describing the occurrence of outliers and developed an approximate Bayesian algorithm for a general case. Previous work has focused on differentiating behavioral and brain responses for the case that learning should increase (in a change-point setting) versus when learning should decrease (in a outlier occurrence setting) (d'Acremont and Bossaerts, 2016; Nassar et al., 2019). In our work, we focused on dissociating signals related to reward from those related to model learning and the use of outliers served as a handle towards this goal.

We see connections between our task and tasks developed recently for studying the role of dopamine in learning (Langdon et al., 2018). For example, Kim et al. (2020) employed a virtual navigation task where mice were teleported to different tracks with same distance to goal. They found that the ramping activity of dopamine neurons codes for an RPE and not sensory surprise (Kim et al., 2020; Mikhael et al., 2019). In another study, Takahashi et al. (2017) administered reward of the same value but different identity (flavor) to rats. This increased the firing rate of some dopamine neurons, suggesting that they respond to errors in reward identity and not only to reward quantity and that dopamine may relay a multi-dimensional prediction error (Stalnaker et al., 2019; Takahashi et al., 2017). Similarly, in an fMRI study (Howard and Kahnt, 2018), the identity of an unexpected odor with same pleasantness could be decoded from midbrain BOLD signals (Howard and Kahnt, 2018; Stalnaker et al., 2019). The focus, the tasks or the nature of the data of the above studies differ from ours, but the common line is the introduction of a (sensory) change, while keeping the value similar. Here, we did not find striatal activation uniquely explained by surprise (SPE or **S**$_{BF}$). However, dopaminergic neurons are known to project to many other regions in the brain apart from the striatum, such as the prefrontal cortex, thus it is hard to tell from our data if dopamine is or is not involved in surprise signaling.

4.2. Behavior is best explained by model-free policy learning

Participants learned fast and all the models that were the most likely descriptions of behavior used eligibility traces, consistent with findings in Lehmann et al. (2019). The winning algorithms come from the family of policy gradient methods, with contributions of a RPE (derived from a critic in the Actor-Critic architecture) and of a surprise signal (derived from the SPE or **S**$_{BF}$ of MB approaches). Policy gradient learning has received less attention in human studies (Ito and Doya, 2011; Li and Daw, 2011; O'Doherty et al., 2004) than the classic value-based approaches (Daw et al., 2011; Gläscher et al., 2010). Since recent studies indicate that the activity of midbrain dopamine neurons seems to be closely related to the initiation of actions and that policy learning is a likely framework to reconcile these observations (Coddington and Dudman, 2019), policy gradient methods are an important topic for further behavioral and fMRI studies.

Policy gradient methods are considered more flexible and can more readily allow behavior to shift to determinism (for a constant policy temperature τ) (Sutton and Barto, 2018), which in our task can be advantageous. Our task has two actions in each of the non-goal states. We show empirically (see subsection A.9) that policy gradient methods allow the preference of one action over the other to grow arbitrarily high. Value-based MF learning, such as SARSA-λ, combined with a softmax policy – typically assumed in human studies – cannot achieve this. Therefore, policy gradient methods capture gradual changes in the exploration strategy of participants in our task that standard value-based learning cannot capture.

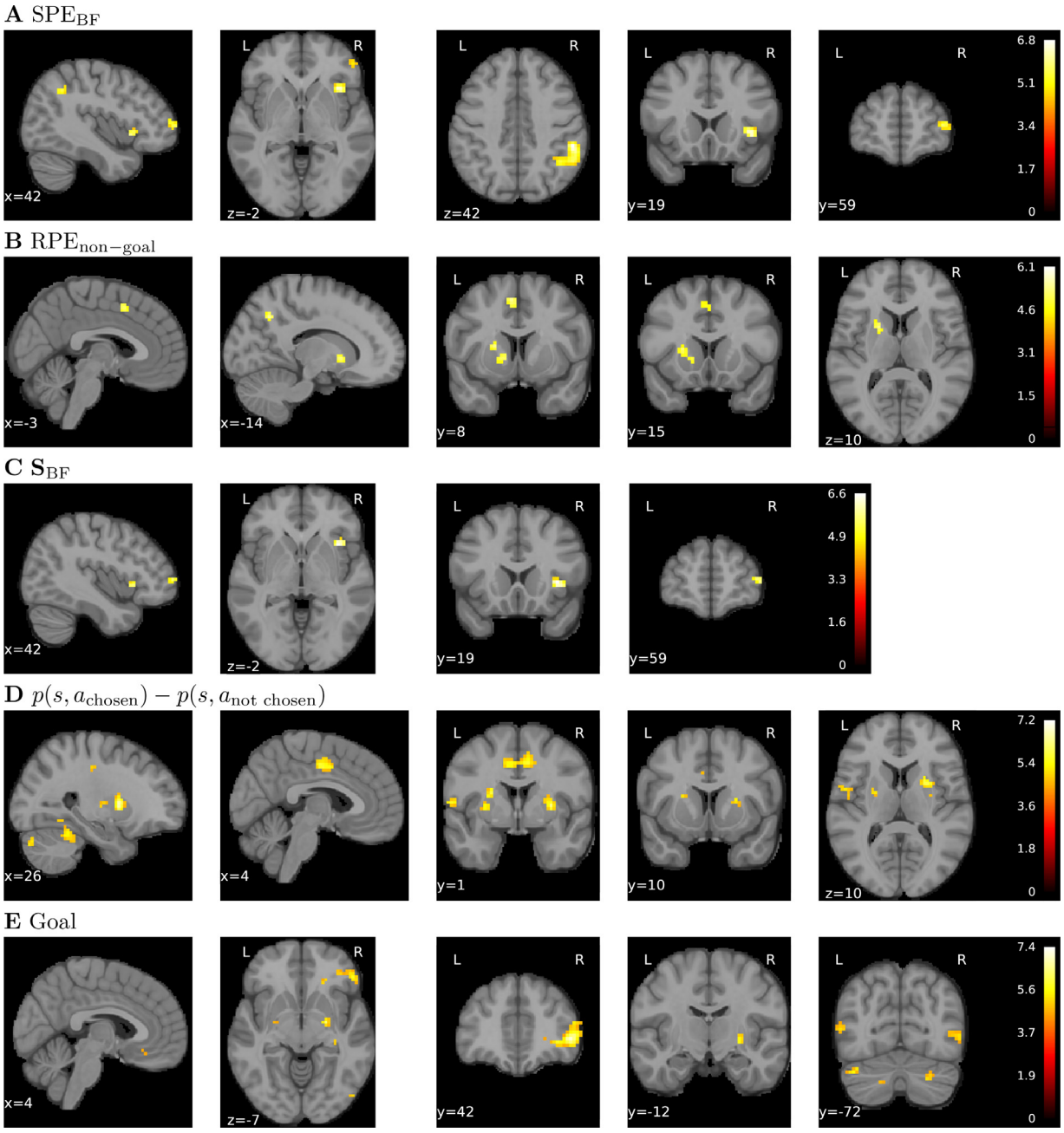


Fig. 7. Neural correlates of learning signals of the Surprise Actor-critic – GLM$_7$. T-statistic maps (21 subjects, random effects whole brain analysis, cluster-wise correction with CDT of $p = 10^{-4}$ FWE-corrected at $p = .05$, nonparametric permutation test with maximum statistic approach) of **A**. $\mathrm{SPE}_{\mathrm{BF}}$ (right insula, right middle frontal gyrus, right angular gyrus, right supramarginal gyrus). The SMA in the case is significant at a CDT of $p = 10^{-3}$ (FWE $p = .05$, Fig. A.13A). **B**. RPE ($\mathrm{RPE}_{\mathrm{non-goal}}$) (striatum – putamen and pallidum –, superior parietal lobule, SMA and right precuneus). The striatal activation was bilateral at a CDT of $p = 10^{-3}$ (FWE $p = .05$, Fig. A.13A). **C**. $\mathbf{S}_{\mathrm{BF}}$ (right insula, right middle frontal gyrus). **D**. Relative policy preferences $p(s, a_{\text{chosen}}) - p(s, a_{\text{not chosen}})$ (middle cingulate gyrus, putamen, fusiform gyrus, middle temporal gyrus). **E**. Reward (goal occurrences). For the statistical maps with the more commonly used CDT of $p = 10^{-3}$ see Fig. A.13.

The fitted learning rate of the Actor-critic algorithms in our model selection was very low, making their MF updates very similar to the ones of the REINFORCE algorithm. The latter relies on gradient ascent to optimize preceding actions, instead of online computations (as, for example, the Actor-critic algorithm with a substantial learning rate). So far there is limited evidence in human RL studies of this type of learning strategy (Li and Daw, 2011). However, it is worth mentioning that all MF algorithms, such as Q-learning, Actor-critic and REINFORCE, can be written as neoHebbian three-factor learning rules (Frémaux and Gerstner, 2016) given a suitable representation of the state-space: two factors are the activity of pre- and postsynaptic neurons (as in standard Hebbian learning) whereas the third factor is a neuromodulator such as dopamine. An eligibility trace with an appropriate time-scale that has already "tagged" the preceding actions gives room for a gradient ascent update to occur when the reward is obtained (Gerstner et al., 2018). The exact implementation of this type of updating in the brain remains unclear and echoes the active research on the neuronal implementation of back-propagation in the brain (Lillicrap et al., 2020; Illing et al., 2021).

As an aside, we note that our behavioral analysis using cross-validation may carry a bias in favor of the assumption that all participants used the same learning algorithm in this task, possibly with different parameter values (e.g. learning rate), whereas learning pro-

cesses may differ across participants (Piray et al., 2019; Rigoux et al., 2014; Stephan et al., 2009). However, our posterior predictive checks for artificial data generated by a Surprise Actor-critic with a *single* set of parameters exhibited the same behavioral patterns and the same variability as in the real data (Fig. 4). This indicates that if a single participant performs 21 different instantiations of the experiment, we would observe the same variability and same patterns of action choices as in our 21 real participants. Moreover, our model and parameter recovery results on the same artificial data were also the same as for the real data; the group of winning models and the variability in the fitted parameters were same (subsection 3.3). Therefore, any observed variability in the results on the 21 real participants does not have to be attributed to different participants using different strategies or different parameters. These results support that the assumption that different participants use similar strategy is not necessarily harmful in this setting. These observations are also consistent with our expectations, since our task was not designed to make clear distinctions between learning algorithms at the behavioral level, but to disentangle learning signals.

4.3. Model-free and model-based neural signatures

Overall, learning the model of the task does not increase significantly the fit to behavior of the respective surprise-modulated MF and hybrid algorithms in our experiment. Nevertheless, these algorithms predicted the brain data better, and we found correlates of MB $\mathrm{SPE}_{\mathrm{BF}}$ and $\mathbf{S}_{\mathrm{BF}}$. This suggests that a MB learning system is active, possibly building an internal model of the task and performing "latent learning" (Bast et al., 2009; Tolman, 1948), but is not (yet) in control or not used for planning, consistent with findings in Xu et al. (2021). Such an interpretation is also consistent with the idea that a "mixture of experts" co-exist and run in parallel in the brain, and the control of the behavior is delegated among them depending on the circumstances and on multiple factors such as the uncertainty of each expert and time constraints (Daw et al., 2005; Geerts et al., 2020; Lee et al., 2014; O'Doherty et al., 2020). In our case, in the Surprise Actor-critic, it is difficult to draw conclusions about the reliability of the MB module due to insufficient constraints of certain MB parameters. Moreover, our experimental design (deterministic graph with strategic surprising events) would not in principle lead to an asymmetry between the reliability of MB and MF. On the other hand, due to the higher number of states and transitions, our task can lead to a significant difference between the computational cost of MF learning and MB planning. We, therefore, hypothesize that the observed dominance of MF contribution to behavior in our experiment is due to the fact that an MF agent can safely solve the task as good as a MB agent but with much less computational resources – see Huys et al. (2015); Kahneman (2011); Xu et al. (2021) for similar observations.

Concerning the neural representation of $\mathrm{SPE}_{\mathrm{BF}}$ and $\mathbf{S}_{\mathrm{BF}}$, we found regions belonging to the salience network (Seeley et al., 2007). The intraparietal sulcus has been found to correlate with SPE and surprise signals in previous studies (Gläscher et al., 2010; Lee et al., 2014; Schad et al., 2020), as well as regions in the lateral prefrontal and orbitofrontal cortex (Doll et al., 2012; Gläscher et al., 2010; O'Doherty et al., 2015; Simon and Daw, 2011). Moreover, the insula and the SMA have been found to be components of the network related to surprise (Fouragnan et al., 2018; Loued-Khenissi et al., 2020). Surprise and its network have been viewed to comprise various roles. One role is the encoding of saliency or how much an observation protrudes among others, driving an attentional mechanism that helps in guiding actions or driving the urge to explore (Fouragnan et al., 2018; Friston, 2010; Friston et al., 2017; Gottlieb and Oudeyer, 2018; Schwartenbeck et al., 2013). A second role is the implementation of a learning signal that mediates the updating of beliefs and behavior (Faraji et al., 2018; Fouragnan et al., 2018; Friston, 2010; Friston et al., 2017; Liakoni et al., 2021). Concerning its role in saliency, representations of surprise signals have been found in lateral parietal cortex and the SMA, whereas the second role has been additionally associated with other brain structures, such the insula and the striatum (Fouragnan et al., 2018). Our imaging results include a mixture of these two components. We interpret SPE as a signal of "state mismatch", present in the brain regardless of how it may be used by downstream structures, i.e. for integrating (Hybrid Actor-critic) or ignoring (Surprise Actor-critic) the surprising information. This signal can then be broadcasted to executive structures responsible for action selection and possibly to deeper structures, such as the locus coeruleus (LC) (Aston-Jones and Cohen, 2005; Aston-Jones et al., 1994; Avery and Krichmar, 2017), that, taking other factors into account, such as aspects of the task at hand, may convert it into neuromodulatory signals of surprise.

For the MF $\mathrm{RPE}_{\mathrm{non-goal}}$ we found signatures in the striatum, the SMA, the ACC (at CDT $p = 10^{-3}$), and parietal regions, such as the superior parietal lobule and the precuneus. The ACC has been reported to be active with errors, with the assessment of outcomes and with value expectation (Chase et al., 2015; Kolling et al., 2016; Vassena et al., 2020; 2014), whereas a multitude of functionalities have been attributed to SMA, among which learning of new associations and movement sequences (Nachev et al., 2008) and reward-related surprise (Vassena et al., 2020). The striatal activity we found is slightly more dorsal than ventral. Dorsal striatum is known to receive dopaminergic input from the substantia nigra and to be involved in motor planning and execution (Takahashi et al., 2008), and it has been hypothesized to implement the actor in an Actor-critic framework (Joel et al., 2002; Takahashi et al., 2008). Consistent to this hypothesis, we additionally found correlated activity with the relative policy preferences in the dorsal striatum.

Importantly, the RPE timelines of the leading algorithms Surprise Actor-critic and Hybrid Actor-critic gave rise to similar brain regions with significant activations. Even more interestingly, the timelines of their SPE that stem from different update rules also gave rise to similar brain regions with significant activations (see Appendix Fig. A.16 and Fig. A.17). Thus, the observed neural representations seem to be robust, and our results point to regions involved in this type of computations, beyond the specific details of each signal and each algorithm.

4.4. Surprise Actor-critic and other measures of surprise

The Bayes Factor Surprise $\mathbf{S}_{\mathrm{BF}}$ (Liakoni et al., 2021) used in our Surprise Actor-critic algorithm is a measure of "puzzlement surprise" (Faraji et al., 2018), expressing a violation in our current knowledge about the world. Other measures of puzzlement surprise are the Shannon Surprise (Shannon, 1948), from the field of information theory, Free Energy (Friston, 2010; Schwartenbeck et al., 2013), which is a variational approximation of Shannon Surprise, the SPE, introduced in decision making tasks (Daw et al., 2011; Gläscher et al., 2010), and the Confidence Corrected Surprise (Faraji et al., 2018). A large body of literature has found evidence for surprise manifestation in pupil dilation (Nassar et al., 2012; Preuschoff et al., 2011) and in electroencephalography (EEG) signals or behavioral indicators, often in oddball experiments (Gijsen et al., 2020; Lieder et al., 2013; Mars et al., 2008; Meyniel et al., 2016; Modirshanechi et al., 2019; Squires et al., 1976). Therefore, although the role of the Bayes Factor surprise in our Surprise Actor-critic algorithm is inspired by a normative Bayesian approach, one can think of variations of the Surprise Actor-critic with other measures of puzzlement surprise.

Puzzlement surprise is, however, fundamentally different from "enlightenment surprise" (Faraji et al., 2018) that focuses on the information gain caused by a surprising event. The classic measure of enlightenment surprise is Bayesian Surprise (Itti and Baldi, 2006; Schmidhuber, 2010; Storck et al., 1995; Sun et al., 2011), also known as information gain (Little and Sommer, 2013; Storck et al., 1995; Sun et al., 2011). Although Bayesian Surprise has repeatedly been used in models of curiosity (Bruckner et al., 2020; Gottlieb and Oudeyer, 2018; Schmidhuber, 1991) and has been shown to have its own neural signatures (d'Acremont et al., 2013; Gijsen et al., 2020; O'Reilly et al., 2013; Ostwald et al., 2012; Visalli et al., 2019), its fundamental difference from

puzzlement surprise, i.e. that is calculated after the belief update, makes it a less appropriate candidate for modulation of learning online.

Finally, a MF RPE can also be viewed as a "surprise" signal and as a mechanism serving the detection of changes (Rouhani and Niv, 2021; Rouhani et al., 2020). However, a change detected by a RPE would be related to differences in perceived MF values and not to sensory aspects of the change or to a low transition probability. Hence, change detection in our experiment based on RPE is, by construction, less effective.

4.5. Future directions

We presented an experimental paradigm for dissociating MF and MB learning signals at the level of brain responses via introducing outlier events of similar value. Our paradigm can be applied in different settings and be combined with other experimental manipulations. Our design differentiates the dynamics of different signals, but it comes at the price that different learning systems may give rise to similar behavior. Combining our task with an additional manipulation, in order to achieve a double dissociation, would be an interesting and important future direction. Also interesting would be the investigation of possible differences in the neural signatures of SPE_{BF} ($\mathbf{S}_{BF}$) after participants have adequately learned the graph structure compared to during the learning process.

Our results indicate that behavioral choices in our task were best explained by MF learning modules. Yet observers exhibited additional delay after surprise trials indicating that MB signals were calculated in the brain, and the neural data showed a better fit for algorithms with a MB module. We considered a large range of RL algorithms, both existing and novel, with different characteristics. It is still possible, however, that the true strategy that participants followed was not among them. A recent study (Silva and Hare, 2020) on the two-stage task pointed out the impact of participants' understanding of the task on behavior. The authors also showed that if a simulated agent is MB but is using a "wrong" model of the task structure, then the apparent best fit for behavior can be a hybrid mixture of MF and MB. Under the idea that there were MB contributions in our task, the question is then, what is the model structure that human subjects used? Can it be that behavior appears MF because we yet do not know the "imperfect" model and updating scheme that humans function with? These are in our view central questions towards understanding human learning behavior, and more theoretical as well as experimental work are needed to address them.

Our motivation has been to study human learning and brain signals in a multi-step, more complex, and presumably more realistic scenario. Our task is, however, still far from realistic situations encountered by biological agents. The design of more experiments involving tasks that are closer to real life, as well as the consideration of large batteries of competing algorithms to explain behavior will be a crucial step in understanding the learning schemes that animals and humans may employ.

Credit authorship contribution statement

Vasiliki Liakoni: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Marco P. Lehmann:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Alireza Modirshanechi:** Conceptualization, Methodology, Validation, Investigation, Writing – review & editing. **Johanni Brea:** Conceptualization, Methodology, Investigation, Writing – review & editing. **Antoine Lutti:** Methodology, Validation, Investigation, Resources, Supervision. **Wulfram Gerstner:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Kerstin Preuschoff:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

This research was supported by the Swiss National Science Foundation (CRSII2-147636 to K.P. and W.G. and 200020_184615 to W.G.) as well as by the European Union Horizon 2020 Framework Program No.785907 (Human Brain Project, SGA2). A.L. was supported by the ROGER DE SPOELBERCH foundation and the Swiss National Science Foundation (320030_184784). The experimental work was carried out on the MRI platform of the Département des Neurosciences Cliniques - Centre Hospitalier Universitaire Vaudois (CHUV), which is generously supported by the ROGER DE SPOELBERCH and Partridge Foundations. We would like to thank Dr. Bogdan Draganski for his help. We also thank Remi Castella, Estelle Dupuis and Dr. Leyla Loued-Khenissi for their help with data acquisition, as well as Dr. Samuel Muscinelli, Maya Jastrzebowska, Dr. Thomas Bolton, Dr. Jonas Richiardi, Dr. Leyla Loued-Khenissi, Prof. Dimitri Van De Ville and Prof. Michael Herzog for useful discussions. Finally, we thank our reviewers for their constructive and useful comments.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.10.1016/j.neuroimage.2021.118780

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G., 2014. Machine learning for neuroimaging with scikit-learn. Front Neuroinform 8 (14).

Anggraini, D., Glasauer, S., Wunderlich, K., 2018. Neural signatures of reinforcement learning correlate with strategy adoption during spatial navigation. Sci Rep 8 (1), 1–14.

Aston-Jones, G., Cohen, J.D., 2005. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu. Rev. Neurosci. 28, 403–450.

Aston-Jones, G., Rajkowski, J., Kubiak, P., Alexinsky, T., 1994. Locus coeruleus neurons in monkey are selectively activated by attended cues in a vigilance task. J. Neurosci. 14 (7), 4467–4480.

Avery, M.C., Krichmar, J.L., 2017. Neuromodulatory systems and their interactions: a review of models, theories, and experiments. Front Neural Circuits 11 (108).

Balaguer, J., Spiers, H., Hassabis, D., Summerfield, C., 2016. Neural mechanisms of hierarchical planning in a virtual subway network. Neuron 90 (4), 893–903.

Balleine, B.W., 2005. Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits. Physiology & behavior 86 (5), 717–730.

Bast, T., Wilson, I.A., Witter, M.P., Morris, R.G.M., 2009. From rapid place learning to behavioral performance: a key role for the intermediate hippocampus. PLoS Biol 7 (4), e1000089.

Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., Rushworth, M.F.S., 2008. Associative learning of social value. Nature 456 (7219), 245–249.

Berger, J.O., Pericchi, L.R., 1996. The intrinsic bayes factor for model selection and prediction. J Am Stat Assoc 91 (433), 109–122.

Brainard, D.H., 1997. The psychophysics toolbox. Spat Vis 10 (4), 433–436.

Bruckner, R., Heekeren, H. R., Ostwald, D., 2020. Belief states and categorical-choice biases determine reward-based learning under perceptual uncertainty.

Chase, H.W., Kumar, P., Eickhoff, S.B., Dombrovski, A.Y., 2015. Reinforcement learning models and their neural correlates: an activation likelihood estimation meta-analysis. Cognitive, affective, & behavioral neuroscience 15 (2), 435–459.

Coddington, L.T., Dudman, J.T., 2019. Learning from action: reconsidering movement signaling in midbrain dopamine neuron activity. Neuron 104 (1), 63–77.

Collins, A.G.E., Cockburn, J., 2020. Beyond dichotomies in reinforcement learning. Nat. Rev. Neurosci. 1–11.

Cushman, F., Morris, A., 2015. Habitual control of goal selection in humans. Proceedings of the National Academy of Sciences 112 (45), 13817–13822.

d'Acremont, M., Bossaerts, P., 2016. Neural mechanisms behind identification of leptokurtic noise and adaptive behavioral response. Cerebral Cortex 26 (4), 1818–1830.

d'Acremont, M., Schultz, W., Bossaerts, P., 2013. The human brain encodes event frequencies while forming subjective beliefs. J. Neurosci. 33 (26), 10887–10897.

Daw, N.D., et al., 2011. Trial-by-trial data analysis using computational models. Decision making, affect, and learning: Attention and performance XXIII 23 (1).

Daw, N.D., 2015. Of goals and habits. Proceedings of the National Academy of Sciences 112 (45), 13749–13750.

Daw, N.D., 2018. Are we of two minds? Nat. Neurosci. 21 (11), 1497–1499.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron 69 (6), 1204–1215.

Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. 8 (12), 1704–1711.

Dayan, P., 1993. Improving generalization for temporal difference learning: the successor representation. Neural Comput 5 (4), 613–624.

Deserno, L., Huys, Q.J.M., Boehme, R., Buchert, R., Heinze, H.-J., Grace, A.A., Dolan, R.J., Heinz, A., Schlagenhauf, F., 2015. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. Proceedings of the National Academy of Sciences 112 (5), 1595–1600.
Dezfouli, A., Lingawi, N.W., Balleine, B.W., 2014. Habits as action sequences: hierarchical action control and changes in outcome value. Philosophical Transactions of the Royal Society of London B: Biological Sciences 369 (1655). 20130482
Doll, B.B., Duncan, K.D., Simon, D.A., Shohamy, D., Daw, N.D., 2015. Model-based choices involve prospective neural activity. Nat. Neurosci. 18 (5), 767.
Doll, B.B., Shohamy, D., Daw, N.D., 2015. Multiple memory systems as substrates for multiple decision systems. Neurobiol Learn Mem 117, 4–13.
Doll, B.B., Simon, D.A., Daw, N.D., 2012. The ubiquity of model-based reinforcement learning. Curr. Opin. Neurobiol. 22 (6), 1075–1081.
Doucet, A., Godsill, S., Andrieu, C., 2000. On sequential monte carlo sampling methods for bayesian filtering. Stat Comput 10 (3), 197–208.
Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M., Dolan, R.J., 2015. Model-based reasoning in humans becomes automatic with training. PLoS Comput Biol 11 (9), e1004463.
Efron, B., Hastie, T., 2016. Computer age statistical inference, volume 5. Cambridge University Press.
Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: why fmri inferences for spatial extent have inflated false-positive rates. Proceedings of the national academy of sciences 113 (28), 7900–7905.
Faraji, M., Preuschoff, K., Gerstner, W., 2018. Balancing new against old information: the role of puzzlement surprise in learning. Neural Comput 30 (1), 34–83.
Fermin, A.S.R., Yoshida, T., Yoshimoto, J., Ito, M., Tanaka, S.C., Doya, K., 2016. Model-based action planning involves cortico-cerebellar and basal ganglia networks. Sci Rep 6 (31378).
Fong, E., Holmes, C.C., 2020. On the marginal likelihood and cross-validation. Biometrika 107 (2), 489–496.
Fouragnan, E., Retzler, C., Philiastides, M.G., 2018. Separate neural representations of prediction error valence and surprise: evidence from an fmri meta-analysis. Hum Brain Mapp 39 (7), 2887–2906.
Frémaux, N., Gerstner, W., 2016. Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. Front Neural Circuits 9 (85).
Friston, K., 2010. The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11 (2), 127–138.
Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., 2017. Active inference: a process theory. Neural Comput 29 (1), 1–49.
Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S.J., 1994. Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 2 (4), 189–210.
Geerts, J.P., Chersi, F., Stachenfeld, K.L., Burgess, N., 2020. A general model of hippocampal and dorsal striatal learning and decision making. Proceedings of the National Academy of Sciences 117 (49), 31427–31437.
Gershman, S.J., Markman, A.B., Otto, A.R., 2014. Retrospective revaluation in sequential decision making: a tale of two systems. Journal of Experimental Psychology: General 143 (1), 182.
Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., Brea, J., 2018. Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules. Front Neural Circuits 12.
Gijsen, S., Grundei, M., Lange, R.T., Ostwald, D., Blankenburg, F., 2020. Neural surprise in somatosensory bayesian learning. BioRxiv.
Gläscher, J., Daw, N., Dayan, P., O'Doherty, J.P., 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66 (4), 585–595.
Gordon, N.J., Salmond, D.J., Smith, A.F.M., 1993. Novel approach to nonlinear/non-gaussian bayesian state estimation. IEE proceedings F (radar and signal processing) 140, 107–113. IET
Gottlieb, J., Oudeyer, P.-Y., 2018. Towards a neuroscience of active sampling and curiosity. Nat. Rev. Neurosci. 19 (12), 758–770.
Gremel, C.M., Costa, R.M., 2013. Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. Nat Commun 4.
Griswold, M.A., Jakob, P.M., Heidemann, R.M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., Haase, A., 2002. Generalized autocalibrating partially parallel acquisitions (grappa). Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 47 (6), 1202–1210.
Hare, T.A., O'doherty, J., Camerer, C.F., Schultz, W., Rangel, A., 2008. Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. J. Neurosci. 28 (22), 5623–5630.
Hastings, W. K., 1970. Monte carlo sampling methods using markov chains and their applications.
Held, L., Ott, M., 2018. On p-values and bayes factors.
Howard, J.D., Kahnt, T., 2018. Identity prediction errors in the human midbrain update reward-identity expectations in the orbitofrontal cortex. Nat Commun 9 (1), 1–11.
Huettel, S.A., Mack, P.B., McCarthy, G., 2002. Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex. Nat. Neurosci. 5 (5), 485–490.
Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., Turner, R., 2002. Image distortion correction in fmri: a quantitative evaluation. Neuroimage 16 (1), 217–240.
Huys, Q.J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., 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 (3), e1002410.
Huys, Q.J.M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S.J., Dayan, P., Roiser, J.P., 2015. Interplay of approximate planning strategies. Proceedings of the National Academy of Sciences 112 (10), 3098–3103.
Illing, B., Gerstner, W., Bellec, G., 2021. Local plasticity rules can learn deep representations using self-supervised contrastive predictions. Thirty-Fifth Conference on Neural Information Processing Systems.
Ito, M., Doya, K., 2011. Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. Curr. Opin. Neurobiol. 21 (3), 368–373.
Itti, L., Baldi, P.F., 2006. Bayesian surprise attracts human attention. Adv Neural Inf Process Syst 547–554.
Joel, D., Niv, Y., Ruppin, E., 2002. Actor–critic models of the basal ganglia: new anatomical and computational perspectives. Neural networks 15 (4–6), 535–547.
Kahneman, D., 2011. Thinking, fast and slow. Macmillan.
Kim, H.R., Malik, A.N., Mikhael, J.G., Bech, P., Tsutsui-Kimura, I., Sun, F., Zhang, Y., Li, Y., Watabe-Uchida, M., Gershman, S.J., et al., 2020. A unified framework for dopamine signals across timescales. Cell 183 (6), 1600–1616.
Kolling, N., Wittmann, M.K., Behrens, T.E.J., Boorman, E.D., Mars, R.B., Rushworth, M.F.S., 2016. Value, search, persistence and model updating in anterior cingulate cortex. Nat. Neurosci. 19 (10), 1280–1285.
Kroemer, N.B., Lee, Y., Pooseh, S., Eppinger, B., Goschke, T., Smolka, M.N., 2019. L-Dopa reduces model-free control of behavior by attenuating the transfer of value to action. Neuroimage 186, 113–125.
Langdon, A.J., Sharpe, M.J., Schoenbaum, G., Niv, Y., 2018. Model-based predictions for dopamine. Curr. Opin. Neurobiol. 49, 1–7.
Lee, S.W., Shimojo, S., O'Doherty, J.P., 2014. Neural computations underlying arbitration between model-based and model-free learning. Neuron 81 (3), 687–699.
Lehmann, M.P., Xu, H.A., Liakoni, V., Herzog, M.H., Gerstner, W., Preuschoff, K., 2019. One-shot learning and behavioral eligibility traces in sequential decision making. Elife 8, e47463.
Li, J., Daw, N.D., 2011. Signals in human striatum are appropriate for policy update rather than value prediction. J. Neurosci. 31 (14), 5504–5511.
Liakoni, V., Modirshanechi, A., Gerstner, W., Brea, J., 2021. Learning in volatile environments with the bayes factor surprise. Neural Comput 1–72.
Lieder, F., Daunizeau, J., Garrido, M.I., Friston, K.J., Stephan, K.E., 2013. Modelling trial-by-trial changes in the mismatch negativity. PLoS Comput. Biol. 9 (2).
Lillicrap, T.P., Santoro, A., Marris, L., Akerman, C.J., Hinton, G., 2020. Backpropagation and the brain. Nat. Rev. Neurosci. 21 (6), 335–346.
Little, D.Y.-J., Sommer, F.T., 2013. Learning and exploration in action-perception loops. Front Neural Circuits 7 (37).
Loued-Khenissi, L., Pfeuffer, A., Einhäuser, W., Preuschoff, K., 2020. Anterior insula reflects surprise in value-based decision-making and perception. Neuroimage 116549.
Lutti, A., Thomas, D.L., Hutton, C., Weiskopf, N., 2013. High-resolution functional mri at 3 t: 3d/2d echo-planar imaging with optimized physiological noise correction. Magn Reson Med 69 (6), 1657–1664.
Mack, M.L., Preston, A.R., Love, B.C., 2013. Decoding the brain?s algorithm for categorization from its neural implementation. Current Biology 23 (20), 2023–2027.
Mars, R.B., Debener, S., Gladwin, T.E., Harrison, L.M., Haggard, P., Rothwell, J.C., Bestmann, S., 2008. Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. J. Neurosci. 28 (47), 12539–12545.
Meyniel, F., Maheu, M., Dehaene, S., 2016. Human inferences about sequences: a minimal transition probability model. PLoS Comput. Biol. 12 (12), e1005260.
Mikhael, J.G., Kim, H.R., Uchida, N., Gershman, S.J., 2019. Ramping and state uncertainty in the dopamine signal. bioRxiv 805366.
Miller, K., Venditto, S. J., 2020. Multi-step planning in the brain.
Modirshanechi, A., Kiani, M.M., Aghajan, H., 2019. Trial-by-trial surprise-decoding model for visual and auditory binary oddball tasks. Neuroimage 196, 302–317.
Moore, A.W., Atkeson, C.G., 1993. Prioritized sweeping: reinforcement learning with less data and less time. Mach Learn 13 (1), 103–130.
Mumford, J.A., Poline, J.-B., Poldrack, R.A., 2015. Orthogonalization of regressors in fmri models. PLoS ONE 10 (4), e0126255.
Nachev, P., Kennard, C., Husain, M., 2008. Functional role of the supplementary and pre-supplementary motor areas. Nat. Rev. Neurosci. 9 (11), 856–869.
Nassar, M.R., Bruckner, R., Frank, M.J., 2019. Statistical context dictates the relationship between feedback-related eeg signals and learning. Elife 8, e46975.
Nassar, M.R., Frank, M.J., 2016. Taming the beast: extracting generalizable knowledge from computational models of cognition. Curr Opin Behav Sci 11, 49–54.
Nassar, M.R., Rumsey, K.M., Wilson, R.C., Parikh, K., Heasly, B., Gold, J.I., 2012. Rational regulation of learning dynamics by pupil-linked arousal systems. Nat. Neurosci. 15 (7), 1040.
Neath, A.A., Cavanaugh, J.E., 2012. The bayesian information criterion: background, derivation, and applications. Wiley Interdiscip. Rev. Comput. Stat. 4 (2), 199–203.
Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15 (1), 1–25.
O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., Dolan, R.J., 2004. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304 (5669), 452–454.
O'Doherty, J. P., Lee, S., Tadayonnejad, R., Cockburn, J., Iigaya, K., Charpentier, C. J., 2020. Why and how the brain weights contributions from a mixture of experts.
O'Doherty, J.P., Lee, S.W., McNamee, D., 2015. The structure of reinforcement-learning mechanisms in the human brain. Curr Opin Behav Sci 1, 94–100.
O'Reilly, J.X., Schüffelgen, U., Cuell, S.F., Behrens, T.E.J., Mars, R.B., Rushworth, M.F.S., 2013. Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. Proceedings of the National Academy of Sciences 110 (38), E3660–E3669.
Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T.T., Kiebel, S.J., Blankenburg, F., 2012. Evidence for neural encoding of bayesian surprise in human somatosensation. Neuroimage 62 (1), 177–188.
Otto, A.R., Gershman, S.J., Markman, A.B., 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., Raio, C.M., Chiang, A., Phelps, E.A., Daw, N.D., 2013. Working-memory capacity protects model-based learning from stress. Proceedings of the National Academy of Sciences 110 (52), 20941–20946.
Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., 2011. Statistical parametric mapping: The analysis of functional brain images. Elsevier.
Pernet, C.R., 2014. Misconceptions in the use of the general linear model applied to functional mri: a tutorial for junior neuro-imagers. Front Neurosci 8 (1).
Peters, J., 2010. Policy gradient methods. Scholarpedia 5 (11), 3698. doi:10.4249/scholarpedia.3698. Revision #137199
Piray, P., Dezfouli, A., Heskes, T., Frank, M.J., Daw, N.D., 2019. Hierarchical bayesian inference for concurrent model fitting and comparison for group studies. PLoS Comput. Biol. 15 (6), e1007043.
Poser, B.A., Versluis, M.J., Hoogduin, J.M., Norris, D.G., 2006. Bold contrast sensitivity enhancement and artifact reduction with multiecho epi: parallel-acquired inhomogeneity-desensitized fmri. Magn Reson Med 55 (6), 1227–1235.
Preuschoff, K., Hart, B.M.t., Einhäuser, W., 2011. Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. Front Neurosci 5 (115).
Razavi, M., Grabowski, T.J., Vispoel, W.P., Monahan, P., Mehta, S., Eaton, B., Bolinger, L., 2003. Model assessment and model building in fmri. Hum Brain Mapp 20 (4), 227–238.
Rigoux, L., Stephan, K.E., Friston, K.J., Daunizeau, J., 2014. Bayesian model selection for group studies - revisited. Neuroimage 84, 971–985.
Rouault, M., Drugowitsch, J., Koechlin, E., 2019. Prefrontal mechanisms combining rewards and beliefs in human decision-making. Nat Commun 10 (1), 1–16.
Rouhani, N., Niv, Y., 2021. Signed and unsigned reward prediction errors dynamically enhance learning and memory. Elife 10, e61077.
Rouhani, N., Norman, K.A., Niv, Y., Bornstein, A.M., 2020. Reward prediction errors create event boundaries in memory. Cognition 203 (104269).
Rust, R.T., Schmittlein, D.C., 1985. A bayesian cross-validated likelihood method for comparing alternative specifications of quantitative models. Marketing Science 4 (1), 20–40.
Särkkä, S., 2013. Bayesian filtering and smoothing, volume 3. Cambridge University Press.
Schad, D.J., Rapp, M.A., Garbusow, M., Nebe, S., Sebold, M., Obst, E., Sommer, C., Deserno, L., Rabovsky, M., Friedel, E., et al., 2020. Dissociating neural learning signals in human sign-and goal-trackers. Nat. Hum. Behav. 4 (2), 201–214.
Schmidhuber, J., 1991. Curious model-building control systems. Proc. international joint conference on neural networks 1458–1463.
Schmidhuber, J., 2010. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). IEEE Trans Auton Ment Dev 2 (3), 230–247.
Schulman, J., Moritz, P., Levine, S., Jordan, M., Abbeel, P., 2015. High-dimensional continuous control using generalized advantage estimation. arXiv:1506.02438
Schwartenbeck, P., FitzGerald, T., Dolan, R., Friston, K., 2013. Exploration, novelty, surprise, and free energy minimization. Front Psychol 4 (710).
Seeley, W.W., Menon, V., Schatzberg, A.F., Keller, J., Glover, G.H., Kenna, H., Reiss, A.L., Greicius, M.D., 2007. Dissociable intrinsic connectivity networks for salience processing and executive control. J. Neurosci. 27 (9), 2349–2356.
Seijen, H.V., Sutton, R.S., 2013. Efficient planning in mdps by small backups. Proc. 30th Int. Conf. Mach. Learn. 1–3.
Shannon, C., 1948. A mathematical theory of communication. Bell System Technical Journal 27 (379–423), 623–656. 20
Silva, C.F.d., Hare, T.A., 2020. Humans primarily use model-based inference in the two-stage task. Nat. Hum. Behav. 1–14.
Simon, D.A., Daw, N.D., 2011. Neural correlates of forward planning in a spatial decision task in humans. The Journal of Neuroscience 31 (14), 5526–5539.
Soch, J., Allefeld, C., 2018. Macs–a new spm toolbox for model assessment, comparison and selection. J. Neurosci. Methods 306, 19–31.
Squires, K.C., Wickens, C., Squires, N.K., Donchin, E., 1976. The effect of stimulus sequence on the waveform of the cortical event-related potential. Science 193 (4258), 1142–1146.
Stalnaker, T.A., Howard, J.D., Takahashi, Y.K., Gershman, S.J., Kahnt, T., Schoenbaum, G., 2019. Dopamine neuron ensembles signal the content of sensory prediction errors. Elife 8.
Stalnaker, T.A., Liu, T.-L., Takahashi, Y.K., Schoenbaum, G., 2018. Orbitofrontal neurons signal reward predictions, not reward prediction errors. Neurobiol Learn Mem 153, 137–143.
Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. Neuroimage 46 (4), 1004–1017.
Storck, J., Hochreiter, S., Schmidhuber, J., 1995. Reinforcement driven information acquisition in non-deterministic environments. Proceedings of the international conference on artificial neural networks, Paris, volume 2 159–164.
Sun, Y., Gomez, F., Schmidhuber, J., 2011. Planning to be surprised: optimal bayesian exploration in dynamic environments. International Conference on Artificial General Intelligence 41–51. Springer
Sutton, R.S., Barto, A.G., 1998. Reinforcement learning: An introduction. A Bradford book. Bradford Book.
Sutton, R.S., Barto, A.G., 2018. Reinforcement learning: An introduction. MIT press.
Takahashi, Y., Schoenbaum, G., Niv, Y., 2008. Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. Front Neurosci 2 (14).
Takahashi, Y.K., Batchelor, H.M., Liu, B., Khanna, A., Morales, M., Schoenbaum, G., 2017. Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. Neuron 95 (6), 1395–1405.
Tanaka, S., Pan, X., Oguchi, M., Taylor, J.E., Sakagami, M., 2015. Dissociable functions of reward inference in the lateral prefrontal cortex and the striatum. Front Psychol 6.
Tartaglia, E.M., Clarke, A.M., Herzog, M.H., 2017. What to choose next? a paradigm for testing human sequential decision making. Front Psychol 8 (312).
Tolman, E.C., 1948. Cognitive maps in rats and men. Psychol Rev 55 (4), 189.
Turner, B.M., Forstmann, B.U., Love, B.C., Palmeri, T.J., Maanen, L.V., 2017. Approaches to analysis in model-based cognitive neuroscience. J Math Psychol 76, 65–79.
Turner, B.M., Forstmann, B.U., Wagenmakers, E.-J., Brown, S.D., Sederberg, P.B., Steyvers, M., 2013. A bayesian framework for simultaneously modeling neural and behavioral data. Neuroimage 72, 193–206.
Vassena, E., Deraeve, J., Alexander, W.H., 2020. Surprise, value and control in anterior cingulate cortex during speeded decision-making. Nat. Hum. Behav. 4 (4), 412–422.
Vassena, E., Krebs, R.M., Silvetti, M., Fias, W., Verguts, T., 2014. Dissociating contributions of acc and vmpfc in reward prediction, outcome, and choice. Neuropsychologia 59, 112–123.
Visalli, A., Capizzi, M., Ambrosini, E., Mazzonetto, I., Vallesi, A., 2019. Bayesian modeling of temporal expectations in the human brain. Neuroimage 202 (116097).
Wang, Y., Pericchi, L., 2020. A bridge between cross-validation bayes factors and geometric intrinsic bayes factors. arXiv:2006.06495
Williams, R.J., 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 8 (3–4), 229–256.
Wilson, R.C., Collins, A.G.E., 2019. Ten simple rules for the computational modelling of behavioral data. Elife 8, e49547.
Wilson, R.C., Niv, Y., 2015. Is model fitting necessary for model-based fmri? PLoS Comput. Biol. 11 (6), e1004237.
Wimmer, G.E., Shohamy, D., 2012. Preference by association: how memory mechanisms in the hippocampus bias decisions. Science 338 (6104), 270–273.
Worsley, K.J., Friston, K.J., 1995. Analysis of fmri time-series revisited?again. Neuroimage 2 (3), 173–181.
Wunderlich, K., Dayan, P., Dolan, R.J., 2012. Mapping value based planning and extensively trained choice in the human brain. Nat. Neurosci. 15 (5), 786–791.
Wunderlich, K., Smittenaar, P., Dolan, R.J., 2012. Dopamine enhances model-based over model-free choice behavior. Neuron 75 (3), 418–424.
Xu, H.A., Modirshanechi, A., Lehmann, M.P., Gerstner, W., Herzog, M.H., 2021. Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. PLoS Comput. Biol. 17 (6), e1009070.
Yu, A.J., Dayan, P., 2005. Uncertainty, neuromodulation, and attention. Neuron 46 (4), 681–692.