EPFL

Thèse n° 10 168

Neural Correlates of Reinforcement Learning: eligibility trace, reward prediction error, novelty and surprise

Présentée le 22 mai 2020

à la Faculté des sciences de la vie
Laboratoire de psychophysique
Programme doctoral en neurosciences

pour l'obtention du grade de Docteur ès Sciences

par

He XU

Acceptée sur proposition du jury

Prof. F. Schürmann, président du jury
Prof. M. Herzog, directeur de thèse
Prof. W. Senn, rapporteur
Prof. F. Schlagenhauf, rapporteur
Prof. B. McCabe, rapporteur

École polytechnique fédérale de Lausanne

2020

Imagination is more important than knowledge
— Albert Einstein

To my parents, Philippe, and Xi Li

Acknowledgements

I came to Switzerland on the day of 2nd September 2012 for my master study. At that time, I did not expect to spend such a long time in this country. After my master thesis, I started my first PhD in Touradj Ebrahimi's lab on 1st March 2015. Soon I realised that it was not a suitable lab for me and decided to quit the position. On 1st September 2015 I started a new position in Michael's lab and continued up to now.

It was an interesting journey of life during the 5 years PhD study. I met my dear husband Philippe Hanhart in the very early days. Thanks to his love and support I am able to overcome a lot of difficulties. He always tries to cheer me up and encourage me when I am stressed or depressed. I started to enjoy my life in Switzerland only after being together with Philippe, and decided to stay here after my study. Xi Li (李曦) was the first Chinese I met after arriving in Ecublens. I still remember that it was my second day in Switzerland and I met her in an English writing class. The class was boring but she is interesting. We baked fake marrons in her kitchen and laughed at each other. We then moved and lived together for one semester, having hotpot every weekends made both of us fat.

There are also some people I need to thank for helping with my PhD study. First to Aaron Clarke, who introduced me with the project and provided a lot of help in my early days. I am sad that he passed away and I really hope he has a good after life. Marco Lehmann who is the former PhD in the lab of LCN, worked on the same Sinergia project with me. We had a lot of discussion on the data analysis and result interpretation in both mine and his thesis. Marc Repnow, our lab technician, has very special personality but is very helpful in statistics and experimental design. Wulfram Gestner, professor in the lab of LCN, who I was always afraid of, provided very useful advises and directions in data analysis and interpretation.

Of course I want to thank my parents who agreed to let me study abroad. I understand it must be difficult for them not being able to see me frequently, but they still supported me as much as they can. I also want to thank Philippe's parents, Marcel, Gilberte and Marie-Francoise who are always very nice to me and make me feel like home here.

It is a very special experience to have my PhD exam online, maybe my public defence will also be like this. Since the beginning of this year I have spent quite some time staying and working at home because of the coronavirus. I hope all my family and friends stay safe and healthy.

Vevey, 20 April 2020 徐赫

Abstract

In reinforcement learning (RL), an agent makes sequential decisions to maximise the reward it can obtain from an environment. During learning, the actual and expected outcomes are compared to tell whether a decision was good or bad. The difference between the actual outcome and expected outcome is the prediction error. The prediction error can be categorised into two types: the reward prediction error (RPE) and the state prediction error (SPE), which can serve as teaching signals in reinforcement learning models.

The reward prediction error (RPE), i.e., the difference between the actual and the expected reward, is one of the crucial variables in model-free reinforcement learning. In humans, the RPE has been shown be generated from the mid-brain dopamine system. Electroencephalogram (EEG) studies have also shown that the RPE can be reflected by a EEG waveform occurring in the frontal-central brain region, between 250 and 400ms after a reward signal is shown. This RPE-related waveform is called the feedback-related negativity (FRN). Most FRN studies use N-armed bandit tasks to study the relationship between FRN amplitude and the RPE. In the N-armed bandit tasks, participants receive the reward immediately after an action is taken. However, everyday reinforcement learning situations come usually with many non-rewarded states and actions until a reward is obtained. The first part of this thesis aims to answer the question whether the FRN still reflects the RPE in complex tasks.

The state prediction error (SPE) measures how much the agent's expectation on state transitions differs before and after an observation. Novelty and surprise are two types of SPE signals that drive learning when the external reward is not yet provided. Novelty measures how frequently an observation occurs, whereas surprise measures how much expectations are violated. EEG studies have showed that novelty can be reflected in different EEG components, such as the N1 and the P300. Novelty can even be observed in human infants when they learn a novel stimulus. RL algorithms with additional novelty-driven module showed good exploration behaviour in learning bandit tasks and simple Markov decision tasks. Surprise, on the other hand, is also used as a learning signal in many surprise-based RL models. The mismatch negativity (MMN) in EEG is generally considered as a neural signature of surprise. However, how novelty and surprise interact and contribute in learning remained un-addressed. In this thesis, I studied the neural correlates of novelty and surprise in a sequential decision-making task, and proposed a model combining both novelty and surprise to explain human learning.

I implemented different sequential decision-making tasks to study four RL signals, which are the eligibility trace, the RPE, novelty and surprise. I showed the evidence of eligibility trace in human learning using pupil dilation measurement. With EEG recording, I confirmed that the RPE is reflected in the amplitude of FRN (time window of 280-390ms after the state onset), for both directly rewarded and non-directly rewarded states. I proposed a new RL model, called SurNoR, using novelty as the intrinsic reward and surprise as the learning modulator, to explain human learning where no external reward is provided. The novelty signal is found to be reflected between 80-130ms after the state onset in EEG waveform. The surprise signal occurs later than the novelty signal, which is reflected between 150-210ms after the state onset. By using the sequential decision-making paradigm, this thesis extends the EEG observations of RPE and SPE signals from simple one-step tasks to complex multi-step decision-making tasks.

Key words: reinforcement learning, reward prediction error (RPE), state prediction error (SPE), surprise, novelty, eligibility trace, sequential decision-making, electroencephalogram (EEG), event-related potentials (ERP)

Résumé

En apprentissage par renforcement (AR), un agent prend des décisions séquentielles pour maximiser la récompense qu'il peut obtenir d'un environement. Pendant l'apprentissage, les résultats actuels et attendus sont comparés pour dire si une décision était bonne ou mauvaise. La différence entre le résultat actuel et le résultat attendu est l'erreur de prédiction. L'erreur de prédiction peut être classée en deux types : l'erreur de prédiction de récompense (EPR) et l'erreur de prédiction d'état (EPE), qui peuvent servir de signaux d'apprentissage dans les modèles d'apprentissage par renforcement.

L'erreur de prédiction de récompense (EPR), c'est-à-dire la différence entre la récompense actuelle et la récompense attendue, est l'une des variables cruciales en apprentissage par renforcement sans modèle. Chez l'humain, il a été montré que l'EPR est générée par le système de dopamine situé dans le mésencéphale. Des études électroencéphalogramme (EEG) ont également montrées que l'EPR peut être reflétée par une onde EEG se produisant dans la région cérébrale frontale-centrale, entre 250 et 400ms après l'affichage d'un signal de récompense. Cette forme d'onde liée à l'EPR est appelée la négativité liée à la rétroaction (NLR). La plupart des études traitant de la NLR utilisent le problème du bandit manchot pour étudier la relation entre l'amplitude de la NLR et l'EPR. Dans le problème du bandit manchot, les participants recoivent la récompense immédiatement après que l'action a été prise. Cependant, les situations d'apprentissage par renforcement quotidiennes comportent généralement de nombreux états et actions non récompensés jusqu'à ce qu'une récompense soit obtenue. La première partie de cette thèse vise à déterminer si la NLR est toujours reliée à l'EPR dans les tâches complexes.

L'erreur de prédiction d'état (EPE) mesure de combien diffère l'attente de l'agent sur les transitions d'état avant et après une observation. La nouveauté et la surprise sont deux types de signaux EPE qui stimulent l'apprentissage lorsque la récompense externe n'est pas encore fournie. La nouveauté mesure la fréquence à laquelle une observation se produit, tandis que la surprise mesure le degré de violation des attentes. Des études EEG ont montrées que la nouveauté peut se refléter dans différentes composantes EEG, tels que N1 et P300. La nouveauté peut même être observée chez les enfants lorsqu'ils apprennent un nouveau stimulus. Les algorithmes AR avec un module supplémentaire basé sur la nouveauté ont montrés un bon comportement d'exploration dans l'apprentissage du problème du bandit manchot et des tâches de décision markovien simples. La surprise, d'autre part, est également

utilisée comme signal d'apprentissage dans de nombreux modèles AR basés sur la surprise. La négativité de discordance (ND) observée en EEG est généralement considérée comme une signature neuronale de surprise. Cependant, la façon dont la nouveauté et la surprise interagissent et contribuent à l'apprentissage est restée sans réponse. Dans cette thèse, j'ai étudié les corrélats neuronaux de la nouveauté et de la surprise dans une tâche de prise de décision séquentielle, et j'ai proposé un modèle combinant à la fois la nouveauté et la surprise pour expliquer l'apprentissage humain.
J'ai mis en œuvre différentes tâches de prise de décision séquentielle pour étudier quatre signaux AR, qui sont la trace d'éligibilité, l'EPR, la nouveauté et la surprise. J'ai montré les preuves de traces d'éligibilité dans l'apprentissage humain en utilisant la mesure de la dilatation des pupilles. Avec l'enregistrement EEG, j'ai confirmé que l'EPR se reflète dans l'amplitude de la NLR (fenêtre temporelle de 280 à 390 ms après le début de l'état), pour les états directement récompensés et non directement récompensés. J'ai proposé un nouveau modèle AR, appelé SurNoR, utilisant la nouveauté comme récompense intrinsèque et la surprise comme modulateur d'apprentissage, pour expliquer l'apprentissage humain où aucune récompense externe n'est fournie. Le signal de nouveauté se reflète entre 80 et 130 ms après le début de l'état dans la forme d'onde EEG. Le signal de surprise survient plus tard que le signal de nouveauté, qui se reflète entre 150 et 210 ms après le début de l'état. En utilisant le paradigme de prise de décision séquentielle, cette thèse étend les observations EEG des signaux EPR et EPE de simples tâches en une étape à des tâches complexes en plusieurs étapes.

Mots clefs : apprentissage par renforcement, erreur de prédiction de récompense (EPR), erreur de prédiction d'état (EPE), suprise, nouveauté, trace d'éligibilité, prise de décision séquentielle, électroencéphalogramme (EEG), potentiel lié à l'événement (PLE)

Contributions

Eligibility Trace in Sequential Decision-Making
In this study, I contributed to:

- Implementing and conducting the experiment
- Analysing the data
- Discussing and interpreting the results
- Writing the manuscript

Neural Correlates of the Reward Prediction Error
In this study, I contributed to:

- Designing, implementing and conducting the experiment
- Analysing the data
- Discussing and interpreting the results
- Writing the manuscript

Neural Correlates of the State Prediction Error
In this study, I contributed to:

- Designing, implementing and conducting the experiment
- Analysing the data
- Discussing and interpreting the results
- Writing the manuscript

Curiosity or Reward
In this study, I contributed to:

- Designing, implementing and conducting the experiment
- Analysing the data
- Discussing and interpreting the results

Contents

List of Figures

1 Introduction

1.1 Reinforcement Learning Theory

Learning is defined as the 'modification of a behavioural tendency by experience' according to the Merriam Webster dictionary (Webster, 2008). Learning can occur in humans, animals, machines, and even plants (Karban, 2015). When we think of human and animal learning, the first idea is that learning occurs from the interactions with the surrounding environment. Reinforcement learning (RL) describes such learning behaviours. In RL, humans and animals that interact with environments are usually referred to as agents. The agent aims to obtain rewards from the environment. Reinforcement learning is different from supervised learning because there is no label for each action telling whether it is good or not. It is also different from unsupervised learning because it is driven by the desire to maximise the reward the agent can obtain from an environment (Klopf, 1972), instead of finding the hidden structure of the environment. The history of RL can be tracked back to the 19th century and many of the RL models are inspired by psychology and neuroscience studies. I will describe a brief history of RL and some classical RL models in this section and RL-related physiological studies in the next section.

1.1.1 A brief history of RL

The family tree of RL has two main branches. One branch started with the psychology of animal learning and the other started from computational-based solutions to the RL problems.

The first and major branch of RL started from the idea of 'trial-and-error learning'. It was a term used by C. L. Morgan in 1894, a British psychologist, to describe animal behaviours when they learn from past experience (Woodworth, Barber, & Schlosberg, 1954). Then in 1911, E.L. Thorndike pointed out that trial-and-error learning is essential for animal learning (Thorndike, 1911). Based on this, Thorndike came up with the 'Law of Effect', stating that behaviours followed by pleasant outcomes tend to be repeated and that behaviours followed by unpleasant outcomes tend to be avoided. In 1927, Oxford University Press translated

Pavlov's famous study of classical conditioning into English (Pavlov, 1927), and used the term 'reinforcement' to describe a similar learning behaviour as the 'Law of Effect'. This might be the origin of the word 'reinforcement' used in the context of 'reinforcement learning'. The 'Law of Effect' encouraged many researchers to implement trial-and-error learning in electro-mechanical machines in the 1940s and 1950s (Deutsch, 1954; Minsky, 1954; Ross, 1933; Shannon, 1952; Turing, 1948; Walter, 1950).

About the same time, the other branch of RL started to grow. This branch has its origin in the optimal control theory in the mid 1950s. The optimal control theory is used for finding a control law for a system that optimises an objective function over time. Richard Bellman and others used two componnets, the state of a dynamical system and a value function, to define an equation to solve the optimal control problems. This equation is the famous *Bellman Equation* used in dynamic programming (Bellman, 1957b). Bellman also introduced the Markovian Decision Processes (MDPs) as a discrete stochastic version of the optimal control problem (Bellman, 1957a). It later became an important part of the RL theory. Since the late 1950s, dynamic programming, which is considered the only feasible way to solve stochastic optimal control problems, was extensively developed (Bryson, 1996 for review).

In the 1960s, a very important sub-branch sprouted on the main branch of animal learning. The Russian scientist M.L. Tsetlin developed a method for solving the N-armed bandit task (Tsetlin, 1973). The N-armed bandit task is an analogy to a slot machine with N levers, and is very widely used in RL experiments nowadays. The problem of N-armed bandit and the methods for solving it were later extended into the field of economics and game theory.

There is also a third branch on the RL family tree, which is supported by the other two main branches. This branch consists of temporal-difference learning (TD-learning) methods which is unique to RL. TD-learning describes learning driven by the difference between temporally successive evaluations of the same event. For example, an apple has a weight at time t, and a new weight at time t+1, the difference between the two weights at the two time points drives TD-learning in the apple weight evaluation. TD-learning is originated from the animal learning psychology. A. Samuel was the first to implement the TD-learning idea in his checkers playing patent (Samuel, 1959).

In the modern age of reinforcement learning, H. Klopf brought together the ideas of trial-and-error learning and TD-learning in 1972 (Klopf, 1972). Later R.S. Sutton developed Klopf's idea linking it to animal learning (Sutton, 1978a, 1978c), where he described learning rules driven by prediction changes in temporally successive events. There was a vast amount of research and methods coming out in the 1980s using the combined idea of trial-and-error learning and TD-learning, including the Actor-Critic methods, TD(0) and TD(λ) methods (Barto, Sutton, & Anderson, 1983; Sutton, 1988; Witten, 1977). In 1989, Chris Watkins published his PhD thesis about Q-learning (Watkins, 1989), which fully brought together the optimal control and the TD-learning branch. After this, the family tree of RL grew broadly and lushed in the fields of neuroscience, machine learning, and artificial intelligence.

After Sutton and Barto published their book 'Reinforcement Learning: An Introduction (1st Edition)' (Sutton & Barto, 1998) in 1998, the interdisciplinary field joining the RL algorithms and RL in neural system has been developed fruitfully. The relationship between TD-learning and the dopaminergic neuron activity was found by many researchers (Barto, 1995; Friston, Tononi, Reeke, Sporns, & Edelman, 1994; Houk, Davis, & Beiser, 1994; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997) and encouraged later in depth research using modern techniques such as EEG and fMRI. The neural correlates of RL signals will be discussed in detail in the next section.

1.1.2 State-of-the-art RL models

In this section, I will introduce the RL concepts and models that are used in this thesis.

RL focuses on reward-directed (or goal-directed) learning from the agent's interaction with an environment. The agent performs actions in the environment and receives outcomes from it. The outcome can be a reward, a punishment, or neutral. The aim of the agent is to maximise the reward it can obtain from the environment. However, the action chosen by the agent does not only affect the immediate reward but also the subsequent rewards, which makes it a difficult problem to solve for classic machine learning models.

There are six main concepts, or elements, in an RL system. They are the *agent*, the *environment*, the *policy*, the *reward signal*, the *value function*, and the *model of the environment*. An *agent* can be a human, an animal or a computer program that interacts with a given environment. An *environment* contains a number of states, which are presented to the agent one at a time. The agent's objective is to maximise the total *reward* it can receive in the long term. After the agent makes an action, the reward signal given by the environment tells immediately whether the outcome is good or bad. Different from the reward signal, a *value function* tells what is good in the long run. The value of a state tells about total amount of accumulated reward in long term the agent can expect, if the agent starts from that state. Mathematically, the value function is defined as:

$$v(s) = E\left[\sum_{k=0}^{\infty} \gamma^k \cdot R_{t+k+1} | S_t = s\right] \tag{1.1}$$

where s is the state that the agent is in at time t, R_t is the reward signal that the agent receives at time t, γ is the discount rate for future reward at present time. If the agent knows the reward R_{t+1} at time $t+1$, this reward is worth $\gamma^0 \cdot R_{t+1} = R_{t+1}$ to the agent. Similarly, the reward R_{t+2} at time $t+2$ is worth $\gamma^1 \cdot R_{t+1}$ to the agent at time t. Usually γ has a value between 0 and 1. When γ approaches 0, the agent considers future values less important to the current decision. When γ approaches 1, the agent considers the future value more important to the current decision.

Equation 1.1 can also be expressed in the form of state-action value $q(s,a)$ when the agent takes action a at state s as:

$$q(s,a) = E\left[\sum_{k=0}^{\infty} \gamma^k \cdot R_{t+k+1} | S_t = s, A_t = a\right] \tag{1.2}$$

After evaluating the state-action value $q(s,a)$, the agent needs to make an action to maximise the expected reward it can get. The agent follows a *policy* π when it makes an action. A *policy* defines the way the agent behaves. For example, the policy can be a lookup table that tells the agent which action to choose at a given state, or a stochastic function telling the probabilities of choosing an action. Under a different policy π', the state value $v(s)$ and state-action value $q(s,a)$ would be different. Thus, we usually use $v_\pi(s)$ and $q_\pi(s,a)$ to represent the values when the agent uses policy π.

There are two main types of RL models, one is model-free and the other is model-based. Model-free methods are closely related to the 'trial-and-error' learning, which learns the value functions of the states in an environment. Different from model-free methods, model-based methods learn a *model of the environment*. The environment model predicts the environment's response to the agent's action. The environment model contains two parts, the state-transition and the reward model. In state-transition models, the agent learns transitions between states via actions. The reward model estimates the expected reward at each state. When taking an action at a given state, the agent first uses the state-transition to predict which states are upcoming, and then uses the reward model to tell the expected reward from future states. The difference between the model-free and model based methods is that the policy of model-free methods is deduced from the value functions, and the policy of model-based methods is deduced from the environment model.

Model-free RL models – Sarsa & Q-Learning

The two model-free RL models used in this thesis are SARSA(λ) and Q(λ). The learning signal in both models is the reward prediction error (RPE), defined as the difference between the actual reward and the predicted reward. Equation 1.3 shows the RPE computed for SARSA(λ) and equation 1.4 for Q(λ):

$$RPE_t = r + \gamma Q(s_{t+1}, a_{t+1}) - Q(s,a) \tag{1.3}$$

$$RPE_t = r + \gamma \cdot \max Q(s_{t+1}, a^*) - Q(s,a) \tag{1.4}$$

where γ is a discounting rate parameter of future rewards. A positive RPE indicates that the tendency of selecting action a at state s should be strengthened, whereas a negative RPE indicates that the tendency should be weakened.

The difference between SARSA(λ) and Q(λ) is that SARSA(λ) computes the RPE after taking the action at time $t+1$ following the current policy (usually called on-policy), while Q(λ) computes the RPE at time t based on the optimal $Q(s', a')$ value at current time t under the current policy (called off-policy).

The Q-values, $Q(s,a)$, represent an estimate of the expected future reward when starting in state s, taking action a. This value function is iteratively improved by applying an update after each step:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha \cdot RPE_t \cdot e_t(s,a) \tag{1.5}$$

The quantity $e(s,a)$ is known as a short-term memory (Sutton & Barto, 1998) which implements a decaying memory trace of past state-action pairs with the following dynamics:

$$e_t(s,a) = \begin{cases} \gamma\lambda e_{t-1}(s,a), & \text{if } (s,a) \text{not visited} \\ 1, & \text{if } (s,a) \text{visited} \end{cases} \tag{1.6}$$

$e(s,a)$ marks an event in memory eligible for undergoing learning changes. At each trial, the eligibility trace for all state-action pairs decay by $\lambda\gamma$, where λ is the trace decay parameter.

The Q values calculated in this way are then used to select an action at each state according to a *softmax* policy:

$$P(s,a) = \frac{\exp\left(Q_t(s,a)/\tau\right)}{\sum_i \exp\left(Q_t(s,i)/\tau\right)} \tag{1.7}$$

where $P(s,a)$ defines the probability of choosing action a at state s, τ is the temperature parameter which controls the tendency of exploration and exploitation, and i presents all possible actions at state s.

These equations define the learning model with up to four free parameters: the learning rate α, the discount rate γ, the eligibility decay rate λ, and the temperature τ.

Model-based RL model – Forward-Learner

The model-based RL model used in this thesis is the 'Forward-Learner' published by Gläscher (Gläscher, Daw, Dayan, & O'Doherty, 2010). It uses the experienced state transitions to update the state transition matrix $T(s,a,s')$ by:

$$SPE_t = 1 - T(s,a,s') \tag{1.8}$$

where s' is the arrival state after taking action a at current state s, and $T(s,a,s')$ is probability of this transition. The matrix T is thus a state transition model of the environment. After observing state s', the transition matrix $T(s,a,s')$ is updated by:

$$T(s,a,s') = T(s,a,s') + \eta \cdot SPE_t \tag{1.9}$$

where the parameter η is the learning rate. For the states s^* that are not observed when transitioning from s to s' taking action a, the transition probabilities are:

$$T(s,a,s^*) = T(s,a,s^*) \cdot (1-\eta) \tag{1.10}$$

In this model-based model, the state-action value $Q_{FWD}(s,a)$ is computed as:

$$Q_{FWD}(s,a) = \sum s' T(s,a,s') \cdot \left[r(s') + \arg\max_{a'} Q_{FWD}(s',a') \right] \tag{1.11}$$

A softmax policy is used for action selection using $Q_{FWD}(s,a)$ in Equation 1.7.

1.2 The Neural Correlates of Reinforcement Learning

1.2.1 The Reward Prediction Error

The development of RL theory is closely related to the animal and human learning research. The relationship between RL theory and animal learning is strengthened after the discovery of dopaminergic neuron activity was found to be explained by the TD errors. This finding initiated the *Reward Prediction Error Hypothesis* (RPEH) of dopamine. This hypothesis proposes that the phasic activity of dopaminergic neurons code for an error between old and new estimates of expected future reward. This error is delivered throughout the brain. In 1996, Montague, Dayan, and Sejnowski published the first study supporting this hypothesis. An experiment ran by W. Schultz (Schultz et al., 1997) in the 1980s and early 1990s, although published in 1997, showed clearly how the TD errors (which are considered as RPEs in this thesis) were aligned with the phasic activity of dopaminergic neurons.

A decade later, Holroyd and Coles (Holroyd & Coles, 2002) published the RL-ERN hypothesis, stating that an error-related component in EEG is generated in the anterior cingulate cortex (ACC) via the mesencephalic dopamine system when a negative reinforcement learning

signal occurs. In this hypothesis, the error related EEG component is called the error-related negativity (ERN). This component is also called the feedback-related negativity (FRN) and the feedback-error-related negativity (fERN). I use the term 'FRN' to refer to the prediction error related EEG component in this thesis. The FRN is observed in frontal-central electrodes and is thought to be generated from the ACC when there are phasic changes in dopamine signals. When a negative prediction error occurs, the phasic decrease in dopamine activity disinhibits ACC neurons, producing a more negative FRN. When a positive prediction error occurs, the phasic increase in dopamine activity inhibits ACC neurons, producing a more positive FRN.

Since the phasic activity of dopaminergic neurons are thought to reflect the RPEs, and FRN amplitudes are thought to reflect the dopamine activity, I proposed that the FRN amplitudes can reflect RPEs in this thesis.

In FRN studies, researchers mostly use probabilistic learning experiments, which are equivalent to the N-armed bandit tasks as introduced in section 1.1.1. Figure 1.1 shows an example of the FRN amplitudes when receiving reward with different probabilities (Walsh & Anderson, 2011b). In the study of (Walsh & Anderson, 2011b), there were three stimuli and each stimuli was associated with different reward probabilities ($P_{reward} = 0\%, 33\%, 66\%$). The FRN occurs in the time window from 200 to 400ms after the stimulus onset. Some studies consider the FRN as the ERP waveform in this time window, others consider FRN as the waveform difference between rewarded (win) and non-rewarded (loss) conditions. In this thesis, I consider FRN as the ERP waveforms in the 250-400ms time window. FRN amplitudes differ in four feedback conditions: improbable loss, probable loss, probable win and improbable win. Using the RL theory, these four conditions can be converted into four scales of RPEs. Improbable loss occurs when expected loss is low while the actual loss is high, which produces a high negative RPE. A probable loss occurs when expected loss is similar to the actual loss, which produces a low negative RPE. A probable win occurs when expected win is similar to the actual win, which produces a low positive RPE. An improbable win occurs when expected win is lower than the actual win, which produces a high positive RPE. By comparing the RPEs in the four conditions to the FRN amplitudes, we find the similar trend that *RPE(improbable loss) < RPE (probable loss) < RPE (probable win) < RPE(improbable win)*, and also *FRN(improbable loss) < FRN(probable loss) < FRN(probable win) < FRN(improbable win)*.

Similar experiments were performed in most FRN studies, see (Walsh & Anderson, 2012) for a review. The topography of the FRN shows a high activity in the prefrontal cortex. EEG source localization indicates that the FRN is generated from the ACC (Bellebaum & Daum, 2008; Cohen, Elger, & Ranganath, 2007; Gehring & Willoughby, 2002; Gruendler, Ullsperger, & Huster, 2011; Hewig, Hecht, et al., 2007; Mathewson, Dywan, Snyder, Tays, & Segalowitz, 2008; Miltner, Braun, & Coles, 1997; Nieuwenhuis, Slagter, Von Geusau, Heslenfeld, & Holroyd, 2005; Potts, Martin, Burton, & Montague, 2006; Ruchsow, Grothe, Spitzer, & Kiefer, 2002; Tucker, Luu, Frishkoff, Quiring, & Poulsen, 2003; Zhou, Yu, & Zhou, 2010), which is compatible with Holroyd and Coles hypothesis. There are rarely any FRN studies using sequential decision-making tasks or Markovian Decision Processes (MDPs) tasks. It could be because the N-armed

bandit task is easy to implement, reward is easy to control, and the RPEs are straightforward to compute in such tasks. To our knowledge, the only study used a two-step decision-making task using FRN analysis is by (Walsh & Anderson, 2011a). However, in this task, the authors did not compared the FRN amplitudes with the RPEs on trial-by-trial basis, but only compared the correlation between the averaged FRN amplitude with the averaged RPE.

In this thesis, by using sequential decision-making paradigm (Tartaglia, Clarke, & Herzog, 2017) and RL methods to compute RPEs at different states, I confirmed that the relationship between FRN amplitude and the RPE holds true not only for simple N-armed bandit task, but also for complex sequential decision-making tasks.

1.2.2 The State Prediction Errors — Surprise and Novelty

Besides the RPE, I also studied two SPE signals, namely surprise and novelty. In the literature, surprise signals are usually studied using an oddball task. In this task, participants are presented with a sequence of stimuli in a repeated pattern, such as *'AAB AAB AAB AAB...'*. After participants get used to the pattern, the sequence is interrupted by an infrequent stimulus and becomes for example *'AAB AAB AAB AA**C** AAB...'*. The stimulus *'AA**C**'* is unexpected and triggers a surprise signal to participants. The EEG waveform that reflects such surprise signal is called the Mismatch Negativity (MMN). The MMN was observed in both auditory oddball tasks (Näätänen, Gaillard, & Mäntysalo, 1978) and visual tasks (Cammann, 1990). Since the experiments in this thesis are done using visual stimuli, I will only focus on the visual Mismatch Negativity (vMMN) here. The features of different visual surprise rises vMMN in different latency and locations. The visual stimulus used to trigger vMMN differ in sizes, shapes, motions, orientation, and contrasts etc. The latency varies from 75ms to 450ms after the stimuli onset, and the vMMN can be observed in frontal and occipital electrodes. For a detailed review of vMMN, see (Pazo-Alvarez, Cadaveira, & Amenedo, 2003).

Another type of oddball task, called the novelty oddball, is used to study brain responses to novel signals. In the novelty oddball task, three different stimuli are presented to subjects. The three stimulus are a stimuli occurs with high probability, a stimuli with low probability and an improbable unexpected 'novel' stimuli which is used to rise the response to novelty. The ERP triggered by this novel stimulus can be observed in frontal electrodes around 300ms after the stimulus onset. The component is defined as the 'P3a' component. After subjects habituate to the stimuli, the 'P3a' amplitude attenuates (Courchesne, Hillyard, & Galambos, 1975; Lynn, 2013; Sokolov, 1990).

The SPE signal, such as surprise and novelty, can be considered to be produced by the belief-updating process. A belief-updating related ERP component is the N1 component, hypothesised by KJ. Friston in 2018 (Friston, Rosch, Parr, Price, & Bowman, 2018). Friston used a multi-hierarchy generative model to present the process of belief updating, and simulated the EEG response in a sentence-reading task. The model learns the associations between words in a sentence. When a word changes in the sentence, it rises two violations, a local violation (the

word itself) and a global violation (the meaning of the sentence) in the learning model. The simulated EEG signal showed two peaks after the word change. One occurs around the N1 latency (100ms after stimulus onset) that reflects the local violation. The other occurs as a late peak at 300ms after stimulus onset (P300), similar to P3a, that reflects the global violation.

The three ERP component, vMMN, P3a, and N1, related to the surprise and novelty in literature, provide the guidance to look for bio-markers of state prediction errors in this thesis.

1.3 A Sequential Decision Making Paradigm

The experimental tasks used in this thesis are adapted from the sequential decision making paradigm proposed by (Tartaglia et al., 2017). In this paradigm, states are represented by clip-art images, actions are presented by grey disks under each image (Figure 1.2). At the beginning of each experiment, participants are informed about the goal state image and are asked to find the goal image for 5 times. For each image, clicking at the same disk leads always to the same subsequent image, i.e., the state-action transitions are deterministic. In other words, the state-action transition matrix, which defines the environment, contains only ones and zeros.

During a trial (Figure 1.3), an image (state) is shown for an interval between 700 to 1700ms (uniformly random) and then the grey disks appear while the image stays on the screen. Disks are shown until participants click on one of them (action). There is no time limit for making an action. After an action, a blank screen is presented with a randomly chosen duration between 700 to 1700ms. Then, the next image is shown and so on.

In this thesis, I designed two types of environmental structures. One is the complex structure shown in Figure 1.4 and the other is the simple structure shown in Figure 1.5.

The complex environmental structure contains three types of states: (1) a goal state, which is the immediately rewarded state; (2) several progressing states (state 1, 2, 3, 4 in Figure 1.4A), which lead participants to the goal state; and (3) several trap states (state 5, 6 in Figure 1.4A). If participants come to one of the trap states, they need to find the correct action that leads them back to the first progressing state, which is furthest away from the goal (state 1 in Figure 1.4A). At each non-goal state, there are three types of actions. (1) One type of action brings participants to the next progressing state, which is closer to the goal (green arrows in Figure 1.4A). (2) One type of action brings participants to one of the trap states (blue arrows in Figure 1.4A), where participants have to find the way to the first progressing state. (3) One type of action let participants stay at the current state (yellow arrows in Figure 1.4A).

Different from the complex environments, the structure of the simple environment only allows participants to follow a given path to the rewarded state. The structure of the simple environment is presented in Figure 1.5. The environment contains a goal state and several non-goal states. The goal state contains an immediate reward. Each non-goal state contains

a number of actions that participants can choose. Among these actions, there is only one action can lead participants to the next state (green arrows in Figure 1.5). Other actions only let participants stay at the current state (yellow arrows in Figure 1.5).

Before starting the experiment, I showed the participants the goal image that they needed to find. Then, participants were presented the other non-goal images that they may encounter the experiment. After seeing the images on the screen, participants clicked the 'start' button to start the experiment. Participants clicked through the environment until they found the goal state. An episode was finished when participants found the goal state.

1.4 Current Research

Every day learning is far more complex than the N-armed bandit task used in previous RL-EEG studies. Humans usually need to make a sequence of actions to obtain a reward. For example, when a person is hungry and wants to eat, he or she first needs to walk to the fridge (first action), then to open the door of the fridge (second action), take the cold food out (third action), and to warm it up (fourth action). Only after these actions are taken in a proper sequence and at the right state, he or she can get the final reward, which is the eatable food.

The question to be answered in this thesis is whether the observations and conclusions drawn from the simple N-armed bandit tasks can be generalised to complex tasks. Recent researchers have started to use sequential tasks to study the neural correlates of RL components, such as the RPE, reward and SPE. Glaescher (2010) used a two-stage decision-making task to dissociate the RPE from the SPE. However, in Glaescher's experiment, participants learned the state transitions and reward mappings in separate stages. Especially in the reward mapping learning phase, participants did not make any actions to obtain reward but only observed associations between reward and states. In this design, although participants made sequential decisions to obtain the reward, they lack the opportunity to learn the environment from scratch. Another two-step task, introduced by Daw et al (Daw, Gershman, Seymour, Dayan, & Dolan, 2011), also aimed to distinguish model-based and model-free RL aspects of human RL. In Daw's task, one state was presented to participants at a time. Participants made an action at the given state and were led to the second state. At the second state, participants made an action again to obtain the reward. The state transitions were stochastic in Daw's two-stage task. The fMRI results showed that both model-based and model-free learning signals can be observed from the striatum and prefrontal cortex.

Both Glaescher and Daw studied the neural sources of RPEs and SPEs using fMRI. However, studies about temporal resolution of the prediction errors occur are still largely lacking. Sambrook (Sambrook, Hardwick, Wills, & Goslin, 2018) adapted Daw's two-stage task to study the time window of prediction errors using EEG recording. Participants made one action at the first state, and then observed the following state and final reward. Although the task contained two states, the fact that participants made only one action made the task similar to a N-armed bandit task.

Previous researches revealed the neural sources of RPEs and SPEs. However, there are two limitations. The first limitation is that the tasks used in previous researches were too simple, where participants only made one or two actions to obtain a reward. Whether the RPEs and SPEs observed can be generalised into complex learning tasks remained unanswered. The second limitation is that the temporal information of the RL signals were not thoroughly studied in sequential learning tasks.

To address the two limitations, I implemented a truly sequential decision-making task based on the paradigm by (Tartaglia et al., 2017). I recorded and analysed the EEG signals when participants performed the task. With the help of classic RL models and the newly developed SurNoR model (see Chapter 4 for details), I identified the time windows of the RPE, surprise and novelty signals in sequential RL tasks.

1.5 Aims of This Thesis

In this thesis, I employed three sequential decision making experiments, aiming to study four RL signals. The four signals I studied in this thesis are: the eligibility trace, the reward prediction error (RPE), surprise and novelty.

The first experiment (Chapter 2) aimed to study the evidence of eligibility trace in human learning. The experiment was in collaboration with Dr. Marco Lehmann in the Lab of Computational Neuroscience. We used an adapted version of Tartaglia's paradigm (Tartaglia et al., 2017) and recorded both pupil dilation and EEG during the experiment. The results of this study showed evidence of the eligibility trace in human behaviours and pupil dilation (Lehmann et al., 2019).

In the second study (Chapter 3), I designed six simple learning environments and two complex environments, aiming to study the relationship between the FRN amplitude and the RPE. Participants explored the environments by making actions at each states, in order to find the rewarded goal state. The RPEs of non-goal states fluctuated during learning, which made it difficult to search for corresponding time window using the RPEs of non-goal states. The RPEs of the goal states decreased monotonically from beginning to end. Thus, I used this monotonic trend to search for the EEG time window of RPEs at the goal states. The time window I found was between 250 to 400ms after the state onset in both simple and complex environments. This time window is very close to the FRN introduced in section 1.2.1. Then I tested if the mean amplitudes in this time window reflect the RPEs of the non-goal states. I computed the RPEs of non-goal states using the SARSA(λ) model. The regression between the estimated RPEs and the mean EEG amplitudes was significant, confirming that the FRN amplitude can reflect the RPEs in sequential decision-making tasks. This results generalised the RL-ERN theory (Holroyd & Coles, 2002, section 1.2.1) from N-armed bandit task to sequential tasks.

The third project (Chapter 4) contained two blocks, aiming to study the time course of surprise and novelty. In the first block, participants were asked to find the goal state in a complex

environment. The environment contained several trap states. If fell into the trap states, participants needed to find the path to exit. The trap states made it difficult for participant to find the goal. However, I observed that participants can learn the structure of the environment and avoid trap states very quickly. Even before seeing the goal state, participants were able to find the correct path to the goal. Participants' behaviour in the un-rewarded stage can neither be explained by model-free nor model-based RL methods. Model-free methods cannot compute the values of states because there were neither reward nor RPE. Thus, a model-free RL agent only makes random actions before seeing the goal. Model-based methods can only learn the state transitions but not the reward mapping without a reward (Blodgett, 1929). Similar to the model-free agent, a model-based agent also makes random actions in this case. I proposed that the learning is driven by the novelty of the states when reward is absent. To find out when novelty is processed in the human brain, I compared the ERP between frequently visited states and rarely visited states. States that were visited more often had low novelty. States that were visited less often had high novelty. The two ERP curves differed significantly in the time window from 80 to 130ms after the state onset. I computed the mean amplitudes in the selected time window, and estimated the novelty using the SurNoR model. The correlations between the EEG amplitudes and estimated novelty were significant, confirming that the EEG amplitude in the time window between 80 and 130ms is a potential marker for the novelty signal. In the second block of the experiment, I swapped the images of two states without informing participants, aiming to rise surprise signal to participants. By comparing the ERP between surprised trials and un-surprised trials, I found that the amplitude in the time window from 150 to 210ms reflected different surprise level. Then I correlated the EEG amplitudes with the estimated surprise. The correlation is significant, indicating that the time window of 150 to 210ms after the state onset is a potential marker for the surprise signal.

Inspired by the results from the second and third experiments, I also designed another experiment to study the relationship between novelty and reward (Chapter 5). In this experiment, I introduced an 'infinite' state, where participants always see new images at this state. By comparing participants novelty-seeking and reward-seeking behaviours, I want to test which signal is stronger in driving learning. This experiment is still undergoing.

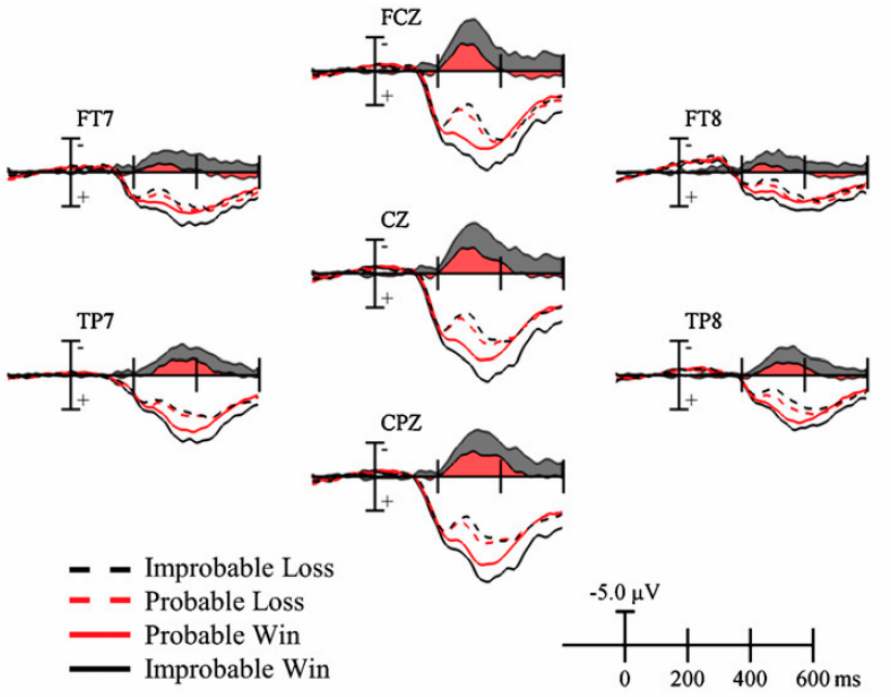


Figure 1.1 – Feedback related negativity (FRN) for improbable and probable wins and losses. Coloured region presents the difference between high probability win/loss and low probability win/loss EEG waveforms. The figure is adapted from the study in (Walsh & Anderson, 2011b).

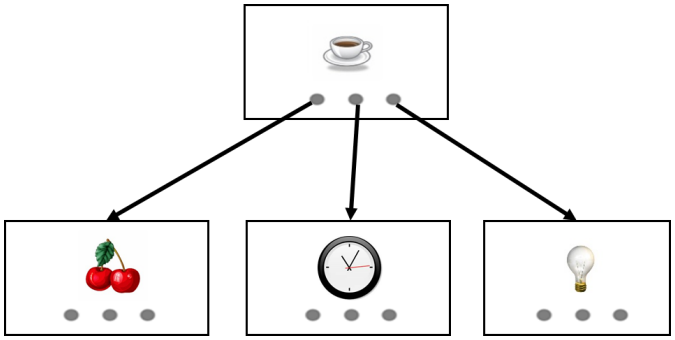


Figure 1.2 – Stimulus presentation in the sequential decision making paradigm. RL states are represented by clip-art images in the centre of the screen. RL actions are represented by grey disks below the images.

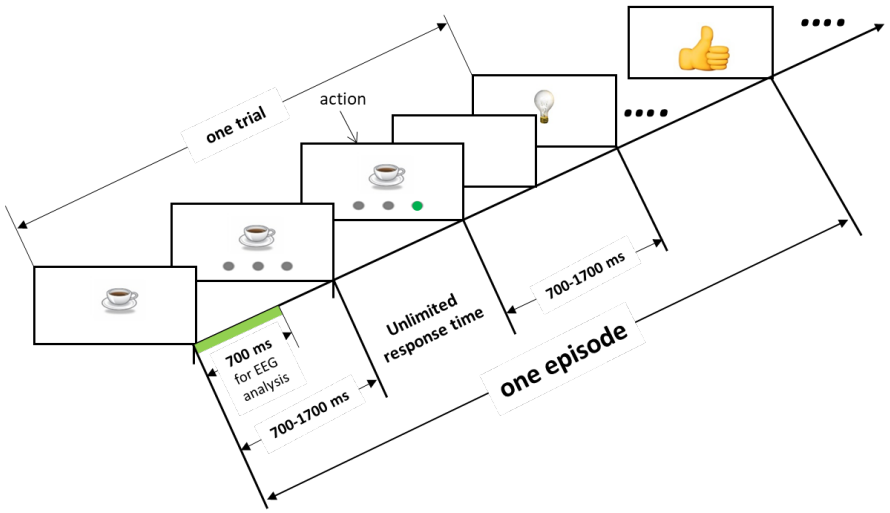


Figure 1.3 – EEG recording during the experiment. An image (state) is presented on the screen. After a random interval of 700-1700ms, grey disks appear, on which participants are asked to click (actions). After an action, a blank screen is shown for a random interval between 700 and 1700ms and then the next state appears. The goal state is a 'thumb-up' image in this example. The green interval indicates the time (0-700ms after the image onset), for which ERP was analysed.

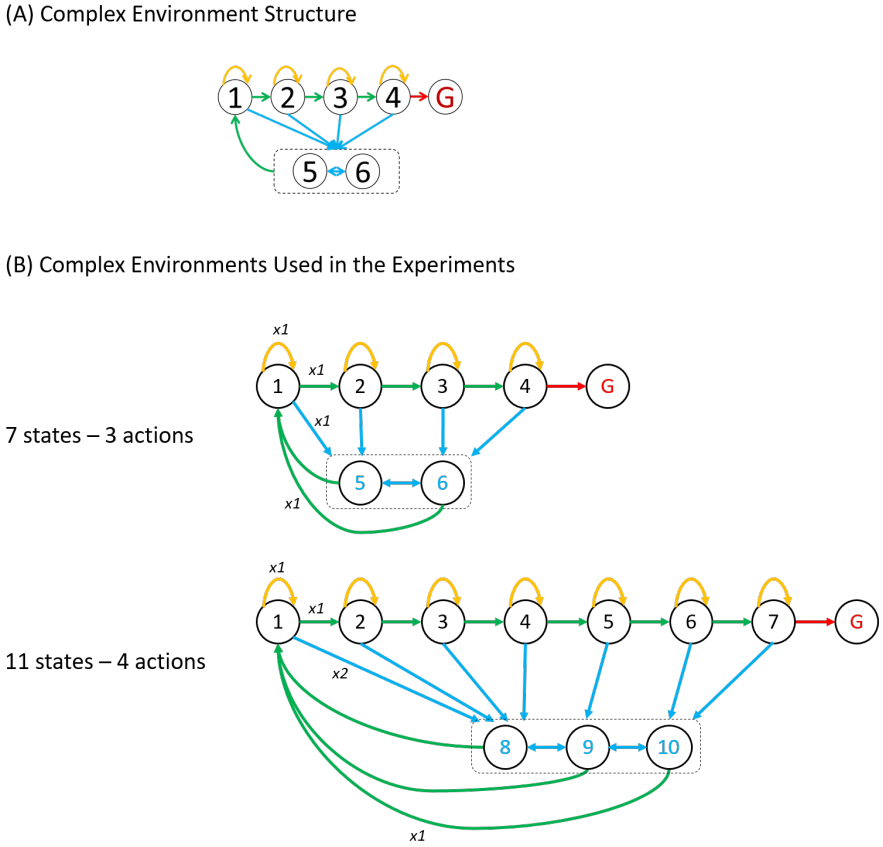


Figure 1.4 – **(A)**The structure of the environment used in the complex experiment. Digits present the non-goal states, red G presents the goal states. Green arrows present the actions that lead participants from one state to the next progressing state. Yellow arrows present actions that let participants stay at the same current state, yellow arrows at state 5, 6 are not shown because of lacking space. Blue arrows present actions that lead participants to one of the trap states. Red arrow present the action that lead participants to the goal state. **(B)** Two complex environments used in the experiments in Chapter 3.

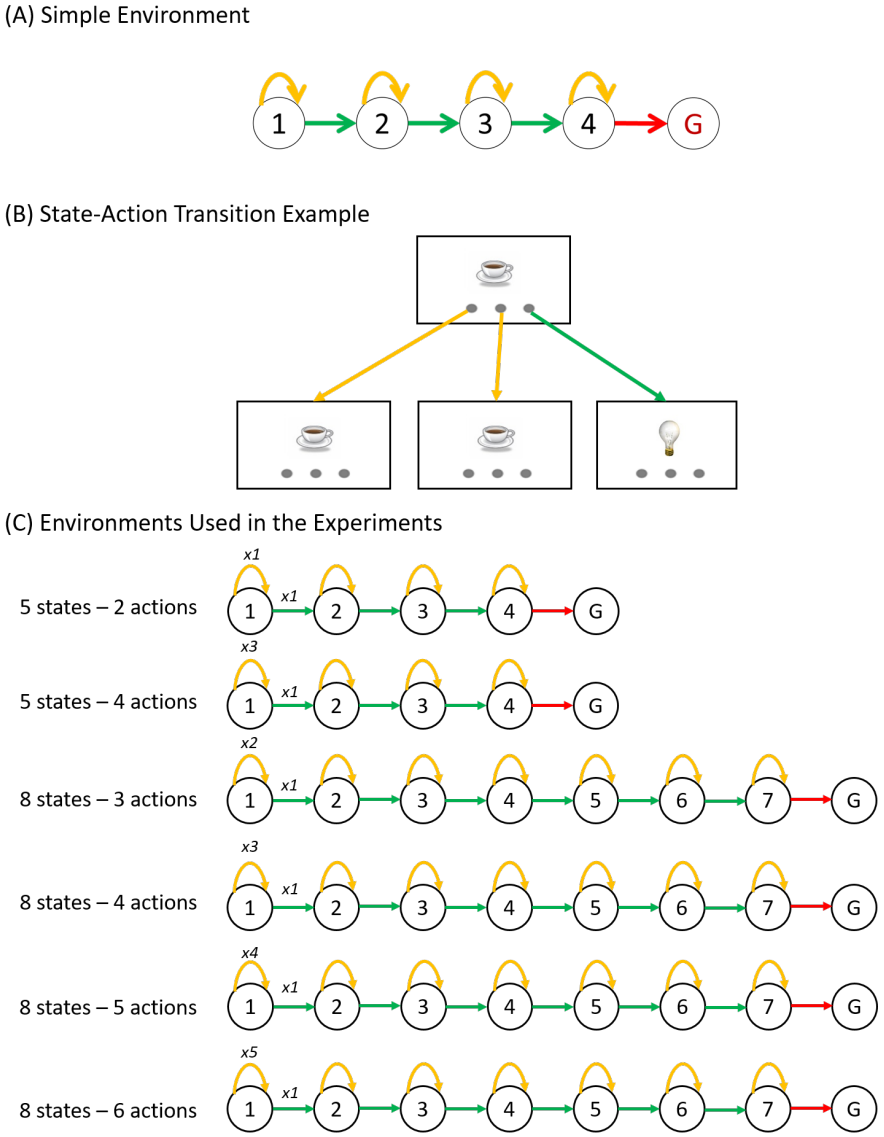


Figure 1.5 – **(A)** The structure of the simple environment. Digits present the non-goal states, red G presents the goal states. Green arrows present the actions that lead participants from one state to the next state. Yellow arrows present actions that lead participants stay at the same current state. Red arrow present the action that lead participants to the goal state. **(B)** An example of state-action transitions is shown in (A). A state is represented by an image and an action is represented by a grey disk. There are three actions at the state 'cup', two of them lead participants to the same state and one leads to a different state. **(C)** Six simple environments used in the experiments in Chapter 3.

2 Eligibility Trace in Sequential Decision-Making

2.1 Preface

Humans and animals interact with the environment to obtain reward. To obtain the maximum amount of reward, they learn from past experience. Reinforcement learning theory describes such learning situations and provides powerful algorithms to model the learning process. When agents (humans, animals or machines) are facing a sequence of decisions before obtaining a reward, a mechanism is needed to map earlier decisions to the final delayed reward. In other words, an agent needs to memorise previously experienced states and actions in order to learn the consequences of past actions. The memory is then used for later learning. RL provides such a memory mechanism, called the eligibility trace, which allows past traces (i.e., experienced states and actions) to be eligible for future decision-making.

Klopf introduced the idea of eligibility traces to RL models in 1972 (Klopf, 1972). Since then, computational models with eligibility traces are well developed (Barto & Sutton, 1981a, 1981b; Barto et al., 1983; Sutton, 1978a, 1978b, 1978c). Physiological experiments also showed evidence of the eligibility trace in synaptic plasticity during learning (Bittner, Milstein, Grienberger, Romani, & Magee, 2017; Fisher et al., 2017; He et al., 2015; Yagishita et al., 2014). Computational models with eligibility traces outperformed those without in explaining human learning (Bogacz, McClure, Li, Cohen, & Montague, 2007; Daw et al., 2011; Tartaglia et al., 2017; Walsh & Anderson, 2011a). However, there is still a lack of direct physiological evidence of eligibility trace in human sequential decision-making.

In this study, we designed a sequential decision making experiment with pupillometry and EEG recordings to show direct evidence for the existence of eligibility traces in human learning. The study is published as *'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'* (see Appendix 1).

2.2 Experimental Design

In TD-learning without eligibility trace (TD(0)), only the last state-action pair is reinforced when the agent obtains a reward after making a sequence of actions. Whereas with eligibility trace, the whole sequence of state-action pairs before reaching the goal is reinforced. The eligibility trace parameter λ controls the memory decay rate for the past state-action sequence. We call the TD-learning models with eligibility trace TD(λ) models. For example, if the state action sequence is '$SA1- > SA2- > SA3- > Goal'$, by using a TD(0) model, only the '$SA3- > Goal'$ transition is reinforced. If we use a TD(λ) model, all the transitions of '$SA1- > SA2- > SA3- > Goal'$ are reinforced to different degrees controlled by the parameter λ.

Based on the difference between TD(0) and TD(λ), we designed an experiment to test if human participants use TD(0) or TD(λ) in sequential decision-making. A special design was used here. In the first episode, no matter which actions the participants took, they were always guided through the state sequence 'S->D2->D1->Goal' (Figure 2.1A). We assumed that participants keep a memory of the experienced state sequence, and tested to which extend the memory traced back. According to TD(0), participants only reinforced her last state-action pair 'D1->Goal'. According to TD(λ), participants also reinforce the previous state-action pairs such as 'D2->D1'. We divided participants into two groups in the second episode. The first group (Figure 2.1 B1) was used to test evidence for TD(0) model. If a participant chose action 'b' (as the example shown in Figure 1B1) at state D1, which was the same action he/she took in the first episode (Figure 2.1A), we can confirm that the transition from state D1 to the goal was reinforced, as predicted by the TD(0) model. The second group (Figure 2.1 B2) was used to test if participants behaviour can be explained by a TD(λ) model. If a participant chose action 'a' (as the example in Figure 2.1 B2) at state D2, which was the same action as in the first episode (Figure 2.1A), we can confirm that the transition from state D2 to D1 was also reinforced. The behaviour cannot be explained by the TD(0).

2.3 Results

The results are presented in the published manuscript *'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'* (see Appendix 1).

2.3.1 Behavioural results

Three conditions were used in the experiment (Figure 2.2A). In the spatial location condition, states were presented by rectangles on different locations of the screen. In the audio condition, states were presented by short sound clip. In the clip-art condition, states were presented by clip-art images in the centre of the screen. We defined the action that participant took from state D1 to the goal as action 'b', and the action that participant took from state D2 to state D1 as action 'a'.

Behavioural performance of action bias is shown in Figure 2.2B. When participants started from state D1 in the second episode (Figure 2.1 B1), we found that participants were more likely to repeat the same action (action 'b') they took in the first episode (Figure 2.2 B1). The bias towards action 'b' confirmed that the transition 'D1->Goal' was reinforced as predicted by TD(0). When participants started from state D1 in the second episode (Figure 2.1 B2), we found that participants were more likely to repeat the same action (action 'a') as they took in the first episode (Figure 2.2 B2). The action selection bias towards action 'a' cannot be explained by TD(0) models, because the transition 'D2->D1' was not reinforced in TD(0) models. TD(λ) explained the action selection bias because with the eligibility trace, the transition 'D2->D1' was reinforced.

2.3.2 Pupil dilation results

In the first episode, when participant saw state D1 (the state before goal state), no reward was obtained yet. In the second episode, participant had already seen the reward (goal state) at the end of the first episode. By comparing the pupil response of state D1 before and after seeing the reward, we can identify the effect of reward. Figure 2.3A shows that after seeing the reward, the pupil response to state D1 was later and the amplitude was higher. We then tested if the pupil response to state D2 ('D2->D1->Goal') is similar to the response to state D1. Figure 2.3B shows that indeed the pupil response to state D2 was later in time and higher in amplitude in the second episode than in the first episode. The statistical t-test between pupil response in the first and second episode at state D2 was significant, confirming the existence of an eligibility trace in human sequential decision-making.

2.3.3 Model fitting

We chose three types of models to fit human behaviours, which are four TD(λ) models, two model-based RL models and two TD(0) models (Figure 2.4). A biased random model was used for baseline comparison. Detailed implementation of all the models tested in Figure 2.4 is described the manuscript (Lehmann et al., 2019).

The second to fourth column in Figure 2.4 shows the model fitting result measured by the Akaike Information Criterion (AIC) for each experimental condition. Lower AIC value indicated better model fitting performance. wAIC in the figure presents the normalised Akaike weights (Gernand & Fenske, 2009), telling the probability of current model being the best model. Higher values of wAIC indicate the corresponded model explained human behaviours the best. wAIC value smaller than 0.01 are not shown in the figure. A Wilcoxon rank-sum test was used to compare the AIC of models. k pairs of individual ranks were used to compare models and to compute the p-values. $p(a)$ presented p-value when comparing each TD(λ) model with the Hybrid model (best performed model without eligibility trace). $p(b)$ presented the p-value when comparing Q-0 model with Q-λ model. $p(c)$ presented the p-value when comparing SARSA-0 model with SARSA-λ model. $p(d)$ presented the p-value when comparing

the biased random model with the Forward Learner model. In the last column, the models were compared aggregating all three conditions using the Wilcoxon rank-sum test. The result showed that TD(λ) models, which were the models with eligibility trace, explained human behaviour better than models without eligibility trace.

2.4 Discussion

In this study, we employed a special experimental design to test if humans use the eligibility trace mechanism in learning a sequential task. Participants explored an environment and obtained a reward at the goal state in the first episode. We tested if participants memorised their past actions at two states (state D1 and D2). According to TD(0) models, which do not utilise an eligibility trace, only the last state-action pair before the goal is reinforced. To the contrary, TD(λ) models predict that not only the last state-action pair before the goal, but also previous state-action pairs, are reinforced. The behaviour, pupil dilation and model fitting results support the hypothesis that humans use eligibility trace when learning a long state-action sequence with delayed reward. Our results confirm the existence of eligibility trace in human learning, which agrees with previous studies (Daw et al., 2011; Gläscher et al., 2010; Niv, Edlund, Dayan, & O'Doherty, 2012; O'Doherty, Cockburn, & Pauli, 2017; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006; Walsh & Anderson, 2011a) and also extends the conclusion to sequential decision-making paradigm.

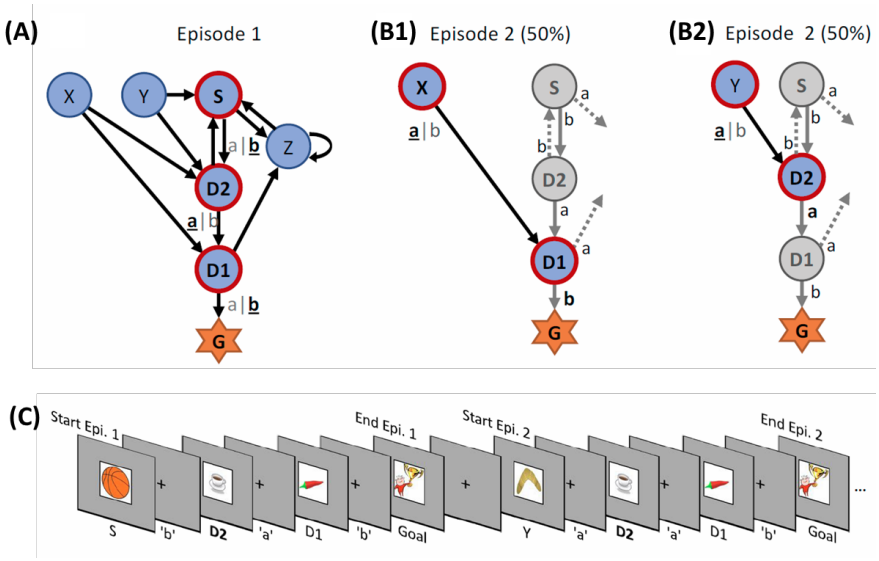


Figure 2.1 – **(A)** The structure of the learning environment in the first episode. Participants started from state S and chose one from the two actions ('a' or 'b'). No matter which action they chose, state D2 was presented after taking the action. Participants chose one of the two actions again at state D2. No matter which action they chose, state D1 was presented after the action was taken. Again, they chose from action 'a' or 'b' at state D1 and the goal (G) was presented after D1. For example, the action sequence taken by a participant in this figure is 'b-a-b' (underlined). However, no matter which action was taken at each state, the state presentation sequence was always 'S-D2-D1-G' to participants in the first episode. **(B)** Participants were divided equally into two groups, one group experienced their second episode as shown in (B1), the other group experienced their second episode as shown in (B2). (B1) Half of the participants started from state X in the second episode. No matter which action ('a' or 'b') they chose, state D1 was presented after X. We tested if participants still chose action 'b' at D1 (as in the first episode) to obtain the goal in this design. (B2) Half of the participants started from state Y in the second episode. No matter which action they chose, state D2 was presented after Y. We tested if participants still chose action 'a' at D2 (as in the first episode) to go to state D1. **(C)** The example of the experimental stimuli in clip-art condition. Starting state S in the first episode was presented by a basketball image. After the image was shown, a fixed point was presented meaning participants needed to choose an action. After the action 'b' was taken, state D2 (a coffee image) was shown. The first episode ended when the goal image was shown. Then the second episode started depends on the two conditions in (B1) and (B2). In the example here, the second episode is in condition (B2). The figure is adapted from (Lehmann et al., 2019).

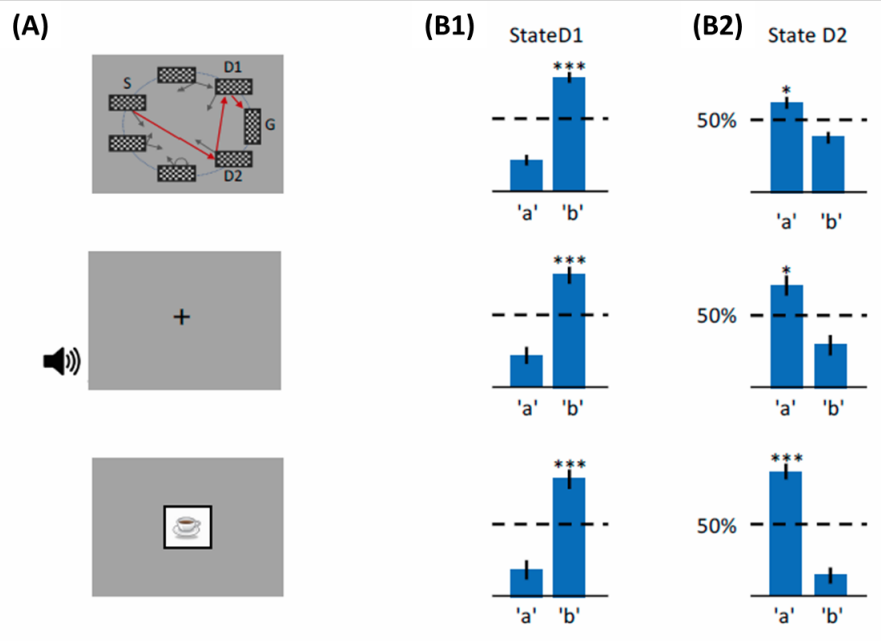


Figure 2.2 – **(A)** The experiment was ran in three different conditions: spatial location condition (states were presented by rectangles in different locations), audio condition (states were presented by different sounds) and clip-art image condition (states were presented by different clip-art image in the middle of the screen). **(B1)** Averaged participants' action selection bias in the second episode corresponding to condition in Figure 1B1. Action 'b' was the action chosen by participants at state D1 to go to the goal state. **(B2)** Averaged participants' action selection bias in the second episode corresponding to condition in Figure 1B2. Action 'a' was the action chosen by participants at state D2 to go to state D1. The figure is adapted from (Lehmann et al., 2019).

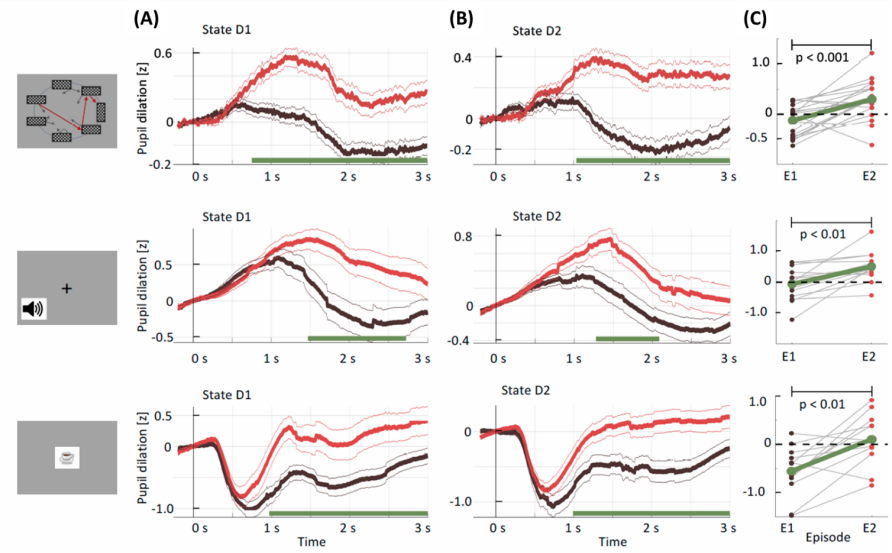


Figure 2.3 – Pupil dilation in different conditions in the first two episodes. Black curve: pupil response in the first episode. Red curve: pupil response in the second episode. Thin lines indicate the pupil signal ± SEM. Green Interval marks the time course where the two curves differed significantly ($p < 0.05$). **(A)** Comparison of pupil response at state D1. Pupil response to D1 was later with higher amplitude in episode 2 (red curve) than in episode 1(black curve). **(B)** Comparison of pupil response at state D2. Pupil response to D2 was later with higher amplitude in episode 2 (red curve) than in episode 1(black curve). Significance was reached at a time t_{min}, which depends on the condition and the state: spatial D1: $t_{min} = 730ms(22,131,85)$; spatial D2: $t_{min} = 1030ms(22,137,130)$; sound D1: $t_{min} = 1470ms(15,34,19)$; sound D2: $t_{min} = 1280ms(15,35,33)$; clip-art D1: $t_{min} = 970ms(12,39,19)$; clip-art D2: $t_{min} = 980ms(12,45,41)$. **(C)** Participant-based comparison between pupil response to state D1 in episode 1 (black dots) and episode 2 (red dots). Each grey line presents one participant. The pupil response differences between the two episodes are significant in all three conditions. P-values are presented in the figure. The figure is adapted from (Lehmann et al., 2019).

	Condition	Spatial		Sound		Clip-art		Aggregated	
	Model	AIC	Rank sum (k=11)	AIC	Rank sum (k=7)	AIC	Rank sum (k=7)	all ranks	
with elig. tr.	Q-λ	**6470.2** $^{p(a)=.003}_{wAIC=1.00}$	**24**	**1489.1** $^{p(a)=.015}_{wAIC=0.23}$	**20**	1234.8 $^{p(a)=.062}_{wAIC=0.27}$	20	**64**	$p < .001$
	Reinforce	**6508.7** $^{p(a)=.016}$	**35**	**1486.8** $^{p(a)=.015}_{wAIC=0.74}$	**10**	1239.2 $^{p(a)=.109}_{wAIC=0.03}$	22	**67**	
	3-step-Q	**6488.8** $^{p(a)=.013}$	**33**	**1494.3** $^{p(a)=.046}_{wAIC=0.02}$	**26**	**1236.6** $^{p(a)=.015}_{wAIC=0.11}$	**16**	**71**	
	SARSA-λ	**6502.4** $^{p(a)=.003}$	**36**	1495.2 $^{p(a)=.140}_{wAIC=0.01}$	30	**1233.2** $^{p(a)=.015}_{wAIC=0.59}$	**16**	**82**	
Model based	Hybrid	6536.6	61	1498.3	43	1271.3	33	137	
	Forward Learner	6637.5	79	1500.6	41	1316.3	48	168	
without elig. tr.	Q-0	6604.0 $^{p(b)=.003}$	60	1518.6 $^{p(b)=.046}$	39	1292.0 $^{p(b)=.015}$	51	150 $^{p(b)<.001}$	
	SARSA-0	6643.3 $^{p(c)=.001}$	68	1520.2 $^{p(c)=.093}$	43	1289.5 $^{p(c)=.015}$	46	157 $^{p(c)<.001}$	
	Biased Random	7868.3 $^{p(d)=.001}$	99	1866.1 $^{p(d)=.015}$	63	1761.1 $^{p(d)=.015}$	63	225 $^{p(d)<.001}$	

Figure 2.4 – Comparison between TD(λ), TD(0) and model-based models when fitting behavioural data. Four TD(λ) models were used to fit behavioural data, which were Q-λ, Reinforce, 3-step-Q and SARSA-λ. Two model-based model were used to fit behavioural data, which were Hybrid and Forward Learner (Gläscher, Daw, Dayan, & O'Doherty, 2010). Two TD(0) models were used, which were Q-0 and SARSA-0. A biased random model was used as the null-model, i.e. the baseline for comparing model performance. Values in each column are explained in the main text. The figure is adapted from (Lehmann et al., 2019).

3 Neural Correlates of the Reward Prediction Error

3.1 Preface

During learning or decision-making, human can adjust their behaviours based on delayed and sparse feedback given by the environment. The feedback can be either reward or punishment. For example in maze running tasks, a player will not choose the path if he or she finds that the road in front is blocked. However, once a player finds the exit of the maze, he or she is very likely to take the same path to the exit if asked to explore the maze again. In this scenario, the blocked road and the exit are both reward signals given by the environment to the player, while the former can be considered as a negative reward and the latter as a positive reward.

Reinforcement learning (RL) theory is well suited in solving this learning scenario. In this chapter, I will focus on model-free learning in RL. Holroyd and Coles came up with the RL-ERN hypothesis in 2002 (Holroyd & Coles, 2002), stating that the amplitude of an EEG component, called the Feedback-Related Negativity (FRN), reflects the RPEs. The hypothesis was later confirmed by many other researches (see (Walsh & Anderson, 2012) for a review). However, most of the learning tasks in these studies are modified versions of the N-armed bandit task. In the N-armed bandit task, participants do not need to make sequential decisions but only one decision to obtain the reward. The simple task cannot represent the learning situations in daily life. Thus, I used a truly sequential task to test if the signal in reward-based learning, such as the RPEs, can be reflected by the FRN amplitude.

3.2 Results

The results are presented in the manuscript to be submitted *"EEG signatures of the Reward-Prediction Error at non-rewarded states. He A. Xu, Marco P. Lehmann, Wulfram Gerstner, and Michael H. Herzog"* (see Appendix 2).

In this study, I designed two sequential decision making experiments based on the paradigm proposed in (Tartaglia et al., 2017). The first experiment uses 2 complex sequential learning

environments (Figure 1.4) to test if the FRN amplitude reflects the RPEs in non-directly rewarded states (non-goal states). The second experiment uses 6 simple environments (Figure 1.5) to test if the FRN amplitude reflects the RPEs in directly rewarded states (goal states).

When participants saw the rewarded states (goal state) for the 1st, 3rd and 5th times, the corresponding RPEs decreased. By comparing the Event Related Potentials (ERPs) of the three conditions, we can find the potential time window where the RPEs are reflected. The result shows that the ERP amplitudes in the time window between 280 and 390ms after the state onset reflect the RPE changing trend (Figure 3.3, 280-360ms for the complex environments, 280-390ms for the simple environments). This time window is in the FRN time range. We then estimated the RPEs of each state visit using the SARSA(λ) model, and analysed the linear regression between the estimated RPE and the mean FRN amplitudes in this window. The participant-by-participant based linear regression (Figure 3.4 and 3.5) confirmed that the FRN amplitudes reflect the RPEs of both non-goal states ($p = 0.02$, $t(11) = 2.5$, $sd = 6.1$) and goal states ($p = 0.03$, $t(13) = 2.3$, $sd = 2.7$).

The sequential tasks also provided an opportunity to investigate how brain reacted to directly rewarded and non-directly rewarded stimuli. To do this, I divided all the trials to two groups in each experiment: the REWARD group and the Non-REWARD group. The REWARD group contained all the trials when participants found the goal states. The Non-REWARD group contained all the trials when participants visited the non-goal states.

First, I compared the ERPs between the REWARD and Non-REWARD groups and found in both experiments there were two time windows that showed the significant difference (Figure 3.6A, B). The EEG source localisation was performed using the standardised low resolution brain electromagnetic tomography (sLoreta) (Pascual-Marqui et al., 2002). The two groups of EEG data (REWARD/ Non-REWARD) were converted into 3D-MNI space for all blocks in both complex and simple experiments. sLoreta estimates the current source density distribution for trials of EEG data across all data points on a dense grid of 6239 voxels at 5 mm spatial resolution.

I compared the current source densities for each observer's ERPs averaged in the selected time windows shown in Figure 3.6. Statistical analysis was applied between groups within each experiment using the implemented statistical nonparametric mapping tool using corrected p-value < 0.05 for significance.

In the late time window (500-570ms) of the complex experiment, I found medial frontal gyrus (BA10, X = -10, Y = 55, Z = -5), the anterior cingulate cortex (BA32, X =-10, Y = 33, Z = -7) and the cuneus (occipital lobe, BA18, X=-10, Y = -96, Z = 15) activated more for REWARD condition than for Non-REWARD condition (Figure 3.6C, $t(10) = 2.76$, $p = 0.02$).

In the late time window (421-559ms) in the simple experiment, I found that the lingual gyrus (occipital lobe, BA18, X = -10, Y = -91, Z = -20), the precuneus (parietal lobe, BA7, X = -10, Y -50, Z = 50), the medial frontal gyrus (BA10, X = -10, Y = 46, Z = 14) and the anterior cingulate cortex

(BA32, X = -10, Y = 45, Z = 5) were highly activated for the REWARD condition (Figure 3.6D). The current source densities were high in the REWARD condition and low in the Non-REWARD condition ($t(12) = 4.32, p < 0.001$), indicating that these sources were highly activated on rewarding feedback.

I also compared the current source densities of the REWARD condition and each of the Non-REWARD conditions in the early windows of both experiments, no significant difference was found between conditions.

3.3 Discussion

Decision masking is a complex process involving the evaluation of the reward, the RPE, and potentially other values. In model-free reinforcement learning, the RPE is the most important variable. In N-armed bandit tasks, the FRN is positively correlated with the RPE (Holroyd & Coles, 2002). In this chapter we wanted to check whether a similar correlation is also true in more interesting situations where decision making is sequential and reward is not delivered immediately. Classic model-free reinforcement learning models propose that the RPE plays an essential role also at states that are not directly rewarded. Hence, we asked the question whether there is evidence for RPEs in EEG signals at non-rewarded states. To address this question, we used a previously developed sequential decision making paradigm, where a goal is found only after a sequence of actions (Clarke et al., 2015; Tartaglia et al., 2017).

We first used two complex environments which contained trap states and loops to test if the FRN-ERP relationship proposed by Holroyd and Coles still holds true. The FRN amplitude reflected the RPEs of the non-goal states in a time window of 280 - 360ms after the state onset. Since the goal state occurred rarely - only 120 times when summed over all participants and epochs in the complex environments, the correlation between FRN amplitudes and RPEs at the goal states was not significant. Thus, we used six simple linear environments with 1-dimensional arrangement of states to test if the FRN amplitudes reflect the RPEs of the goal states. Indeed, in the time window of 280-390ms after the state onset the correlation was significant. Both time windows are very close to the FRN window.

Contrary to most studies in reinforcement learning, we used a deep sequential decision making task, where only one of many states was rewarded. Sambrook et al., (Sambrook et al., 2018) used a two-step task. They found that the RPEs of the intermediate state was also reflected in the EEG around 200-400ms. In an fMRI study, Glaescher, Daw and Dayan (Gläscher et al., 2010) used a 2-step design and found that that the sources of RPE are in the Ventral Striatum, which is line with the proposal by Holroyd and Coles (Holroyd & Coles, 2002) that the FRN sources of the RPE are in the ACC. In contrast to such a 2-step design, some of our participants spent more than 100 steps in loops of the environment before they saw the first goal image.

We also found that the prefrontal cortex (PFC), the anterior cingulate (ACC) and the primary visual cortex showed higher activity when participants received the rewarding stimulus than

when they received the non-rewarding stimulus. The previous two regions (the PFC and the ACC) are usually observed in fMRI and EEG studies as the brain regions related with reward processing (Badgaiyan & Posner, 1998; Bellebaum & Daum, 2008; Cohen et al., 2007; Doñamayor, Schoenfeld, & Münte, 2012; Gehring & Willoughby, 2002; Gruendler et al., 2011; Haruno & Kawato, 2006; McClure, Berns, & Montague, 2003; Nieuwenhuis et al., 2005; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Tucker et al., 2003). Nevertheless, the time window where we found brain activities differing in rewarding and non-rewarding signals is relatively late (later than 400ms after the state onset). Regarding the visual processing time, which takes usually 80ms after the stimuli onset, the visual cortex activity is unlikely due to the visual signal processing. We propose that this high activity in the primary visual cortex is due to the top-down control from the frontal cortex to the visual cortex during reward processing. When comparing the time window between the two environments, we found that the high-activity time window in complex environment is later than in the simple environment. We suggest that that in the complex environment, participants need to apply higher cognitive load to solve the task, which makes the top-down control arriving later than that in the simple task. However, this hypothesis needs to be tested with tasks of different complexity. If for the same environmental structure, the visual cortex shows higher activity for the rewarding stimuli in the later window in the more complex task, we could say that indeed this activity is due to the top-down control. Our hypothesis is also supported by other studies. The primary visual cortex is reported as a part of the reward processing network in animals and humans (Anderson, 2017; Arsenault, Nelissen, Jarraya, & Vanduffel, 2013; Roelfsema & Ooyen, 2005; Rombouts, Bohte, Martinez-Trujillo, & Roelfsema, 2015; Shuler & Bear, 2006).

There are some caveats. Our results, as all results in the field, are based on correlations, which limit conclusions to some extent. For example, humans may compute RPEs but do not use them for learning. Or RPE may be used as a confidence measure rather than as an action choice variable. Second, we computed RPE with SARSA. We do not however claim that humans use a SARSA like mechanism because many other algorithms, including unknown ones, may deliver similar results. Third, our results show evidence that humans make use of model free RL components. However, this does mean that humans do not use model based learning, which they most likely do. We currently explore model-based exploration in very similar environments.

Taken together, our results suggest that the FRN reflects the RPE (or related measures) in deep, sequential decision making paradigms in both rewarded and non-rewarded states.

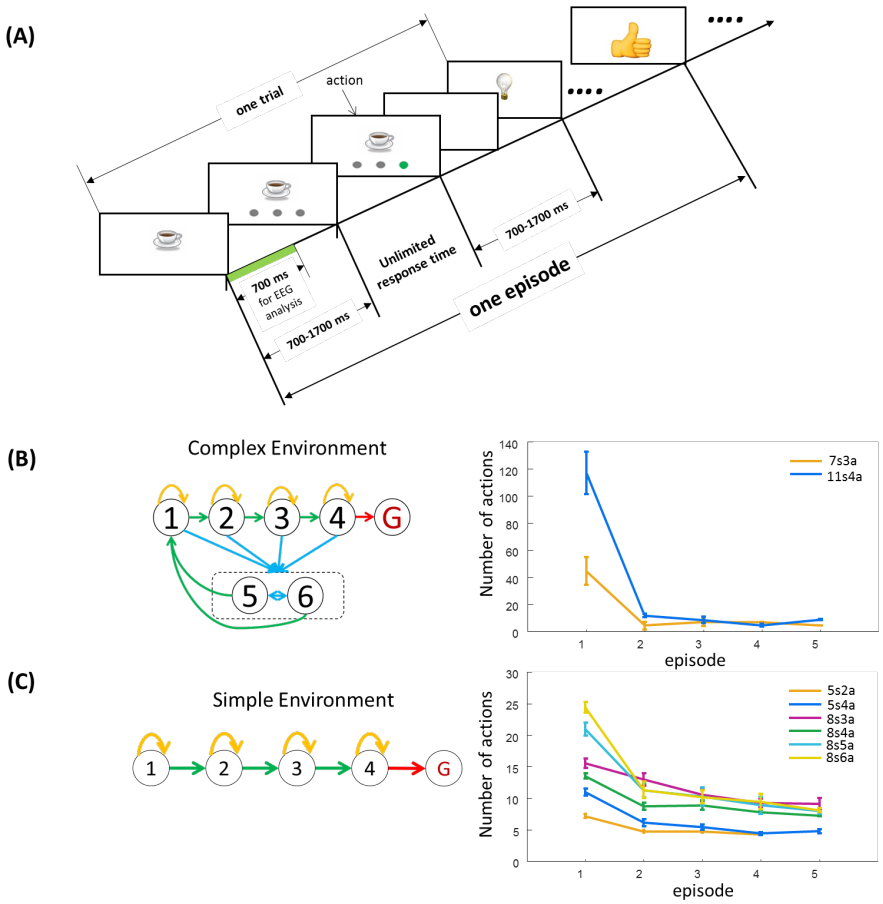


Figure 3.1 – **(A)** Sequential Decision Making Paradigm. An image (state) is presented on the screen. After a random interval of 700-1700ms, grey disks appear, on which participants are asked to click (actions). After an action, a blank screen is shown for a random interval between 700 and 1700ms and then the next state appears. The goal state is a 'thumb-up' image in this example. The green interval indicates the time (0-700ms after the image onset), for which ERP was analysed. **(B)** Structure of the complex environment. Non-goal states are indicated by numbers while the goal state is presented by the red G. 's' indicates 'states', 'a' indicates 'actions'. For example, '7s3a' means that the environment has 7 states (including the goal state) and each state comes with three actions. Arrows present the outcomes of the actions. There were three groups of states: (i) the goal state (red G), (ii) progressing states (states 1-4) and (iii) trap states (states 5-6). In order to find the goal state as fast as possible, participants needed to avoid the trap states. For each non-goal state, there was only one action (green arrows), which led participants to the next state; one other action (yellow arrows) led participants back to the current state. Actions that led participants to states 5-6 are shown in blue (see methods for details). Performance was determined as the number of actions participants needed to find the goal state. Performance is shown on the right as a function of the number of episodes finished. Points connected by lines indicate the means and bars indicate the standard error. **(C)** Structure of the simple environment. There were only two types of actions at each non-goal states: one action led participants to the next state (green arrows), all other actions let participant stay at the current state (yellow arrows). The task is much easier because participants either stayed or moved towards the goal states. Performance is shown on the right.

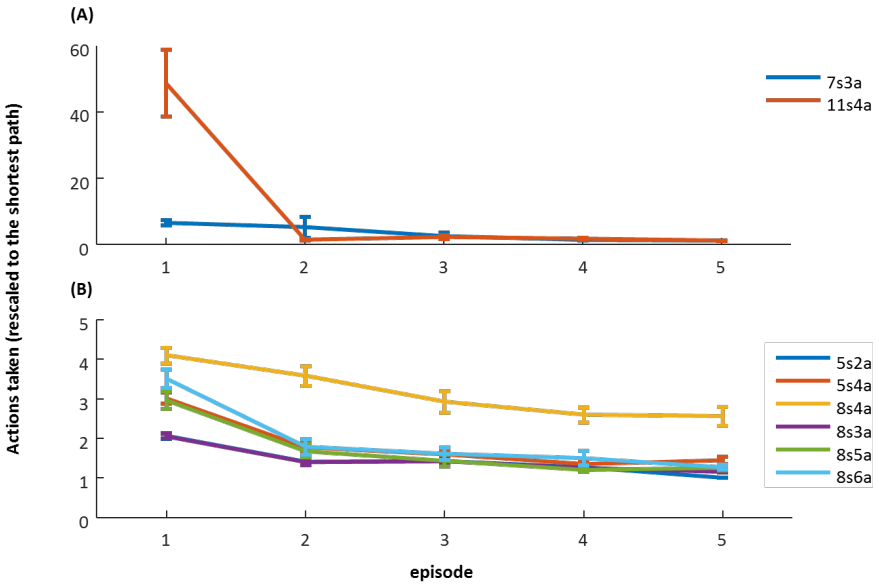


Figure 3.2 – Behavioural performance (data from Figure 3.1B, C) re-scaled to the shortest path in each episode. The y-axis presents performance, which is calculated as the ratio between the number of actions participants took to finish an episode and the minimum number of actions needed. A y-value of 1 indicates that the participants used the shortest path. **(A)** In the complex environment with 11 states, the first episode started in state 6 and the second episode always started in state 9; in the environment with 7 states, the first episode always started in state 6 and the second episode in state 4 (detailed environment structure see Figure 1.4). **(B)** In the simple environment with eight states, the second episode always started in state 1, for the environment with 4 actions and in state 1 for the one with three actions (detailed environment structure see Figure 1.5). Please note the difference in the y-axis scales of (A) and (B).

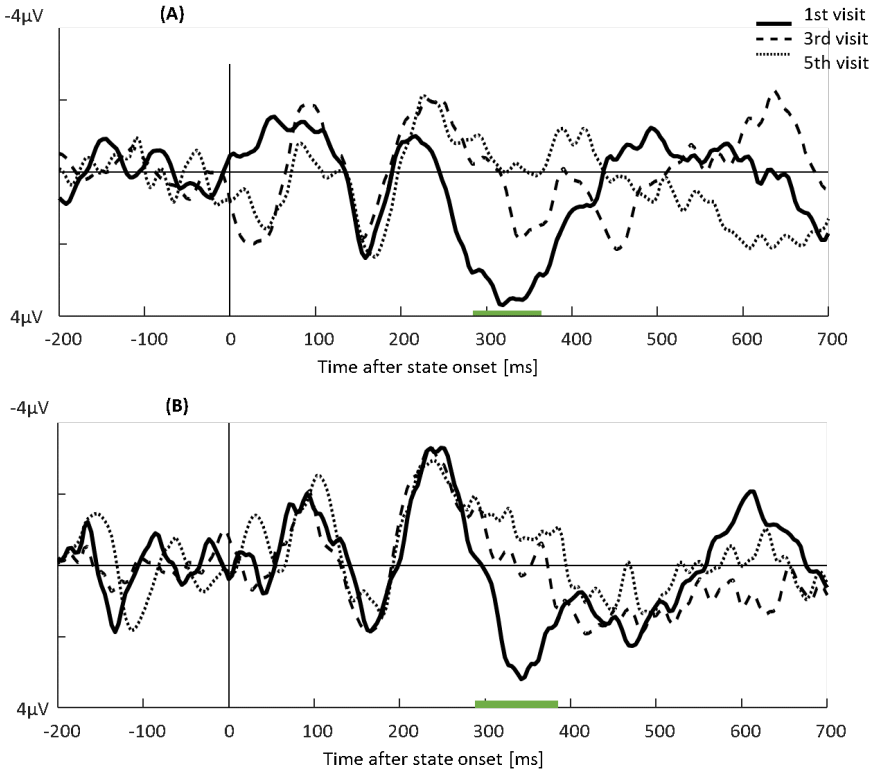


Figure 3.3 – ERPs for the 1st, 3rd and 5th goal visit. 0 on the x-axis indicates the image onset. Negative values are plotted up by convention. Green lines indicate significant differences between the ERPs of the 1st, 3rd, and 5th visit to the goal with a monotonic trend of the RPEs. **(A)** Complex environments. ERPs were significantly different between 280-360ms ($F(2,33) = 4.84, p = 0.014$). **(B)** Simple environments. ERPs were significantly different between 280-390ms ($F(2,39) = 5.39, p = 0.008$).

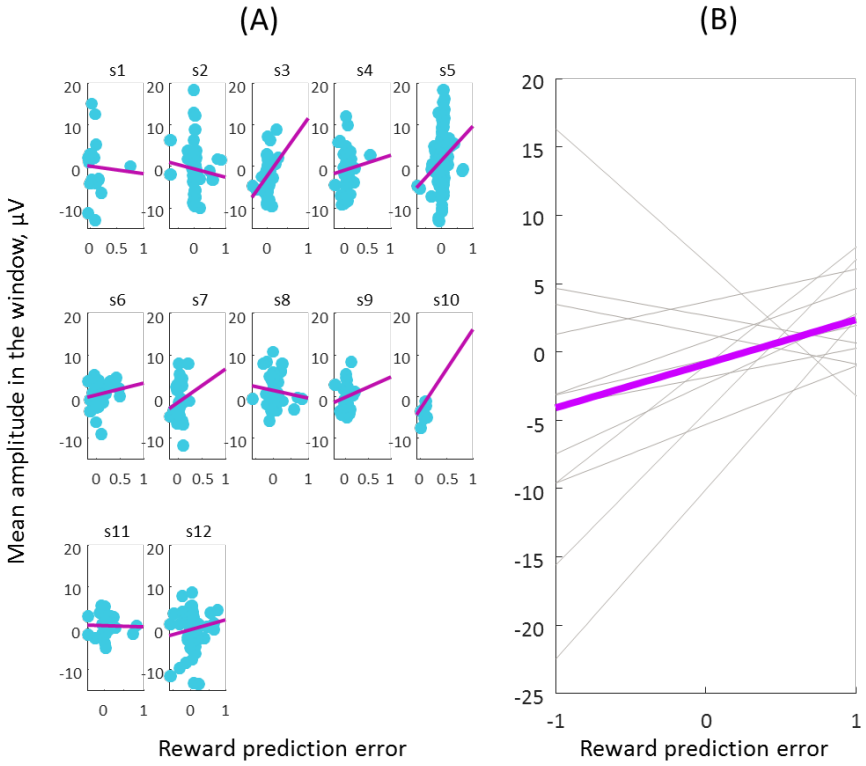


Figure 3.4 – Linear regressions between the RPEs and the mean ERP amplitudes for each participant at the non-goal states in the complex experiments. X-axis presents the estimated RPE from SARSA(λ) model, y-axis presents the mean amplitude of the ERP in selected time window. **(A)** The regression of each participants. Each dot represents one trial when the participant visited the non-goal states. **(B)** The averaged regression between the RPEs and ERP amplitudes for all participants. The regression coefficients are significantly different from zero ($p = 0.02$, $t(11) = 2.5$, $sd = 6.1$, mean coefficient = 3.2). Each grey line presents the regression of one participant (purple lines in A).

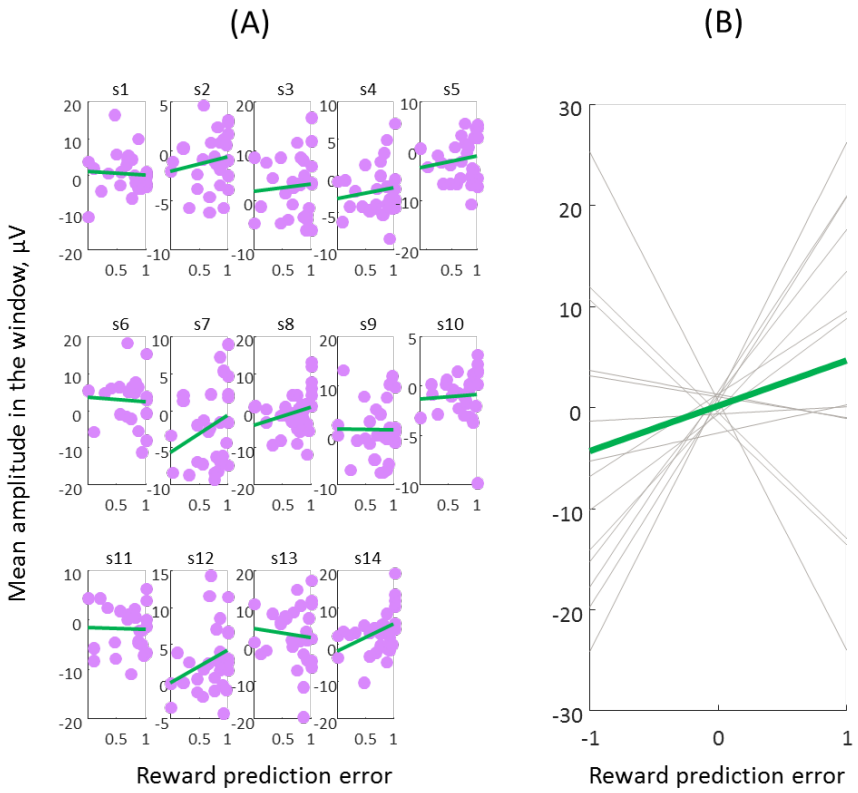


Figure 3.5 – Linear regressions between the RPEs and the mean ERP amplitudes for each participant at the goal states in the simple experiments. X-axis presents the estimated RPE from SARSA(λ) model, y-axis presents the mean amplitude of the ERP in selected time window. **(A)** The regression of each participants. Each dot represents one trial when the participant visited the goal state. **(B)** The averaged regression between the RPEs and ERP amplitudes for all participants. The regression coefficients are significantly different from zero ($p = 0.03$, $t(13) = 2.3$, $sd = 2.7$, mean coefficient = 1.7). Each grey line presents the regression of one participant (green lines in A).

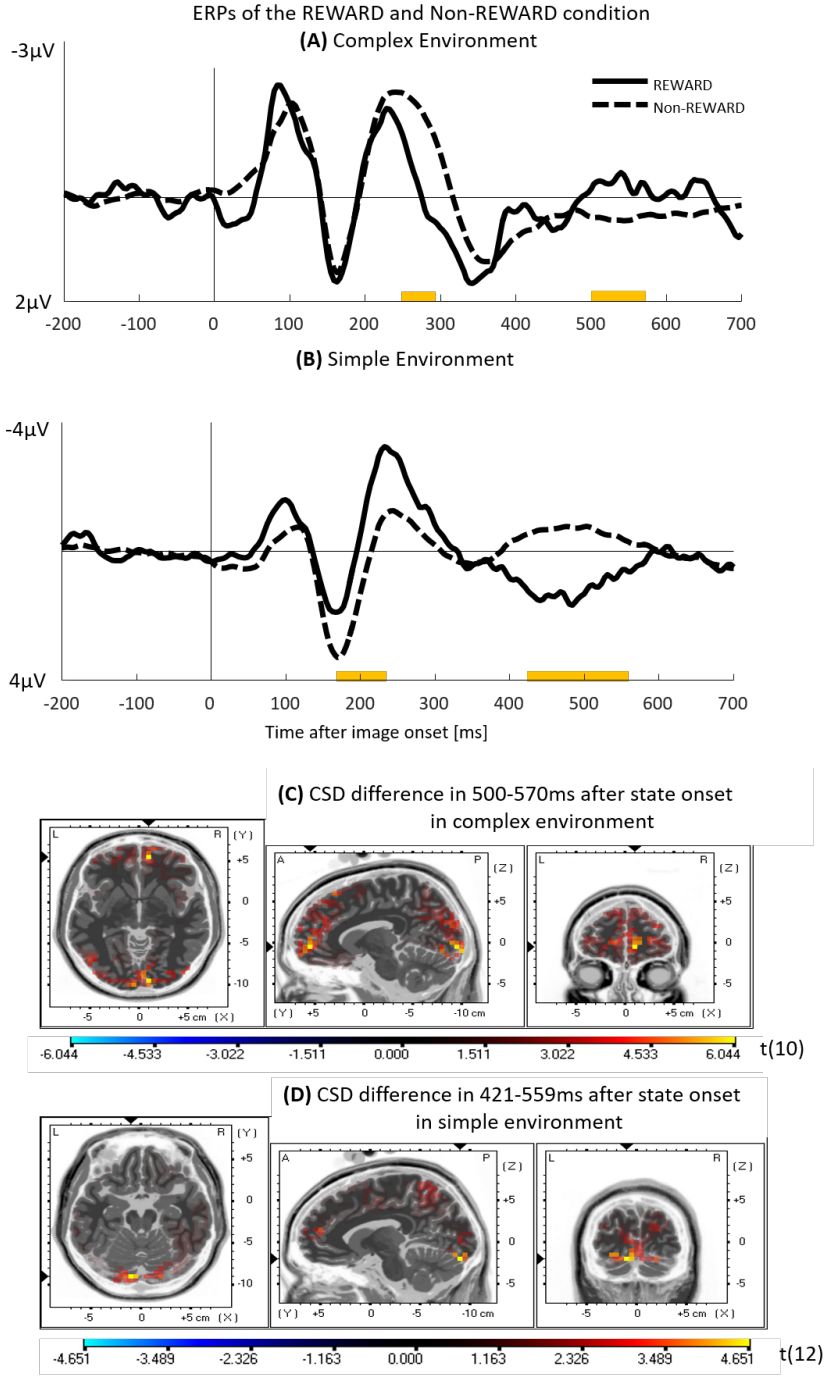


Figure 3.6 – **(A)** In the complex experiment, the ERPs of REWARD and Non-REWARD conditions differ significantly in an early window at 261-291ms ($t(24) = 2.84, p = 0.009$) and a late window at 500-570ms ($t(24) = 2.15, p = 0.04$). **(B)** In the simple experiment, the ERPs of REWARD and Non-REWARD conditions differ significantly in an early window at 185-236ms ($t(26) = 2.29, p = 0.03$) and a late window at 421-559ms ($t(26) = 4.04, p < 0.001$). **(C)** In the window 500-570ms after feedback onset in the complex experiment, Middle Frontal Gyrus (BA10) and Cuneus (BA18) are more activated for REWARD condition compared to Non-REWARD condition. **(D)** In the window 421-559ms after state onset in the simple experiment, Middle Frontal Gyrus (BA10), Precuneus (BA7) and Cuneus (BA18) are more activated for REWARD condition compared to Non-REWARD condition.

4 Neural Correlates of the State Prediction Error

4.1 Preface

Model-based RL models upadate the environment model using the state prediction error (SPE). In this chapter, I will focus on two types of SPE signals, which are novelty and surprise.

Novelty and surprise are different (Barto, Mirolli, & Baldassarre, 2013). For example, when you enter a train in a foreign country, it is likely that all passengers are novel to you, but this is not surprising. However, when you see Roger Federer on the same train you may be very surprised even though Federer is not novel to you. Previous studies showed that, humans, even in the infancy age are able to explore in the environment driven by the novelty of events (Reynolds, 2015). Novelty-seeking behaviour has been interpreted in the theory of reinforcement learning as steps towards building a model of the world (Sutton & Barto, 2018). Surprise, on the other hand, is triggered when the agent finds an observation not matching the prediction of its expectation. Agents, such as humans, can adjust their environment model based on the surprise (Behrens, Woolrich, Walton, & Rushworth, 2007; Holland, 1997; Krugel, Biele, Mohr, Li, & Heekeren, 2009; Nassar et al., 2012; Pearce & Hall, 1980; Wilson, Boumphrey, & Pearce, 1992). Surprise is usually quantified by two different approaches. The first approach models surprise as prediction errors, such as reward prediction or risk prediction errors (Hayden, Heilbronner, Pearson, & Platt, 2011; Pearce & Hall, 1980; Preuschoff & Bossaerts, 2007; Roesch, Esber, Li, Daw, & Schoenbaum, 2012). The reward prediction error was discussed in the previous chapter (Chapter 3) as a component of model-free RL models. The other approach models surprises as the Bayesian updating of beliefs about current environment model (Adams & MacKay, 2007; Angela & Dayan, 2005; Behrens et al., 2007; Kolossa, Fingscheidt, Wessel, & Kopp, 2013; Kolossa, Kopp, & Fingscheidt, 2015; Mathys, Daunizeau, Friston, & Stephan, 2011; Meyniel, Maheu, & Dehaene, 2016). Here I estimated surprise using the second approach.

In this chapter, I designed a sequential decision making task to study surprise and novelty signals in human RL. The experiment contained two blocks. The structure of the two blocks are shown in Figure 4.1. The purpose of designing block 1 was to study how human learn a complex task using the novelty signal. The purpose of block 2 was to trigger surprise and

study how humans adapted to the changes in the environment. I used a newly developed RL model, called the SurNoR model (SUprise-NOvelty-Reward), to estimate the surprise and novelty. With the help of the model, I found the corresponding EEG time windows of surprise and novelty signal.

4.2 Results

The results are presented in the manuscript in preparation *'Model-Building by Exploration: Surprise and Novelty in Reward-based Learning. He A. Xu, Marco P. Lehmann, Alireza Modirshanechi, Wulfram Gerstner, and Michael H. Herzog'* (see Appendix 3).

4.2.1 Novelty and Surprise Estimation using SurNoR Model

SurNoR stands for SURprise-NOvelty-Reward. SurNoR learns a model of the environment using the SMiLe rule (Faraji, Preuschoff, & Gerstner, 2018), uses the novelty signal as an internal reward to explore, and uses the external reward signal to exploit the environment.

The structure of the SurNoR model is shown in Figure 4.3. The model consists of three phases. The first phase is the *Novelty Estimation* phase. When the agent visits state S_t and obtains reward R_t at time t, it first computes the novelty of the state as:

$$N^t(s) = -log(\frac{C^t{}_s + 1}{t + |\mathbf{S}|}) \tag{4.1}$$

$|\mathbf{S}|$ is the total number of states in the environment (i.e. 11 for this experiment), $C^t{}_s$ is the count of how many times state s is encountered up to time t. Equation 4.1 computes the logarithm of the empirical frequency of encountering state s. In the SurNoR model, novelty is used as an internal reward for exploration. Figure 4.4 shows the estimated novelty of each state over time steps and the averaged novelty of each state in the 1st episode of the 1st block at one time step.

In the second phase, the novelty of state s (N_t) together with S_t and R_t are passed to a model-based module and a model-free module. The model-based module learns the state-action transitions and reward distribution of the environment and updates the state-action transitions using the SMiLe rule (Faraji et al., 2018). The SMiLe rule computes the surprise of each $state - action \rightarrow nextstate$ transition. Figure 4.5 shows the estimated surprise of the state transitions in the 1st episode of the 2nd block. Surprise can affect the learning speed in the model-based module: the higher the surprise, the faster the agent learns; the lower the surprise, the slower the agent learns. The model-free module learns the state-value functions and reward functions of the environment, and updates the state-value functions using the Q-learning (Watkins, 1989). We propose that since humans use both model-based and model-free models to learn (Daw et al., 2011; Gläscher et al., 2010), the surprise should also modulate the learning rate in the model-free module. The surprise-modulated parameter

γ_t is transmitted from the model-based module to the model-free module and modulates the learning rate of the Q-learning model:

$$\alpha_t = \alpha + \gamma_t \Delta\alpha \tag{4.2}$$

Note that in both model-based and model-free modules, the novelty is considered as an internal reward in addition to the external reward R_t given by the environment. After processing the information of S_t, R_t, and N_t, the model-based module produces an estimated state-action value ${Q^t}_{MB}$ telling the preference of taking which action at state S_t. The model-free module also produces an estimated ${Q^t}_{MF}$. ${Q^t}_{MB}$ and ${Q^t}_{MF}$ are the inputs for the third phase *Policy* to determine the final action.

The third phase, which is the *Policy* phase, takes into account both ${Q^t}_{MB}$ and ${Q^t}_{MF}$ for action selection. The final state-action value ${Q^t}_{sunor}$ is computed as a weighted sum:

$${Q^t}_{sunor} = \omega_{MB} \times {Q^t}_{MB} + \omega_{MF} \times {Q^t}_{MF} \tag{4.3}$$

where $\omega_{MB} = 1 - \omega_{MF}$.

We used 3-fold cross validation to fit the SurNoR model to participants behavioural data. Figure 4.6 shows that the SurNoR model outperforms the other models. The model comparison criteria used here the *model posterior probability* (Stephan, Tittgemeyer, Knösche, Moran, & Friston, 2009), showing how much evidence that the model explains human behaviour. The model implementation and fitting details are provided in the Appendix 3. Here we only discuss parameters that explains participants learning, such as the model-free learning rate (α_t) and the weight of model-free Q-values (ω_{MF}) as shown in Figure 4.7.

Estimated learning rate α_t. As equation 4.2 shows, the learning rate α_t of the Q-learning model is composed of two components: a constant rate α and a surprise-modulated rate $\Delta\alpha$. Since the surprise varies with the time step t, the overall learning rate α_t also changes. The best fit value of the constant learning rate is $\alpha = 0.000 \pm 0.005$, which is very close to 0. The best fit surprise-modulated $\Delta\alpha$ is 0.65 ± 0.01. These results indicate that when surprise is low, the model-free learning rate is close to 0, meaning participants update little on their learned knowledge. When surprise is high, the model-free learning rate is higher and participants update the learned knowledge more.

Estimated model-free learning weight ω_{MF}. Figure 4.7 shows that in the 1st episode of both blocks, the weight of model-free learning $\hat{\omega_{MF}}$ is lower than 0.5, indicating that it is the model-based learning module dominates in action selection. In the 2nd to 5th episodes of both blocks, $\hat{\omega_{MF}}$ is close to 1.0 (0.95 ± 0.03) meaning that the model-free module dominates in action selection. The difference in the parameters shows that participants rely more on model-based learning before obtaining the reward, which leads to exploration behaviour. After obtaining the reward, participants rely more on model-free learning, which leads to exploitation behaviour.

4.2.2 Neural Correlates of the SPE

In the 1st episode of the 1st block, participants took on average 117 actions to find the goal state ($mean = 117.0, std = 54.2, se = 15.6$). Performance was strongly improved in episode 2 compared to episode 1 and quickly reached optimal performance (Figure 4.2A). The learning curve in the 1st episode of the 1st block (Figure 4.2B) showed that optimal actions were taken at the progressing states (state 1-7) but not at the trap states (state 8-10). The closer the progressing state is to the goal, the faster the optimal action was found.

After the 5th episode of the 1st block, image 3 and image 7 were swapped, participants started the 2nd block. The performance in the 1st episode of the 2nd block was significantly improved compared to the 1st episode of the 1st block ($t(11) = 2.55$, $p = 0.02$, Figure 4.2A), i.e., participants needed less actions to find the goal image (Figure 4.2A) in the 1st episode of the 2nd block than in the 1st episode of the 1st block.

Marker of novelty

At the beginning of the 1st episode of the 1st block, participants do not have any model of the external world. They make actions randomly when seeing a non-goal state for the first time, hence they end up in a trap state with a probability equal to 0.5 (Figure 4.1, blue arrows). Every time when they come out of the trap state, they start from state 1 and continued again to look for the goal state. During the 1st episode, the trap states are the most frequently visited states and the states close to the goal states are least frequently visited. Based on the number of visits to the states, we grouped the states into two conditions: High-Novelty condition and Low-Novelty condition. The High-Novelty condition contains the states that were the least visited, which are states 5, 6, 7 in Figure 4.1A. The Low-Novelty condition contains the states that were the most visited, which are states 8, 9, 10 in Figure 4.1A.

To search for the EEG time window where novelty is reflected, we averaged the ERPs of the states with high novelty values (states 5, 6, 7, High-Novelty condition) and the ones with low novelty values (states 8, 9, 10, Low-Novelty condition) in the 1st episode of the 1st block. Figure 4.4A shows the ERPs are significantly different in the two conditions in a time interval from 80 to 110ms after the state onset ($p = 0.01$, $t(16) = -2.13$, $sd = 1.28$). The average ERPs in two conditions removed physiological and instrumental noise and improved the signal-to-noise ratio. However, it also removed the participant-specific information. We used a sliding window method to search for potential time window of novelty on a trial-by-trial and participant-by-participant basis. Linear regression on a on a trial-by-trial and participant-by-participant basis between the mean amplitudes in the interval of 80-130ms after the state onset and the estimated Novelty from the computational model is significant ($p = 0.02$, $t(10) = 2.68$, $sd = 0.46$, mean slope = 0.37). In summary, our results demonstrate that N1 component (80-130ms) of EEG is a potential marker for novelty, tested both by using on-averaged ERPs comparison and trial-by-trial analysis.

Marker of surprise

The SurNoR model partitions state transitions into 3 groups based on their effects on learning:

(1) The ones that are experienced for the 1st time (mostly in the 1st episode of the 1st block),

(2) The ones that are already experienced once and have not been change since then (i.e. the learned ones), and

(3) The ones corresponding to the transitions from or to the swapped states (mostly in the 2nd episode of the 2nd block).

The 2nd group is considered as un-surprising transitions, while the 1st and the 3rd groups contain the surprising transitions - mild surprise for 1st group and large surprise for the 3rd one. Therefore, we can group all trials to two surprise conditions: surprised condition and un-surprised condition. By comparing the ERPs averaged in both conditions, we found that the time interval from 150 to 300ms after the state onset was a potential interval that reflects the magnitude of surprise. We then extracted the mean amplitudes in the interval of interest and regressed the amplitudes with estimated surprise from the SurNoR model on a trial-by-trial and participant-by-participant basis for all trials in the 1st episodes of both blocks. The regression between mean amplitudes in the time interval 150-300ms and estimated surprise was not significant ($p = 0.86, t(18) = -0.17, sd = 2.61$).

However, to see whether the bio-marker of surprise is hidden in the interval, we used a sliding window of 50ms (10ms per step) to test the regression between the mean amplitudes in the sliding window and the surprise computed by the SurNoR model. The ensemble window was determined using the earliest time point of the first sliding window and the latest time point of the last sliding window, whose mean amplitudes correlated significantly with surprise. We found that the regression is significant between surprise and the mean amplitude in the interval of 150-210ms after the state onset ($p = 0.03, t(10) = -2.50, sd = 0.30,$ mean slope $= -0.23$).

The results indicate that the interval from 150 to 210ms after the state onset is a potential bio-marker for surprise.

4.3 Discussion

When the external reward is sparse and delayed, surprise and novelty can be considered as internal feedback for learning. In this study we built a learning model (SurNoR) to solve the learning situation where surprise, novelty, reward are all involved in learning . The SurNoR model outperformed the other models in explaining participants behaviours. One of the important factor that makes SurNoR different from other models is that SurNoR considers novelty as an intrinsic reward. Previous studies (Beaufour, Le Bihan, Hamon, & Thiébot, 2001; Bevins et al., 2002; Bódi et al., 2009; Bunzeck, Dayan, Dolan, & Duzel, 2010) have shown that

the dopamine level affects both novelty and reward processing. When a state of high reward appears, the dopamine level increases. Similarly, when a state of high novelty appears, the dopamine level increases. These studies provide physiological evidence for the SurNoR model.

Furthermore, we found that the novelty signal is reflected in EEG recording around 80-130ms after the state onset, and that the surprise signal is reflected around 150-210ms after the state onset. In the SurNoR model, the novelty of a state is defined as the global probability of seeing that state, and the surprise of a state-action transition is defined as the changes in the local transition probability (details see Appendix 3). In corresponding to the EEG result, the global signal (novelty) occurs earlier than the local signal (surprise). This finding is in line with previous findings. In (Maheu, Dehaene, & Meyniel, 2019), although the study used MEG recording in an oddball task, the authors showed that the brain activity around 60-130ms is sensitive to global changes in the stimuli. In (Meyniel et al., 2016), the authors built a Bayesian inference model to explain physiological data from previous researchers. Meyniel's results showed that the P300 component in EEG is a bio-marker that reflects local transition probability changes. Here in our study, we confirmed the previous findings in Maheu's and Meyniel's work, and generalise the conclusions to a truly sequential decision making task.

However, if we observe the ERP curves of surprising trials and un-surprising trials, we can see the two curves also differ around 300ms after the state onset 4.8. We also tested the linear regression between surprise and the mean ERP amplitude in the time window between 300ms and 400ms, but no significance was shown. There could be three reasons to explain this results. First, based on our SurNoR model, the surprise has three levels. We only compared the highest surprising trials and lowest surprising trials in 4.8, so it is possible that the 300-400ms time window only reflect the extreme surprise cases. Second, it could be the surprise estimated using our SurNoR model only explained partially the surprise-generating mechanism. There could be other surprise measurement that explains better the curve difference in the 300-400ms time window. Third, other than simple linear regression, there could be other relationships between the ERP amplitudes and the estimated surprise, for example the second order correlation. Unfortunately we did not test the higher order correlation between the two signals. The three reasons to explain the results in the time window of 300-400ms could lead the future research directions.

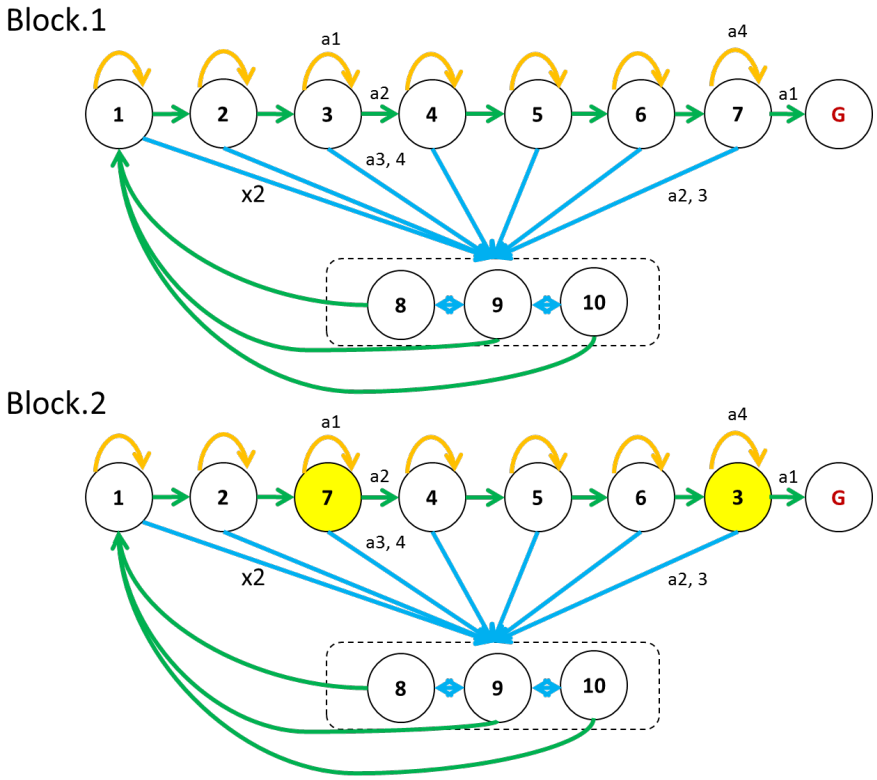


Figure 4.1 – The environmental structure used in the SPE experiment. Non-goal states are presented by numbers; the goal state is presented as the red G; trap states are highlighted by dashed rectangles. Actions are presented by arrows. Green arrows show the actions that bring participants closer to the goal state. Yellow arrows show actions that let participants stay at the current state. Blue arrows show the actions that bring participants to the trap states. There are 11 states in the environment, including (1) the goal state (red G), (2) seven progressing states (images 1, 2, 3, 4, 5, 6, 7), and (3) three trap states (images 8, 9, 10). Participants choose out of 4 possible actions: one action (green arrow) brings them to the next progressing state, two actions (blue arrow) bring them to one of the trap states, one action (yellow arrow) let them stay at the current state. In Block-1, participants performed 5 episodes. After 5 episodes, Block-2 started where image 3 and image 7 were swapped but the actions remained the same as before.

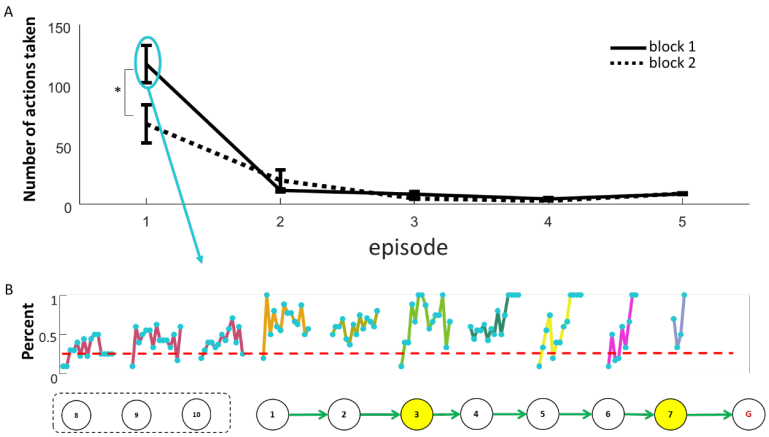


Figure 4.2 – **(A)** Participants' performance in the first (solid line) and second (dashed line) block of the experiment. The x-axis shows the number of episodes. The y-axis shows the number of actions participants needed to find the goal. Performance is measured by the number of actions taken to finish each episode. **(B)** Learning curves in the first episode in the first block (before swapping the images). Each blue point of the learning curve represents one visit to a state. Chance level (dashed red line) of choosing the best action is 1/4 (green arrows in Figure 4.1). The learning curve is averaged over all participants. Fewer visits to a state (fewer blue points) means participants learned the correct action faster.

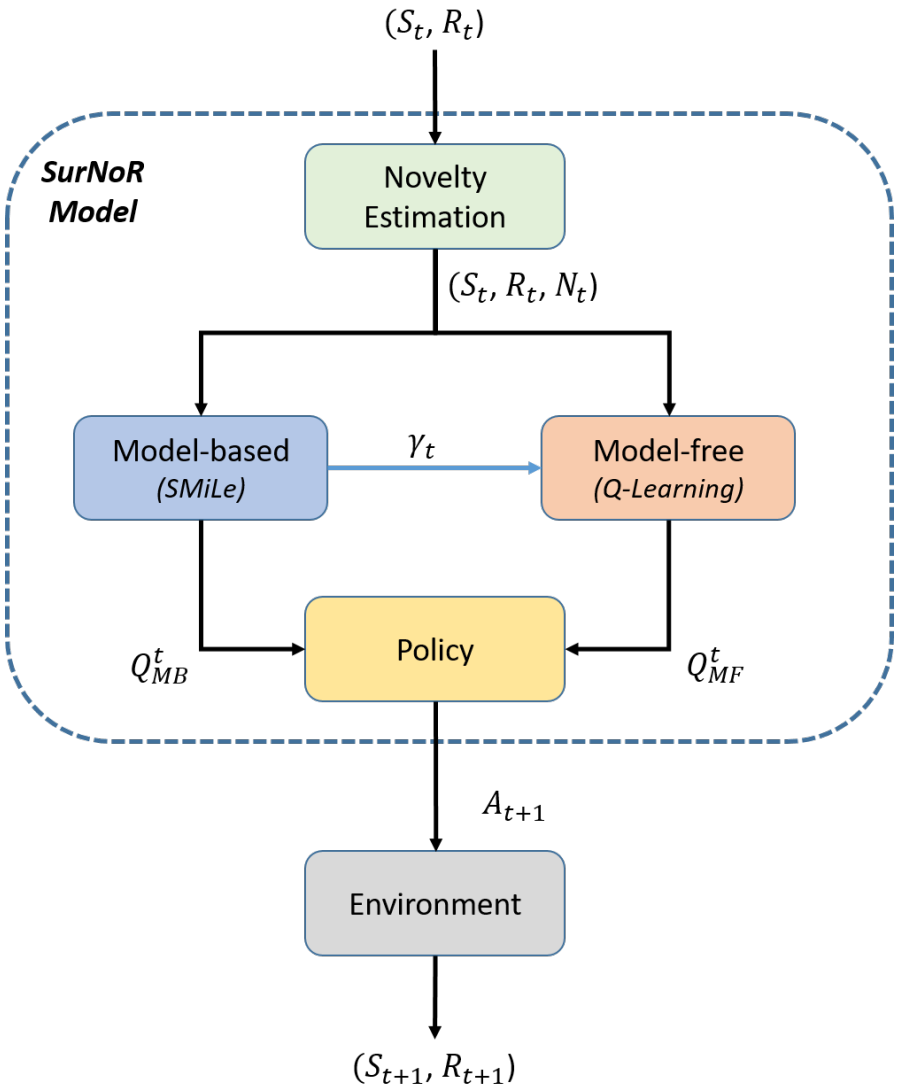


Figure 4.3 – The structure of the SurNoR model. At time t, the agent receives reward R_t at state S_t. Then the SuNoR model estimates the novelty of state S_t at time t based on equation 4.1. The tuple (S_t, R_t, N_t) is then passed separately to both the model-based module (implemented using the SMiLe model) and the model-free module (implemented using the Q-learning model). γ_t is the learning rate used in the SMiLe mode (equation **??**), which is passed to the Q-learning model and modulates the learning rate in the Q-learning model. After updating the SMiLe model and the Q-learning model, both models provides their own estimation of the Q-values ($Q^t{}_{MB}$ and $Q^t{}_{MF}$). The Q-values provide information on which action is preferred to be chosen in the next time step $t+1$. Both Q-values are passed into the policy module for action selection. In the policy module, $Q^t{}_{MB}$ and $Q^t{}_{MF}$ are weighted using equation 4.3, and a final action A_{t+1} is chosen based on the weighted sum.

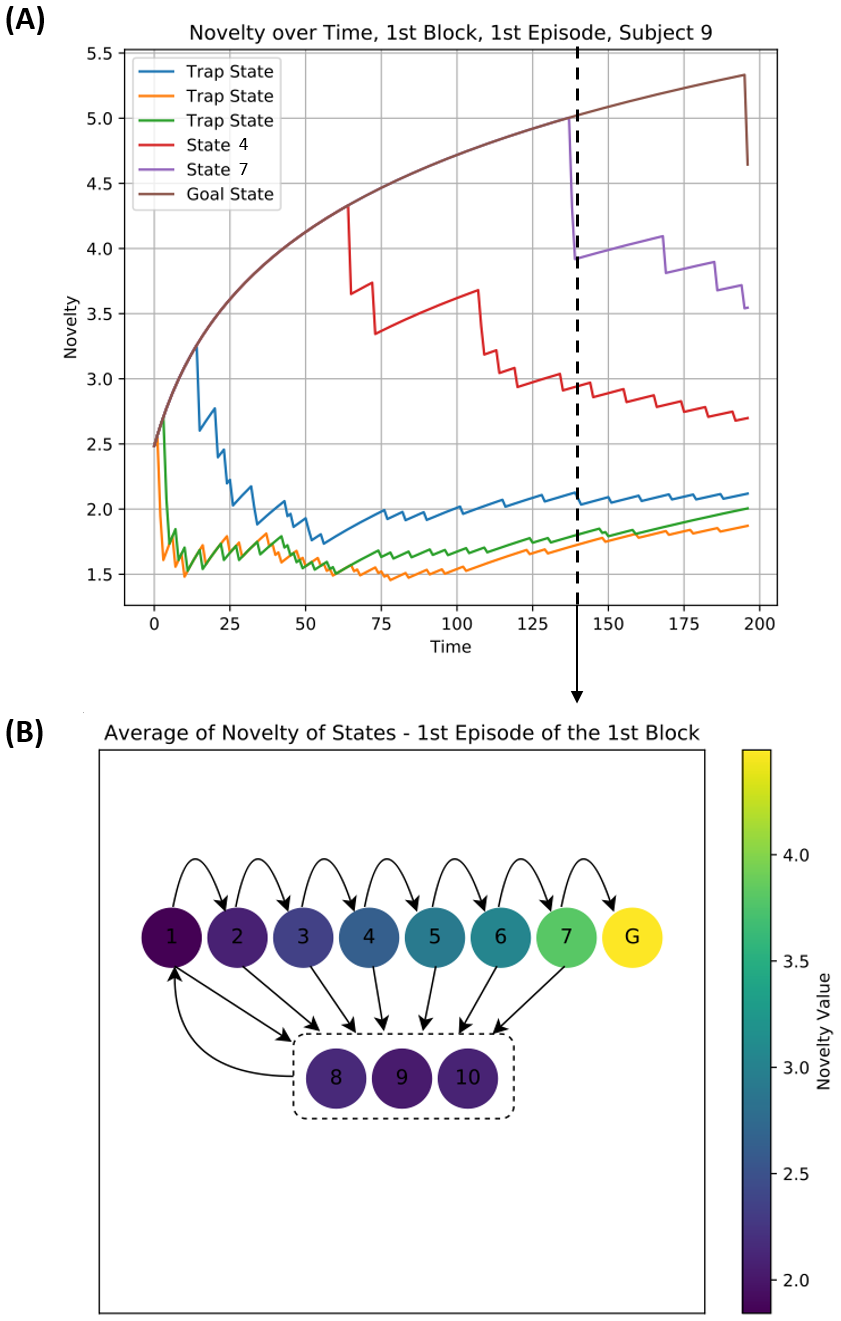


Figure 4.4 – An example of the estimated novelty of each state using the SurNoR model. **(A)** Novelty time-series during the 1st episode of the 1st block: Data is for a single subject. **(B)** Novelty heat-map at the end of the 1st episode of the 1st block (marked with dashed line in (A)). The values are averaged over all subjects.

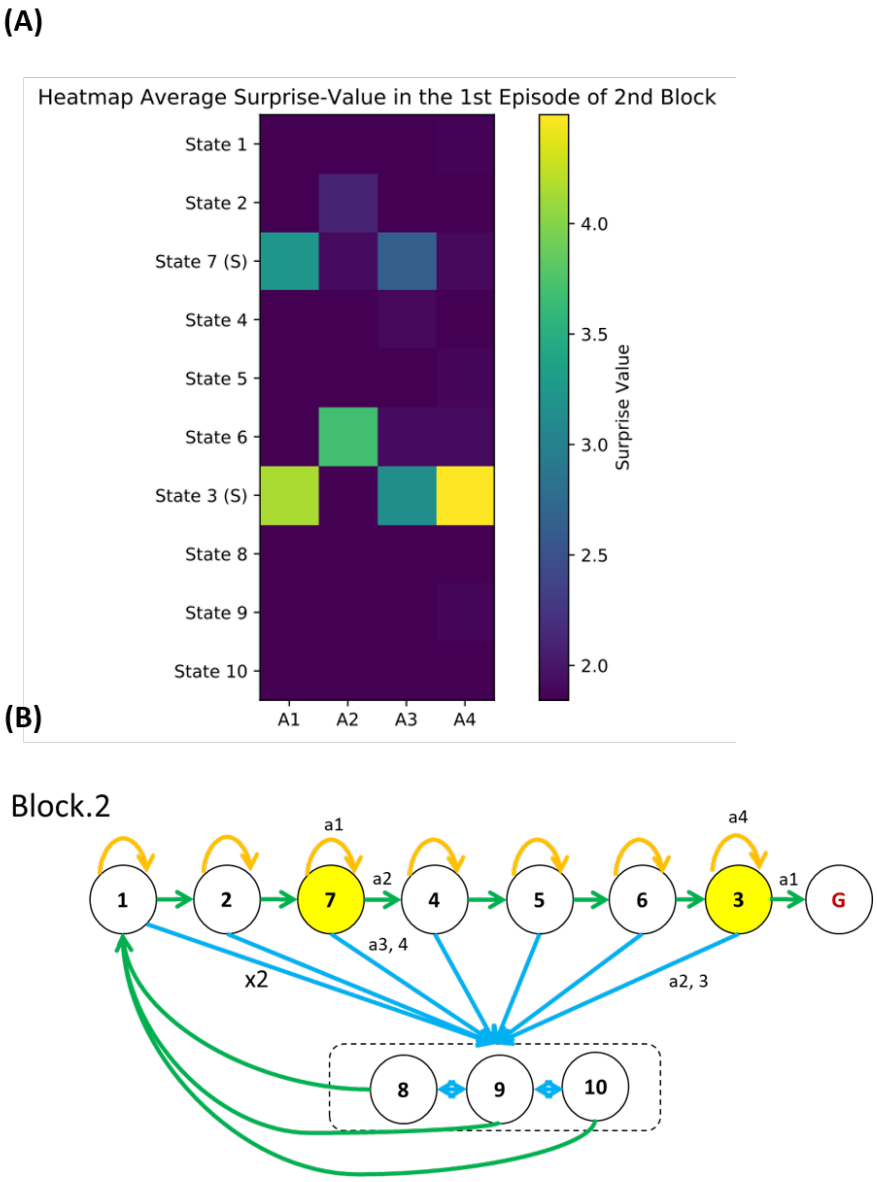


Figure 4.5 – An example of estimated surprise of each state-action transition using the SurNoR model in the 1st episode of the 2nd block. **(A)** Surprise heat-map averaged over all participants in the 1st episode of the 2nd block. **(B)** Environment used in the 2nd block, yellow-marked states are the swapped states that trigger the highest surprise.

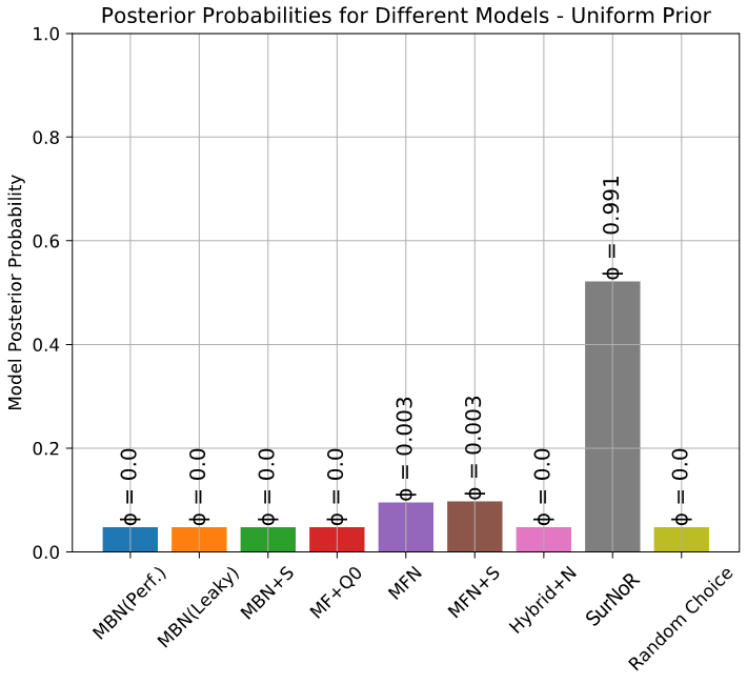


Figure 4.6 – Comparison between the fitting results for the different models in both blocks. The model performance is measured by the model posterior probability using uniform priors. The SurNoR model outperformed the other models. *MBN(Perf.)*: Model-Based model with novelty estimation, perfect integration was used to update the model. *MBN(Leaky)*: Model-Based model with novelty estimation, leaky integration was used to update the model. *MBN+S*: Model-Based model with novelty and surprise estimation, surprise was used to update the model. *MF+Q0*: Model-Free model, Q-learning model was used. *MFN*: Model-Free model with novelty estimation. *MFN+S*: Model-Free model with novelty estimation, the learning rate of the model-free model was modulated by surprise. *Hybrid+N*: Hybrid model using both model-based, model-free models and novelty estimation. *Random Choice*: A model that makes actions randomly.

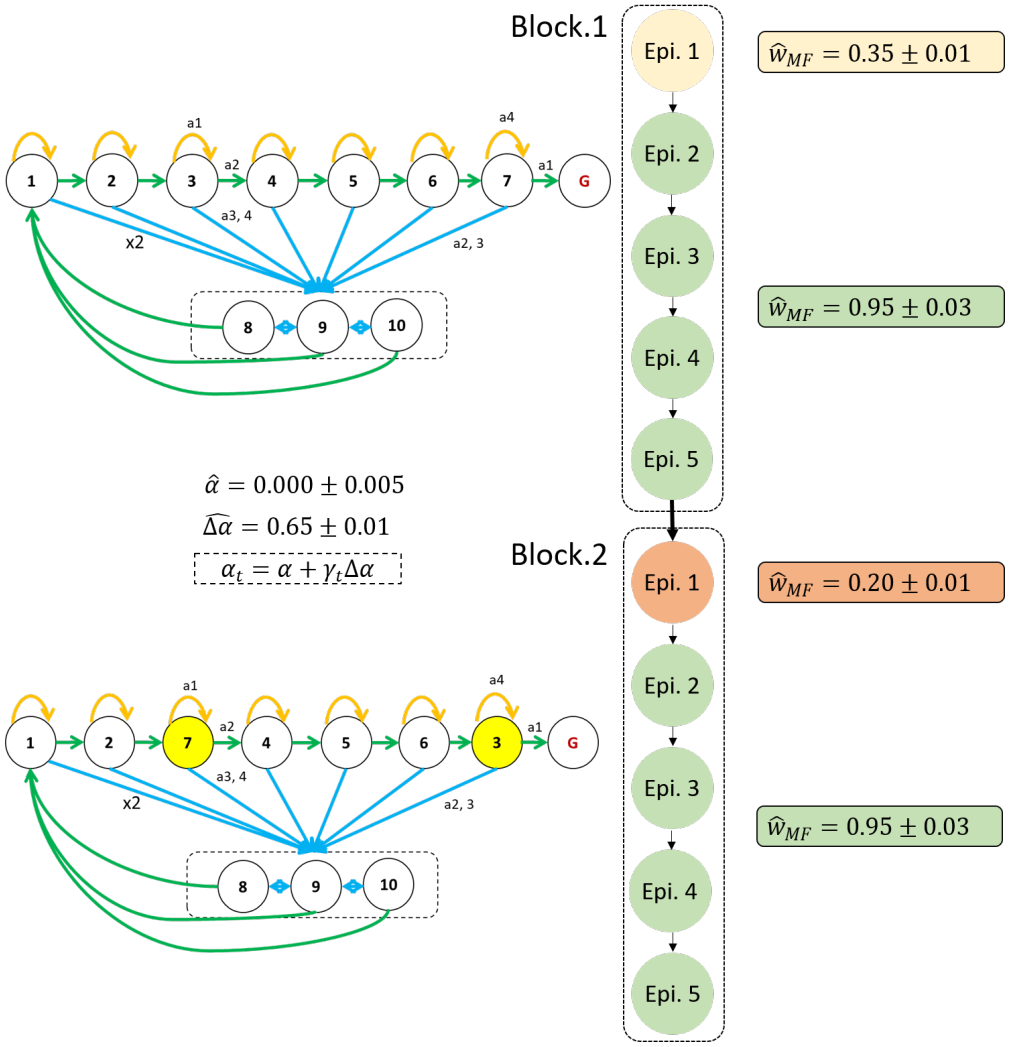


Figure 4.7 – Best fitting parameter value and its confidence interval of model-free learning rate, suprise-modulated learning rate and the weights of model-free Q-values of the SurNoR model.

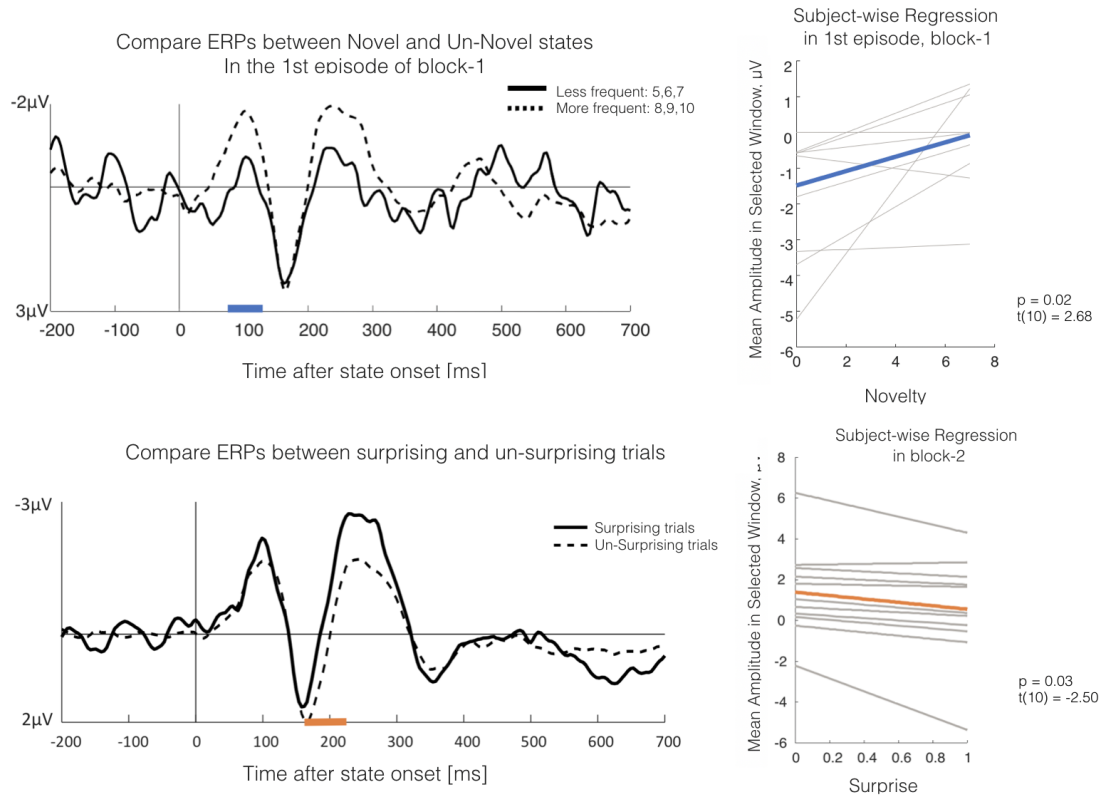


Figure 4.8 – EEG time windows of novelty and surprise. **Uper Left** ERPs compared between High-Novelty condition and Low-Novelty condition in the 1st episode of the 1st block. Blue internal marks the time window in which the novelty is reflected by the ERP amplitude. **Uper Right** Linear regression between the novelty and ERP amplitude in the blue interval, on participant-by-participant basis. Each grey line presents the regression of one participant. The blue line is the averaged linear regression of all participants. **Lower Left** ERPs compared between surprising and un-surprising trials in the first episodes of the 2nd blocks. Orange interval marks the time window in which the surprise is reflected by the ERP amplitude. **Lower Right** Linear regression between the surprise and ERP amplitude in the orange interval. Each grey line presents the regression of one participant. The orange line is the averaged linear regression of all participants.

5 Curiosity or Reward?

5.1 Preface

As shown in Chapter 3 and 4, both reward and novelty can drive learning. Here in this chapter, I investigated whether reward or novelty, is stronger.

I designed an environment with a special 'infinite' state (Figure 5.1). In this environment, there are eleven states including ten non-goal states and one goal state. Participants can choose from three actions at each non-goal state. Among the ten non-goal states, there are six *progressing* states (states 2, 3, 4, 5, 6, 7) and two *trap* states (states 9, 10) similar to the environments used in previous chapters. A new type of state, which is called the *'infinite'* state (state 8), is added (state 8 in Figure 5.1). When participants enter state 5 for the first time, no matter which action they chose, they will be led to state 8 (red arrows from state 5 to state 8 in Figure 5.1 and 5.2). Among the three actions of state 8, the left two actions always lead participants to state 8 presented with different images (yellow arrows in Figure 5.2). The rightmost action always leads participants back to state 5 (red arrows in Figure 5.2). There are in total of 50 different images for state 8. Every time when participants visit state 8, a new image is presented. Hence, each of the 50 image is relatively new to participants compared to the other images. My hypothesis is that, if novelty dominates learning, participants prefer to stay in state 8 searching for new images during the whole environment; if reward dominates learning, participants would avoid going to state 8 and go for the goal state after seeing the goal once.

To test whether novelty or reward dominates learning, I first ran an experiment with 23 participants using the new environment in Figure 5.1. In this experiment, participants were told that there was only one goal. Participants were asked to find the goal for 5 times within 30 minutes.

Then, I ran a second experiment with 9 new participants using the same environment. However, this time I wanted to test if the magnitude of the reward can affect participants' preference for reward and novelty. Participants were told that there were three different goal states, re-

warded by 10CHF, 15CHF and 20CHF. Participants are paid according to which goal state they find. As in the first experiment, participants were asked to find the goal state for 5 times within 30 minutes. What participants did not know is that there was only one goal state, meaning that for the 5 episodes, they can only find the same goal for 5 times. The probability of presenting which goal state was determined by a random number generator before they start the experiment. The probability of seeing the 10CHF as goal is 60%, 15CHF for 30% and 20CHF for 10%.

Here I only present the preliminary results of this study. Further analysis needs to be done and models to be built.

5.2 Results

In the first experiment, participants were told that there is only one goal in the environment. Figure 5.3A shows that participants visited each non-goal states, especially the 'infinite state' (state 8) many times in the first episode before they found the goal state. This behaviour indicates that participants explored the environment in the first episode. Then the number of visits to each state reduced in the second episode, indicating that participants explored less than in the first episode. Figure 5.3B shows the learning curves at each state across each visit from episode 1 to episode 3. For some states (state 5,8,9,10), the number of visits was too long to be plotted. Thus I only plot the first 25 visits for these states. The correct action is the action that leads participants to the next progressing state. The learning curve of state 5 in episode 1 to 3 shows that participants tend not to chose the action, that lead them to the next progressing state (state 6) but rather go to the trap state or state 8.

In the second experiment, I manipulated the magnitude of the reward to test if participants are more eager to look for reward or novelty when there are multiple rewards in the environment. Results (Figure 5.4A) show that the number of visit to state 8 from episode 1 to episode 3 does not decrease as in the first experiment (Figure 5.3A). This result indicates that after knowing there are multiple rewards in the environment, participants were more eager to explore. Learning curve at state 8 (Figure 5.4B) shows that participants tended to stay at state 8 to search for reward. Moreover, if we compare the number of visits to state 8 in the three different reward condition (Figure 5.5A), we find that the exploration behaviour is the most prominent when participants found the least amount of reward (10CHF). The learning curve at state 8 (Figure 5.5B) also shows that participants who obtained the reward of 10CHF tend to stay more at state 8 and look for new rewards.

5.3 Discussion

If participants are driven by novelty, they would prefer to stay longer at the 'infinite' state (state 8) and look for new images no matter what reward they can obtain. However, if participants are driven by reward, they would spend less time in seeing new images at state 8 after seeing a

high reward; and they would spend more time at state 8 looking for new rewards after they see a low reward. The result of the first experiment (Figure 5.3A) shows that participants have reduced the number of visits to the 'infinite' state in the second episode, indicating that they reduced novelty-seeking but increased reward-seeking behaviours. However, when participants are told about the existence of multiple rewards in the second experiment, the novelty-seeking behaviour is not reduced after the first episode Figure 5.5A).

The preliminary results show that reward, especially multiple rewards, can motivate novelty-seeking in learning. Combined with our novelty study in Chapter 4, in which novelty is considered as an intrinsic reward, we can form a closed loop for the relationship between reward and novelty. The existence of reward motivates the novelty-seeking behaviour, which aims to obtain the reward. In the study of (Marvin & Shohamy, 2016), the researchers propose that curiosity (seeking for novelty) can be the motivation to obtain reward. However, no studies have shown how the magnitude of reward affects novelty-seeking behaviours and to which extend. New experiments is being carried on to answer this question.

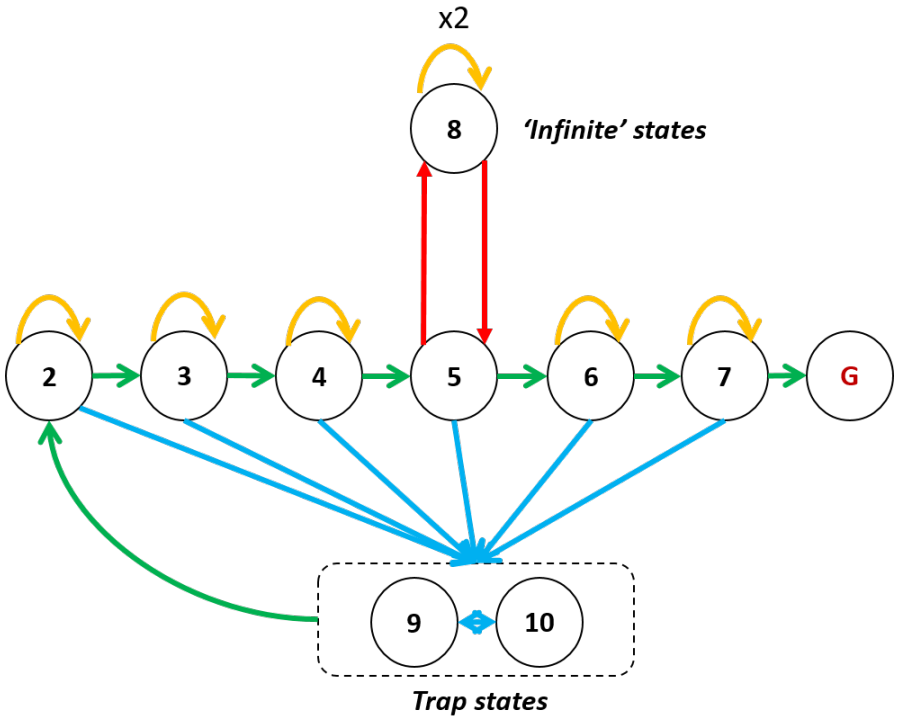


Figure 5.1 – Environment structure. Non-goal states are presented by digits, the goal state is presented by the red G. There are three types of non-goal states: **(i)** six *progressing states*, which are the states 2-7; **(ii)** two *trap states*, which are the states 9 and 10; **(iii)** one *infinite state* which is state 8. Every time when participants visit state 8, a new image is presented. There are in total 50 images presenting state 8. At each non-goal state except state 5 and 8, there are three actions. One action leads participants to the next *progressing state* (green arrows), one action leads participants to one of the *trap states* (blue arrows), one action leads participants to stay at the current state (yellow arrows). At state 5, one action leads participants to state 8 (red arrow), one action leads participants to state 6 (green arrow) and one action leads participants to one of the trap state (blue arrow). At state 8, one action leads participants to state 5 (red arrow) and the two others lead participants to new images (yellow arrow).

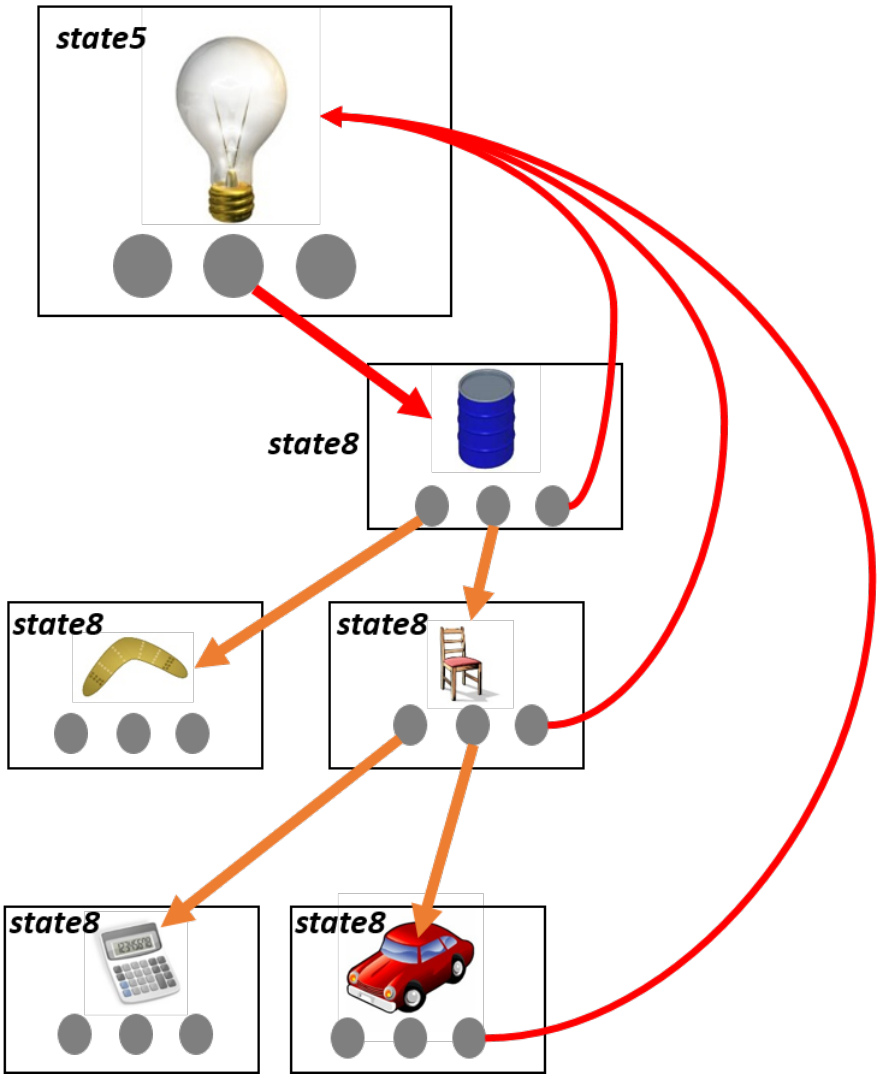


Figure 5.2 – In this example, state 5 is presented by a bulb image. The first time when participants visit state 5, no matter which action they chose, they are led to state 8 (red arrow). One of the other two actions at state 5 leads participants to the next progressing state, and the other leads participants to one of the trap states. When participants visit state 8, two actions let them stay at state 8 while the state 8 is now presented by a different image. The third action (rightmost action) always leads participants back to state 5.

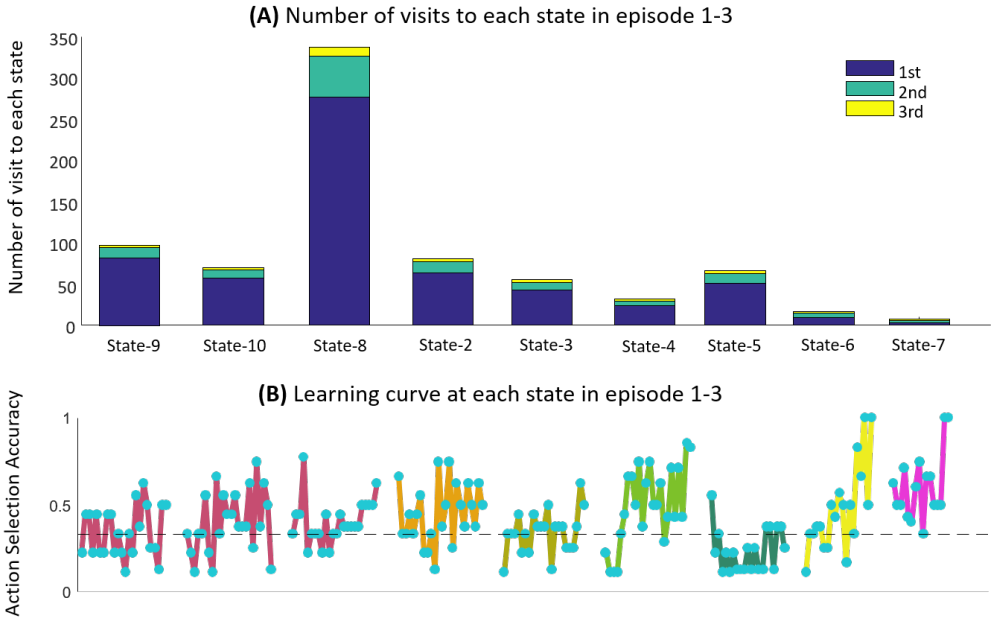


Figure 5.3 – **(A)** Averaged number of visit to each state for all participants in episodes 1 to 3 in the first experiment, when participants know there is only one goal state. **(B)** The learning curve at each state in episode 1 to 3. Learning curve is plotted as the action selection accuracy at each visit to a state, averaged over all participants. The action that leads participants to the next progressing state is considered as the correct action, thus the chance level of choosing the correct action is 1/3. The correct action at state 5 is considered as the action leading to state 6. The correct action at state 8 is considered as the action to state 5. Each blue dot presents one visit to the state. For lacking of space reason, only the first 25 visits to a state are plotted here.

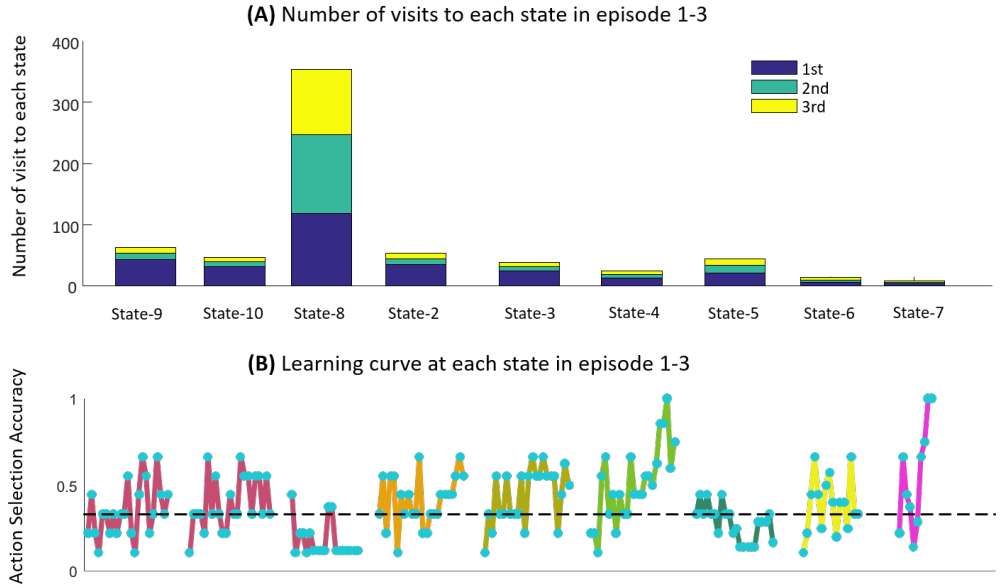


Figure 5.4 – **(A)** The averaged number of visit to each state across all participants in episode 1 to 3 in the second experiment, when participants know there are three goal states to find. **(B)** The learning curve at each state in episode 1 to 3. Learning curve is measured by the action selection accuracy averaged over all participants. The action that leads participants to the next progressing state is considered as the correct action, thus the chance level of choosing the correct action is 1/3. The correct action at state 5 is considered as the action to state 6. The correct action at state 8 is considered as the action to state 5. Each blue dot presents one visit to the state. For lacking of space reason, only the first 25 visits to a state are plotted here.

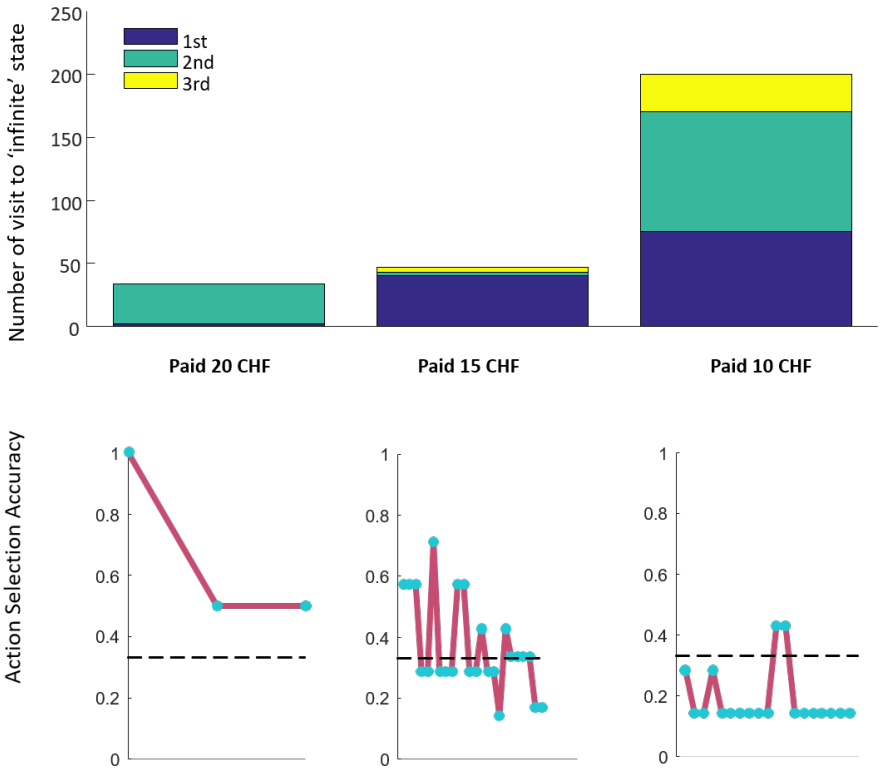


Figure 5.5 – **(A)** The number of visits to state 8 (infinite state) in each payment condition for episode 1s to 3. **(B)** The learning curve at state 8 in the three payment conditions. Learning curve is measured by the action selection accuracy averaged over all participants. The correct action at state 8 is considered as the action to state 5.

6 General Discussion

Humans and animals make decisions all the time in order to fulfil their living needs in their living environment. Machines and computer algorithms make decisions, in order to achieve the goals defined in an artificial environment by the programmer or designer. Reinforcement learning (RL) theory hosts a variety of solutions of decision making situations. There are two main types of RL models, model-free models and model-based models. In this thesis, I studied the physiological evidences of four important RL components: the eligibility trace, RPE, surprise and novelty.

In the first part (Chapter 2 and 3), I focused on model-free learning. Holroyd and Coles proposed a RL-ERN theory in 2002 to explain the effect of reward prediction on an EEG component (Holroyd & Coles, 2002), called the **Feedback Related Negativity** (FRN). The FRN occurs in the frontal-central electrodes between 250-400ms after the feedback onset. According to the RL-ERN theory, dopaminergic neurons produces the RPE and broadcast it to different brain regions. The ACC receives the broadcast RPE, generates the FRN, and plans for the next action. Many researches confirmed the RL-ERN theory by showing that the amplitude of the FRN reflects the RPE in one-step decision making tasks (N-armed bandit tasks). However, whether the FRN amplitudes reflects RPEs in multi-step decision making tasks was never addressed.

To study the relationship between the FRN and the RPE in sequential decision making tasks, we need to answer two questions. First, how to estimate the RPEs at each state (including non-rewarded and rewarded states) in a sequential task? Second, is the FRN-RPE relationship the same for both non-rewarded and rewarded states? In Chapter 2, we answered the first question by testing different RL models in a sequential decision making paradigm (Tartaglia et al., 2017). With this paradigm, we were able to produce longer decision making sequence than the N-armed bandit tasks used in previous researches. We found that models with the implementation of the eligibility trace, such as TD(λ), SARSA(λ) and Q(λ), explained human behaviours better than the models without eligibility trace. Furthermore, the pupil dilation results showed physiological evidence of the eligibility trace in human learning. These results confirm that the eligibility trace is an essential component in human sequential decision

making. The results of this study contribute to the discussion of the eligibility trace on a synaptic level (Bittner et al., 2017; Fisher et al., 2017; Gerstner, Lehmann, Liakoni, Corneil, & Brea, 2018; He et al., 2015; Yagishita et al., 2014).

Using the RL models with eligibility trace, we were able to estimate the RPE more accurately in a sequential task. In chapter 3, we used two sequential decision making experiments to test the FRN-RPE correlation in non-rewarded and rewarded states. We first used the decreasing RPE trend at the rewarded state to locate the RPE-related FRN window. Then we tested the regressions between the FRN amplitudes and RPEs on a participant-by-participant basis. The results of this study confirm that the amplitude of the FRN component reflects the RPEs for both non-rewarded and rewarded states. Our study also provides evidence that humans compute RPEs even at the non-directly rewarded states. Although recent studies (Daw et al., 2011; Gläscher et al., 2010; Sambrook et al., 2018) have used two-step decision making tasks to show that humans use both model-based and model-free models, our study revealed that humans are able to compute the RPEs even when they need to make sequences varying from 10 to 100 steps.

The study of neural correlates of the RPE extends the RL-ERN theory from simple one-step decision making tasks (such as N-armed bandit task) to complex multi-step tasks. In previous FRN studies, researchers used variations of the N-armed bandit tasks and reported different cognitive signals that affect FRN. For example, the FRN amplitude was reported to reflect the reward probability (Eppinger, Mock, & Kray, 2009; Hajcak, Holroyd, Moser, & Simons, 2005; Hewig et al., 2006; Potts et al., 2006; Potts, Martin, Kamp, & Donchin, 2011), the reward magnitude (Bellebaum, Polezzi, & Daum, 2010; Bunzeck et al., 2010; Goyer, Woldorff, & Huettel, 2008; Hajcak, Moser, Holroyd, & Simons, 2006; Holroyd, Larsen, & Cohen, 2004), predictive cues (Baker & Holroyd, 2008; Dunning & Hajcak, 2007; Eppinger et al., 2009; Nieuwenhuis et al., 2002), experienced rewarding stimuli (Cohen & Ranganath, 2007; van der Helden, Boksem, & Blom, 2009; Yasuda, Sato, Miyawaki, Kumano, & Kuboki, 2004), etc. All of these aspects can be described as changes in the reward prediction error. In 2011, Walsh and Anderson published the first paper, to our knowledge, studying the relationship between FRN amplitude and the RPE in a two-step decision-making task. They confirmed that the FRN amplitudes reflect the RPEs in the two-step task, on the averaged analysis but not on trial-by-trial basis. Besides, as Walsh and Anderson (2012) pointed out in their review paper, it is difficult to scale participant's perceived reward across different studies, and even within the same study. Here by using a sequential task, we were able to scale the reward of each state using the state value function, because the state value function itself is computed as the scaled future reward (equation 1.1). In conclusion, our sequential FRN-RPE study provided two advantages over previous research: the first is to integrate and extend previous findings to confirm that the FRN amplitude reflects the RPE in both simple and complex tasks on a trial-by-trial basis; the second is to provide a powerful method to scale participant's perceived reward adaptively.

The inverse solution of the ERP waveform in rewarded states and non-rewarded states showed the brain regions involved in reward processing. The regions we found in this study are the

prefrontal cortex and the anterior cingulate cortex, which is in line with previous findings (Bellebaum & Daum, 2008; Walsh & Anderson, 2011b). We also found that some regions in the primary visual cortex are activated when receiving a reward stimuli. Previous studies (Anderson, 2017; Arsenault et al., 2013; Roelfsema & Ooyen, 2005; Rombouts et al., 2015; Shuler & Bear, 2006) have shown that the visual system is part of the value-driven attention network, and the neurons in the occipital cortex have selective plasticity for rewarding stimuli. The study of (Anderson, 2017) suggested that the early visual cortex were highly activated for rewarding stimuli because features of rewarding stimuli ARE stored in V1. However, this hypothesis was not confirmed by our experiment. We show that the early visual cortex is activated more than 400ms after the feedback onset, which is far beyond visual processing (Thorpe, Fize, & Marlot, 1996). We tested the linear regression between the state values and RPEs with the ERP amplitudes in selected time window and found no significance. The result indicates that the RPE/Value are not computed in the visual system but rather transferred to it. Thus, this activation could be created because of top-down feedback from the reward processing system. However, this hypothesis needs to be further tested. One possibility is to design several tasks using the same environmental structure. If the primary visual cortex activity is due to the top-down control, the time window showing such activity will be later for more complex tasks.

In the second part of the thesis (Chapter 3,4,5), I studied the components of model-based learning. The model-based model learns the state-transitions and reward function. When the model-based model takes an action at a state, it observes the next state. If the observed state is different from what is expected, a state prediction error (SPE) occurs. The model updates the learned state-transitions using the SPE.

In chapter 4, I studied two types of SPE signals, which are surprise and novelty. In this SPE study, I designed a 2-block sequential decision making experiment. The environment contains several trap states, which participants need to learn to avoid. We found that participants spent far more steps in the 1st episode than in the later episodes, which means that participants explored and learned the environment well in the 1st episode and exploited reward in the later episodes. We built a computational model, called **SurNoR**, to explain participants behaviours. We propose that participant's exploration behaviour is driven by novelty, meaning that they seek for novel states, and try to avoid un-novel states. In such a manner, participants are able to avoid the trap states, which they have visited frequently. In the 2nd block of the experiment, 2 images are swapped without informing participants. When participants observed the unexpected "state-action-next state" transition, a large surprise signal was triggered. The learning rate of the SurNoR model was modulated by the magnitude of surprise. When there is a high surprise, the model learns faster to be able to adapt to the environment. Compared to when there is no surprise, the model does not need to update too much because everything was learned before. Thus the model can learn the environment in an adaptive manner. Comparing with models without novelty-seeking mechanism and surprise-modulated learning, the SurNoR model performed the best in explaining human behaviour. This result indicates that both novelty-seeking and surprise-modulated learning are essential in learning complex

environments.

The SurNoR model is a hybrid model that contains both model-free learning and model-based learning modules. Previous behavioural experiments have shown that humans use both model-free and model-based learning. fMRI and EEG studies showed evidences of the RPE signals in the brain (Badgaiyan & Posner, 1998; Bellebaum & Daum, 2008; Cohen et al., 2007; Doñamayor et al., 2012; Gehring & Willoughby, 2002; Gruendler et al., 2011; Haruno & Kawato, 2006; McClure et al., 2003; Nieuwenhuis et al., 2005; O'Doherty et al., 2003; Tucker et al., 2003) and the SPE (Fabiani & Friedman, 1995; Gläscher et al., 2010; Opitz, Mecklinger, Friederici, & von Cramon, 1999) signals in the brain. However, the unanswered question is how model-free and model based learning can be integrated in one framework and to determine under which conditions either one prevails. Gläscher and Daw (Gläscher et al., 2010) proposed a hybrid model combining a straightforward model-based learner (Forward-Learner) and a Q(0)-learner into one model. The trade-off between the two learners was determined by a free parameter changes over time (Camerer & Hua Ho, 1999). But Gläscher's model is fairly simple and only computes the SPE as state-action-state transition difference before and after an observation (see Chapter 1.1.2 for details). It does not have the ability to navigate through complex environments because it does not have the mechanism to avoid the 'trap' states. Besides, Gläscher's model does not have an adaptive learning rate, which means that the model learns environmental changes slowly. Another hybrid model proposed by Lee and O'Doherty (Lee, Shimojo, & O'Doherty, 2014) is called the arbitration model. The arbitration model contains three layers of computation. The first layer is composed of model-based/model-free models to estimate the SPEs and RPEs. The second layer computes the reliabilities of each model based on the prediction error. The lower the prediction error, the higher reliability. For example, if the SPE is closer to 0 than the RPE, the model-based model is more reliable. The third layer balances the weight of model-based/model-free models based on the two estimated reliabilities. In this model, the model-based/model-free balance is controlled by their prediction errors but not a free parameter. However, it still faces the issue of not being able to avoid 'trap' states in our environment because the SPEs of the 'trap' states decrease fast towards 0, making the arbitration model a model-based model. Without a intrinsic desire of leaving the 'trap' state, such as novelty, the arbitration model will likely stay in the 'trap' states forever. Compare with previously proposed hybrid RL models, our SurNoR model is able to learn complex environments using novelty as intrinsic motivation, and to adapt fast to environmental changes. Furthermore, the SurNoR model can balance the exploration/exploitation behaviours by changing the weights of the model-based/model-free learning module. The SurNoR parameters indicates that participants rely on model-based learning before they find the external reward, which leads to the exploration behaviour. After obtaining the external reward, participants switch to model-free learning, which leads to the exploitation behaviour.

With the SurNoR model, we can estimate the novelty signal for each state visit, and the surprise signal for each state transition. By regressing the estimated novelty and surprise signal with recorded EEG amplitudes, we found that the novelty is reflected in the time window between

80-130ms after the state onset, and the surprise is reflected in the time window between 150-210ms after the state onset. The novelty of a state is defined as the empirical probability of seeing the state in the SurNoR model, and is considered as a global inference signal over the state distribution of the environment. According to Dehaene, Meyniel, Wacongne, Wang and Pallier (2015), when human observe a sequence, the mental process can be described in different abstraction levels. The most basic level is to count the item frequency no matter in which order it is (Armstrong, Frost, & Christiansen, 2017; Garrido, Kilner, Stephan, & Friston, 2009; Grill-Spector, Henson, & Martin, 2006; Näätänen, Paavilainen, Rinne, & Alho, 2007; Santolin & Saffran, 2018). This item frequency level corresponds to the novelty signal in our SurNoR model. Maheu et al. (2019) found that in a sequence observation task, the frequency of items is reflected in an early post-stimulus time window around 60-130ms, which is in line with our findings. Different from the basic level where the item frequency is computed, the higher level in Dehaene's theory is to estimate the transition probabilities between items, where the order of the item sequence matters. This transition probability level corresponds to the surprise signal in our case. Many studies have observed that human learn the transition probability between events or items, and use it to make decisions (Domenech & Dreher, 2010; Higashi, Minami, & Nakauchi, 2017; Maheu et al., 2019; Meyniel & Dehaene, 2017; Meyniel, Schlunegger, & Dehaene, 2015; Mittag, Takegata, & Winkler, 2016). Maheu et al (2019) reported that the mid-latency brain waves (160-320ms in their study) reflect the transition probability changes in frontal and central brain regions in MEG recordings. Similarly, in Meyniel et al. (2016), the violation of expectations, i.e., the changes in the learned transition probability, is reflected by the P300 EEG component, which is close to the surprise-related time window found in our study. The novelty- and surprise-corresponded time window found in our study is in line with previous studies and extend the observation from simple observation task to complex decision making tasks.

In the SurNoR model we considered novelty as an intrinsic reward to motivate learning, it is important to know how strong it reacts to an extrinsic reward. In chapter 5, I designed an environment with an 'infinite' state, aiming to test whether human participants are driven more by the intrinsic or the extrinsic reward. Preliminary results show that when multiple rewards exist in an environment, novelty-seeking behaviour is elicited. Previous studies have shown similar results. In a study of Marvin and Shohamy (2016), curiosity (novelty in our case) was observed to be a motivation to obtain rewards, and the positive information, such as gaining novelty and reward, can enhance the formation of long-term memory. Physiological evidences from animals and humans show that the novelty-seeking behaviour is affected by the dopamine level (Beaufour et al., 2001; Bevins et al., 2002; Bódi et al., 2009), which could be an explanation on why reward and novelty works not in competition but in cooperation. Since our studying is still undergoing, further work needs to be done to elucidate the relationship between novelty and reward.

In this thesis, from Chapter 2 to Chapter 4, I studied three different RL component which are the eligibility trace, the RPE and the SPE. In the EEG study of the RPE and the SPE, I found that the three prediction error signals (RPE, novelty and surprise) are distributed in

different time windows. Novelty occurs in the earliest time window (80-130ms), followed by surprise (150-210ms) and RPE comes the latest (280-380ms). The order of the three signals gives us a hint about the information process and cognitive load required for each signal. Novelty is a signal relatively easy to compute compared to the other two, which allows it to be processed early and fast. Surprise, on the other hand, is computed by comparing the learned and observed transitions, which requires more cognitive load than computing novelty. The RPE is computed as the difference between expected reward and actual reward. The expected reward at a state is considered as the sum of the reward offered by the environment at this state and the discounted future rewards from this state. Thus, in the computation of RPE, memory retrieval is needed for estimating the reward. This memory retrieval process makes the RPE computation longer and requires more cognitive load than the other two signals.

According to previous studies, dopaminergic neuron activity affects not only reward processing, but also novelty processing (Beaufour et al., 2001; Bevins et al., 2002; Bódi et al., 2009). These results support our SurNoR model of treating novelty as intrinsic reward. Unfortunately, in our SPE experiment design, novelty is closely associated with reward, which means the states with high reward values also have high novelty. We provide a potential solution to dissociate novelty and reward in order to study the two signals separately in the experiment with 'infinite' state. Our EEG result shows that even though both novelty and reward are affected by dopamine, human brain processes the two signals separately. However, it remains unknown how the two signals are transmitted in the brain.

In this thesis, we used and tested many classic RL models such as the model-free models SARSA, Q-learning, the model-based model Forward-Learner, and the newly proposed hybrid model SurNoR, etc. However, as George Box said *"all models are wrong, but some are useful"*, the RL models we used to explain human behaviours in this thesis could be wrong, but they can give us information on the possible mechanisms humans use in learning. Humans may compute the RPEs as the SARSA model does, they may compute novelty and surprise as the SurNoR does, they may also use some unknown reinforcing signals to achieve the same outcome. Algorithms that are not used in this thesis, and algorithms that are not developed yet may produce similar results and explain behaviours and EEG waveform better. The current and future study of reinforcing signals, such as the RPE, novelty, surprise and others, can help us understanding better the learning mechanism, and thus to help solving problems in education, self-development, social interaction and machine learning.

Bibliography

Adams, R. P., & MacKay, D. J. C. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

Anderson, B. A. (2017). Reward processing in the value-driven attention network: reward signals tracking cue identity and location. *Social cognitive and affective neuroscience*, *12*(3), 461–467.

Angela, J. Y., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692.

Armstrong, B. C., Frost, R., & Christiansen, M. H. (2017). The long road of statistical learning research: past, present and future.

Arsenault, J. T., Nelissen, K., Jarraya, B., & Vanduffel, W. (2013). Dopaminergic reward signals selectively decrease fmri activity in primate visual cortex. *Neuron*, *77*(6), 1174–1186.

Badgaiyan, R. D., & Posner, M. I. (1998). Mapping the cingulate cortex in response selection and monitoring. *Neuroimage*, *7*(3), 255–260.

Baker, T. E., & Holroyd, C. B. (2008). Which way do I go? Neural activation in response to feedback and spatial processing in a virtual T-maze. *Cerebral Cortex*, *19*(8), 1708–1722.

Barto, A. G. (1995). Adaptive critics and the basal ganglia, models of information processing in the basal ganglia. *Houk J C, Davis J L, Beiser D G*.

Barto, A. G., Mirolli, M., & Baldassarre, G. (2013). Novelty or surprise? *Frontiers in psychology*, *4*, 907.

Barto, A. G., & Sutton, R. S. (1981a). *Goal seeking components for adaptive intelligence: an initial assessment* (tech. rep. No. AFWAL-TR-81-1070). Air Force Wright Aeronautical Laboratories/Avionics Laboratory, Wright-Patterson AFB.

Barto, A. G., & Sutton, R. S. (1981b). Landmark learning: an illustration of associative search. *Biological cybernetics*, *42*(1), 1–8.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5), 834–846.

Beaufour, C. C., Le Bihan, C., Hamon, M., & Thiébot, M.-H. (2001). Extracellular dopamine in the rat prefrontal cortex during reward-, punishment-and novelty-associated behaviour. effects of diazepam. *Pharmacology Biochemistry and Behavior*, *69*(1-2), 133–142.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, *10*(9), 1214.

Bellebaum, C., & Daum, I. (2008). Learning-related changes in reward expectancy are reflected in the feedback-related negativity. *European Journal of Neuroscience*, *27*(7), 1823–1835.

Bellebaum, C., Polezzi, D., & Daum, I. (2010). It is less than you expected: the feedback-related negativity reflects violations of reward magnitude expectations. *Neuropsychologia*, *48*(11), 3343–3350.

Bellman, R. (1957a). A Markovian decision process. *Journal of mathematics and mechanics*, 679–684.

Bellman, R. (1957b). Dynamic programming. *Princeton, USA: Princeton University Press*, *1*(2), 3.

Bevins, R. A., Besheer, J., Palmatier, M. I., Jensen, H. C., Pickett, K. S., & Eurek, S. (2002). Novel-object place conditioning: behavioral and dopaminergic processes in expression of novelty reward. *Behavioural brain research*, *129*(1-2), 41–50.

Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S., & Magee, J. C. (2017). Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science*, *357*(6355), 1033–1036.

Blodgett, H. C. (1929). The effect of the introduction of reward upon the maze performance of rats. *University of California publications in psychology*.

Bódi, N., Kéri, S., Nagy, H., Moustafa, A., Myers, C. E., Daw, N., ... Gluck, M. A. (2009). Reward-learning and the novelty-seeking personality: a between-and within-subjects study of the effects of dopamine agonists on young parkinson's patients. *Brain*, *132*(9), 2385–2395.

Bogacz, R., McClure, S. M., Li, J., Cohen, J. D., & Montague, P. R. (2007). Short-term memory traces for action bias in human reinforcement learning. *Brain research*, *1153*, 111–121.

Bunzeck, N., Dayan, P., Dolan, R. J., & Duzel, E. (2010). A common mechanism for adaptive scaling of reward and novelty. *Human brain mapping*, *31*(9), 1380–1394.

Camerer, C., & Hua Ho, T. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, *67*(4), 827–874.

Cammann, R. (1990). Is there a mismatch negativity (MMN) in visual modality? *Behavioral and Brain Sciences*, *13*(2), 234–235.

Clarke, A. M., Friedrich, J., Tartaglia, E. M., Marchesotti, S., Senn, W., & Herzog, M. H. (2015). Human and machine learning in non-markovian decision making. *PloS One*, *10*(4), e0123105.

Cohen, M. X., Elger, C. E., & Ranganath, C. (2007). Reward expectation modulates feedback-related negativity and EEG spectra. *Neuroimage*, *35*(2), 968–978.

Cohen, M. X., & Ranganath, C. (2007). Reinforcement learning signals predict future decisions. *Journal of Neuroscience*, *27*(2), 371–378.

Courchesne, E., Hillyard, S. A., & Galambos, R. (1975). Stimulus novelty, task relevance and the visual evoked potential in man. *Electroencephalography and clinical neurophysiology*, *39*(2), 131–143.

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.

Deutsch, J. A. (1954). A machine with insight. *Quarterly Journal of Experimental Psychology*, *6*(1), 6–11.

Domenech, P., & Dreher, J.-C. (2010). Decision threshold modulation in the human brain. *Journal of Neuroscience*, *30*(43), 14305–14317.

Doñamayor, N., Schoenfeld, M. A., & Münte, T. F. (2012). Magneto-and electroencephalographic manifestations of reward anticipation and delivery. *Neuroimage*, *62*(1), 17–29.

Dunning, J. P., & Hajcak, G. (2007). Error-related negativities elicited by monetary loss and cues that predict loss. *Neuroreport*, *18*(17), 1875–1878.

Eppinger, B., Mock, B., & Kray, J. (2009). Developmental differences in learning and error processing: evidence from ERPs. *Psychophysiology*, *46*(5), 1043–1053.

Fabiani, M., & Friedman, D. (1995). Changes in brain activity patterns in aging: the novelty oddball. *Psychophysiology*, *32*(6), 579–594.

Faraji, M., Preuschoff, K., & Gerstner, W. (2018). Balancing new against old information: the role of puzzlement surprise in learning. *Neural computation*, *30*(1), 34–83.

Fisher, S. D., Robertson, P. B., Black, M. J., Redgrave, P., Sagar, M. A., Abraham, W. C., & Reynolds, J. N. (2017). Reinforcement determines the timing dependence of corticostriatal synaptic plasticity in vivo. *Nature communications*, *8*(1), 334.

Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2018). Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews*, *90*, 486–501.

Friston, K. J., Tononi, G., Reeke, G. N., Sporns, O., & Edelman, G. M. (1994). Value-dependent selection in the brain: simulation in a synthetic neural model. *Neuroscience*, *59*(2), 229–243.

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: a review of underlying mechanisms. *Clinical neurophysiology*, *120*(3), 453–463.

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, *295*(5563), 2279–2282.

Gernand, L., & Fenske, N. (2009). Understanding AIC and BIC in model selection. *Handreichungen zum Vortrag vom*, *20*, 1–18.

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. *Frontiers in neural circuits*, *12*.

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.

Goyer, J. P., Woldorff, M. G., & Huettel, S. A. (2008). Rapid electrophysiological brain responses are influenced by both valence and magnitude of monetary rewards. *Journal of cognitive neuroscience*, *20*(11), 2058–2069.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in cognitive sciences*, *10*(1), 14–23.

Gruendler, T. O., Ullsperger, M., & Huster, R. J. (2011). Event-related potential correlates of performance-monitoring in a lateralized time-estimation task. *PloS One*, *6*(10), e25591.

Hajcak, G., Holroyd, C. B., Moser, J. S., & Simons, R. F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology*, *42*(2), 161–170.

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological psychology*, *71*(2), 148–154.

Haruno, M., & Kawato, M. (2006). Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *Journal of neurophysiology*, *95*(2), 948–959.

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., & Platt, M. L. (2011). Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *Journal of Neuroscience*, *31*(11), 4178–4187.

He, K., Huertas, M., Hong, S. Z., Tie, X., Hell, J. W., Shouval, H., & Kirkwood, A. (2015). Distinct eligibility traces for LTP and LTD in cortical synapses. *Neuron*, *88*(3), 528–538.

Hewig, J., Hecht, H. et al. (2007). Decisionmaking in blackjack: an electrophysiological analysis. cereb cortex 17: 865.

Hewig, J., Trippe, R., Hecht, H., Coles, M. G., Holroyd, C. B., & Miltner, W. H. (2006). Decision-making in blackjack: an electrophysiological analysis. *Cerebral Cortex*, *17*(4), 865–877.

Higashi, H., Minami, T., & Nakauchi, S. (2017). Variation in event-related potentials by state transitions. *Frontiers in human neuroscience*, *11*, 75.

Holland, P. C. (1997). Brain mechanisms for changes in processing of conditioned stimuli in pavlovian conditioning: implications for behavior theory. *Animal Learning & Behavior*, *25*(4), 373–399.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, *109*(4), 679.

Holroyd, C. B., Larsen, J. T., & Cohen, J. D. (2004). Context dependence of the event-related brain potential associated with reward and punishment. *Psychophysiology*, *41*(2), 245–253.

Houk, J. C., Davis, J. L., & Beiser, D. G. (1994). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In *Models of information processing in the basal ganglia* (pp. 249–270). MITP.

Karban, R. (2015). Plant learning and memory. *Plant sensing and communication*, 31–44.

Klopf, A. H. (1972). *Brain function and adaptive systems: a heterostatic theory* (tech. rep. No. AFCRL-72-0164). Air Force Cambridge Research Laboratories, Air Force Systems Command.

Kolossa, A., Fingscheidt, T., Wessel, K., & Kopp, B. (2013). A model-based approach to trial-by-trial p300 amplitude fluctuations. *Frontiers in human neuroscience*, *6*, 359.

Kolossa, A., Kopp, B., & Fingscheidt, T. (2015). A computational analysis of the neural bases of bayesian inference. *Neuroimage*, *106*, 222–237.

Krugel, L. K., Biele, G., Mohr, P. N., Li, S.-C., & Heekeren, H. R. (2009). Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proceedings of the National Academy of Sciences*, *106*(42), 17951–17956.

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.

Lynn, R. (2013). *Attention, arousal and the orientation reaction: international series of monographs in experimental psychology*. Elsevier.

Maheu, M., Dehaene, S., & Meyniel, F. (2019). Brain signatures of a multiscale process of sequence learning in humans. *Elife*, *8*, e41541.

Marvin, C. B., & Shohamy, D. (2016). Curiosity and reward: valence predicts choice and information prediction errors enhance learning. *Journal of Experimental Psychology: General*, *145*(3), 266.

Mathewson, K. J., Dywan, J., Snyder, P. J., Tays, W. J., & Segalowitz, S. J. (2008). Aging and electrocortical response to error feedback during a spatial learning task. *Psychophysiology*, *45*(6), 936–948.

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience*, *5*, 39.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339–346.

Meyniel, F., & Dehaene, S. (2017). Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences*, *114*(19), E3859–E3868.

Meyniel, F., Maheu, M., & Dehaene, S. (2016). Human inferences about sequences: a minimal transition probability model. *PLoS computational biology*, *12*(12), e1005260.

Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The sense of confidence during probabilistic learning: a normative account. *PLoS computational biology*, *11*(6), e1004305.

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a "generic" neural system for error detection. *Journal of cognitive neuroscience*, *9*(6), 788–798.

Minsky, M. L. (1954). *Theory of neural-analog reinforcement systems and its application to the brain model problem*. Princeton University.

Mittag, M., Takegata, R., & Winkler, I. (2016). Transitional probabilities are prioritized over stimulus/pattern probabilities in auditory deviance detection: memory basis for predictive sound processing. *Journal of Neuroscience*, *36*(37), 9572–9579.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of neuroscience*, *16*(5), 1936–1947.

Näätänen, R., Gaillard, A. W., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta psychologica*, *42*(4), 313–329.

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical neurophysiology*, *118*(12), 2544–2590.

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. *Nature neuroscience*, *15*(7), 1040.

Nieuwenhuis, S., Ridderinkhof, K. R., Talsma, D., Coles, M. G. H., Holroyd, C. B., Kok, A., & Van der Molen, M. W. (2002). A computational account of altered error processing in older age: dopamine and the error-related negativity. *Cognitive, Affective, & Behavioral Neuroscience*, *2*(1), 19–36.

Nieuwenhuis, S., Slagter, H. A., Von Geusau, N. J. A., Heslenfeld, D. J., & Holroyd, C. B. (2005). Knowing good from bad: differential activation of human cortical areas by positive and negative outcomes. *European Journal of Neuroscience*, *21*(11), 3161–3168.

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551–562.

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward, and decision making. *Annual review of psychology*, *68*, 73–100.

O'Doherty, J. P., Dayan, P., Friston, K. J., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337.

Opitz, B., Mecklinger, A., Friederici, A. D., & von Cramon, D. Y. (1999). The functional neuroanatomy of novelty processing: integrating ERP and fMRI results. *Cerebral Cortex*, *9*(4), 379–391.

Pascual-Marqui, R. D. et al. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find Exp Clin Pharmacol*, *24*(Suppl D), 5–12.

Pavlov, I. P. (1927). *Conditional reflexes: an investigation of the physiological activity of the cerebral cortex*. Oxford Univ. Press.

Pazo-Alvarez, P., Cadaveira, F., & Amenedo, E. (2003). MMN in the visual modality: a review. *Biological psychology*, *63*(3), 199–236.

Pearce, J. M., & Hall, G. (1980). A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review*, *87*(6), 532.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042.

Potts, G. F., Martin, L. E., Burton, P., & Montague, P. R. (2006). When things are better or worse than expected: the medial frontal cortex and the allocation of processing resources. *Journal of cognitive neuroscience*, *18*(7), 1112–1119.

Potts, G. F., Martin, L. E., Kamp, S.-M., & Donchin, E. (2011). Neural response to action and reward prediction errors: comparing the error-related negativity to behavioral errors and the feedback-related negativity to reward prediction violations. *Psychophysiology*, *48*(2), 218–228.

Preuschoff, K., & Bossaerts, P. (2007). Adding prediction risk to the theory of reward learning. *Annals of the New York Academy of Sciences*, *1104*(1), 135–146.

Reynolds, G. D. (2015). Infant visual attention and object recognition. *Behavioural brain research*, *285*, 34–43.

Roelfsema, P. R., & Ooyen, A. v. (2005). Attention-gated reinforcement learning of internal representations for classification. *Neural computation*, *17*(10), 2176–2214.

Roesch, M. R., Esber, G. R., Li, J., Daw, N. D., & Schoenbaum, G. (2012). Surprise! Neural correlates of Pearce–Hall and Rescorla–Wagner coexist within the brain. *European Journal of Neuroscience*, *35*(7), 1190–1200.

Rombouts, J. O., Bohte, S. M., Martinez-Trujillo, J., & Roelfsema, P. R. (2015). A learning rule that explains how rewards teach attention. *Visual Cognition*, *23*(1-2), 179–205.

Ross, T. (1933). Machines that think. *Scientific American*, *148*(4), 206–208.

Ruchsow, M., Grothe, J., Spitzer, M., & Kiefer, M. (2002). Human anterior cingulate cortex is activated by negative feedback: evidence from event-related potentials in a guessing task. *Neuroscience letters*, *325*(3), 203–206.

Sambrook, T. D., Hardwick, B., Wills, A. J., & Goslin, J. (2018). Model-free and model-based reward prediction errors in eeg. *NeuroImage*, *178*, 162–171.

Samuel, A. J. (1959). Aerosol dispensers and like pressurized packages. US Patent 2,904,229. Google Patents.

Santolin, C., & Saffran, J. R. (2018). Constraints on statistical learning across species. *Trends in Cognitive Sciences*, *22*(1), 52–63.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Shannon, C. E. (1952). Creative thinking. *Typescript. Reproduced*.

Shuler, M. G., & Bear, M. F. (2006). Reward timing in the primary visual cortex. *Science*, *311*(5767), 1606–1609.

Sokolov, E. N. (1990). The orienting response, and future directions of its development. *The Pavlovian journal of biological science*, *25*(3), 142–150.

Stephan, K. E., Tittgemeyer, M., Knösche, T. R., Moran, R. J., & Friston, K. J. (2009). Tractography-based priors for dynamic causal models. *Neuroimage*, *47*(4), 1628–1638.

Sutton, R. S. (1978a). A unified theory of expectation in classical and instrumental conditioning.

Sutton, R. S. (1978b). Learning theory support for a single channel theory of the brain.

Sutton, R. S. (1978c). Single channel theory: a neuronal theory of learning. *Brain Theory Newsletter*, *4*, 72–75.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, *3*(1), 9–44.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. MIT Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: an introduction*. MIT press.

Tartaglia, E. M., Clarke, A. M., & Herzog, M. H. (2017). What to choose next? a paradigm for testing human sequential decision making. *Frontiers in psychology*, *8*, 312.

Thorndike, E. L. (1911). Animal intelligence. Darien, CT, Hafner Publishing.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *nature*, *381*(6582), 520.

Tsetlin, M. L. (1973). Automaton theory and modeling of biological systems.

Tucker, D. M., Luu, P., Frishkoff, G., Quiring, J., & Poulsen, C. (2003). Frontolimbic response to negative feedback in clinical depression. *Journal of abnormal psychology*, *112*(4), 667.

Turing, A. M. (1948). Intelligent machinery. NPL. Mathematics Division.

van der Helden, J., Boksem, M. A. S., & Blom, J. H. G. (2009). The importance of failure: feedback-related negativity predicts motor learning efficiency. *Cerebral Cortex*, *20*(7), 1596–1603.

Walsh, M. M., & Anderson, J. R. (2011a). Learning from delayed feedback: neural responses in temporal credit assignment. *Cognitive, Affective, & Behavioral Neuroscience*, *11*(2), 131–143.

Walsh, M. M., & Anderson, J. R. (2011b). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences*, *108*(47), 19048–19053.

Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience & Biobehavioral Reviews*, *36*(8), 1870–1884.

Walter, W. G. (1950). An imitation of life. *Scientific American*, *182*(5), 42–45.

Watkins, C. J. C. H. (1989). Learning from delayed rewards.

Webster, M. (2008). Webster's all-in-one dictionary and thesaurus. Springfield.

Wilson, P. N., Boumphrey, P., & Pearce, J. M. (1992). Restoration of the orienting response to a light by a change in its predictive accuracy. *The Quarterly Journal of Experimental Psychology Section B*, *44*(1b), 17–36.

Witten, I. H. (1977). An adaptive optimal controller for discrete-time markov environments. *Information and control*, *34*(4), 286–295.

Woodworth, R. S., Barber, B., & Schlosberg, H. (1954). *Experimental psychology*. Oxford and IBH Publishing.

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, *345*(6204), 1616–1620.

Yasuda, A., Sato, A., Miyawaki, K., Kumano, H., & Kuboki, T. (2004). Error-related negativity reflects detection of negative reward prediction error. *Neuroreport*, *15*(16), 2561–2565.

Zhou, Z., Yu, R., & Zhou, X. (2010). To do or not to do? Action enlarges the FRN and P300 effects in outcome evaluation. *Neuropsychologia*, *48*(12), 3606–3613.

He XU

Rue des Moulins 16,
1800, Vevey

he.ayu.xu@gmail.com
(+41) 78-826-06-48

EDUCATION

École polytechnique fédérale de Lausanne, Lausanne
Ph.D Student
Laboratory of Psychophysics 2015-2020

École polytechnique fédérale de Lausanne, Lausanne
Master of Science
Computer Science 2012-2015

Queen Mary University of London, London
Beijing University of Posts and Telecommunications, Beijing
Bachelor of Science
Telecommunication Engineering with Management 2008-2012

TECHNICAL SKILLS

Languages: Python, Java, R, Matlab
Database: MySQL, SQLite
Expertise: Anaconda, Jupyter Notebook, TensorFlow, Matlab, EEG, HTML, CSS
Skills: Data Analysis, Machine Learning, Statistics, Data Structures

EXPERIENCE

Bio-Rad Laboratories, Immunohematology Division, Cressier
IT Project Manager at Customer Service Management 02.2014-08.2014

Organiser of Brain Mind Symposium "Neural Implementations of Learning Models", Lasuanne
https://bmisymposia.epfl.ch/2017winter 01.2017 - 11.2017

PROJECTS

Study of immersive-level on peripheral signals during video and audio perception, Lasuanne
Multimedia Signal Processing Group 03.2015-09.2015

Functional Connectivity study of EEG signals during perceiving pleasant and unpleasant odours, Lasuanne
Multimedia Signal Processing Group 09.2014-02.2015

Neurophysiological Origin and Meaning of Muscle Synergies, Lasuanne
Translational Neural Engineering Laboratory 09.2013-02.2014

Incorporate Cylinder Quality Measures in MCCbased crime scene fingerprint matching, Lasuanne
LIDIAP 02.2013-06.2013

Study of mobile sensor data in relation with music patterns, Lasuanne
Human Media Achieved research group 02.2012-06.2012

CERTIFICATION

Google Certificate of CodeU Program by Google 08.2017
Google Developer Scholarship by Udacity & Google 12.2017

Appendix

.1 One-shot learning and behavioral eligibility traces in sequential decision making. eLife, 8, e47643

Lehmann, M.P., Xu, H.A., Liakoni, V., Herzog, M.H., Gerstner, W., & Preuschoff, K. (2019)

.2 EEG signatures of the Reward Prediction Error at non-rewarded states. (to be submitted)

Xu, H.A., Lehmann, M.P., Gerstner, W., Herzog, M.H.

.3 Model building by exploration: surprise and novelty in reward-based learning (in preparation)

Xu, H.A., Lehmann, M.P., Modirshanechi, A., Gerstner, W., Herzog, M.H.

eLIFE elifesciences.org

RESEARCH ARTICLE

One-shot learning and behavioral eligibility traces in sequential decision making

Marco P Lehmann[1,2]*, He A Xu[3], Vasiliki Liakoni[1,2], Michael H Herzog[3†], Wulfram Gerstner[1,2†], Kerstin Preuschoff[4†]

[1]Brain-Mind-Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; [2]School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; [3]Laboratory of Psychophysics, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; [4]Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland

Abstract In many daily tasks, we make multiple decisions before reaching a goal. In order to learn such sequences of decisions, a mechanism to link earlier actions to later reward is necessary. Reinforcement learning (RL) theory suggests two classes of algorithms solving this credit assignment problem: In classic temporal-difference learning, earlier actions receive reward information only after multiple repetitions of the task, whereas models with eligibility traces reinforce entire sequences of actions from a single experience (one-shot). Here, we show one-shot learning of sequences. We developed a novel paradigm to *directly* observe which actions and states along a multi-step sequence are reinforced after a single reward. By focusing our analysis on those states for which RL with and without eligibility trace make qualitatively distinct predictions, we find direct behavioral (choice probability) and physiological (pupil dilation) signatures of reinforcement learning with eligibility trace across multiple sensory modalities.

***For correspondence:** marco.lehmann@alumni.epfl.ch

†These authors contributed equally to this work

Competing interests: The authors declare that no competing interests exist.

Funding: See page 22

Received: 05 April 2019
Accepted: 01 November 2019
Published: 11 November 2019

Reviewing editor: Thorsten Kahnt, Northwestern University, United States

Introduction

In games, such as chess or backgammon, the players have to perform a sequence of many actions before a reward is received (win, loss). Likewise in many sports, such as tennis, a sequence of muscle movements is performed until, for example, a successful hit is executed. In both examples, it is impossible to immediately evaluate the goodness of a single action. Hence the question arises: How do humans learn sequences of actions from delayed reward?

Reinforcement learning (RL) models (***Sutton and Barto, 2018***) have been successfully used to describe reward-based learning in humans (***Pessiglione et al., 2006***; ***Gläscher et al., 2010***; ***Daw et al., 2011***; ***Niv et al., 2012***; ***O'Doherty et al., 2017***; ***Tartaglia et al., 2017***). In RL, an action (e.g. moving a token or swinging the arm) leads from an old state (e.g. configuration of the board, or position of the body) to a new one. Here, we grouped RL theories into two different classes. The first class, containing classic Temporal-Difference algorithms (such as *TD-0* ***Sutton, 1988***) cannot support one-shot learning of long sequences, because multiple repetitions of the task are needed before reward information arrives at states far away from the goal. Instead, one-shot learning requires algorithms that keep a memory of past states and actions making them eligible for later, that is delayed reinforcement. Such a memory is a key feature of the second class of RL theories – called *RL with eligibility trace* –, which includes algorithms with explicit eligibility traces (***Sutton, 1988***; ***Watkins, 1989***; ***Williams, 1992***; ***Peng and Williams, 1996***; ***Singh and Sutton, 1996***)

and related reinforcement learning models (**Watkins, 1989**; **Moore and Atkeson, 1993**; **Blundell et al., 2016**; **Mnih et al., 2016**; **Sutton and Barto, 2018**).

Eligibility traces are well-established in computational models (**Sutton and Barto, 2018**), and supported by synaptic plasticity experiments (**Yagishita et al., 2014**; **He et al., 2015**; **Bittner et al., 2017**; **Fisher et al., 2017**; **Gerstner et al., 2018**). However, it is unclear whether humans show one-shot learning, and a direct test of predictions that are manifestly different between the classes of RL models with and without eligibility trace has never been performed. Multi-step sequence learning with delayed feedback (**Gläscher et al., 2010**; **Daw et al., 2011**; **Walsh and Anderson, 2011**; **Tartaglia et al., 2017**) offers a way to directly compare the two, because the two classes of RL models make *qualitatively* different predictions. Our question can therefore be reformulated more precisely: Is there evidence for RL with eligibility trace in the form of one-shot learning? In other words, are actions and states more than one step away from the goal, reinforced after a single rewarded experience? And if eligibility traces play a role, how many states and actions are reinforced by a single reward?

To answer these questions, we designed a novel sequential learning task to directly observe which actions and states of a multi-step sequence are reinforced. We exploit that after a single reward, models of learning without eligibility traces (our null hypothesis) and with eligibility traces (alternative hypothesis) make qualitatively distinct predictions about changes in action-selection bias and in state evaluation (**Figure 1**). This qualitative difference in the second episode (i.e. after a single reward) allows us to draw conclusions about the presence or absence of eligibility traces independently of specific model fitting procedures and independently of the choice of physiological correlates, be it EEG, fMRI, or pupil responses. We therefore refer to these qualitative differences as 'direct' evidence.

We measure changes in action-selection bias from behavior and changes in state evaluation from a physiological signal, namely the pupil dilation. Pupil responses have been previously linked to decision making, and in particular to variables that reflect changes in state value such as expected reward, reward prediction error, surprise, and risk (**O'Doherty et al., 2003**; **Jepma and Nieuwenhuis, 2011**; **Otero et al., 2011**; **Preuschoff et al., 2011**). By focusing our analysis on those states for which the two hypotheses make distinct predictions after a *single* reward ('one-shot'), we find direct behavioral and physiological signatures of reinforcement learning with eligibility trace. The observed one-shot learning sheds light on a long-standing question in human reinforcement learning (**Bogacz et al., 2007**; **Daw et al., 2011**; **Walsh and Anderson, 2011**; **Walsh and Anderson, 2012**; **Weinberg et al., 2012**; **Tartaglia et al., 2017**).

Results

Since we were interested in one-shot learning, we needed an experimental multi-step action paradigm that allowed a comparison of behavioral and physiological measures between episode 1 (before any reward) and episode 2 (after a single reward). Our learning environment had six states plus a goal G (**Figures 1** and **2**) identified by clip-art images shown on a computer screen in front of the participants. It was designed such that participants were likely to encounter in episode 2 the same states D1 (one step away from the goal) and/or D2 (two steps away) as in episode 1 (**Figure 1 (a)**). In each state, participants chose one out of two actions, 'a' or 'b', and explored the environment until they discovered the goal G (the image of a reward) which terminated the episode. The participants were instructed to complete as many episodes as possible within a limited time of 12 min (Materials and methods).

The first set of predictions applied to the state D1 which served as a control if participants were able to learn, and assign value to, states or actions. Both classes of algorithms, with or without eligibility trace, predicted that effects of learning after the first reward should be reflected in the action choice probability during a subsequent visit of state D1 (**Figure 1 (b)**). For estimated effect size, see subsection Q-lambda model predictions in 'Methods. Furthermore, any physiological variable that correlates with variables of reinforcement learning theories, such as action value Q, state value V, or TD-error, should increase at the second encounter of D1. To assess this effect of learning, we measured the pupil dilation, a known physiological marker for learning-related signals (**O'Doherty et al., 2003**; **Jepma and Nieuwenhuis, 2011**; **Otero et al., 2011**; **Preuschoff et al., 2011**). The advantage of our hypothesis-driven approach was that we did not need to make assumptions about the

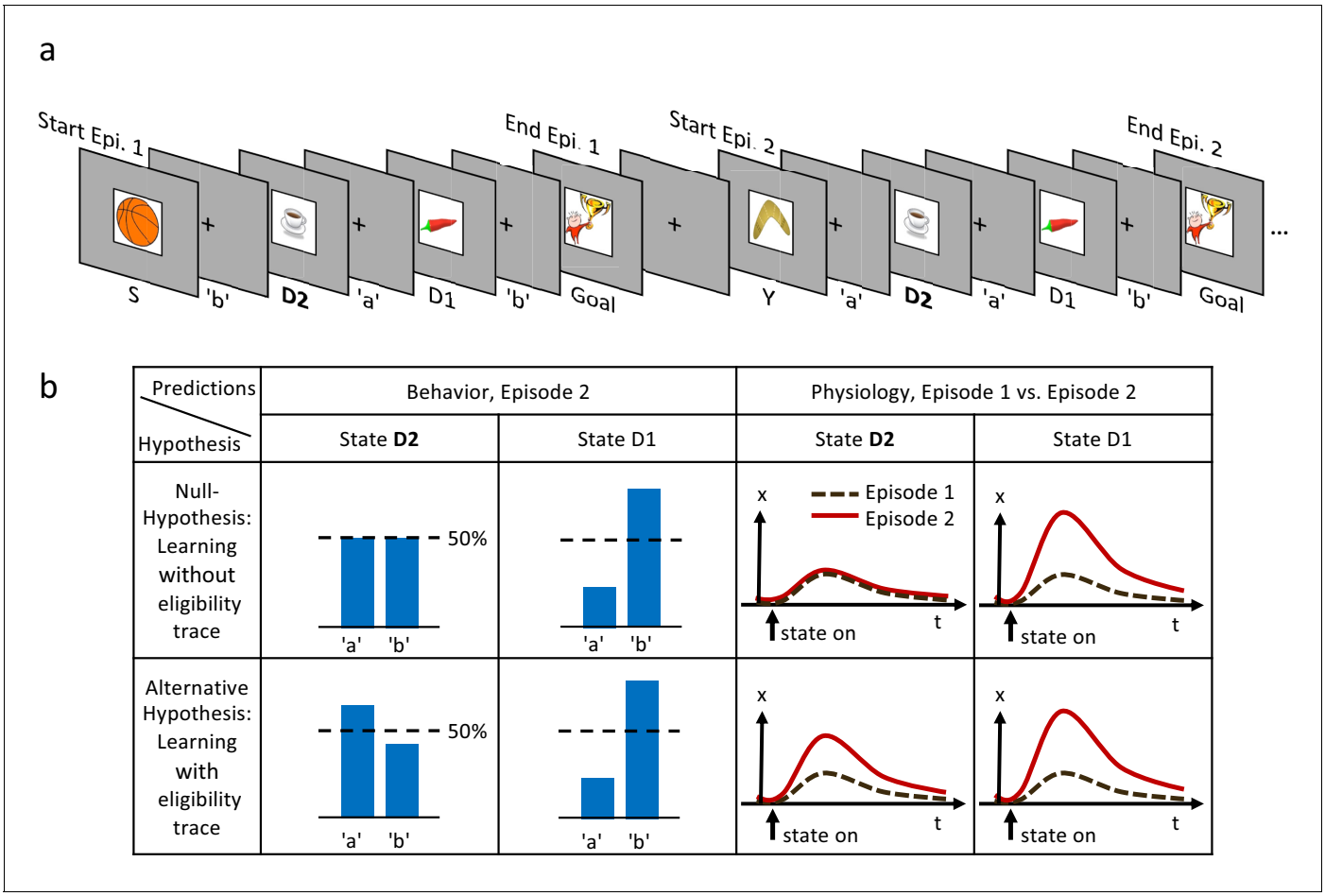


Figure 1. Experimental design and hypothesis. (**a**) Typical state-action sequences of the first two episodes. At each state, participants execute one of two actions, 'a' or 'b', leading to the next state. Here, the participant discovered the goal state after randomly choosing three actions: 'b' in state S (Start), 'a' in D2 (two actions from the goal), and 'b' in D1 (one action from the goal). Episode 1 terminated at the rewarding goal state. Episode 2 started in a new state, Y. Note that D2 and D1 already occurred in episode 1. In this example, the participant repeated the actions which led to the goal in episode 1 ('a' at D2 and 'b' at D1). (**b**) Reinforcement learning models make predictions about such behavioral biases, and about learned properties (such as action value Q, state value V or TD-errors, denoted as x) presumably observable as changes in a physiological measure (e.g. pupil dilation). Null Hypothesis: In RL without eligibility traces, only the state-action pair immediately preceding a reward is reinforced, leading to a bias at state D1, but not at D2 (50%-line). Similarly, the state value of D2 does not change and therefore the physiological response at the D2 in episode 2 (solid red line) should not differ from episode 1 (dashed black line). Alternative Hypothesis: RL with eligibility traces reinforces decisions further back in the state-action history. These models predict a behavioral bias at D1 and D2, and a learning-related physiological response at the onset of these states after a single reward. The effects may be smaller at state D2 because of decay factors in models with eligibility traces.

neurophysiological mechanisms causing pupil changes. Comparing the pupil dilation at state D1 in episode 1 to episode 2 (***Figure 1(b)***, null hypothesis *and* alternative), provided a baseline for the putative effect.

Our second set of predictions concerned state D2. RL without eligibility trace (null hypothesis) such as *TD-0*, predicted that the action choice probability at D2 during episode 2 should be at 50 percent, since information about the reward at the goal state G cannot 'travel' two steps. However, the class of RL with eligibility trace (alternative hypothesis) predicted an increase in the probability of choosing the correct action, that is the one leading toward the goal (For estimated effect size, see subsection Q-lambda model predictions in *Methods*). The two hypotheses also made different predictions about the pupil response to the onset of state D2. Under the null hypothesis, the evaluation of the state D2 could not change after a single reward. In contrast, learning with eligibility trace predicted a change in state evaluation, presumably reflected in pupil dilation (***Figure 1(b)***).

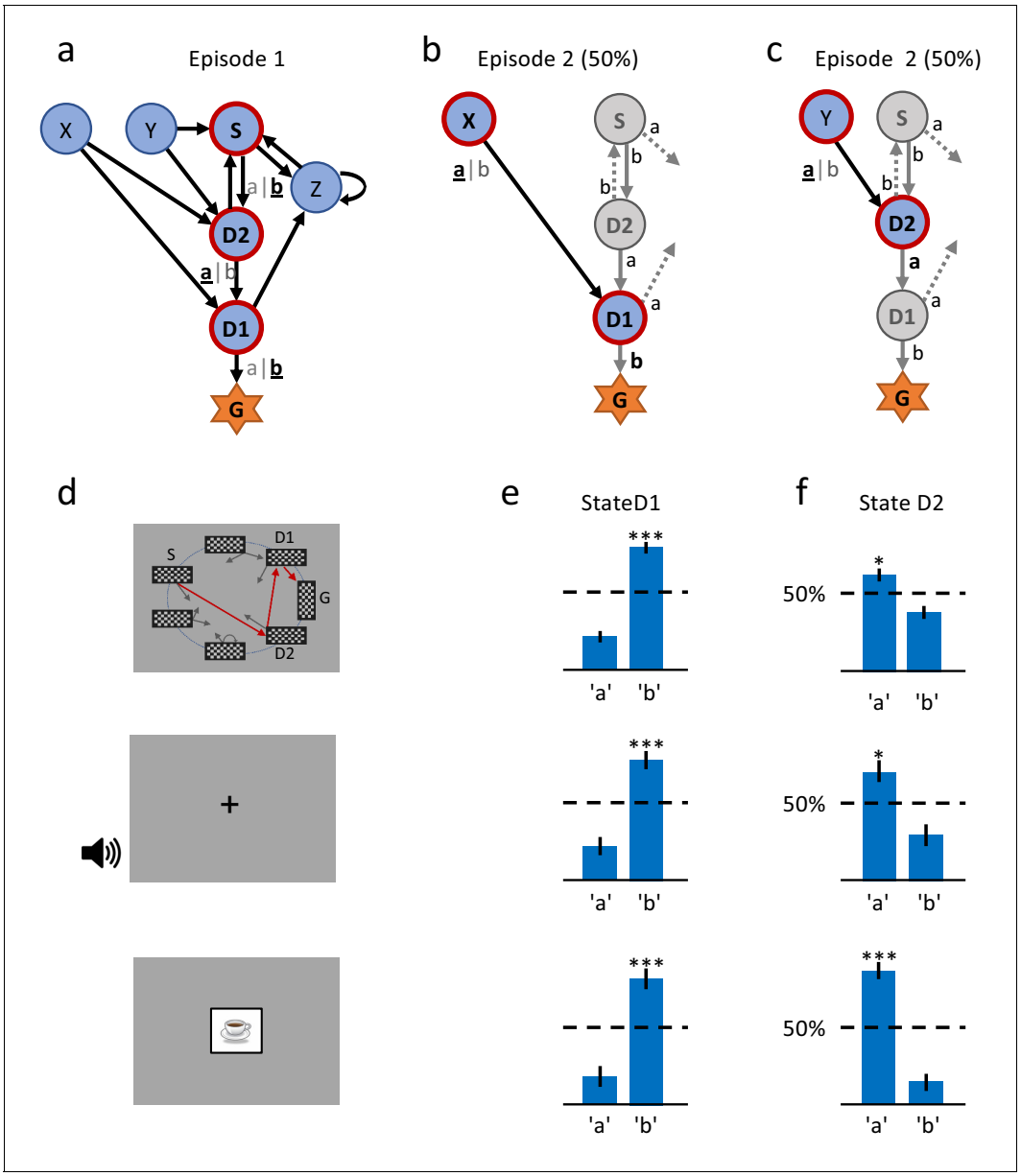


Figure 2. A single delayed reward reinforces state-action associations. (**a**) Structure of the environment: six states, two actions, rewarded goal G. Transitions (arrows) were predefined, but actions were attributed to transitions *during* the experiment. Unbeknownst to the participants, the first actions always led through the sequence S (Start), D2 (two steps before goal), D1 (one step before goal) to G (Goal). Here, the participant chose actions 'b', 'a', 'b' (underlined boldface). (**b**) Half of the experiments, started episode 2 in X, always leading to D1, where we tested if the action rewarded in episode 1 was repeated. (**c**) In the other half of experiments, we tested the decision bias in episode 2 at D2 ('a' in this example) by starting from Y. (**d**) The same structure was implemented in three conditions. In the *spatial* condition (22 participants, *top* row in Figures (**d**), (**e**) and (**f**)), each state is identified by a fixed location (randomized across participants) of a checkerboard, flashed for a 100 ms on the screen. Participants only see one checkerboard at a time; the red arrows and state identifiers S, D2, D1, G are added to the figure to illustrate a first episode. In the *sound* condition (15 participants, *middle* row), states are represented by unique short sounds. In the *clip-art* condition (12 participants, *bottom* row), a unique image is used for each state. (**e**) Action selection bias in state D1, in episode 2, averaged across all participants. (**f**) In all three conditions the action choices at D2 were significantly different from chance level (dashed horizontal line) and biased toward the actions that have led to reward in episode 1. Error bars: SEM, $^{*}p<0.05$, $^{***}p<0.001$. For clarity, actions are labeled 'a' and 'b' in (**e**) and (**f**), consistent with panels (**a**) - (**c**), even though actual choices of participants varied.

Participants could freely choose actions, but in order to maximize encounters with states D1 and D2, we assigned actions to state transitions 'on the fly'. In the first episode, all participants started in state S (***Figure 1 (a)*** and ***2(a)***) and chose either action 'a' or 'b'. Independently of their choice and unbeknownst to the participants, the first action brought them always to state D2, two steps away from the goal. Similarly, in D2, participants could freely choose an action but always transitioned to D1, and with their third action, to G. These initial actions determined the assignment of state-action pairs to state transitions for all remaining episodes in this environment. For example, if, during the first episode, a participant had chosen action 'a' in state D2 to initiate the transition to D1, then action 'a' brought this participant in all future encounters of D2 to D1, whereas action 'b' brought her from D2 to Z (***Figure 2***). In episode 2, half of the participants started from state Y. Their first action always brought them to D2, which they had already seen once during the first episode. The other half of the participants started in state X and their first action brought them to D1 (***Figure 2 (b)***). Participants who started episode 2 in state X started episode 3 in state Y and vice versa. In episodes 4 to 7, the starting states were randomly chosen from {S, D2, X, Y, Z}. After seven episodes, we considered the task as solved, and the same procedure started again in a new environment (see Materials and methods for the special cases of repeated action sequences). This task design allowed us to study human learning in specific and controlled state sequences, without interfering with the participant's free choices.

Behavioral evidence for one-shot learning

As expected, we found that the action taken in state D1 that led to the rewarding state G was reinforced after episode 1. Reinforcement was visible as an action bias toward the correct action when D1 was seen again in episode 2 (***Figure 2 (e)***). This action bias is predicted by many different RL algorithms including the early theories of ***Rescorla and Wagner (1972)***.

Importantly, we also found a strong action bias in state D2 in episode 2: participants repeated the correct action (the one leading toward the goal) in 85% of the cases. This strong bias is significantly different from chance level 50% ($p<0.001$; ***Figure 2 (f)***), and indicates that participants learned to assign a positive value to the correct state-action pair after a *single exposure* to state D2 and a *single reward* at the end of episode 1. In other words, we found evidence for one-shot learning in a state two steps away from goal in a multi-step decision task.

This is compatible with our alternative hypothesis, that is the broad class of RL 'with eligibility trace', (***Sutton, 1988***; ***Watkins, 1989***; ***Williams, 1992***; ***Moore and Atkeson, 1993***; ***Peng and Williams, 1996***; ***Singh and Sutton, 1996***; ***Mnih et al., 2016***; ***Blundell et al., 2016***; ***Sutton and Barto, 2018***) that keep explicit or implicit memories of past state-action pairs (see Discussion). However, it is not compatible with the null hypothesis, that is RL 'without eligibility trace'. In both classes of algorithms, action biases or values that reflect the expected future reward are assigned to states. In RL 'without eligibility trace', however, value information collected in a single action step is shared only between neighboring states (for example between states G and D1), whereas in RL 'with eligibility trace' value information can reach state D2 after a single episode. Importantly, the above argument is both fundamental and qualitative in the sense that it does not rely on any specific choice of parameters or implementation details of an algorithm. Our finding can be interpreted as a signature of a behavioral eligibility trace in human multi-step decision making and complements the well-established synaptic eligibility traces observed in animal models (***Yagishita et al., 2014***; ***He et al., 2015***; ***Bittner et al., 2017***; ***Fisher et al., 2017***; ***Gerstner et al., 2018***).

We wondered whether the observed one-shot learning in our multi-step decision task depended on the choice of stimuli. If clip-art images helped participants to construct an imaginary story (e.g. with the method of loci; ***Yates, 1966***) in order to rapidly memorize state-action associations, the effect should disappear with other stimuli. We tested participants in environments where states were defined by acoustic stimuli (2nd experiment: *sound* condition) or by the spatial location of a black-and-white rectangular grid on the grey screen (3rd experiment: *spatial*condition; see ***Figure 2*** and Materials and methods). Across all conditions, results were qualitatively similar (***Figure 2 (f)***): not only the action directly leading to the goal (i.e. the action in D1) but also the correct action in state D2 were chosen in episode 2 with a probability significantly different from random choice. This behavior is consistent with the class of RL with eligibility trace, and excludes all algorithms in the class of RL without eligibility trace.

Even though results are consistent across different stimuli, we cannot exclude that participants simply memorize state-action associations independently of the rewards. To exclude a reward-independent memorization strategy, we performed a control experiment in which we tested the action-bias at state D2 (see ***Figure 3***) in the absence of a reward. In a design similar to the *clip-art* condition (***Figure 1 (a)***), the participants freely chose actions that moved them through a defined, non-rewarded, sequence of states (namely S-D2-D1-N-Y-D2, see ***Figure 3 (b)*** during the first episode. By design of the control experiment, participants reach the state D2 twice before they encounter any reward. Upon their second visit of state D2, we measured whether participants repeated the same action as during their first visit. Such a repetition bias could be explained if participants tried to memorize and repeat state-action associations even in the absence of a reward between the two visits. In the control experiment we observed a weak non-significant (p=0.45) action-repetition bias of only 56% (***Figure 3 (c)*** in contrast to the main experiment (with a reward between the first and second encounter of state D2) where we observed a repetition bias of 85%. These results indicate that earlier rewards influence the action choice when a state is encountered a second time.

Reinforcement learning with eligibility trace is reflected in pupil dilation

We then investigated the time-series of the pupil diameter. Both, the null and the alternative hypothesis predict a change in the evaluation of state D1, when comparing the second with the first encounter. Therefore, if the pupil dilation indeed serves as a proxy for a learning-related state evaluation (be it Q-value, V-value, or TD-error); we should observe a difference between the pupil response to the onset of state D1 before (episode 1) and after (episode 2) a single reward.

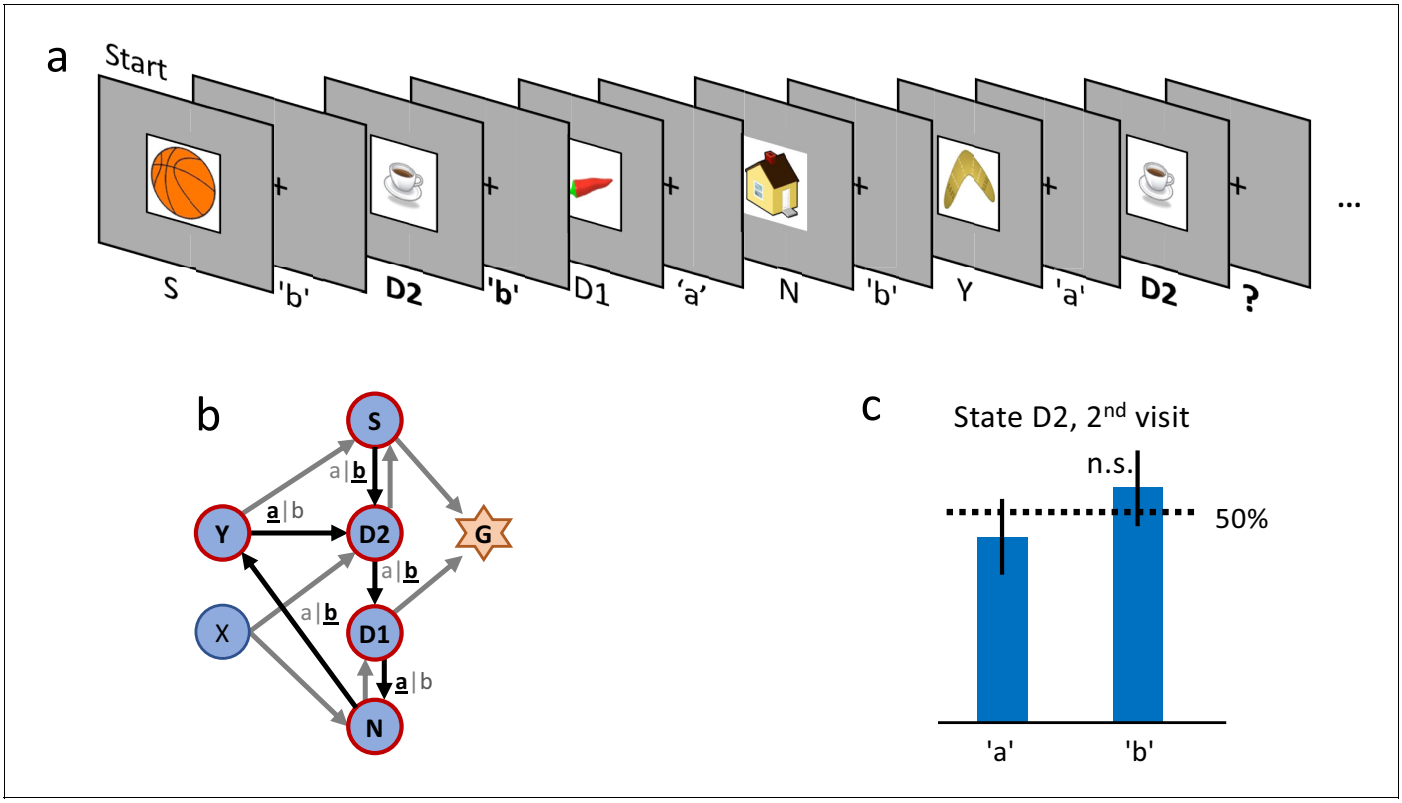


Figure 3. Control experiment without reward. (**a**) Sequence of the first six state-action pairs in the first control experiment. The state D2 is visited twice and the number of states between the two visits is the same as in the main experiment. The original goal state has been replaced by a non-rewarded state N. The control experiment focuses on the behavior during the second visit of state D2, further state-action pairs are not relevant for this analysis. (**b**) The structure of the environment has been kept as close as possible to the main experiment (***Figure 2 (a)***). (**c**) Ten participants performed a total of 32 repetitions of this control experiment. Participants show an average action-repetition bias of 56%. This bias is not significantly different from the 50% chance level ($p = 0.45$) and much weaker than the 85% observed in the main experiment (***Figure 2 (f)***).

We extracted (Materials and methods) the time-series of the pupil diameter, focused on the interval [0s, 3s] after the onset of states D2 or D1, and averaged the data across participants and environments (***Figure 4***, black traces). We observed a significant change in the pupil dilatory response to stimulus D1 between episode 1 (black curve) and episode 2 (red curve). The difference was computed per time point (paired samples t-test); significance levels were adjusted to control for false discovery rate (FDR, ***Benjamini and Hochberg, 1995***) which is a conservative measure given the temporal correlations of the pupillometric signal. This result suggests that participants change the evaluation of D1 after a single reward and that this change is reflected in pupil dilation.

Importantly, the pupil dilatory response to the state D2 was also significantly stronger in episode 2 than in episode 1. Therefore, if pupil diameter is correlated with the state value V, the action value Q, the TD-error, or a combination thereof, then the class of *RL without eligibility trace* must be excluded as an explanation of the pupil response (i.e. we can reject the null hypothesis in ***Figure 1***).

However, before drawing such a conclusion we controlled for correlations of pupil response with other parameters of the experiment. First, for visual stimuli, pupil responses changed with stimulus luminance. The rapid initial contraction of the pupil observed in the *clip-art* condition (bottom row in

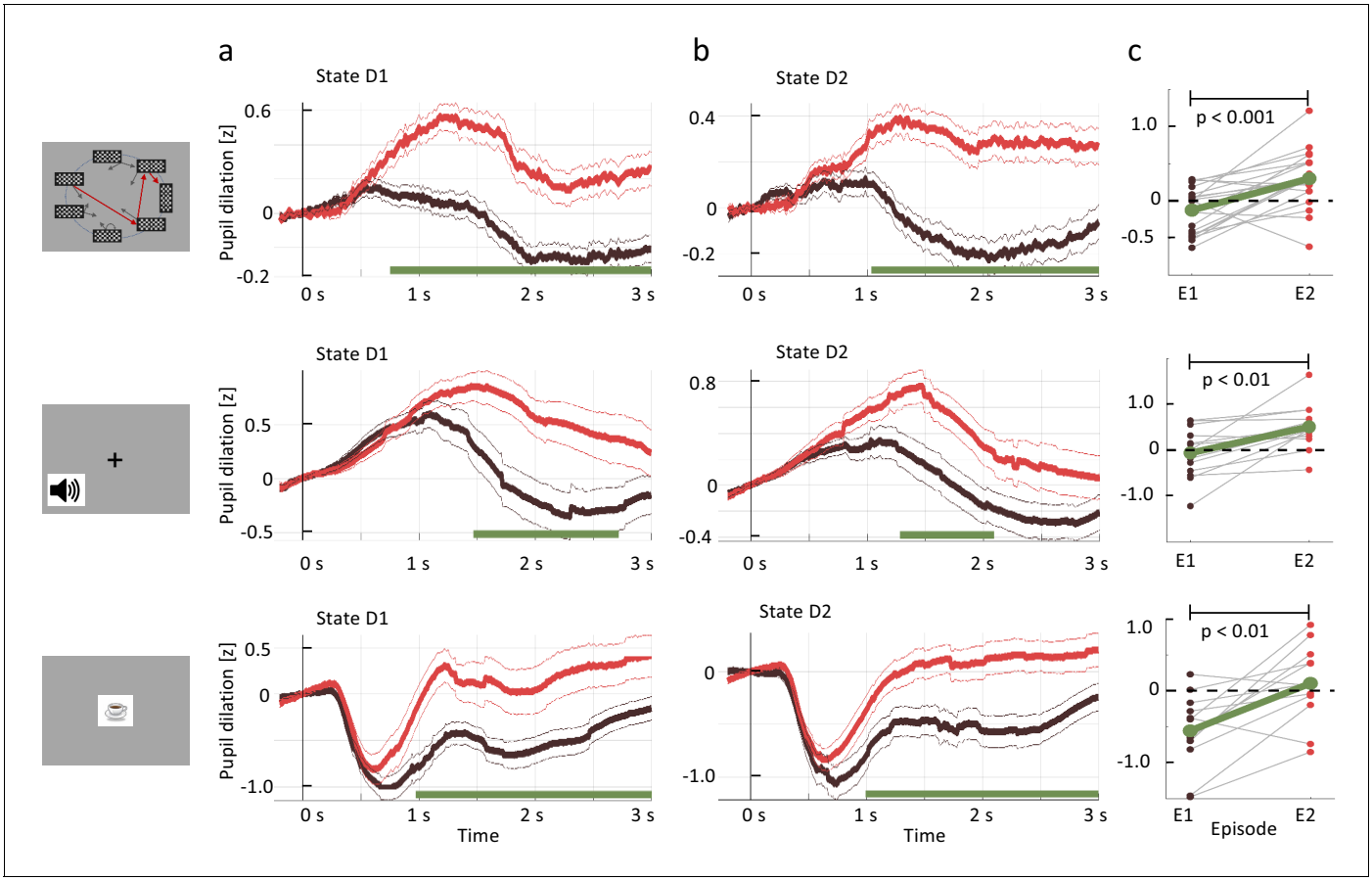


Figure 4. Pupil dilation reflects one-shot learning. (**a**) Pupil responses to state D1 are larger during episode 2 (red curve) than during episode 1 (black). (**b**) Pupil responses to state D2 are larger during episode 2 (red curve) than during episode 1 (black). Top row: *spatial*, middle row: *sound*, bottom row: *clip-art* condition. Pupil diameter averaged across all participants in units of standard deviation (z-score, see Materials and methods), aligned at stimulus onset and plotted as a function of time since stimulus onset. Thin lines indicate the pupil signal ± SEM. Green lines indicate the time interval during which the two curves differ significantly ($p<FDR_{\alpha}=0.05$). Significance was reached at a time t_{min}, which depends on the condition and the state: *spatial* D1:$t_{min}=730\,\mathrm{ms}$ (22, 131, 85); *spatial* D2: $t_{min}=1030\,\mathrm{ms}$ (22, 137,130) *sound* D1: $t_{min}=1470\,\mathrm{ms}$ (15, 34, 19); *sound* D2: $t_{min}=1280\,\mathrm{ms}$ (15, 35, 33); *clip-art* D1: $t_{min}=970\,\mathrm{ms}$ (12, 39, 19); *clip-art* D2: $t_{min}=980\,\mathrm{ms}$ (12, 45, 41); (Numbers in brackets: number of participants, number of pupil traces in episode 1 or 2, respectively). (**c**) Participant-specific mean pupil dilation at state D2 (averaged over the interval (1000 ms, 2500 ms)) before (black dot) and after (red dot) the first reward. Grey lines connect values of the same participant. Differences between episodes are significant (paired t-test, p-values indicated in the Figure).

Figure 4) was a response to the 300 ms display of the images. In the *spatial* condition, this initial transient was absent, but the difference in state D2 between episode 1 and episode 2 were equally significant. For the *sound* condition, in which stimuli were longer on average (Materials and methods), the significant separation of the curves occurred slightly later than in the other two conditions. A paired t-test of differences showed that, across all three conditions, pupil dilation changes significantly between episodes 1 and 2 (*Figure 4(c)*; paired t-test, $p<0.001$ for the *spatial* condition, $p<0.01$ for the two others). Since in all three conditions luminance is identical in episodes 1 and 2, luminance cannot explain the observed differences.

Second, we checked whether the differences in the pupil traces could be explained by the novelty of a state during episode 1, or familiarity with the state in episode 2 (*Otero et al., 2011*), rather than by reward-based learning. In a further control experiment, a different set of participants saw a sequence of states, replayed from the main experiment. In order to ensure that participants were focusing on the state sequence and engaged in the task, they had to push a button in each state (freely choosing either 'a' or 'b'), and count the number of states from start to goal. Stimuli, timing and data analysis were the same as in the main experiment. The strong difference after $1000\,ms$ in state D2, that we observed in *Figure 4 (b)*, was absent in the control experiment (*Figure 5*) indicating that the significant differences in pupil dilation in response to state D2 cannot be explained by novelty or familiarity alone. The findings in the control experiment also exclude other interpretations of correlations of pupil diameter such as memory formation in the absence of reward.

In summary, across three different stimulus modalities, the single reward received at the end of the first episode strongly influenced the pupil responses to the same stimuli later in episode 2. Importantly, this effect was observed not only in state D1 (one step before the goal) but also in state D2 (two steps before the goal). Furthermore, a mere engagement in button presses while observing a sequence of stimuli, as in the control experiment, did not evoke the same pupil responses as the main task. Together these results suggested that the single reward at the end of the first episode triggered increases in pupil diameter during later encounters of the same state. The increases observed in state D1 are consistent with an interpretation that pupil diameter reflects state value V, action value Q, or TD error - but do not inform us whether Q-value, V-value, or TD-error are estimated by the brain using RL with or without eligibility trace. However, the fact that very similar changes are also observed in state D2 excludes the possibility that the learning-related contribution to the pupil diameter can be predicted by *RL without eligibility trace*.

While our experiment was not designed to identify whether the pupil response reflects TD-errors or state values, we tried to address this question based on a model-driven analysis of the pupil traces. First, we extracted all pupil responses after the onset of non-goal states and calculated the TD-error (according to the best-fitting model, *Q-λ*, see next section) of the corresponding state transition. We found that the pupil dilation was much larger after transitions with high TD-error compared to transitions with zero TD-error (*Figure 6 (a)* and Materials and methods). Importantly, these temporal profiles of the pupil responses to states with high TD-error had striking similarities across the three experimental conditions, whereas the mean response time course was different across the three conditions (*Figure 6 (c)*. This suggests that the underlying physiological process causing the TD-error-driven component in the pupil responses was invariant to stimulation details. Second, a statistical analysis including data with low, medium, and high TD-error confirmed the correlation of pupil dilation with TD error (see subsection regression analysis in methods). Third, a further qualitative analysis revealed that TD-error, rather than value itself, was a factor modulating pupil dilation (*Figure 6 (b)*.

Estimation of the time scale of the behavioral eligibility trace using reinforcement learning models

Given the behavioral and physiological evidence for *RL with eligibility trace*, we wondered whether our findings are consistent with earlier studies (*Bogacz et al., 2007*; *Daw et al., 2011*; *Tartaglia et al., 2017*) where several variants of reinforcement learning algorithms were fitted to the experimental data. We considered algorithms with and (for comparison) without eligibility trace. Eligibility traces $e_n(s,a)$ can be modeled as a memory of past state-action pairs (s,a) in an episode. At the beginning of each episode all twelve eligibility trace values (two actions for each of the six decision states) were set to $e_n(s,a)=0$. At each discrete time step n, the eligibility of the current state-

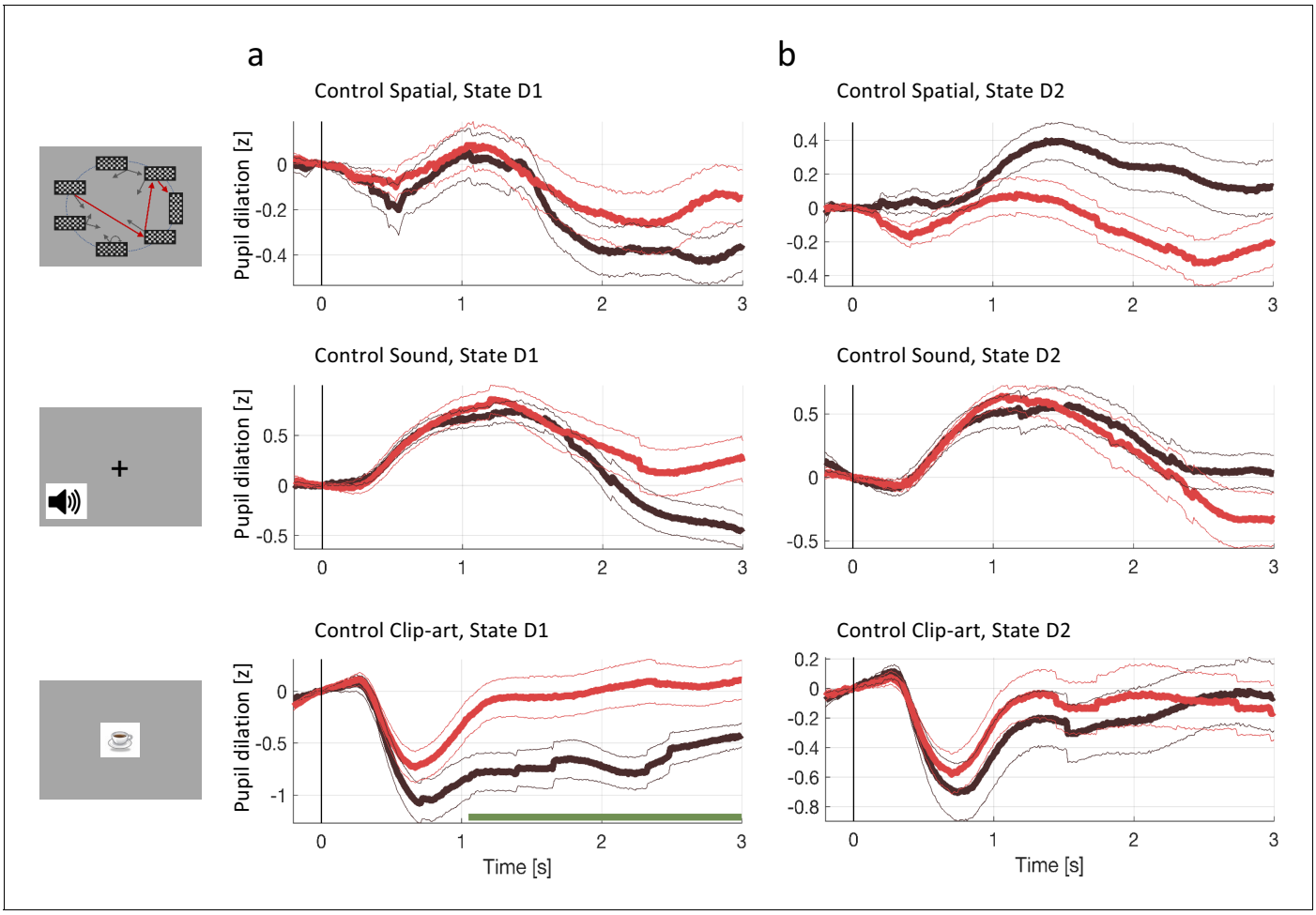


Figure 5. Pupil dilation during the second control experiment. In the second control experiment, different participants passively observed state sequences which were recorded during the main experiment. Data analysis was the same as for the main experiment. (**a**) Pupil time course after state onset ($t = 0$) of state D1 (before goal). (**b**) State D2 (two before goal). Black traces show the pupil dilation during episode one, red traces during episode two. At state D1 in the *clip-art* condition, the pupil time course shows a separation similar to the one observed in the main experiment. This suggest that participants may recognize the clip-art image that appears just before the final image. Importantly in state D2, the pupil time course during episode 2 is qualitatively different from the one in the main experiment (***Figure 4***).

action pair was set to 1, while that of all others decayed by a factor $\gamma\lambda$ according to ***Singh and Sutton (1996)***

$$e_n(s,a) = \begin{cases} 1 & \text{if } s = s_n, a = a_n \\ \gamma\lambda e_{n-1}(s,a) & \text{otherwise.} \end{cases} \quad (1)$$

The parameter $\gamma \in (0,1)$ exponentially discounts a distal reward, as commonly described in neuroeconomics (***Glimcher and Fehr, 2013***) and machine learning (***Sutton and Barto, 2018***); the parameter $\lambda \in [0,1]$ is called the decay factor of the eligibility trace. The limit case $\lambda = 0$ is interpreted as no memory and represents an instance of *RL without eligibility trace*. Even though the two parameters γ and λ appear as a product in ***Equation 1*** so that the decay of the eligibility trace depends on both, they have different effects in spreading the reward information from one state to the next (cf. ***Equation 3*** in Materials and methods). After many trials, the V-values of states, or Q-values of actions, approach final values which only depend on γ, but not on λ. Given a parameter γ>0, the choice of λ determines how far value information spreads in a single trial. Note that for $\lambda = 0$ (*RL without eligibility trace*); ***Equation 1*** assigns an eligibility $e_n = 1$ to state D1 in the first episode at the moment of the transition to the goal (while the eligibility at state D2 is 0). These values of eligibility

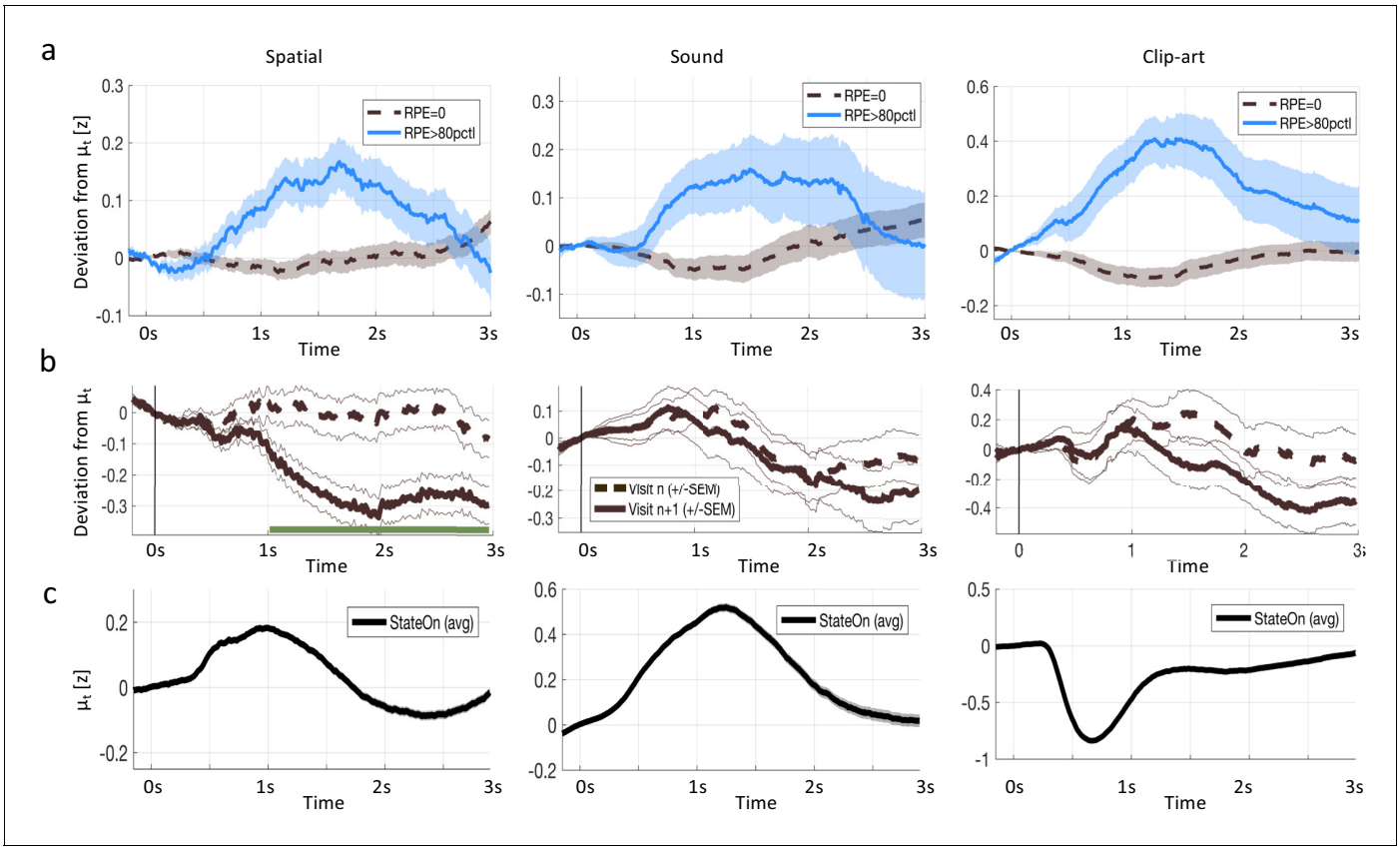


Figure 6. Reward prediction error (RPE) at non-goal states modulates pupil dilation. Pupil traces (in units of standard deviation) from all states except G were aligned at state onset ($t = 0ms$) and the mean pupil response μ_t was subtracted (see Materials and methods). (**a**) The deviation from the mean is shown for states where the model predicts $RPE = 0$ (black, dashed) and for states where the model predicts $RPE \geq 80^{th}$ percentile (solid, blue). Shaded areas: ± SEM. Thus the pupil dilation reflects the RPE predicted by a reinforcement learning model that spreads value information to nonrewarded states via eligibility traces. (**b**) To qualitatively distinguish pupil correlations with RPE from correlations with state values $V(s)$, we started from the following observation: the model predicts that RPE decreases over the course of learning (due to convergence), while the state values $V(s)$ increase (due to spread of value information). We wanted to observe this qualitative difference in the pupil dilations of subsequent visits of the *same* state. We selected pairs of visits n and $n+1$ for which the RPE decreased while $V(s)$ increased and extracted the pupil measurements of the two visits (again, mean μ_t is subtracted). The dashed, black curves show the average pupil trace during the n^{th} visit of a state. The solid black curves correspond to the next visit $(n+1)$ of the same state. In the spatial condition, the two curves significantly ($p<FDR_\alpha = 0.05$) separate at $t>1s$ (indicated by the green line). All three conditions show the same trend (with strong significance in the spatial condition), compatible with a positive correlation of pupil response with RPE, but not with state value $V(s)$. (**c**) The mean pupil dilation μ_t is different in each condition, whereas the learning related deviations from the mean (in (**a**) and (**b**)) have similar shapes.

traces lead to a spread of reward information from the goal to state D1, but not to D2, at the end of the first episode in models without eligibilty trace (cf. ***Equation 3*** and subsection *Q-λ model predictions* in methods), hence the qualitative argument for episodes 1 and 2 as sketched in ***Figure 1***.

We considered eight common algorithms to explain the behavioral data: Four algorithms belonged to the class of *RL with eligibility traces*. The first two, *SARSA-λ* and *Q-λ* (see Materials and methods, ***Equation 3***) implement a memory of past state-action pairs by an eligibility trace as defined in ***Equation 1***; as a member of the Policy-Gradient family, we implemented a variant of *Reinforce* (***Williams, 1992***; ***Sutton and Barto, 2018***), which memorizes all state-action pairs of an episode. A fourth algorithm with eligibility trace is the 3-step Q-learning algorithm (***Watkins, 1989***; ***Mnih et al., 2016***; ***Sutton and Barto, 2018***), which keeps memory of past states and actions over three steps (see Discussion and Materials and methods). From the model-based family of RL, we chose the *Forward Learner* (***Gläscher et al., 2010***), which memorizes not state-action pairs, but learns a state-action-next-state model, and uses it for offline updates of action-values. The *Hybrid Learner* (***Gläscher et al., 2010***) combines the *Forward Learner* with *SARSA-0*. As a control, we chose

two algorithms belonging to the class of *RL without eligibility traces* (thus modeling the null hypothesis): *SARSA-0* and *Q-0*.

We found that the four RL algorithms with eligibility trace explained human behavior better than the *Hybrid Learner*, which was the top-scoring among all other RL algorithms. Cross-validation confirmed that our ranking based on the Akaike Information Criterion (AIC, ***Akaike, 1974***; see Materials and methods) was robust. According to the Wilcoxon rank-sum test, the probability that the *Hybrid Learner* ranks better than one of the three RL algorithms with explicit eligibility traces was below 14% in each of the conditions and below 0.1% for the aggregated data ($p<0.001$, ***Table 1*** and Materials and methods). The models *Q-λ* and *SARSA-λ* with eligbility trace performed each significantly better than the corresponding models *Q-0* and *SARSA-0* without eligbility trace.

Since the ranks of the four RL algorithms with eligibility traces were not significantly different, we focused on one of these, viz. *Q-λ*. We wondered whether the parameter λ that characterizes the decay of the eligibility trace in ***Equation 1*** could be linked to a time scale. To answer this question, we proceeded in two steps. First, we analyzed the human behavior in discrete time steps corresponding to state transitions. We found that the best fitting values (maximum likelihood, see Materials and methods) of the eligibility trace parameter λ were 0.81 in the *clip-art*, 0.96 in the *sound*, and 0.69 in the *spatial* condition (see ***Figure 7***). These values are all significantly larger than zero (p<0.001) indicating the presence of an eligibility trace consistent with our findings in the previous subsections.

In a second step, we modeled the same action sequence in continuous time, taking into account the measured inter-stimulus interval (ISI) which was the sum of the reaction time plus a random delay of 2.5 to 4 seconds after the push-buttons was pressed. The reaction times were similar in the *spatial-* and *clip-art* condition, and slightly longer in the *sound* condition with the following 10%, 50%

Table 1. Models with eligibility trace explain behavior significantly better than alternative models.
Four reinforcement learning models with eligibility trace (*Q-λ*, REINFORCE, *SARSA-λ*, *3-step-Q*); two model-based algorithms (*Hybrid*, *Forward Learner*), two RL models without eligibility trace (*Q-0*, *SARSA-0*), and a null-model (*Biased Random*, Materials and methods) were fitted to the human behavior, separately for each experimental condition (*spatial, sound, clip-art*). Models with eligibility trace ranked higher than those without (lower Akaike Information Criterion, AIC, evaluated on all participants performing the condition). $wAIC$ indicates the *normalized Akaike weights* (***Burnham and Anderson, 2004***), values < 0.01 are not added to the table. Note that only models with eligibility trace have $wAIC>0.01$. The ranking is stable as indicated by the sum of k rankings (column *rank sum*) on test data, in k-fold crossvalidation (Materials and methods). P-values refer to the following comparisons: P(a): Each model in the *with eligibility trace* group was compared with the best model *without eligibility trace* (*Hybrid* in all conditions); models for which the comparison is significant are shown in bold. P(b): *Q-0* compared with *Q-λ*. P(c): *SARSA-0* compared with *SARSA-λ*. P(d): *Biased Random* compared with the second last model, which is *Forward Learner* in the *clip-art* condition and *SARSA-0* in the two others. In the *Aggregated* column, we compared the same pairs of models, taking into account all ranks across the three conditions. All algorithms with eligibility trace explain the human behavior better (p(e)<.001) than algorithms without eligibility trace. Differences among the four models with eligibility trace are not significant. In each comparison, k pairs of individual ranks are used to compare pairs of models and obtain the indicated p-values (Wilcoxon rank-sum test, Materials and methods).

	Condition	Spatial		Sound		Clip-art		Aggregated
Model		AIC	Rank Sum (k = 11)	AIC	Rank Sum (k = 7)	AIC	Rank Sum (k = 7)	all ranks
With elig tr.	Q-λ	**$6470.2^{p(a)=.003}_{wAIC=1.00}$**	**24**	**$1489.1^{p(a)=.015}_{wAIC=0.23}$**	**20**	$1234.8^{p(a)=.062}_{wAIC=0.27}$	20	**$64^{p(e)<.001}$**
	Reinforce	**$6508.7^{p(a)=.016}$**	**35**	**$1486.8^{p(a)=.015}_{wAIC=0.74}$**	**10**	$1239.2^{p(a)=.109}_{wAIC=0.03}$	22	**$67^{p(e)<.001}$**
	3-step-Q	**$6488.8^{p(a)=.013}$**	**33**	**$1494.3^{p(a)=.046}_{wAIC=0.02}$**	**26**	**$1236.6^{p(a)=.015}_{wAIC=0.11}$**	**16**	**$71^{p(e)<.001}$**
	SARSA-λ	**$6502.4^{p(a)=.003}$**	**36**	$1495.2^{p(a)=.040}_{wAIC=0.01}$	30	**$1233.2^{p(a)=.015}_{wAIC=0.59}$**	**16**	**$82^{p(e)<.001}$**
Model based	Hybrid	6536.6	61	1498.3	43	1271.3	33	$137^{p(e)<.001}$
	Forward Learner	6637.5	79	1500.6	41	1316.3	48	168
Without elig tr.	Q-0	$6604.0^{p(b)=.003}$	60	$1518.6^{p(b)=.046}$	39	$1292.0^{p(b)=.015}$	51	$150^{p(b)<.001}$
	SARSA-0	$6643.3^{p(c)=.001}$	68	$1520.2^{p(c)=.093}$	43	$1289.5^{p(c)=.015}$	46	$157^{p(c)<.001}$
	Biased Random	$7868.3^{p(d)=.001}$	99	$1866.1^{p(d)=.015}$	63	$1761.1^{p(d)=.015}$	63	$225^{p(d)<.001}$

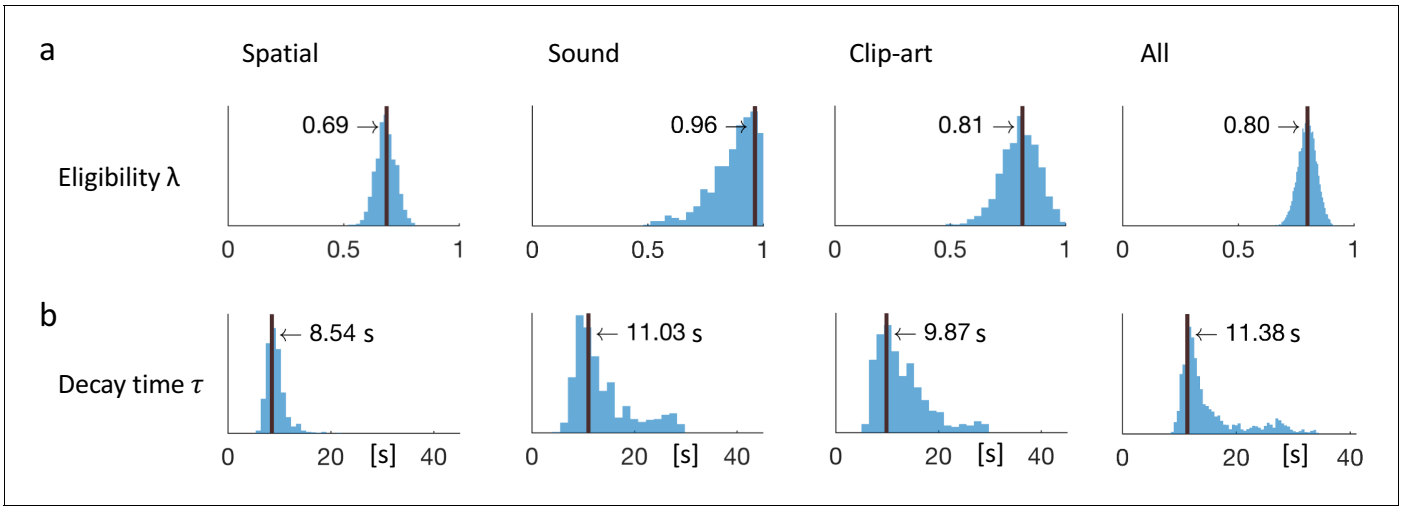


Figure 7. Eligibility for reinforcement decays with a time-scale τ in the order of 10 s. The behavioral data of each experimental condition constrain the free parameters of the model Q-λ to the ranges indicated by the blue histograms (see methods) (**a**) Distribution over the eligibility trace parameter λ in ***Equation 1*** (discrete time steps). Vertical black lines indicate the values that best explain the data (maximum likelihood, see Materials and methods). All values are significantly different from zero. (**b**) Modeling eligibility in continuous time with a time-dependent decay (Materials and methods, ***Equation 5***), instead of a discrete per-step decay. The behavioral data constrains the time-scale parameter τ to around 10 s. Values in the column *All* are obtained by fitting λ and τ to the aggregated data of all conditions.

and 90% percentiles: *spatial*: [0.40, 1.19, 2.73], *clip-art*: [0.50, 1.11, 2.57], *sound*: [0.67, 1.45, 3.78] seconds. In this continuous-time version of the eligibility trace model, both the discount factor γ and the decay factor λ were integrated into a single time constant τ that describes the decay of the memory of past state-action associations in continuous time. We found maximum likelihood values for τ around 10s (***Figure 7***), corresponding to 2 to 3 inter-stimulus intervals. This implies that an action taken 10s before a reward was reinforced and associated with the state in which it was taken – even if one or several decisions happened in between (see Discussion).

Thus eligibility traces, that is memories of past state-action pairs, decay over about 10s and can be linked to a reward occurring during that time span.

Discussion

Eligibility traces provide a mechanism for learning temporally extended action sequences from a single reward (one-shot). While one-shot learning is a well-known phenomenon for tasks such as image recognition (***Standing, 1973***; ***Brady et al., 2008***) and one-step decision making (***Duncan and Shohamy, 2016***; ***Greve et al., 2017***; ***Rouhani et al., 2018***) it has so far not been linked to Reinforcement Learning (RL) with eligibility traces in multi-step decision making.

In this study, we asked whether humans use eligibility traces when learning long sequences from delayed feedback. We formulated mutually exclusive hypotheses, which predict directly observable changes in behavior and in physiological measures when learning with or without eligibility traces. Using a novel paradigm, we could reject the null hypothesis of learning without eligibility trace in favor of the alternative hypothesis of learning with eligibility trace.

Our multi-step decision task shares aspects with earlier work in the neurosciences (***Pessiglione et al., 2006***; ***Gläscher et al., 2010***; ***Daw et al., 2011***; ***Walsh and Anderson, 2011***; ***Niv et al., 2012***; ***O'Doherty et al., 2017***), but overcomes their limitations (i) by using a recurrent graph structure of the environment that enables relatively long episodes (***Tartaglia et al., 2017***), and (ii) by implementing an 'on-the-fly' assignment rule for state-action transitions during the first episodes. This novel design allows the study of human learning in specific and controlled conditions, without interfering with the participant's free choices.

A difficulty in the study of eligibility traces, is that in the relatively simple tasks typically used in animal (***Pan et al., 2005***) or human (***Bogacz et al., 2007***; ***Gureckis and Love, 2009***; ***Daw et al.,***

2011; Walsh and Anderson, 2011; Weinberg et al., 2012; Tartaglia et al., 2017) studies, the two hypotheses make qualitatively different predictions only during the first episodes: At the end of the first episode, algorithms in the class of *RL without eligibility trace* update only the value of state D1 (but not of D2, see Figure 1, Null hypothesis). Then, this value of D1 will drive learning at state D2 when the participants move from D2 to D1 during episode 2. In contrast, algorithms in the class of *RL with eligibility trace*, update D2 already during episode one. Therefore, only during episode 2, the behavioral data permits a clean, qualitative dissociation between the two classes. On the other hand, the fact that for most episodes, the differences are not qualitative, is the reason why eligibility trace contributions have typically been statistically inferred from many trials through model selection (Pan et al., 2005; Bogacz et al., 2007; Gureckis and Love, 2009; Daw et al., 2011; Walsh and Anderson, 2011; Tartaglia et al., 2017). Here, by a specific task design and a focus on episodes 1 and 2, we provided directly observable, qualitative, evidence for learning with eligibility traces from behavior and pupil data without the need of model selection.

In the quantitative analysis, RL models with eligibility trace explained the behavioral data significantly better than the best tested RL models without. There are, however, in the reinforcement learning literature, several alternative algorithms that would also account for one-shot learning but do not rely on the explicit eligibility traces formulated in Equation 1. First, n-step reinforcement learning algorithms (Watkins, 1989; Mnih et al., 2016; Sutton and Barto, 2018) compare the value of a state not with that of its direct neighbor but of neighbors that are n steps away. These algorithms are closely related to eligibility traces and in certain cases even mathematically equivalent (Sutton and Barto, 2018). Second, reinforcement learning algorithm with storage of past sequences (Moore and Atkeson, 1993; Blundell et al., 2016; Mnih et al., 2016) enable the offline replay of the first episode so as to update values of states far away from the goal. While these approaches are formally different from eligibility traces, they nevertheless implement the idea of eligibility traces as memory of past state-action pairs (Crow, 1968; Frémaux and Gerstner, 2015), albeit in a different algorithmic framework. For example, prioritized sweeping with small backups (Seijen and Sutton, 2013) is an offline algorithm that is, if applied to our deterministic environment after the end of the first episode, equivalent to both episodic control (Brea, 2017) and an eligibility trace. Interestingly, the two model-based algorithms (*Forward Learner* and *Hybrid*) would in principle be able to explain one-shot learning since reward information is spread, after the first episode, throughout the model, via offline Q-value updates. Nevertheless, when behavioral data from our experiments were fitted across all seven episodes, the two model-based algorithms performed significantly worse than the RL models with explicit eligibility traces. Since our experimental design does not allow us to distinguish between these different algorithmic implementations of closely related ideas, we put them all in the class of RL with eligibility traces.

Importantly, RL algorithms with explicit eligibility traces (Sutton, 1988; Williams, 1992; Peng and Williams, 1996; Izhikevich, 2007; Frémaux and Gerstner, 2015) can be mapped to known synaptic and circuit mechanisms (Yagishita et al., 2014; He et al., 2015; Bittner et al., 2017; Fisher et al., 2017; Gerstner et al., 2018). A time scale of the eligibility trace of about 10s in our experiments is in the range of, but a bit longer than those observed for dopamine modulated plasticity in the striatum (Yagishita et al., 2014), serotonin and norepinephrine modulated plasticity in the cortex (He et al., 2015), or complex-spike plasticity in hippocampus (Bittner et al., 2017), but shorter than the time scales of minutes reported in hippocampus (Brzosko et al., 2017). The basic idea for the relation of eligibility traces as in Equation 1 to experiments on synaptic plasticity is that choosing action a in state s leads to co-activation of neurons and leaves a trace at the synapses connecting those neurons. A later phasic neuromodulator signal will transform the trace into a change of the synapses so that taking action a in state s becomes more likely in the future (Crow, 1968; Izhikevich, 2007; Sutton and Barto, 2018; Gerstner et al., 2018). Neuromodulator signals could include dopamine (Schultz, 2015), but reward-related signals could also be conveyed, together with novelty or attention-related signals, by other modulators (Frémaux and Gerstner, 2015).

Since in our paradigm the inter-stimulus interval (ISI) was not systematically varied, we cannot distinguish between an eligibility trace with purely time-dependent, exponential decay, and one that decays discretely, triggered by events such as states or actions. Future research needs to show whether the decay is event-triggered or defined by molecular characteristics, independent of the experimental paradigm.

Our finding that changes of pupil dilation correlate with reward-driven variables of reinforcement learning (such as value or TD error) goes beyond the changes linked to state recognition reported earlier (**Otero et al., 2011**; **Kucewicz et al., 2018**). Also, since non-luminance related pupil diameter is influenced by the neuromodulator norepinephrine (**Joshi et al., 2016**) while reward-based learning is associated with the neuromodulator dopamine (**Schultz, 2015**), our findings suggest that the roles, and regions of influence, of neuromodulators could be mixed (**Frémaux and Gerstner, 2015**; **Berke, 2018**) and less well segregated than suggested by earlier theories.

From the qualitative analysis of the pupillometric data of the main experiment (**Figure 5**), together with those of the control experiment (**Figure 5**), we concluded that changes in pupil dilation reflected a learned, reward-related property of the state. In the context of decision making and learning, pupil dilation is most frequently associated with violation of an expectation in the form of a reward prediction error or stimulus prediction error as in an oddball-task (**Nieuwenhuis et al., 2011**). However, our experimental paradigm was not designed to decide whether pupil diameter correlates stronger with state values or TD-errors. Nevertheless, a more systematic analysis (see Materials and methods and **Figure 6**) suggests that correlation of pupil dilation with TD-errors is stronger than correlation with state values.

Conclusion

Eligibility traces are a fundamental factor underlying the human capability of quick learning and adaptation. They implement a memory of past state-action associations and are a crucial element to efficiently solve the credit assignment problem in complex tasks (**Izhikevich, 2007**; **Sutton and Barto, 2018**; **Gerstner et al., 2018**). The present study provides both qualitative and quantitative evidence for one-shot sequence-learning with eligibility traces. The correlation of the pupillometric signals with an RL algorithm with eligibility traces suggests that humans not only exploit memories of past state-action pairs in behavior but also assign reward-related values to these memories. The consistency and similarity of our findings across three experimental conditions suggests that the underlying cognitive, or neuromodulatory, processes are independent of the stimulus modality. It is an interesting question for future research to actually identify the neural implementation of these memory traces.

Materials and methods

Experimental conditions

We implemented three different experimental conditions based on the same Markov Decision Process (MDP) of **Figure 2(a)**. The conditions only differed in the way the states were presented to the participants. Furthermore, in order to collect enough samples from early trials, where the learning effects are strongest, participants did not perform one long experiment. Instead, after completing seven episodes in the same environment, the experiment paused for 45 s while participants were instructed to close and relax their eyes. Then the experiment restarted with a new environment: the transition graph was reset, a different, unused, stimulus was assigned to each state, and the participant had to explore and learn the new environment. We instructed the participants to reach the goal state as often as possible within a limited time (12 min in the *sound* and *clip-art* condition, 20 min in the *spatial* condition). On average, they completed 48.1 episodes (6.9 environments) in the *spatial* condition , 19.4 episodes (2.7 environments) in the *sound* condition and 25.1 episodes (3.6 environments) in the *clip-art* condition.

In the *spatial* condition, each state was defined by the location (on an invisible circle) on the screen of a 100 $\times$ 260 pixels checkerboard image, flashed for 100 ms, **Figure 2(d)**. The goal state was represented by the same rectangular checkerboard, but rotated by 90˚. The checkerboard had the same average luminance as the grey background screen. In each new environment, the states were randomly assigned to locations and the checkerboards were rotated (states: 260 $\times$ 100 pixels checkerboard, goal: 100 $\times$ 260).

In the *sound* condition, each state was represented by a unique acoustic stimulus (tones and natural sounds) of 300 ms to 600 ms duration. New, randomly chosen, stimuli were used in each environment. At the goal state an applause was played. An experimental advantage of the *sound* condition

is that a change in the pupil dilation cannot stem from a luminance change but must be due to a task-specific condition.

In the *clip-art* condition, each state was represented by a unique 100 × 100 pixel clip-art image that appeared for 300 ms in the center of the screen. For each environment, a new set of images was used, except for the goal state which was always the same (a person holding a trophy) in all experiments.

The screen resolution was 1920 × 1080 pixels. In all three conditions, the background screen was grey with a fixation cross in the center of the screen. It was rotated from + to × to signal to the participants when to enter their decision by pressing one of two push-buttons (one in the left and the other in the right hand). No lower or upper bound was imposed on the reaction time. The next state appeared after a random delay of 2.5s to 4s after the push-buttons was pressed. Prior to the actual learning task, they performed a few trials to check they all understood the instructions. While the participants performed the *sound*- and *clip-art* conditions, we recorded the pupil diameter using an SMI iViewX high speed video-based eye tracker (recorded at 500 Hz, down-sampled to 100 Hz for the analysis by averaging over five samples). From participants performing the *spatial* condition, we recorded the pupil diameter using a 60 Hz Tobii Pro tracker. An eye tracker calibration protocol was run for each participant. All experiments were implemented using the Psychophysics Toolbox (***Brainard, 1997***).

The number of participants performing the task was: *sound*: 15; *clip-art*: 12; *spatial*: 22 participants; Control *sound*: 9; Control *clip-art*: 10; Control *spatial*: 12. The participants were recruited from the EPFL students pool. They had normal or corrected-to-normal vision. Experiments were conducted in accordance with the Helsinki declaration and approved by the ethics commission of the Canton de Vaud (164/14 Titre: Aspects fondamentaux de la reconnaissance des objets : protocole général). All participants were informed about the general purpose of the experiment and provided written, informed consent. They were told that they could quit the experiment at any time they wish.

Pupil data processing

Our data processing pipeline followed recommendations described in ***Mathôt et al. (2017)***. Eye blinks (including 100 ms before, and 150 ms after) were removed and short blocks without data (up to 500 ms) were linearly interpolated. In all experiments, participants were looking at a fixation cross which reduces artifactual pupil-size changes (***Mathôt et al., 2017***). For each environment, the time-series of the pupil diameter during the seven episodes was extracted and then normalized to zero-mean, unit variance. This step renders the measurements comparable across participants and environments. We then extracted the pupil recordings at each state from 200 ms before to 3000 ms after each state onset and applied subtractive baseline correction where the baseline was taken as the mean in the interval $(-100ms, +100ms]$. Taking the $+100ms$ into account does not interfere with event-specific effects because they develop only later (>220 ms according to ***Mathôt et al., 2017***); but a symmetric baseline reduces small biases when different traces have different slopes around t = 0 ms. We considered event-locked pupil responses with z-values outside ±3 as outliers and excluded them from the main analysis. We also excluded pupil traces with less than 50% eye-tracker data within the time window of interest, because very short data fragments do not provide information about the characteristic time course of the pupil trace after stimulus onset. As a control, ***Figure 8*** shows that the conclusions of our study are not affected if we drop the two conditions and include all data.

Action assignment in the Markov Decision Process

Actions in the graph of ***Figure 2*** were assigned to transitions during the first few actions as explained in the main text. However, our learning experiment would become corrupted if participants would discover that in the first episode any three actions lead to the goal. First, such knowledge would bypass the need to actually learn state-action associations, and second, the knowledge of 'distance-to-goal' implicitly provides reward information even before seeing the goal state. We avoided the learning of the latent structure by two manipulations: First, if in episode 1 of a new environment a participant repeated the exact same action sequence as in the previous environment, or if they tried trivial action sequences (a-a-a or b-b-b); the assignment of the third action led from state D1 to Z, rather than to the Goal. This was the case in about 1/3 of the first episodes (*spatial*: 48/173,

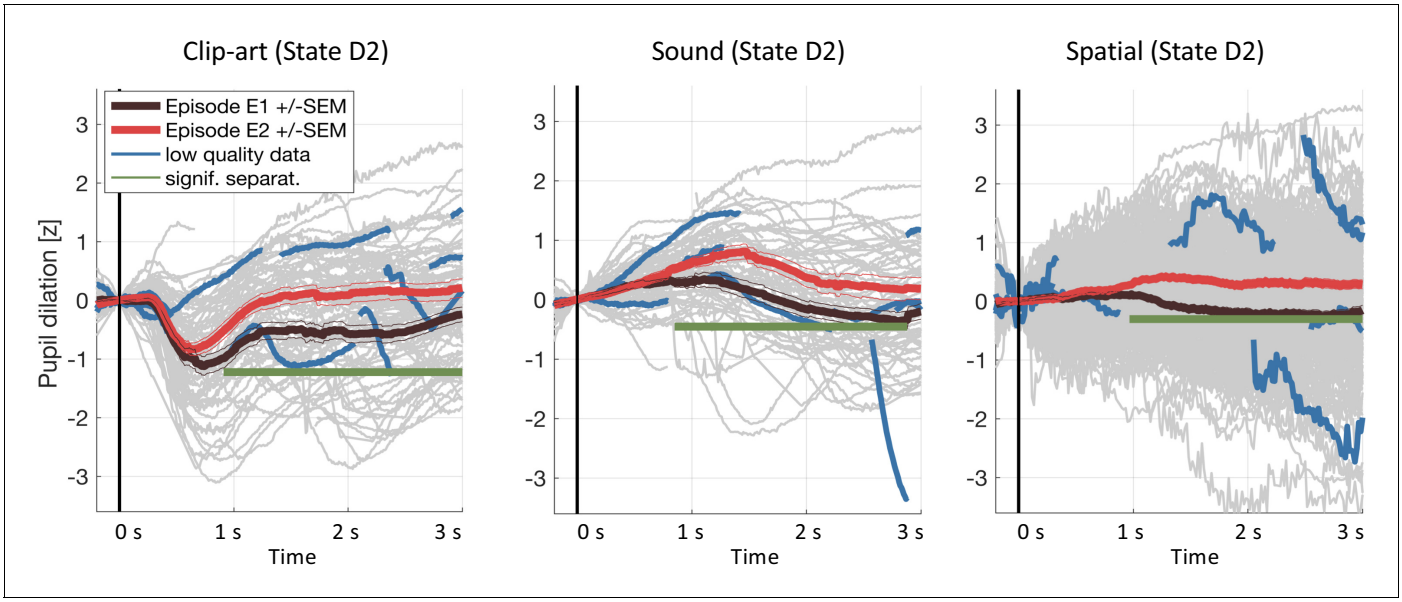


Figure 8. Results including low-quality pupil traces. We repeated the pupil data analysis at the crucial state D2including all data (including traces with less than 50% of data within the 3s window and with z-values outside ±3). Gray curves in the background show all recorded pupil traces. The highlighted blue curves show a few, randomly selected, low-quality pupil traces. Including these traces does not affect the result.

sound: 20/53 *clip-art*: 23/49). The manipulation further implied that participants had to make decisions against their potential left/right bias. Second, an additional state H (not shown in ***Figure 2***) was added in episode 1 in some environments (*spatial*: 23/173, *sound*: 6/53 *clip-art*: 8/49). Participants then started from H (always leading to S) and the path length to goal was four steps. Interviews after the experiment showed that no participant became aware of the experimental manipulation and, importantly, they did not notice that they could reach the goal with a random action sequence in episode 1.

Reinforcement Learning models

For the RL algorithm $Q-\lambda$ (see Algorithm 1); four quantities are important: the reward r; the value $Q(s,a)$ of a state-action association such as taking action 'b' in state D2; the value $V(s)$ of the state itself, defined as the larger of the two Q-values in that state, that is $V(s)=\max_{\tilde{a}}Q(s,\tilde{a})$; and the TD-error (also called Reward Prediction Error or RPE) calculated at the end of the n^{th} action after the transition from state s_n to s_{n+1}

$$\mathrm{RPE}(n\rightarrow n+1)=r_{n+1}+\gamma\cdot V(s_{n+1})-Q(s_n,a_n) \tag{2}$$

Here, γ is the discount factor and $V(s)$ is the estimate of the discounted future reward that can maximally be collected when starting from state s. Note that RPE is different from reward. In our environment a reward occurs only at the transition from state D1 to state G whereas reward prediction errors occur in episodes 2–7 also several steps before the reward location is reached.

The table of values $Q(s,a)$ is initialized at the beginning of an experiment and then updated by combining the RPE and the eligibility traces $e_n(s,a)$ defined in the main text (***Equation 1***);

$$Q(s,a)\leftarrow Q(s,a)+\alpha\cdot RPE(n)\cdot e_n(s,a)\,, \tag{3}$$

where α is the learning rate. Note that *all* Q-values are updated, but changes in $Q(s_n,a_n)$ are proportional to the eligibility of the state-action pair $e_n(s,a)$. In the literature the table $Q(s,a)$ is often initialized with zero, but since some participants pressed the left (or right) button more often than the other one, we identified for each participant the preferred action a_{pref} and initialized $Q(s,a_{pref})$ with a small bias b, adapted to the data.

Action selection exploits the Q-values of **Equation 3** using a softmax criterion with temperature T:

$$p(s,a) = \frac{exp(Q(s,a)/T)}{\sum_{\tilde{a}} exp(Q(s,\tilde{a})/T)} \tag{4}$$

As an alternative to the eligibility trace defined in **Equation 1**, where the eligibility decays at each discrete time-step, we also modeled a decay in continuous time, defined as

$$e_t(s,a) = exp\left(-\frac{t - B(s,a)}{\tau}\right) \quad \text{if } t > B(s,a) \tag{5}$$

and zero otherwise. Here, t is the time stamp of the current discrete step, and $B(s,a)$ is the time stamp of the last time a state-action pair (s,a) has been selected. The discount factor γ in **Equation 2** is kept, while in **Equation 5** a potential discounting is absorbed into the single parameter τ.

Our implementation of *Reinforce* followed the pseudo-code of *REINFORCE: Monte-Carlo Policy-Gradient Control* (*without baseline*) (**Sutton and Barto, 2018**), Chapter 13.3) which updates the action-selection probabilities at the end of each episode. This requires the algorithm to keep a (non-decaying) memory of the complete state-action history of each episode. We refer to **Peng and Williams (1996)**, **Gläscher et al. (2010)** and **Sutton and Barto (2018)** for the pseudo-code and in-depth discussions of all algorithms.

Parameter fit and model selection

The main goal of this study was to test the null-hypothesis 'RL without eligibility traces' from the behavioral responses at states D1 and D2 (**Figure 2(e) and (f)**). By the design of the experiment, we collected relatively many data points from the early phase of learning, but only relatively few episodes in total. This contrasts with other RL studies, where participants typically perform longer experiments with hundreds of trials. As a result, the behavioral data we collected from each single participant is not sufficient to reliably extract the values of the model-parameters on a participant-by-participant basis. To find the most likely values of model parameters, we therefore pooled the behavioral recordings of all participants into one data set D.

Each learning model m is characterized by a set of parameters $\theta^m = (\theta_1^m, \theta_2^m, ...)$. For example, our implementation of the Q-λ algorithm has five free parameters: the eligibility trace decay λ; the learning rate α; the discount rate γ; the softmax temperature T; and the bias b for the preferred action. For each model m, we were interested in the posterior distribution $P(\theta^m|D)$ over the free parameters θ^m, conditioned on the behavioral data of all participants D. This distribution was approximated by sampling using the Metropolis-Hastings Markov Chain Monte Carlo (MCMC) algorithm (**Hastings, 1970**). For sampling, MCMC requires a function $f(\theta^m, D)$ which is proportional to $P(\theta^m|D)$. Choosing a uniform prior $P(\theta^m) = const$, and exploiting that $P(D)$ is independent of θ^m, we can directly use the model likelihood $P(D|\theta^m)$:

$$P(\theta^m|D) = \frac{P(D|\theta^m)P(\theta^m)}{P(D)} \propto P(D|\theta^m) := f(\theta^m, D). \tag{6}$$

We calculated the likelihood $P(D|\theta^m)$ of the data as the joint probability of all action selection probabilities obtained by evaluating the model (**Equations 1, 2, 3, and 4** in the case of $Q(\lambda)$) given a parameter sample θ^m. The log likelihood (LL) of the data under the model is

$$LL(D|\theta^m) = \sum_{p=1}^{N} \sum_{j=1}^{E_p} \sum_{t=1}^{T_j} log(p(a_t|s_t;\theta^m)), \tag{7}$$

where the sum is taken over all participants p, all environments j, and all actions a_t a participant has taken in the environment j.

For each model, we collected $100'000$ parameter samples (burn-in: 1500; keeping only every 10^{th} sample; 50 random start positions; proposal density: Gaussian with $\sigma = 0.004$ for temperature T and bias b, and $\sigma = 0.008$ for all other parameters). From the samples we chose the $\hat{\theta}^m$ which maximizes the log likelihood (LL), calculated the AIC_m and ranked the models accordingly. The AIC_m of each model is shown in **Table 1**, alongside with the Akaike weights $wAIC_m$. The latter can be interpreted

as the probability that the model m is the best model for the data (**Burnham and Anderson, 2004**). Note that the parameter vector $\hat{\theta}^m$ could be found by a hill-climbing algorithm toward the optimum, but such an algorithm does not give any indication about the uncertainty. Here, we obtained an approximate conditional posterior distribution $p(\theta_i^m|D,\hat{\theta}_{j\neq i}^m)$ for each component i of the parameter vector θ^m (cf. **Figure 9**). We estimated this posterior for a given parameter i by selecting only the 1% of all samples falling into a small neighborhood: $\hat{\theta}_j^m - \epsilon_j^m \leq \theta_j \leq \hat{\theta}_j^m + \epsilon_j^m, i \neq j$. We determined ϵ_j^m such that along each dimension j, the same percentage of samples was kept (about 22%) and the overall number of samples was 1000.

One problem using the AIC for model selection stems from the fact that there are considerable behavioral differences across participants and the AIC model selection might change for a different set of participants. This is why we validated the model ranking using k-fold cross-validation. The same procedure as before (fitting, then ranking according to AIC) was repeated K times, but now we used only a subset of participants (training set) to fit $\hat{\theta}_k^m$ and then calculated the LL_k^m and the AIC_k^m on the remaining participants (test set). We created the K folds such that each participant appears in exactly one test set and in $K-1$ training sets. Also, we kept these splits fixed across models, and evaluated each model on the same split into training and test set. In each fold k, the models were sorted with respect to AIC_k^m, yielding K lists of ranks. In order to evaluate whether the difference between two models is significant, we compared their ranking in each fold (Wilcoxon rank-sum test on K matched pairs, p-values shown in **Table 1**). The cross-validation results were summarized by summing the K ranks (**Table 1**). The best rank sum a model could obtain is K, and is obtained if it achieved the first rank in each of the K folds.

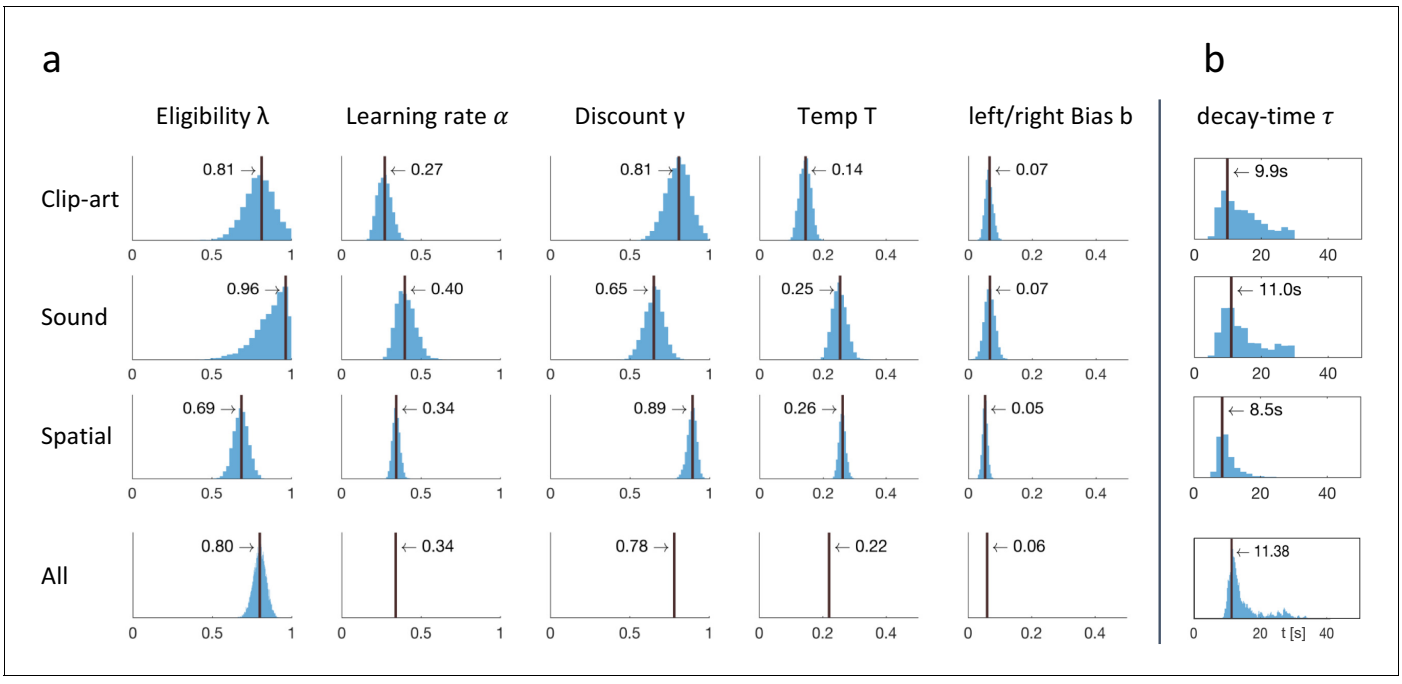


Figure 9. Fitting results: behavioral data constrained the free parameters of Q-λ. (**a**) For each experimental condition a distribution over the five free parameters is estimated by sampling. The blue histograms show the approximate conditional posterior for each parameter (see Materials and methods). Vertical black lines indicate the values of the five-parameter sample that best explains the data (maximum likelihood, ML). The bottom row (All) shows the distribution over λ when fitted to the aggregated data of all conditions, with other parameters fixed to the indicated value (mean over the three conditions). (**b**) Estimation of a time dependent decay (τ instead of λ) as defined in **Equation 5**.

$Q-\lambda$ model predictions

Algorithm 1 Q-λ (and related models):
For *SARSA-λ* we replace the expression $\max_{\tilde{a}} Q(s_{n+1},\tilde{a})$ in line 9 by $Q(s_{n+1},a_{n+1})$ where a_{n+1} is the action taken in the next state s_{n+1}. For *Q-0* and *SARSA-0* we set λ to zero.

1: Algorithm Parameters: learning rate $\alpha \in (0,1]$, discount factor $\gamma \in [0,1]$, eligibility trace decay factor $\lambda \in [0,1]$, temperature $T \in (0,\infty)$ of softmax policy p, bias $b \in [0,1]$ for preferred action $a_{pref} \in \mathbf{A}$.

2: Initialize $Q(s,a)=0$ and $e(s,a)=0$ for all $s \in \boldsymbol{S},\ a \in \boldsymbol{A}$
For preferred action $a_{pref} \in \boldsymbol{A}$ set $Q(s,a_{pref})=b$

3: for each episode do

4: Initialize state $s_n \in \boldsymbol{S}$

5: Initialize step $n=1$

6: while s_n is not terminal do

7: Choose action $a_n \in \mathbf{A}$ from s_n with softmax policy p derived from Q

8: Take action a_n, and observe $r_{n+1} \in \mathbb{R}$ and $s_{n+1} \in \mathbf{S}$

9: $RPE(n \rightarrow n+1) \leftarrow r_{n+1} + \gamma \max_{\tilde{a}} Q(s_{n+1},\tilde{a}) - Q(s_n,a_n)$

10: $e_n(s_n,a_n) \leftarrow 1$

11: for all $s \in \mathbf{S},\ a \in \mathbf{A}$ do

12: $Q(s,a) \leftarrow Q(s,a) + \alpha RPE(n \rightarrow n+1) e_n(s,a)$

13: $e_{n+1}(s,a) \leftarrow \gamma\lambda e_n(s,a)$

14 $n \leftarrow n+1$

The *Q-λ* model (see Algorithm 1), and related models like *ARSA-λ*, have previously been used to explain human data. We used those published results, in particular the parameter values from ***Gläscher et al. (2010)***, ***Daw et al. (2011)*** and ***Tartaglia et al. (2017)***, to estimate the effect size, as well as the reliability of the result. The published parameter values have a high variance: they differ across participants and across tasks. We therefore simulated different agents, each with its own parameters, sampled independently from a uniform distribution in the following ranges: $\alpha \in (0.1,0.5]$, $\lambda \in [0.5,1]$, $\gamma \in [0.5,1]$, $T \in [0.125,1]$ (corresponding to an inverse temperature $1/T \in [1,8]$), and $b=0$. We then simulated episodes 1 and 2 of the experiment, applied the $Q-\lambda$ model to calculate the action-selection bias (***Equation 4***) when the agents visit states $D1$, $D2$ and also S (see ***Figure 10(c)*** during episode 2, and sampled a binary decision (action 'a' or action 'b') according to the model's bias. In the same way as in the main behavioral experiment, each agent repeated the experiment four times and we estimated the empirical action-selection bias as the mean of the (simulated) behavioral data over all repetitions of all agents. This mean value depends on the actual realizations of the random variables and its uncertainty is higher when fewer samples are available. We therefore repeated the simulation of $N=10$ agents 1000 times and plotted the distribution of the empirical means in ***Figure 10(d)***. The same procedure was repeated for $N=20$ agents, showing a smaller standard deviation. The simulations showed a relatively large (simulated) effect size at states $D1$ and $D2$. Furthermore, as expected, the action bias decays as a function of the delay between the action and the final reward in episode 1. We then compared the $Q-\lambda$ model with a member of the class of *RL without eligibility trace*. When the parameter λ, which controls the decay of the eligibility trace, is set to 0, $Q-\lambda$ turns into $Q-0$ (Q-L*earning without eligibility trace* and we can use it to compare the two classes of RL without changing other parameters. Thus, we repeated the simulation for this case ($\lambda=0$, $N=20$) which shows the model predictions under our null hypothesis. ***Figure 10(d)*** shows the qualitative difference between the two classes of RL.

Regression analysis

The reward prediction error (RPE, ***Equation 2***) used for a comparison with pupil data was obtained by applying the algorithm *Q-λ* with the optimal (maximum likelihood) parameters. We chose *Q-λ* for regression because, first, it explained the behavior best across the three conditions and, second, it evaluates the outcome of an action at the onset of the next state (rather than at the selection of the next action as in *SARSA-λ*) which enabled us to compare the model with the pupil traces triggered at the onset of the next state.

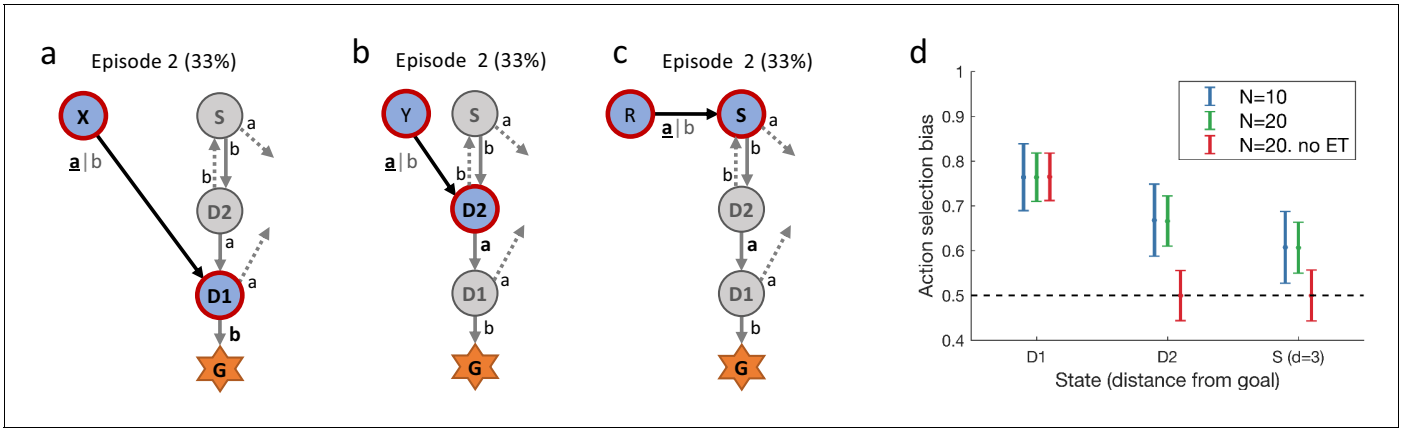


Figure 10. Simulated experiment. (Q-λ model). (**a**) and (**b**): Task structure (same as in ***Figure 2***). Simulated agents performed episodes 1 and 2 and we recorded the decisions at states D1 and D2 in episode 2. (**c**): Additionally, we also simulated the model's behavior at state S, by extending the structure of the (simulated) experiment with a new state R, leading to S. (**d**): We calculated the action-selection bias at states D1, D2 and S during episode 2 from the behavior of $N = 10$ (blue) and $N = 20$ (green) simulated agents. The effect size (observed during episode 2 and visualized in panel (**d**)) decreases when (in episode 1) the delay between taking the action and receiving the reward increases. It is thereby smallest at state S. When setting the model's eligibility trace parameter λ to 0(red, no ET), the effect at state D1 is not affected (see ***Equation 1***) while at D2 and S the behavior was not reinforced. Horizontal dashed line: chance level 50%. Errorbars: standard deviation of the simulated effect when estimating 1000 times the mean bias from $N = 10$ and $N = 20$ simulated agents with individually sampled model parameters.

In a first, qualitative, analysis, we split data of all state transitions of all articipants into two groups: all the state transitions where the model predicts an RPE of zero and the twenty percent of state transitions where the model predicts the largest RPE (***Figure 6(a)***. We found that the pupil responses looked very different in the two groups, across all three modalities.

In a second, rigorous, statistical analysis, we tested whether pupil responses were correlated with the RPE across all RPE values, not just those in the two groups with zero and very high RPE. In our experiment, only state G was rewarded; at nongoal states, the RPE depended solely on learned Q-values ($r_{n+1} = 0$ in ***Equation 2***). Note that at the first state of each episode the RPE is not defined. We distinguished these three cases in the regression analysis by defining two events 'Start' and 'Goal', as well as a parametric modulation by the reward prediction error at intermediate states. From ***Figure 5***, we expected significant modulations in the time window $t \in (500ms, 2500ms)$ after stimulus onset. We mapped t to $t' = (t - 1500ms)/1000ms$ and used orthogonal Legendre polynomials $P_k(t')$ up to order $k = 5$ (***Figure 11***) as basis functions on the interval $-1 \leq t' \leq 1$. We use the indices p for participant and n for the n^{th} state-on event. With a noise term ϵ and μ_t for the overall mean pupil dilation at t, the regression model for the pupil measurements y is

$$y_{p,n+1,t} = \mu_t + \sum_{k=0}^{5} RPE_p(n \rightarrow n+1) \times P_k(t') \times \beta_k + \epsilon_{p,n+1,t}\,, \tag{8}$$

where the participant-independent parameters β_k were fitted to the experimental data (one independent analysis for each experimental condition). The models for 'tart state' and 'oal state' are analogous and obtained by replacing the real valued $RPE_{p,n}$ by a 0/1 indicator for the respective events. By this design, we obtained three uncorrelated regressors with six parameters each.

Using the regression analysis sketched here, we quantified the qualitative observations suggested by (***Figure 6***) and found a significant parametric modulation of the pupil dilation by reward prediction errors at non-goal states (***Figure 11***). The extracted modulation profile reached a maximum at around 1–1.5 s (1300 ms in the *clip-art*, 1100 ms in the *sound* and 1400 ms in the *spatial* condition); with a strong mean effect size (β_0 in ***Figure 11***) of 0.48 ($p<0.001$), 0.41 ($p = 0.008$) and 0.35 ($p<0.001$), respectively.

We interpret the pupil traces at the start and the end of each episode (***Figure 11***) as markers for additional cognitive processes beyond reinforcement learning which could include correlations with cognitive load (***Beatty, 1982***; ***Kahneman and Beatty, 1966***), recognition memory (***Otero et al.,***

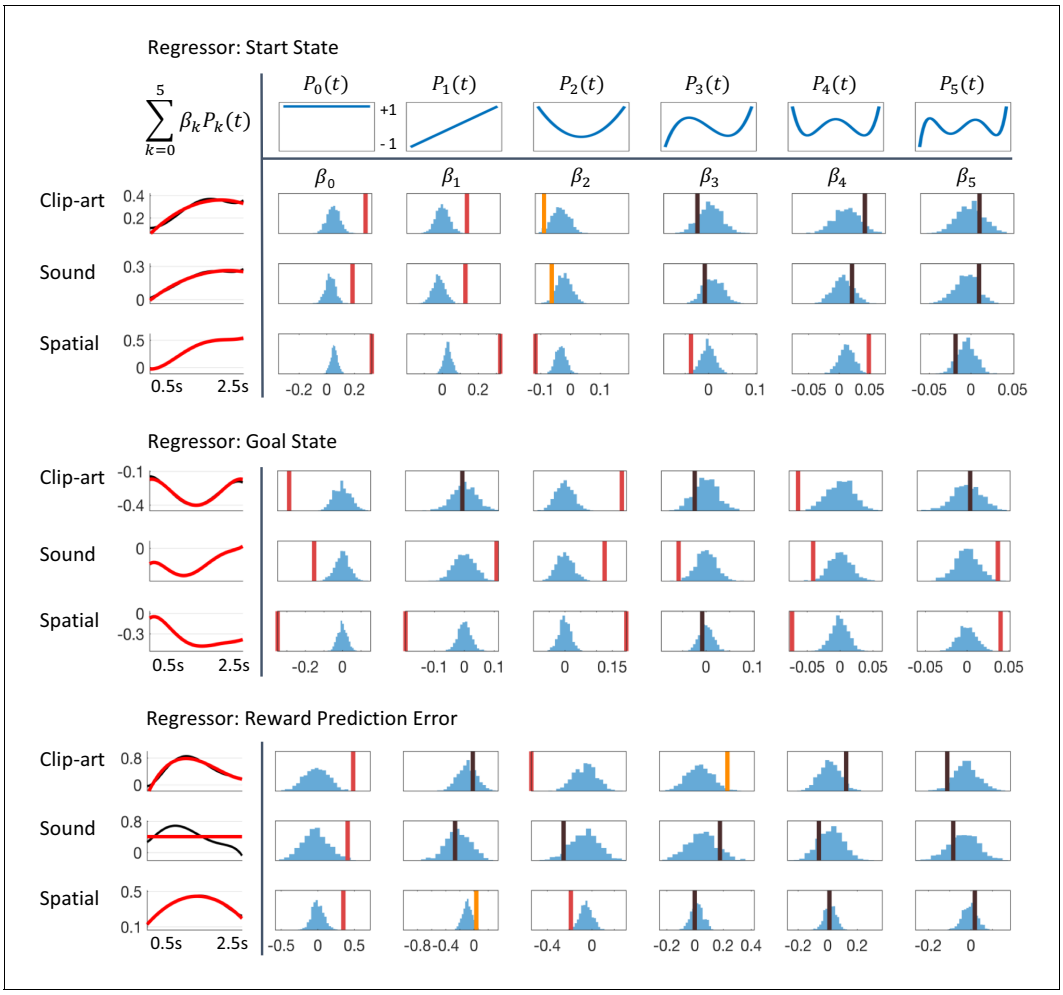


Figure 11. Detailed results of regression analysis and permutation tests. The regressors are *top*: Start state event, *middle*: Goal state event and *bottom*: Reward Prediction Error. We extracted the time course of the pupil dilation in (500 ms, 2500 ms) after state onset for each of the conditions, *clip-art*, *sound* and *spatial*, using Legendre polynomials $P_k(t)$ of orders k = 0 to k = 5 (top row) as basis functions. The extracted weights β_k (cf. ***Equation 8***) are shown in each column below the corresponding Legendre polynomial as vertical bars with color indicating the level of significance (red, statistically significant at p<0.05/6 (Bonferroni); orange, p<0.05; black, not significant). Blue histograms summarize shuffled samples obtained by 1000 permutations. Black curves in the leftmost column show the fits with all six Legendre Polynomials, while the red curve is obtained by summing only over the few Legendre Polynomials with significant β. Note the similarity of the pupil responses across conditions.

2011), attentional effort (***Alnæs et al., 2014***), exploration (***Jepma and Nieuwenhuis, 2011***), and encoding of memories (***Kucewicz et al., 2018***).

Acknowledgements

This research was supported by Swiss National Science Foundation (no. CRSII2 147636 and no. 200020 165538), by the European Research Council (grant agreement no. 268 689, MultiRules), and by the European Union Horizon 2020 Framework Program under grant agreement no. 720270 and no. 785907 (Human Brain Project, SGA1 and SGA2)

Additional information

Funding

Funder	Grant reference number	Author
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung	CRSII2 147636 (Sinergia)	Marco P Lehmann He A Xu Vasiliki Liakoni Michael H Herzog Wulfram Gerstner Kerstin Preuschoff
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung	CRSII2 200020 165538	Marco P Lehmann Vasiliki Liakoni Wulfram Gerstner
Horizon 2020 Framework Programme	Human Brain Project (SGA2) 785907	Michael H Herzog Wulfram Gerstner
H2020 European Research Council	268 689 MultiRules	Wulfram Gerstner
Horizon 2020 Framework Programme	Human Brain Project (SGA1) 720270	Michael H Herzog Wulfram Gerstner

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Marco P Lehmann, Conceptualization, Software, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing; He A Xu, Software, Formal analysis, Validation, Investigation, Methodology, Writing—review and editing; Vasiliki Liakoni, Software, Investigation, Methodology, Writing—review and editing; Michael H Herzog, Conceptualization, Supervision, Funding acquisition, Methodology, Project administration, Writing—review and editing; Wulfram Gerstner, Conceptualization, Supervision, Funding acquisition, Investigation, Methodology, Writing—original draft, Project administration, Writing—review and editing; Kerstin Preuschoff, Conceptualization, Supervision, Funding acquisition, Methodology, Writing—original draft, Project administration, Writing—review and editing

Author ORCIDs

Marco P Lehmann https://orcid.org/0000-0001-5274-144X
Vasiliki Liakoni https://orcid.org/0000-0002-2599-1424

Ethics

Human subjects: Experiments were conducted in accordance with the Helsinki declaration and approved by the ethics commission of the Canton de Vaud (164/14 Titre: Aspects fondamentaux de la reconnaissance des objets : protocole général). All participants were informed about the general purpose of the experiment and provided written, informed consent. They were told that they could quit the experiment at any time they wish.

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.47463.sa1
Author response https://doi.org/10.7554/eLife.47463.sa2

Additional files

Supplementary files

- Transparent reporting form

Data availability

The datasets generated during the current study are available on Dryad (https://doi.org/10.5061/dryad.j7h6f69).

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Lehmann M, Xu HA, Liakoni V, Herzog MH, Gerstner W, Preuschoff K	2019	Data from: One-shot learning and behavioral eligibility traces in sequential decision making	https://doi.org/10.5061/dryad.j7h6f69	Dryad Digital Repository, 10.5061/dryad.j7h6f69

References

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**: 716–723. DOI: https://doi.org/10.1109/TAC.1974.1100705

Alnæs D, Sneve MH, Espeseth T, Endestad T, van de Pavert SHP, Laeng B. 2014. Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *Journal of Vision* **14**:1. DOI: https://doi.org/10.1167/14.4.1

Beatty J. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin* **91**:276–292. DOI: https://doi.org/10.1037/0033-2909.91.2.276, PMID: 7071262

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* **57**:289–300. DOI: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berke JD. 2018. What does dopamine mean? *Nature Neuroscience* **21**:787–793. DOI: https://doi.org/10.1038/s41593-018-0152-y, PMID: 29760524

Bittner KC, Milstein AD, Grienberger C, Romani S, Magee JC. 2017. Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science* **357**:1033–1036. DOI: https://doi.org/10.1126/science.aan3846

Blundell C, Uria B, Pritzel A, Li Y, Ruderman A, Leibo JZ, Rae J, Wierstra D, Hassabis D. 2016. Model-free episodic control. *arXiv*. https://arxiv.org/abs/1606.04460.

Bogacz R, McClure SM, Li J, Cohen JD, Montague PR, Read Montague P. 2007. Short-term memory traces for action bias in human reinforcement learning. *Brain Research* **1153**:111–121. DOI: https://doi.org/10.1016/j.brainres.2007.03.057

Brady TF, Konkle T, Alvarez GA, Oliva A. 2008. Visual long-term memory has a massive storage capacity for object details. *PNAS* **105**:14325–14329. DOI: https://doi.org/10.1073/pnas.0803390105, PMID: 18787113

Brainard DH. 1997. The Psychophysics Toolbox. *Spatial Vision* **10**:433–436. DOI: https://doi.org/10.1163/156856897X00357

Brea J. 2017. Is prioritized sweeping the better episodic control? *arXiv*. https://arxiv.org/abs/1606.04460.

Brzosko Z, Zannone S, Schultz W, Clopath C, Paulsen O. 2017. Sequential neuromodulation of hebbian plasticity offers mechanism for effective reward-based navigation. *eLife* **6**:e27756. DOI: https://doi.org/10.7554/eLife.27756, PMID: 28691903

Burnham KP, Anderson DR. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research* **33**:261–304. DOI: https://doi.org/10.1177/0049124104268644

Crow TJ. 1968. Cortical synapses and reinforcement: a hypothesis. *Nature* **219**:736–737. DOI: https://doi.org/10.1038/219736a0, PMID: 5667068

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. 2011. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**:1204–1215. DOI: https://doi.org/10.1016/j.neuron.2011.02.027, PMID: 21435563

Duncan KD, Shohamy D. 2016. Memory states influence value-based decisions. *Journal of Experimental Psychology: General* **145**:1420–1426. DOI: https://doi.org/10.1037/xge0000231

Fisher SD, Robertson PB, Black MJ, Redgrave P, Sagar MA, Abraham WC, Reynolds JNJ. 2017. Reinforcement determines the timing dependence of corticostriatal synaptic plasticity in vivo. *Nature Communications* **8**:334. DOI: https://doi.org/10.1038/s41467-017-00394-x, PMID: 28839128

Frémaux N, Gerstner W. 2015. Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Frontiers in Neural Circuits* **9**:85. DOI: https://doi.org/10.3389/fncir.2015.00085, PMID: 26834568

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. *Frontiers in Neural Circuits* **12**:53. DOI: https://doi.org/10.3389/fncir.2018.00053, PMID: 30108488

Gläscher J, Daw N, Dayan P, O'Doherty JP. 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**:585–595. DOI: https://doi.org/10.1016/j.neuron.2010.04.016, PMID: 20510862

Glimcher PW, Fehr E. 2013. *Neuroeconomics: Decision Making and the Brain*. Elsevier Inc. DOI: https://doi.org/10.1016/C2011-0-05512-6

Greve A, Cooper E, Kaula A, Anderson MC, Henson R. 2017. Does prediction error drive one-shot declarative learning? *Journal of Memory and Language* **94**:149–165. DOI: https://doi.org/10.1016/j.jml.2016.11.001, PMID: 28579691

Gureckis TM, Love BC. 2009. Short-term gains, long-term pains: how cues about state aid learning in dynamic environments. *Cognition* **113**:293–313. DOI: https://doi.org/10.1016/j.cognition.2009.03.013, PMID: 19427635

Hastings WK. 1970. Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57**: 97–109. DOI: https://doi.org/10.1093/biomet/57.1.97

He K, Huertas M, Hong SZ, Tie X, Hell JW, Shouval H, Kirkwood A. 2015. Distinct eligibility traces for LTP and LTD in cortical synapses. *Neuron* **88**:528–538. DOI: https://doi.org/10.1016/j.neuron.2015.09.037, PMID: 26593091

Izhikevich EM. 2007. *Dynamical Systems in Neuroscience : The Geometry of Excitability and Bursting*. MIT Press.

Jepma M, Nieuwenhuis S. 2011. Pupil diameter predicts changes in the exploration-exploitation trade-off: evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience* **23**:1587–1596. DOI: https://doi.org/10.1162/jocn.2010.21548, PMID: 20666595

Joshi S, Li Y, Kalwani RM, Gold JI. 2016. Relationships between pupil diameter and neuronal activity in the locus coeruleus, Colliculi, and cingulate cortex. *Neuron* **89**:221–234. DOI: https://doi.org/10.1016/j.neuron.2015.11.028, PMID: 26711118

Kahneman D, Beatty J. 1966. Pupil diameter and load on memory. *Science* **154**:1583–1585. DOI: https://doi.org/10.1126/science.154.3756.1583, PMID: 5924930

Kucewicz MT, Dolezal J, Kremen V, Berry BM, Miller LR, Magee AL, Fabian V, Worrell GA. 2018. Pupil size reflects successful encoding and recall of memory in humans. *Scientific Reports* **8**:4949. DOI: https://doi.org/10.1038/s41598-018-23197-6, PMID: 29563536

Mathôt S, Fabius J, Van Heusden E, Van der Stigchel S. 2017. Safe and sensible baseline correction of pupil-size data. *PeerJ Preprints*. https://peerj.com/preprints/2725.

Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K. 2016. Asynchronous methods for deep reinforcement learning. Proceedings of the 33rd International Conference on Machine Learning, PMLR 48 1928–1937. http://proceedings.mlr.press/v48/mniha16.html.

Moore AW, Atkeson CG. 1993. Prioritized sweeping: reinforcement learning with less data and less time. *Machine Learning* **13**:103–130. DOI: https://doi.org/10.1007/BF00993104

Nieuwenhuis S, De Geus EJ, Aston-Jones G. 2011. The anatomical and functional relationship between the P3 and autonomic components of the orienting response. *Psychophysiology* **48**:162–175. DOI: https://doi.org/10.1111/j.1469-8986.2010.01057.x

Niv Y, Edlund JA, Dayan P, O'Doherty JP. 2012. Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *Journal of Neuroscience* **32**:551–562. DOI: https://doi.org/10.1523/JNEUROSCI.5498-10.2012

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. 2003. Temporal difference models and reward-related learning in the human brain. *Neuron* **38**:329–337. DOI: https://doi.org/10.1016/S0896-6273(03)00169-7, PMID: 12718865

O'Doherty JP, Cockburn J, Pauli WM. 2017. Learning, reward, and decision making. *Annual Review of Psychology* **68**:73–100. DOI: https://doi.org/10.1146/annurev-psych-010416-044216, PMID: 27687119

Otero SC, Weekes BS, Hutton SB. 2011. Pupil size changes during recognition memory. *Psychophysiology* **48**: 1346–1353. DOI: https://doi.org/10.1111/j.1469-8986.2011.01217.x

Pan WX, Schmidt R, Wickens JR, Hyland BI. 2005. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. *Journal of Neuroscience* **25**:6235–6242. DOI: https://doi.org/10.1523/JNEUROSCI.1478-05.2005, PMID: 15987953

Peng J, Williams RJ. 1996. Incremental multi-step Q-learning. *Machine Learning* **22**:283–290. DOI: https://doi.org/10.1007/BF00114731

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* **442**:1042–1045. DOI: https://doi.org/10.1038/nature05051, PMID: 16929307

Preuschoff K, 't Hart BM, Einhäuser W. 2011. Pupil dilation signals surprise: evidence for noradrenaline's Role in Decision Making. *Frontiers in Neuroscience* **5**:1–12. DOI: https://doi.org/10.3389/fnins.2011.00115, PMID: 21994487

Rescorla RA, Wagner AR. 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: *Classical Conditioning II: Current Research and Theory*. Appleton Century Crofts.

Rouhani N, Norman KA, Niv Y. 2018. Dissociable effects of surprising rewards on learning and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **44**:1430–1443. DOI: https://doi.org/10.1037/xlm0000518

Schultz W. 2015. Neuronal reward and decision signals: from theories to data. *Physiological Reviews* **95**:853–951. DOI: https://doi.org/10.1152/physrev.00023.2014, PMID: 26109341

Seijen HV, Sutton R. 2013. Planning by prioritized sweeping with small backups. Proceedings of the 30th International Conference on Machine Learning.

Singh SP, Sutton RS. 1996. Reinforcement learning with replacing eligibility traces. *Machine Learning* **22**:123–158. DOI: https://doi.org/10.1007/BF00114726

Standing L. 1973. Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology* **25**:207–222. DOI: https://doi.org/10.1080/14640747308400340, PMID: 4515818

Sutton RS. 1988. Learning to predict by the methods of temporal differences. *Machine Learning* **3**:9–44. DOI: https://doi.org/10.1007/BF00115009
Sutton RS, Barto AG. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
Tartaglia EM, Clarke AM, Herzog MH. 2017. What to choose next? A paradigm for testing human sequential decision making. *Frontiers in Psychology* **8**:1–11. DOI: https://doi.org/10.3389/fpsyg.2017.00312, PMID: 28326050
Walsh MM, Anderson JR. 2011. Learning from delayed feedback: neural responses in temporal credit assignment. *Cognitive, Affective, & Behavioral Neuroscience* **11**:131–143. DOI: https://doi.org/10.3758/s13415-011-0027-0, PMID: 21416212
Walsh MM, Anderson JR. 2012. Learning from experience: event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience & Biobehavioral Reviews* **36**:1870–1884. DOI: https://doi.org/10.1016/j.neubiorev.2012.05.008, PMID: 22683741
Watkins C. 1989. Learning from delayed rewards. Cambridge University.
Weinberg A, Luhmann CC, Bress JN, Hajcak G. 2012. Better late than never? the effect of feedback delay on ERP indices of reward processing. *Cognitive, Affective, & Behavioral Neuroscience* **12**:671–677. DOI: https://doi.org/10.3758/s13415-012-0104-z, PMID: 22752976
Williams RJ. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* **8**:229–256. DOI: https://doi.org/10.1007/BF00992696
Yagishita S, Hayashi-Takagi A, Ellis-Davies GC, Urakubo H, Ishii S, Kasai H. 2014. A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* **345**:1616–1620. DOI: https://doi.org/10.1126/science.1255514, PMID: 25258080
Yates FA. 1966. *Art of Memory*. Routledge and Kegan Paul.

EEG signatures of the Reward-Prediction Error at non-rewarded states

He A. Xu, Marco P. Lehmann, Wulfram Gerstner, and Michael H. Herzog

Ecole Polytechnique Fédérale de Lausanne, Brain-Mind Institute, 1015 Lausanne EPFL

Abstract

The reward prediction error (RPE), i.e., the difference between the actual and the expected reward, is one of the crucial variables in model-free reinforcement learning. In humans, RPE has been shown in one-step decision tasks to be correlated with the feedback-related negativity (FRN), a frontal-central EEG signal (Holroyd & Coles, 2002). Previous FRN studies used N-armed bandit tasks where participants receive reward immediately after an action, contrary to everyday tasks where many actions are needed before a reward occurs. Here, we employed a sequential decision-making paradigm and show that FRN amplitudes reflect the RPE also at non-rewarded states. In our task, participants had to make many decisions until a goal was found. Each decision led to an action that brought the participant from one discrete state, characterized by an image, to another state, characterized by a different image. Based on the predicted qualitative signature of the RPE at the goal state, we extracted the EEG signal in the time window of 280-400ms after state onset. We then fitted the behavioural data of the participants with the reinforcement learning model SARSA(λ). We found that the RPE predicted by the model correlated significantly with the FRN for non-rewarded states. Hence, the FRN reflects the RPE not just in rewarded states, but also at non-rewarded states far from the goal.

1. Introduction

In chess, a series of moves has to be made until a sparse reward (win, loss, draw) is issued, which makes it difficult to immediately evaluate the value of a single move. Reinforcement learning (RL) deals with such types of situations. In a typical RL situation, an agent moves through an environment, which has several states. Some states come with rewards, others do not. At each state ***s***, the agent makes an action ***a***, which brings the agent from the state ***s*** to another state ***s'***. The goal of the agent is to move through the environment to maximise the total reward.

A particular successful class of RL models are model-free models, i.e., the agent does not learn an explicit map of the environment but rather chooses the action at each state based on so-called Q-values, which summarize how successful these actions were in the past (Sutton & Barto, 1998). In model-free RL models, one of the most crucial components is the reward

prediction error (RPE), which is the difference between the actual reward and the expected reward. If the actual reward is higher than the expected reward, a positive reward prediction error occurs, indicating that the decision has led to a more rewarding state than expected. By contrast, if the actual reward is lower than the expectation, a negative reward prediction error is produced, which indicates that the past decision was a 'bad' decision.

Many studies have shown evidence for model free RL models in animals (Montague, 1996; Schultz, Dayan, & Montague, 1997) and humans. Neurophysiological studies found markers for the RPE in brain areas such as the anterior cingulate (Bellebaum & Daum, 2008; Gehring & Willoughby, 2002; Gruendler, Ullsperger, & Huster, 2011; Tucker, Luu, Frishkoff, Quiring, & Poulsen, 2003), the posterior cingulate (Badgaiyan & Posner, 1998; Cohen & Ranganath, 2007; Doñamayor, Marco-Pallarés, Heldmann, Schoenfeld, & Münte, 2011; Nieuwenhuis, Slagter, von Geusau, Heslenfeld, & Holroyd, 2005), the ventral segmental area, the ventral stratum (Haruno & Kawato, 2006; McClure, Berns, & Montague, 2003; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003) and the basal ganglia (Carlson, Foti, Mujica-Parodi, Harmon-Jones, & Hajcak, 2011; Cohen, Cavanagh, & Slagter, 2011; Martin, Potts, Burton, & Montague, 2009).

EEG studies have shown that the amplitude of a frontal-central component, called the feedback-related negativity (FRN), correlates well with the RPE (Miltner et al. 1997; Gehring & Willoughby 2002; Holroyd & Coles 2002). The FRN occurs between 200ms and 400ms after stimulus presentation. In all these EEG studies, participants were tested in N-armed bandit tasks, where the reward is delivered immediately after an action. These tasks are similar to slot machines in a casino. There is one starting state with N possible actions that lead or do not lead to a reward, and end the game. Hence, reward (or no reward) is obtained after each single action. The expected reward for an action can simply be estimated as the average over the past rewards for that action. However, as mentioned above, in everyday reinforcement learning situations, there is a sequence of several non-rewarded states until a reward is found. Hence, an obvious question is whether RPEs at non-rewarded state are reflected by similar electrophysiological signatures. Here, we tested human participants in a deep sequential decision making task. In this paradigm, states are represented by images on a computer screen and actions are represented by grey disks at the bottom of the screen (Figure 1A; Tartaglia et al. 2017). We recorded 128-channel EEG to test whether we find FRN like signals for rewarded and non-rewarded states.

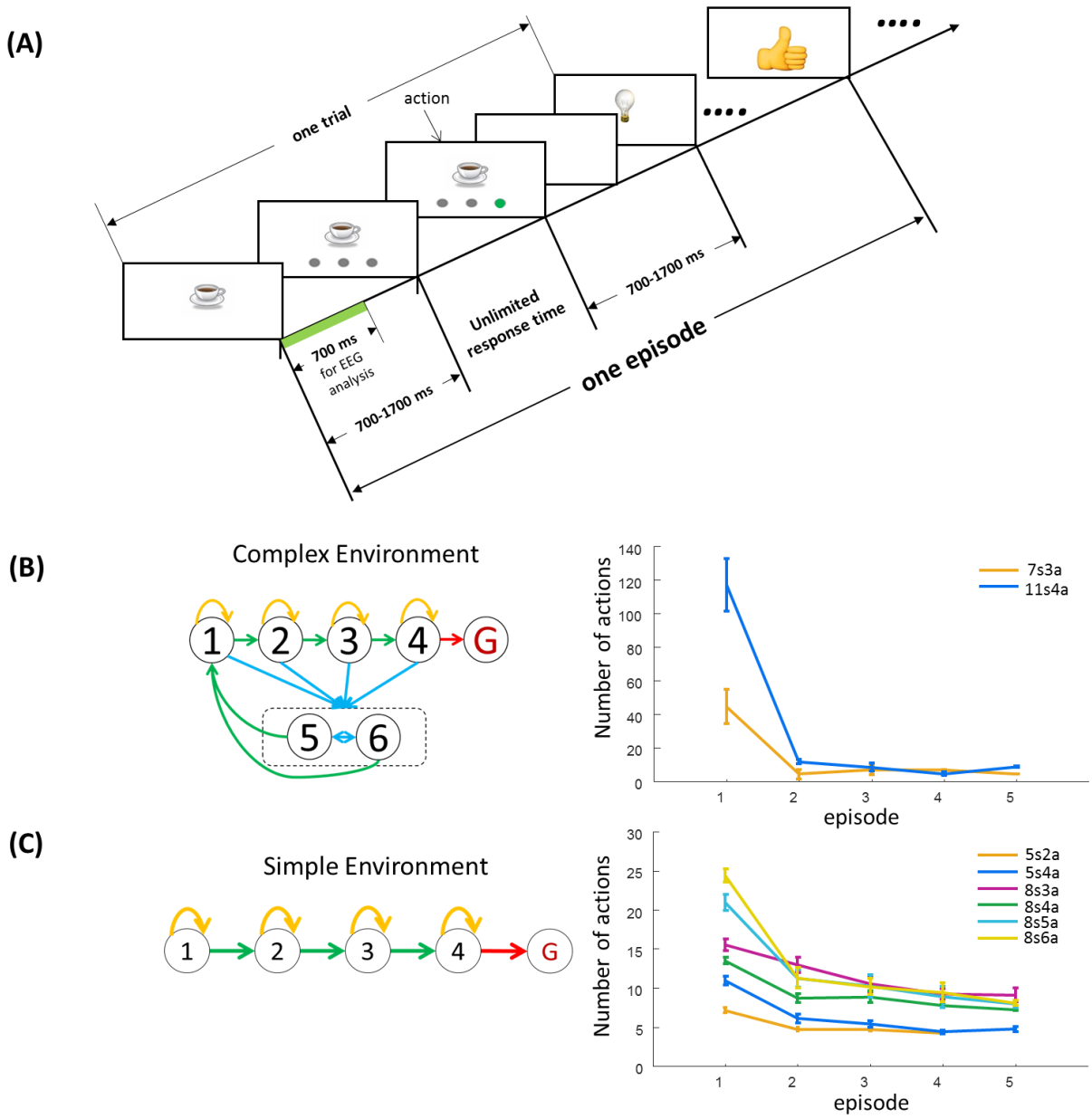


Figure 1 (A). Sequential Decision Making Paradigm. An image (state) is presented on the screen. After a random interval of 700-1700ms, grey disks appear, on which participants are asked to click (actions). After an action, a blank screen is shown for a random interval between 700 and 1700ms and then the next state appears. The goal state is a 'thumb-up' image in this example. The green interval indicates the time (0-700ms after the image onset), for which ERP was analysed. **(B).** Structure of the complex environment. Non-goal states are indicated by numbers while the goal state is presented by the red G. 's' indicates 'states', 'a' indicates 'actions'. For example, '7s3a' means that the environment has 7 states (including the goal state) and each state comes with three actions. Arrows present the outcomes of the actions. There were three groups of states: (i) the goal state (red G), (ii) progressing states (states 1-4) and (iii) trap states (states 5-6). In order to find the goal state as fast as possible, participants needed to

avoid the trap states. For each non-goal state, there was only one action (green arrows), which led participants to the next state; one other action (yellow arrows) led participants back to the current state. Actions that led participants to states 5-6 are shown in blue (see methods for details). Performance was determined as the number of actions participants needed to find the goal state. Performance is shown on the right as a function of the number of episodes finished. Points connected by lines indicate the means and bars indicate the standard error. **(C)**. Structure of the simple environment. There were only two types of actions at each non-goal states: one action led participants to the next state (green arrows), all other actions let participant stay at the current state (yellow arrows). The task is much easier because participants either stayed or moved towards the goal states. Performance is shown on the right.

2. General Materials and Methods

2.1 Experimental set up

Stimuli were generated using the Psychophysics Toolbox (ver 3, Brainard 1997) for Matlab R2011b (Windows OS) and presented on a Phillips 201B4 monitor (screen resolution of 1,980 × 1080 pixels and a refresh rate of 100 Hz).

2.2 Participants

14 paid participant took part in the first experiment with two complex environments (Figure 1B). Two participants quit during the experiment. Hence, we analysed data for 12 participants (5 females, aged 20-26 years, mean = 22.8, sd = 1.7). Another 14 paid participants took part in the second experiment with six simple environments (Figure 1C, 7 females, aged 20–25 years, mean = 22.5, sd = 1.6). All participants were right-handed, and as determined by self-report, naïve to the purpose of the experiments.

All participants had normal or corrected-to-normal visual acuity. All participants gave informed consent in accordance with the protocol 384/2011 “Commission cantonale d’éthique de la recherche sur l’être humain”.

2.3 Experiments

2.3.1 Stimuli and general procedure

In the first experiment, we used two complex environments (Figure 1B). In the second experiment we employed six simple environments (Figure 1C). Participants were presented a clip art image and a number of grey disks below the image (Figure 1A). Clicking on one of the disks (action) led to a subsequent image. Participants clicked through the images until they found the goal image, which ended an episode. Before the experiment, we showed participants the goal image and told them that this was the goal image. In the complex environment, participants started an episode at a randomly chosen non-goal image. In the simple environment, participants always started an episode with the same image (state 1 in Figure 1C).

For each image, clicking at the same disk led always to the same subsequent image, i.e., the transitions were deterministic. In other words, the state-action transition matrix, which defines the environment, contains only ones and zeroes.

During a trial (Figure 1A), an image was shown during an interval of 700 to 1700ms (uniform random) and then the grey disks appeared while the image stayed on the screen. Disks were shown until participants clicked on one of them (action). There was no time limit for making an action. After an action, a blank screen was presented with a duration of 700 to 1700ms randomly chosen. Then, the next image was shown and so on. EEG was recorded during the entire experiment but we analysed only the interval between 200ms before to 700ms after the onset of the image. The 200ms before the state onset were used for EEG baseline correction. The 700ms after the state onset (green interval in Figure 1A) were used for the analysis of the Event-Related Potentials (ERP).

The two complex environments and the six simple environments all differed in the number of states and actions (Figure 1B, 1C). Participants performed 5 episodes with each environment. For each new environment, a new set of images was used. The order of the environments was the same for all participants. After each environment, participants could have a 3-minute break before they started the next block.

2.3.2 Complex Environments

The structure of the complex environments (Figure 1B) contained short and long loops. There was always a unique shortest path from every non-goal state to the goal state. There were three types of states: (i) the goal state, (ii) states that were on the shortest path to the goal state (progressing states), (iii) states that were farthest away from the goal state (trap states).

From each progressing state, exactly one action led participants to the next progressing state and exactly one action led participants stay at the same state. The other actions led them to a trap state. For each trap state, there was only one action that brought participants back to state 1 (the first progressing state in Figure 1B). All the other actions led participants stay within the group of trap states (stayed at the current trap state or went to another trap state). Two environments were tested. The first environment contained 7 states and 3 actions for each state (one goal state, two trap states and four progressing states; among the three actions, one let participants stay at the same state, one led participants to the next progressing state and one led participants to a trap state). The second environment contained 11 states and 4 actions for each state (one goal state, three trap states and seven progressing states; among the four actions, one let participants stay at the same state, one led participants to the next progressing state and each of the remaining two led participants to a trap state).

2.3.3 The Simple Environment

The structure of the simple environments is shown in Figure 1C. For all environments, there was only one shortest path from a non-goal state to the goal state. For each state, only one action led participants to the next state, while all other actions led back to the current state. Hence, participants could only move forward to the goal or stay at a state. We used 6 environments: 5 states-2 actions (5s2a), 5 states-4 actions (5s4a), 8 states-3 actions (8s3a), 8 states-4 actions (8s4a), 8 states-5 actions (8s5a), 8 states-6 actions (8s6a). The goal state was included in the count of the number of states.

2.4 EEG recording and pre-processing

EEG signals were recorded using BioSemi equipment with 128 electrodes at a 2048Hz sampling rate. Data were band pass filtered from 0.1Hz to 40Hz and down sampled to 256Hz. Common average referencing was applied for re-referencing. Bad channels were visually inspected and interpolated using the EEGLAB toolbox (Delorme & Makeig, 2004). Eye movements and electromyography (EMG) artefacts were removed by using independent component analysis (ICA). Trials in which the change in voltage at any channel exceeded 35 µV per sampling point were discarded. For each trial, an epoch was extracted from 200ms before to 700ms after the state onset. The interval from 200ms to 0ms before the state onset was used for baseline correction. Prefrontal Event-Related Potentials (ERPs) were computed by averaging the EEG data (green interval in Figure 1A) of selected prefrontal electrodes (Fz, F1, F2, AFz, FCz) for ERP analysis.

2.5 Data analysis

2.5.1 Computing the RPE from behavioural data

In order to obtain an estimate of the RPE in each state (image), we fitted the SARSA(λ) algorithm (Sutton & Barto, 1998) to the human behavioural data (see Suppl. Materials 2) using the methods of Lehmann et al. (Lehmann et al., 2017).

2.5.2 Determination of the time course of the RPE using goal states

In each environment, participants performed a total of 5 episodes. The very first time the participants found the goal image, the expected reward was low while the actual reward was high. Therefore a high positive RPE occurred. We suppose that in the next episode, participants had a higher reward expectation when reaching the goal image, so the RPE is smaller at the second encounter. Hence, the RPE for the goal state decreases as the goal is visited more often. This qualitative observation is independent of model assumptions and is true for both SARSA and Q-learning whatever the choice of the eligibility parameter λ (see Suppl. Materials 2).

We used this rationale to find the time interval for the RPE. To this end, we determined ERPs from prefrontal electrodes when participants found the goal image the first, third, and fifth time. We averaged the ERP amplitudes in a sliding time window of 50ms (shifted in 10ms-steps) and searched for a rank order of ERP amplitudes with either (i) *ERP (first visit) > ERP (third visit) > ERP (fifth visit)* or (ii) *ERP (first visit) < ERP (third visit) < ERP (fifth visit)*. We considered both options since we were looking for a qualitative correlation, i.e., we did not want to exclude the possibility that the correlation had a negative sign. The continuous segment that started at the earliest time point of the first sliding window fulfilling the above condition and ended at the latest time point of the last sliding window fulfilling the above condition was defined as the interval of interest.

There are many model-free reinforcement learning algorithms (Beeler, Daw, Frazier, & Zhuang, 2010; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Daw, Niv, & Dayan, 2005; Gershman & Daw, 2017; Glaescher, Daw, Dayan, & O'Doherty, 2010; Lehmann et al., 2017; Niv et al., 2015; Niv, Edlund, Dayan, & O'Doherty, 2012; O'Doherty, Cockburn, & Pauli, 2017; O'Doherty et al., 2003; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006). Within this family of reinforcement learning algorithms, we chose SARSA with an eligibility trace λ because this algorithm explained behavioural data well in earlier experiments (Gershman & Daw, 2017; O'Doherty et al., 2003). We used a probabilistic fitting algorithm (Lehmann et al., 2017) to extract possible combinations of parameters that explained the observed behaviour (Suppl. Materials 2). For each environment, we then fitted the behaviour of all participants with a single SARSA model with a set of parameters reflecting the mean values of the parameter distributions.

2.5.3 Trial-by-trial analysis of the RPE for non-goal & goal states

Within the time interval of interest extracted using the methods of the preceding subsection, we evaluated correlations between EEG amplitudes and RPEs at non-goal and goal states. To this end, we predicted the RPEs for each trial and each participant using the parameter set of the SARSA(λ) model fitted on the behavioural data (see Suppl. Materials 2) and, then, correlated the RPEs with the ERPs from the prefrontal electrodes.

3. Results

For the complex environments, participants needed between 19 to 213 actions to find the goal state the first time (Figure 1B; mean = 44.5, std = 36.6, se= 10.1 for the environment with 7 states; mean = 117.0, std = 54.2, se = 15.6 for environment with 11 states). In the simple experiment, participants needed between 7 and 31 actions before they reached the goal state for the first time (Figure 1C).

The time participants spent during the first episode increased with the number of states and number of available actions for all environments and was longest for the environment with 11 states (mean = 117.0, std = 54.2) and shortest for environment with 5 states (5s2a, mean = 8.1, std = 1.1). Importantly, most participants found a much shorter path in subsequent episodes in all environments (Figure 1B) indicating that they understood the aim of the task. In some cases, participants took the shortest path straight to goal already in the second episode (Figure 2A). In simple environments, participants did not always use the shortest path to goal (Figure 2B), which does not affect our analysis because we only focused on the goal state in these environments.

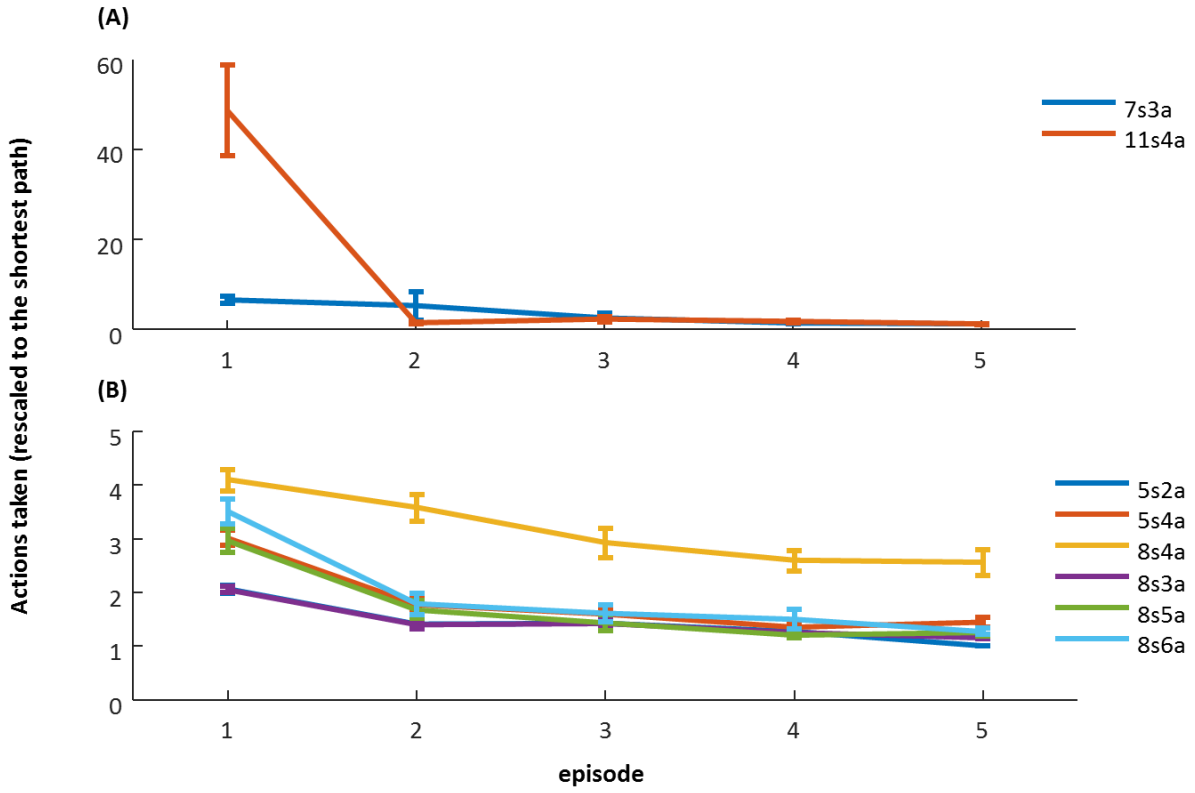


Figure 2. Behavioural performance (data from Fig. 1B, C) rescaled to the shortest path in each episode. The y-axis presents performance, which is calculated as the ratio between *the number of actions participants took to finish an episode* and *the minimum number of actions needed*. A y-value of 1 indicates that the participants used the shortest path. (A) In the complex environment with 11 states, the first episode started in state 6 and the second episode always started in state 9; in the environment with 7 states, the first episode always started in state 6 and the second episode in state 4 (detailed environment structure see Suppl. Materials 3). (B) In the simple environment with eight states, the second episode always started in state 1, for the environment with 4 actions and in state 1 for the one with three actions (detailed

environment structure see Suppl. Materials 3). Please note the difference in the y-axis scales of (A) and (B).

For the analysis, we used only the RPEs, but not participant or environment (see Suppl. Materials 1), as the predictor variables, and computed linear regressions between the RPEs and ERP amplitudes for each participant.

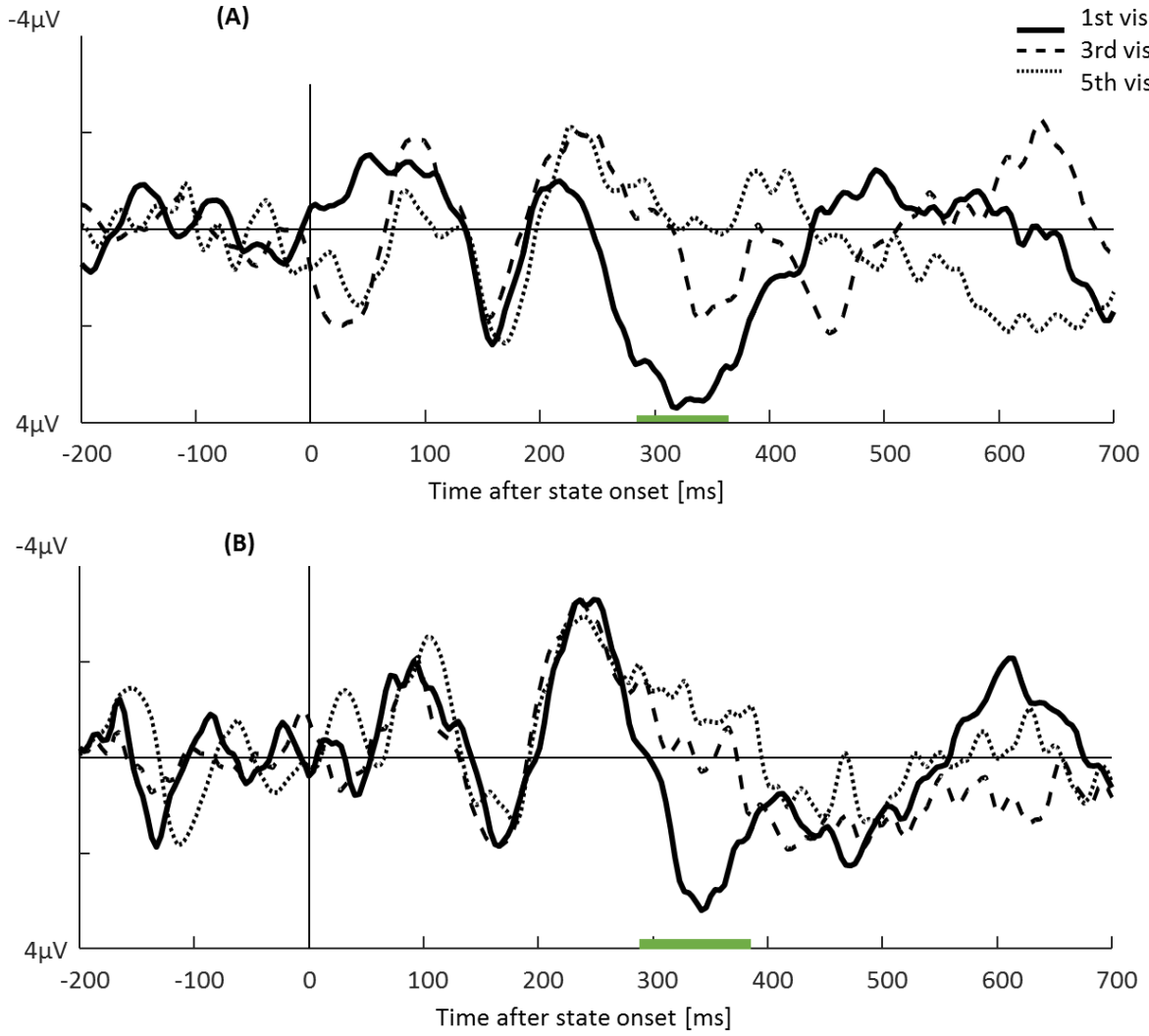


Figure 3. ERPs for the 1st, 3rd and 5th goal visit. 0 on the x-axis indicates the image onset. Negative values are plotted up by convention. Green lines indicate significant differences between the ERPs of the 1st, 3rd, and 5th visit to the goal with a monotonic trend of the RPEs. **(A)** Complex environments. ERPs were significantly different between 280-360ms ($F(2,33) = 4.84$, $p = 0.014$). **(B)** Simple environments. ERPs were significantly different between 280-390ms ($F(2,39) = 5.39$, $p=0.008$).

3.1 Experiment 1: EEG vs. RPE at the non-goal states

In order to study whether FRN amplitudes reflect the RPEs at non-goal states, we used the complex environments, which contained many states 'far away' from the goal. We searched for periods where ERP amplitudes at the goal state decreased or increased monotonically as a function of episode number. We found a significant monotonic trend in the time window of 280-360ms after the onset of the goal image (Figure 3A, $F(2,33) = 4.84$, $p = 0.014$). This time window, extracted by a model-independent qualitative argument, is the one that we take as our interval of interest in the following.

The main focus of this experiment was on the non-goal states. We initialized all Q-values at 0 in model fitting (details in Supp. Materials 2). Thus, the RPEs of non-goal states were also all 0s in the first episode. Since 0-values do not provide any useful information for the regression between ERP amplitudes and RPEs we discarded all non-goal trials in the first episode. In the second and subsequent episodes, Figure 2A shows that most participants used a near-shortest path, meaning that they most often chose the best action at each non-goal state. In this case, the RPE estimations using SARSA and Q-learning are equivalent (see Suppl. Materials 2).

We tested linear regressions between the RPEs (predicted by SARSA(λ))and ERP amplitudes in the time window of 280-360ms for all non-goal state trials in episode 2 to 5 on a participant-by-participant basis. A t-test on the linear regression coefficients was significant indicating that the RPEs predicted the ERPs of the non-goal states (Figure 4B, $p = 0.02$, $t(11) = 2.5$, sd = 6.1, mean coefficient = 3.2).

In the same time window, we also tested linear regressions between the RPEs and ERP amplitudes for all goal state trials on a participant-by-participant basis. A t-test on the regression coefficients was close to significant (Figure 4A, $p = 0.09$, $t(11) = 1.8$, sd = 5.9, mean coefficient = 3.2). The lack of significance is likely due to a lack of power because in the complex environments goal images occur very rarely. Each participant found the goal image 5 times in a given environment, thus there was a total of only 120 encounters of the goal (12 participants, 2 environments) in the complex environments.

3.2. Experiment 2: EEG vs. RPE at the goal state

Since, in experiment 1, the small number of goal encounters indicated only a weak trend ($p= 0.08$) when comparing the FRN and the RPE for the goal image, we used six simpler environments to test more systematically the regression between the FRN amplitudes and the RPEs at the goal images. Again, participants found the goal 5 times in each environment, which added up to a total of 420 encounters of the goal in this second experiment. The simple environment consisted of a one-dimensional string of states, with the goal state at the end. Since backward movements were impossible in the simple environment, the goal state was the

most important state when learning the environment. We wanted to test if the RPEs at the goal state were reflected by the FRN amplitudes similar to what had been observed in N-armed bandit tasks (review, see Walsh & Anderson, 2012). Using the same methods as in experiment 1, we found a time window from 280-390ms in which the ERP amplitudes decreased monotonically (Figure 3B, $F(2,39) = 5.39$, $p=0.008$). This window was very close to the window of 280-360ms we found in experiment 1. The small shift of the windows between experiment 1 and experiment 2, may be due to the lower cognitive load of the simple environments compared to complex ones. Similar shifts have been observed when comparing different N-armed bandit tasks (Hajcak, Holroyd, Moser, & Simons, 2005; Hajcak, Moser, Holroyd, & Simons, 2007; Holroyd, Krigolson, Baker, Lee, & Gibson, 2009; Kreussel et al., 2012; Walsh & Anderson, 2011).

Next, we tested, on a participant-by-participant basis, whether the ERP amplitudes in this time window correlated with the RPEs of the goal state. A t-test on the linear regression coefficients between the RPEs and the ERPs of the goal images was significant (Figure 4C, $p = 0.03$, $t(13) = 2.3$, sd = 2.7, mean coefficients = 1.7), indicating that the RPEs had an effect on the ERPs between 280-390ms, as expected from the analogy to N-armed bandit tasks.

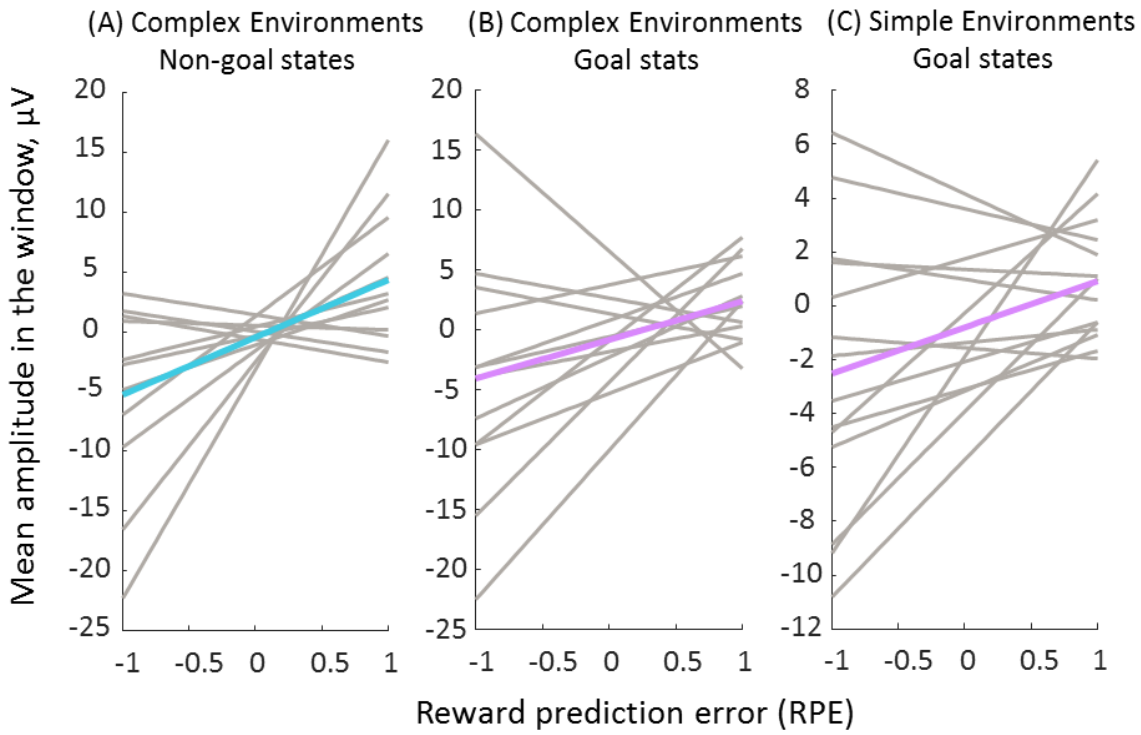


Figure 4. Linear regressions of the RPEs and the mean ERP amplitudes for individual participants (grey lines). Coloured lines present the averaged regression line. (A) Regressions at non-goal states in experiment 1. (B) Regressions at goal states in experiment 1. (C) Regressions at goal states in experiment 2.

4. Discussion

Decision masking is a complex process involving the evaluation of the reward, the RPE, and potentially other values. In model-free reinforcement learning, the RPE is the most important variable. In N-armed bandit tasks, the FRN is positively correlated with the RPE (Holroyd & Coles, 2002b). In this paper we wanted to check whether a similar correlation is also true in more interesting situations where decision making is sequential and reward is not delivered immediately. Classic model-free reinforcement learning models propose that the RPE plays an essential role also at states that are not directly rewarded. Hence, we asked the question whether there is evidence for RPEs in EEG signals at non-rewarded states. To address this question, we used a previously developed sequential decision making paradigm, where a goal is found only after a sequence of actions (Clarke et al., 2015; Tartaglia et al., 2017).

We first used two complex environments which contained trap states and loops to test if the FRN-ERP relationship proposed by Holroyd and Coles (2002b) still holds true. The FRN

amplitude reflected the RPEs of the non-goal states in a time window of 280- 360ms after the state onset. Since the goal state occurred rarely - only 120 times when summed over all participants and epochs in the complex environments, the correlation between FRN amplitudes and RPEs at the goal states was not significant. Thus, we used six simple linear environments with 1-dimensional arrangement of states to test if the FRN amplitudes reflect the RPEs of the goal states. Indeed, in the time window of 280-390ms after the state onset the correlation was significant. Both time windows are very close to the FRN window.

Walsh and Anderson (2011) found that FRN amplitudes changed according to winning conditions in an N-arm bandit task, and that amplitudes scales with RPEs. The larger the RPE, the more positive the amplitude. Eppinger (2009) showed that FRN amplitudes diminished as a positive reward was given more times. We found that a similar trend is present also for non-goal states.

Contrary to most studies in reinforcement learning, we used a deep sequential decision making task, where only one of many states was rewarded. Sambrook et al., (2018) used a two-step task. They found that the RPEs of the intermediate state was also reflected in the EEG around 200-400ms. In an fMRI study, Glaescher, Daw and Dayan (2010) used a 2-step design and found that that the sources of RPE are in the Ventral Striatum, which is line with the proposal by Holroyd and Coles (2002) that the FRN sources of the RPE are in the ACC. In contrast to such a 2-step design, some of our participants spent more than 100 steps in loops of the environment before they saw the first goal image.

There are some caveats. Our results, as all results in the field, are based on correlations, which limit conclusions to some extent. For example, humans may compute RPEs but do not use them for learning. Or RPE may be used as a confidence measure rather than as an action choice variable. Second, we computed RPE with SARSA. We do not however claim that humans use a SARSA like mechanism because many other algorithms, including unknown ones, may deliver similar results. Third, our results show evidence that humans make use of model free RL components. However, this does mean that humans do not use model based learning, which they most likely do. We currently explore model-based exploration in very similar environments.

Taken together, our results suggest that the FRN reflects the RPE (or related measures) in deep, sequential decision making paradigms in both rewarded and non-rewarded states.

Acknowledgements

This research was supported by Swiss National Science Foundation (no. 200020_184615) and by the European Union Horizon 2020 Framework Program under grant agreement no. 785907 (HumanBrain Project, SGA2). PLUS GRANTS OF HERZOG lab.

References

Badgaiyan, R. D., & Posner, M. I. (1998). Mapping the cingulate cortex in response selection and monitoring. *NeuroImage*, *7*(3), 255–260. https://doi.org/10.1006/nimg.1998.0326

Beeler, J. A., Daw, N., Frazier, C. R. M., & Zhuang, X. (2010). Tonic dopamine modulates exploitation of reward learning. *Frontiers in Behavioral Neuroscience*. https://doi.org/10.3389/fnbeh.2010.00170

Bellebaum, C., & Daum, I. (2008). Learning-related changes in reward expectancy are reflected in the feedback-related negativity. *European Journal of Neuroscience*, *27*(7), 1823–1835. https://doi.org/10.1111/j.1460-9568.2008.06138.x

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Carlson, J. M., Foti, D., Mujica-Parodi, L. R., Harmon-Jones, E., & Hajcak, G. (2011). Ventral striatal and medial prefrontal BOLD activation is correlated with reward-related electrocortical activity: A combined ERP and fMRI study. *NeuroImage*, *57*(4), 1608–1616. https://doi.org/10.1016/j.neuroimage.2011.05.037

Clarke, A. M., Friedrich, J., Tartaglia, E. M., Marchesotti, S., Senn, W., & Herzog, M. H. (2015). Human and machine learning in non-Markovian decision making. *PLoS ONE*, *10*(4). https://doi.org/10.1371/journal.pone.0123105

Cohen, M. X, & Ranganath, C. (2007). Reinforcement Learning Signals Predict Future Decisions. *Journal of Neuroscience*, *27*(2), 371–378. https://doi.org/10.1523/JNEUROSCI.4421-06.2007

Cohen, Michael X., Cavanagh, J. F., & Slagter, H. A. (2011). Event-related potential activity in the basal ganglia differentiates rewards from nonrewards: Temporospatial principal components analysis and source localization of the feedback negativity: Commentary. *Human Brain Mapping*. https://doi.org/10.1002/hbm.21358

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*. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*. https://doi.org/10.1038/nn1560

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Doñamayor, N., Marco-Pallarés, J., Heldmann, M., Schoenfeld, M. A., & Münte, T. F. (2011).

Temporal dynamics of reward processing revealed by magnetoencephalography. *Human Brain Mapping*, *32*(12), 2228–2240. https://doi.org/10.1002/hbm.21184

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, *295*(5563), 2279–2282. https://doi.org/10.1126/science.1066893

Gershman, S. J., & Daw, N. D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-122414-033625

Gl??scher, 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. https://doi.org/10.1016/j.neuron.2010.04.016

Gruendler, T. O. J., Ullsperger, M., & Huster, R. J. (2011). Event-related potential correlates of performance-monitoring in a lateralized time-estimation task. *PLoS ONE*, *6*(10). https://doi.org/10.1371/journal.pone.0025591

Hajcak, G., Holroyd, C. B., Moser, J. S., & Simons, R. F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology*, *42*(2), 161–170. https://doi.org/10.1111/j.1469-8986.2005.00278.x

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2007). It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology*, *44*(6), 905–912. https://doi.org/10.1111/j.1469-8986.2007.00567.x

Haruno, M., & Kawato, M. (2006). Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *Journal of Neurophysiology*, *95*, 948–959. https://doi.org/10.1152/jn.00382.2005

Holroyd, C. B., Krigolson, O. E., Baker, R., Lee, S., & Gibson, J. (2009). When is an error not a prediction error? An electrophysiological investigation. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(1), 59–70. https://doi.org/10.3758/CABN.9.1.59

Holroyd, Clay B, & Coles, M. G. H. (2002a). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. https://doi.org/10.1037//0033-295X.109.4.679

Holroyd, Clay B, & Coles, M. G. H. (2002b). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. https://doi.org/10.1037//0033-295X.109.4.679

Kreussel, L., Hewig, J., Kretschmer, N., Hecht, H., Coles, M. G. H., & Miltner, W. H. R. (2012). The influence of the magnitude, probability, and valence of potential wins and losses on the amplitude of the feedback negativity. *Psychophysiology*. https://doi.org/10.1111/j.1469-

8986.2011.01291.x

Lehmann, M., Xu, H., Liakoni, V., Herzog, M., Gerstner, W., & Preuschoff, K. (2017). Evidence for eligibility traces in human learning. *BioRxiv*. https://doi.org/10.1038/s41467-018-06213-1

Martin, L. E., Potts, G. F., Burton, P. C., & Montague, P. R. (2009). Electrophysiological and hemodynamic responses to reward prediction violation. *NeuroReport*, *20*(13), 1140–1143. https://doi.org/10.1097/WNR.0b013e32832f0dca

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339–346. https://doi.org/10.1016/S0896-6273(03)00154-5

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-Related Brain Potentials Following Incorrect Feedback in a Time-Estimation Task: Evidence for a "Generic" Neural System for Error Detection. *Journal of Cognitive Neuroscience*, *9*(6), 788–798. https://doi.org/10.1162/jocn.1997.9.6.788

Montague, P. R. (1996). A Framework for Mesencephalic Predictive Hebbian Learning, *76*(5), 1936–1947.

Nieuwenhuis, S., Slagter, H. A., von Geusau, N. J. A., Heslenfeld, D. J., & Holroyd, C. B. (2005). Knowing good from bad: differential activation of human cortical areas by positive and negative outcomes. *European Journal of Neuroscience*, *21*(11), 3161–3168. https://doi.org/10.1111/j.1460-9568.2005.04152.x

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.5498-10.2012

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, Reward, and Decision Making. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-010416-044216

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337. https://doi.org/10.1016/S0896-6273(03)00169-7

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*. https://doi.org/10.1038/nature05051

Pinheiro, J. C., & Bates, D. M. (2000). Linear Mixed-Effects Models: Basic Concepts and Examples. In *Mixed-Effects Models in S and S-PLUS* (pp. 3–56). https://doi.org/10.1007/0-387-22747-4_1

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Sutton, R. S., & Barto, A. G. (1998). Introduction to Reinforcement Learning. *Learning*, *4*(1996), 1–5. https://doi.org/10.1.1.32.7692

Sutton, R. S., Barto, A. G., & Book, A. B. (1998). Reinforcement Learning : An Introduction. *Learning*.

Tartaglia, E. M., Clarke, A. M., & Herzog, M. H. (2017). What to choose next? A paradigm for testing human sequential decision making. *Frontiers in Psychology*, *8*(MAR), 1–11. https://doi.org/10.3389/fpsyg.2017.00312

Tucker, D. M., Luu, P., Frishkoff, G., Quiring, J., & Poulsen, C. (2003). Frontolimbic Response to Negative Feedback in Clinical Depression. *Journal of Abnormal Psychology*, *112*(4), 667–678. https://doi.org/10.1037/0021-843X.112.4.667

Walsh, M. M., & Anderson, J. R. (2011). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences*, *108*(47), 19048–19053. https://doi.org/10.1073/pnas.1117189108

Walsh, Matthew M., & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2012.05.008

Supplementary Materials

1. Linear-mixed Model Details

Since each participant made the same actions multiple times, these responses are not independent from each other. Thus we used a linear mixed model (Pinheiro & Bates, 2000) to account for repeated measures. To fit the linear mixed model, we used the amplitudes of the ERPs as the response variable and the RPEs as the predictor variable for the fixed effect. In order to ascertain that the relationship between the RPEs and the ERP amplitudes are not caused by possible individual differences, we added a random effect for "participants" to characterize variations due to individual differences and another random effect for "participants * environments" to account for possible interactions between the factors "participants" and "environment". We tested 3 linear mixed models: (1) a model without random effects; (2) a model with "participants" as a random effect; (3) a model with "participants * environment" as a random effect. We used the log-likelihood ratio test (Pinheiro & Bates, 2000) to tell whether a model is significantly better than another one.

The analysis of the behavioural data showed that adding two random effects, 'participant' and 'environment', did not significantly improve the model fit (Table 1).

Experiment 1						
Fixed effect	Random effect		Log likelihood	LRT	DF	p-value
RPEs			-6484.3		2220	
RPEs	Participants		-6442.9	82.8	2220	1
RPEs	Participants * blocks		-6483.3	2	2220	1
Experiment 2						
Fixed effect	Random effect		Log likelihood	LRT	DF	p-value
RPEs			-9237.8		3586	
RPEs	Participants		-9170.6	134.4	3586	1
RPEs	Participants * environments		-9236.5	2.6	3586	1

LRT = Log likelihood ratio

DF = Degree of freedom
Participants*blocks: potential interactions between the factor "participants" and the factor "blocks"

2. SARSA(λ) Model Detail

The learning signal in SARSA(λ) is the reward prediction error (RPE), defined as the difference between the actual reward and the predicted reward (Equation 1).

Equation 1

$$RPE_t = r + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$$

where γ is a discount parameter to determine the value of future rewards. The actual reward is denoted by *r*. The predicted reward is the difference between the action value $Q(s_t, a_t)$ in state s_t and the discounted action value $\gamma Q(s_{t+1}, a_{t+1})$ in state s_{t+1}. Positive RPE indicate that the tendency of selecting action a at state s should be strengthened, negative RPE indicate that the tendency should be weakened.

The Q-values, Q(s, a), represent an estimate of the expected future reward when starting in state s, taking action a. This value function is iteratively improved for all state-action pairs by applying an update after each step:

Equation 2

$$Q_{t+1}(s, a) = Q_t(s, a) + \alpha \times RPE_t \times e_t(s, a)$$

The quantity $e_t(s, a)$ is known as an eligibility trace (Sutton & Barto, 1998) and implements a decaying short-term memory trace of past state-action pairs with the following dynamics:

Equation 3

$$e_t(s, a) = \begin{cases} \gamma\lambda e_{t-1}(s, a) & if\ (s, a)\ not\ visited at\ time\ t \\ 1 & if\ (s, a)\ visited\ at\ time\ t \end{cases}$$

The value $e_t(s, a)$ marks an event in memory eligible for undergoing learning. At each trial, the eligibility traces for all state-action pairs decay by $\gamma\lambda$, where λ is the trace decay parameter.

The Q values are then used to select an action at each state according to a softmax policy:

Equation 4

$$P(s, a) = \frac{\exp(Q_t(s, a)/\tau)}{\sum_i \exp(Q_t(s, i)/\tau)}$$

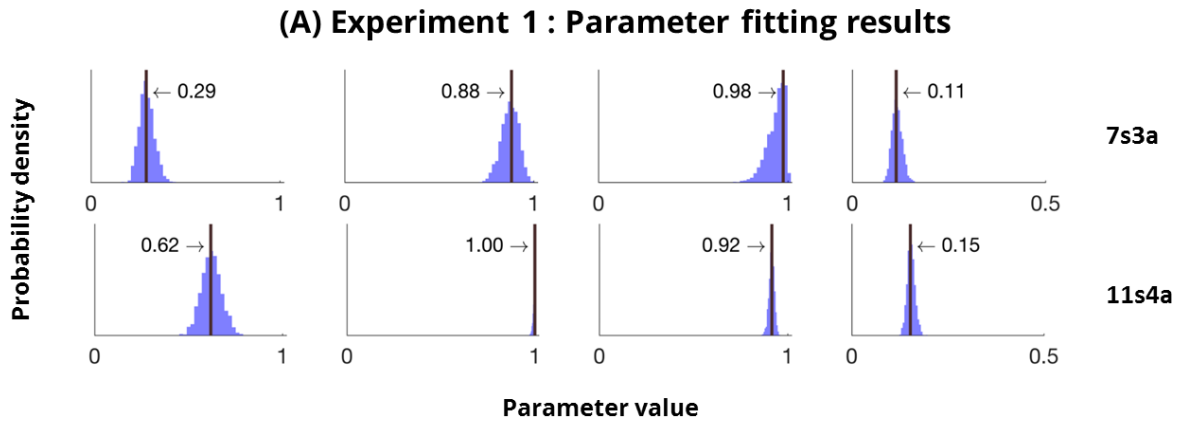
where $P(s, a)$ defines the probability of choosing action a at state s, τ is the temperature parameter which controls the tendency of exploration and exploitation, i presents all possible actions at state s.

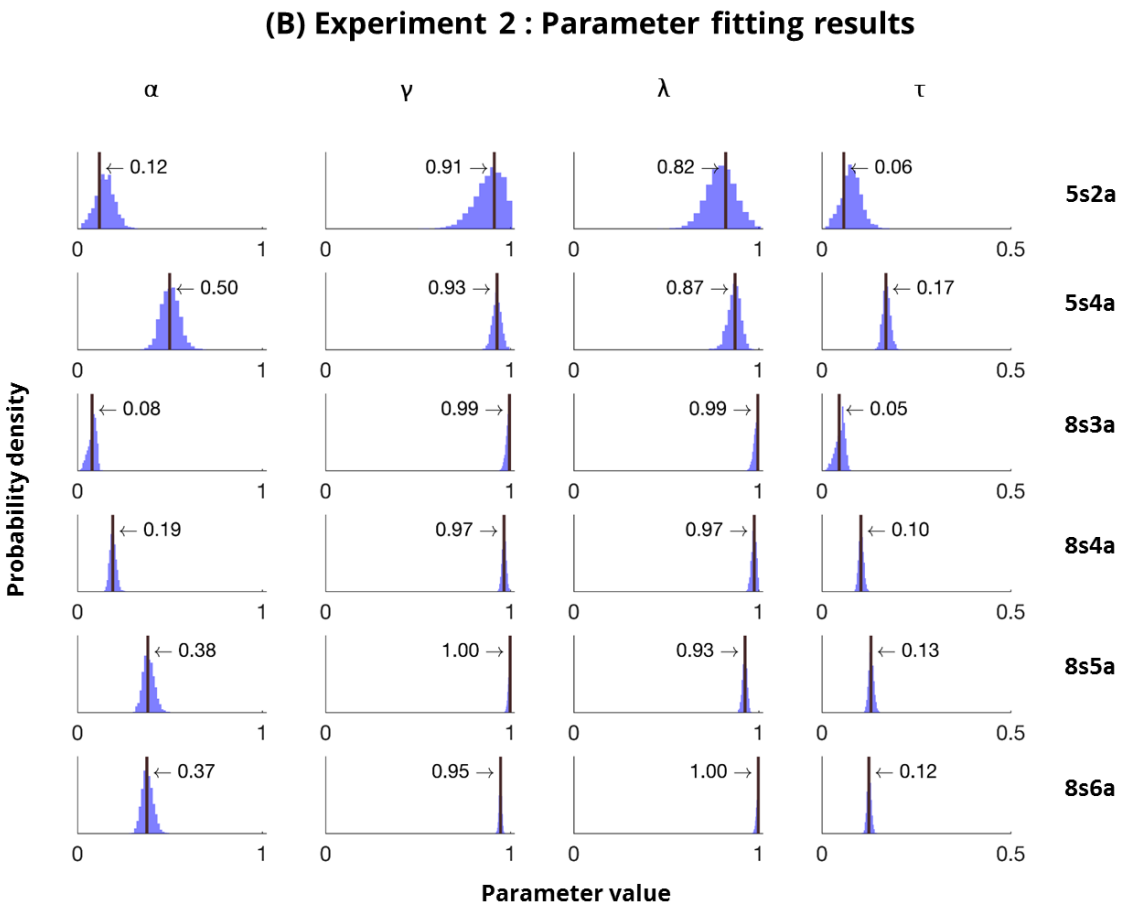
The above equations define a learning model with four free parameters: the learning rate α, the discount rate γ, the eligibility decay rate λ and the temperature τ.

Different choices of parameters yield different action selection probabilities (Equation 4). We are interested in those values that are in best agreement (in a maximum likelihood sense) with the behavioural data. Specifically, following Lehmann et al. (2019) we fit the free parameters using the Metropolis-Hastings Markov Chain Monte Carlo (MCMC). This method has the advantage of giving us not just the most likely values but actually a distribution over the parameters (Supplementary Figure 1).

Q-values of all non-goal states were initialized to 0 at the beginning of each block.

To predict the PRE in a given block of experiments, we used the mean values resulting from fitting the behavioural parameters.





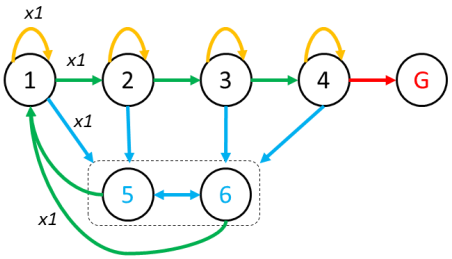
Supplementary Figure 1 Parameter fitting results for both experiments. . Each row presents the fitting results for one environment. Each column presents the estimated distribution for each parameter in the SARSA(λ) model estimated by the MCMC method. The vertical lines are the mean values that were used for the calculation of the RPE. **(A)** Parameter fitting for experiment 1. **(B)** Parameter fitting for experiment 2.

SARSA computes the Q-value as the value of the actual action selected in the next state, while Q-learning computes the Q-value as the value of the optimal action planned in the next state. To test if the two algorithms give similar estimation, we computed the correlation between the RPEs estimated by SARSA and Q-learning. The correlations is significant for the non-goal states in the complex environment ($r = 0.62$, $p < 0.001$), and is also significant for the goal states in the simple environment ($r = 0.65$, $p < 0.001$). The result confirms that the two algorithms give similar estimation on the RPEs.

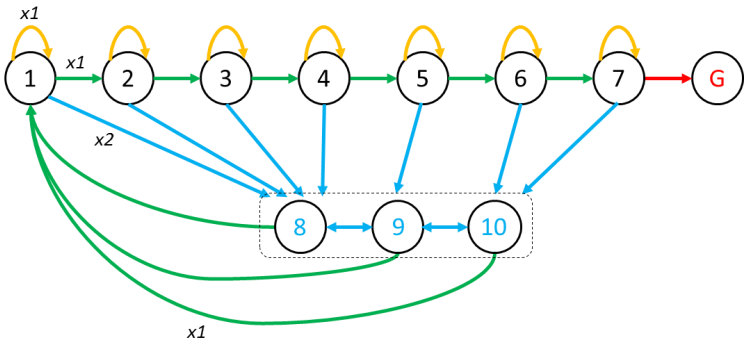
3. Environment structure used in Experiment 1 and 2

Complex environments

7 states – 3 actions

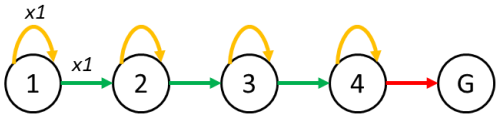


11 states – 4 actions

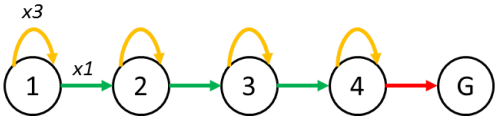


Simple environments

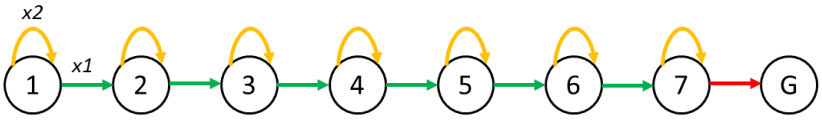
5 states – 2 actions



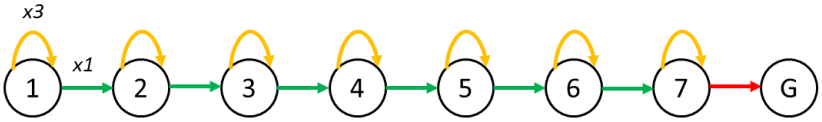
5 states – 4 actions



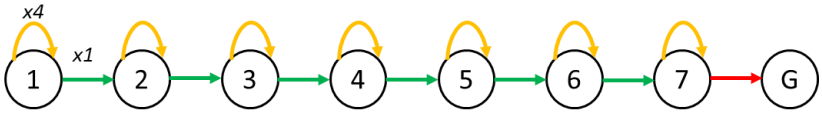
8 states – 3 actions



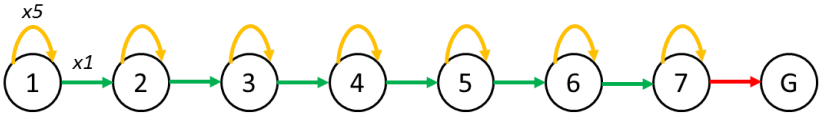
8 states – 4 actions



8 states – 5 actions



8 states – 6 actions



Model-Building by Exploration:

Surprise and Novelty in Reward-based Learning

He A. Xu[(1,*)], Marco P. Lehmann[(1,2,*)], Alireza Modirshanechi[(1,2,*)], Wulfram Gerstner[(1,2)], and Michael H. Herzog[(1)]

(1) Brain-Mind Institute, School of Life Sciences

(2) School of Computer and Communication Sciences

Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne EPFL

* These authors contributed equally

October 2019

1 Introduction

Humans make efforts to receive rewards such as money or praise of parents and peers. But even in the absence of any rewards, children and adults may want to explore a novel toy or a novel city. These exploratory actions driven by novelty seeking behavior are useful to 'understand' the environment and have been interpreted in the theory of reinforcement learning as steps towards building a model of the world [1]. World models in reinforcement learning summarize implicit knowledge such as 'if I open the door to the kitchen, I expect to see a fridge'. However, since the world is much more complex than any model of it, there will occasionally be a mismatch between the expectation arising from the model and the actual observation, e.g., the location of the fridge is empty because it needs repairing. Such mismatches generate the feeling of surprise - and are the basis of jokes [2]. In this study we ask whether moments of reward, novelty, and surprise are correlated with the event-related potential (ERP) in the EEG. And if so, whether all three have the same EEG signature or not.

In reward-based experimental paradigms, the theory of reinforcement learning successfully predicts behavior as well as brain signals [3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. In reward-free situations, exploratory human behavior can be explained by the natural desire to seek novel events [13, 14, 15, 16, 17, 18, 19]. Novelty plays the role of a motivational signal [20, 21] and acts as an intrinsic reward for reinforcement learning [22, 23, 24].

Surprise is fundamentally different from novelty: for example, an image that we see for the first time can be novel but not surprising whereas an image that we see for the tenth time can be surprising but not novel. Surprise signals manifesting themselves physiologically in pupil dilation and the EEG [25, 26, 27, 28, 29, 30] are triggered by the violation of an expectation generated by the brain's model of the world. Whereas a reward prediction error is a mismatch between expected reward and actual reward, surprise is a mismatch between an expected observation and actual observation – even in a reward-free environment.

Behavioral experiments [31, 32, 33] and theories [34, 35] suggest that surprise helps humans to adapt quickly to changes in the environment. Similar to the reward prediction error, surprise is believed to modulate synaptic plasticity [36, 37, 38], potentially through the release of specific neurotransmitters such as acetylcholine and norepinephrine.

While there is some agreement that novelty and surprise are two separate notions [2, 39], it is debated how these two notions can be formally defined and distinguished from each other theoretically or in behavioral tasks [20, 15, 39]. Here, we adapt a sequential decision making paradigm [40] so as to separate contributions of surprise, novelty, and reward to human behavioral choices. We use a model-based reinforcement learning approach in a novel computational model which uses surprise to modulate learning of the model of the environment, and novelty for exploration in the absence of external rewards. We show that the three different notions (i.e. surprise, novelty, and reward) manifest themselves on different time-scales in the ERP.

2 Results

In order to distinguish between signatures of novelty, surprise, and reward, we designed a behavioral experiment in an artificial environment consisting of 11 states and 4 possible actions at each state (Fig. 3). States were presented by images on a computer screen and actions were presented by four grey disks below the image. Before the experiment, participants were told that they were supposed to find the shortest path to a given goal image. In each state, participants chose an action (by clicking on one of the grey disks) which brought them to the next image and then chose the next action. The episode ended when the goal image was found. Unknown to the participants, the states can be classified into three types: progressing states (states 1-7 in Fig. 3), trap states (state 8-10 in Fig. 3), and a goal state (red G in Fig. 3). Progressing states are "good" states that potentially bring participants closer to the goal whereas trap states are "bad" states ("go back to start") that are off the direct path to the goal. At each state, one of the 4 possible actions (green arrow in Fig. 3) brought them to the next progressing state, two actions (blue arrow in Fig. 3) brought them to one of the trap states, and one action (yellow arrow in Fig. 3) made them stay at the current state.

The experiment was organized in two blocks of 5 episodes each. During the 1st episode of the 1st block, participants explored the environment until they found the goal. They then continued for another 4 episodes in the same environment. Thereafter two states (state 3 and 7 in Fig. 3) were swapped, without announcing it to participants, and participants continued for another 5 episodes with the novel layout of the environment (2nd block, Fig. 3). EEG was recorded during the entire experiment, but we only analysed the period shown in Fig. 4 (green interval).

2.1 Computational algorithm and Behavioral Analysis

To navigate in such a complex environment, we assume that subjects build an internal estimation ('world-model') of the lay-out of the environment, i.e. the probabilities of transitions from a given state to another state when performing a given action. In our algorithm, action selection combines aspects of novelty-seeking, so as to explore the environment, with model-based reinforcement learning, so as to exploit known good actions.

The novelty of a state is subject-specific and decreases, in our algorithm, with the number of times the participant has encountered this state in the recent past; see Supplementary Materials. At the beginning of the first epoch in block 1, all states have identical novelty. Because participants often fall into one of the trap states, their novelty decreases rapidly (Fig. 1B.) Because participants rarely visit a state close to the goal during the first episode, the novelty of those states increases over time (Fig. 1B) so that, before the end of the first episode, the novelty is highest for states in the proximity of the goal (Fig. 1C). This observation suggests that seeking novel states will effectively take a subject closer to the goal - even *before* the subject knows where the goal is located.

In our algorithm, learning the world-model, i.e., the lay-out of the environment, is controlled by surprise. Surprise is subject-specific [34, 35] and measures how "unexpected" the next image (state s_{t+1}) is given that the subject chooses action a in the previous state s_t. Whether an event is surprising or not depends on the belief of the subject (his current world-model) which summarizes the knowledge extracted from his previous experiences in this environment; see Supplementary Materials. Our surprise measure indicates that swapping states 3 and 7 before the start of the second block leads to highly surprising events when participants encounter state 3 or 7 in the first

epoch of the second block or when they transit from state 3 to state 7 to another state (new Fig2).

We employed surprise-based learning for building the world-model and our novelty-seeking strategy for exploration in a reward-based learning. Since our approach combines 'Surprise', 'Novelty', and 'Reward' we refer to it as SurNoR-learning. We wanted to check whether the SurNoR-algorithm (Supplementary Materials, Algo1) is capable of explaining human behavioral choices in our experiment. We therefore fitted the parameters of the algorithm to the behavioral data of all 12 subjects using an empirical Bayesian approach with 3-fold cross-validation, see Supplementary Materials. We found that the SurNoR-algorithm predicted the correct action in the first episode of the first block with an accuracy of 42 $\pm$4% across the more than 1500 action choices made by the 12 participants. This fraction is significantly higher than a model based on random action choices (that would predict 25% because there are four different actions) or a model-free reinforcement learning model without a novelty preference (i.e., with all state-values initialized at zero).

Similarly, in the first episode of block 2 when participants were lost because of the swapping between states 3 and 7, the SurNoR-algorithm with novelty seeking was predicting $48 \pm 10\%$ of the actions of the 12 participants, again a value significantly above chance. In the remaining episodes 2-5 of the two blocks, the SurNoR-algorithm predicted $87 \pm 16\%$ of the action choices. Most of these actions moved participants closer to the goal.

In the SurNoR-algorithm, unsurprising events do not lead to a change of the world-model, whereas surprising events induce large improvements in the world-model, and hence in the action preference To quantify the importance of surprise for adaptation of the world model, we compared the SurNoR-algorithm with two alternative approaches. The first approach ('perfect integrator') uses optimal Bayesian integration under the assumption of stationary statistics. The second approach ('leaky integrator') is a heuristic modification of the perfect integrator so as to allow for changes in the environment. A Bayesian model selection approach [41] (see Supplementary Materials) indicates that the SurNoR-algorithm outperforms the alternatives (i.e. leaky and perfect integrator) with an Exceedance Probability = 0.96. See Supplementary Materials for more details. Hence, the results of the statistical model selection show that the notions of both novelty and surprise are necessary to explain human behavior in our reward-based learning task.

2.2 N1 is a potential bio-marker for Novelty

At the beginning of the 1st episode of the 1st block, participants do not have any model of the external world. They made actions randomly when seeing a non-goal state for the first time, hence they ended up in a trap state with a probability equal to 0.5. Every time when they came out of the trap state, they started from state 1 and continued again to look for the goal state. During the first episode, the trap states are the most frequently visited states and the states close to the goal states are least frequently visited.

To search for the EEG time window where Novelty is reflected, we averaged the ERPs of the states with high novelty values (state 5, 6, 7,HIGH-NOVELTY condition) and the ones with low novelty values (state 8, 9, 10, LOW-NOVELTY condition) in the 1st episode of the 1st block. Fig. 5 shows the ERPs are significantly different in the two conditions in a time interval from 80 to 110ms after the state onset ($p = 0.01$, $t(16) = -2.13$, $sd = 1.28$). The average ERPs in two conditions removed physiological and instrumental noise and improved the signal-to-noise ratio. However, it also removed the participant-specific information. We used a sliding window method to search for potential time course of Novelty on a trial-by-trial and participant-by-participant basis

124

(Details in Method). Linear regression on a on a trial-by-trial and participant-by-participant basis between the mean amplitudes in the interval of 80-130ms after the state onset and the estimated Novelty from the computational model (see subsection 2.1) is significant ($p = 0.02$, $t(10) = 2.68$, $sd = 0.46$, mean slope $= 0.37$). In summary, our results demonstrate that N1 component of EEG is a potential bio-marker for novelty, tested both by using on-averaged ERPs comparison and trial-by-trial analysis.

2.3 Biomarker for Surprise

According to our computational model fitted to the behavioral data, subjects learned the transitions in almost one shot, e.g. see Fig. 1.C. Therefore, we can partition transitions to 3 groups based on their effects on learning: (1) The ones that are experienced for the 1st time (mostly in the 1st episode of the 1st block), (2) The ones that are already experienced once and have not been change since then (i.e. the learned ones), and (3) The ones corresponding to the transitions from or to the swapped states (mostly in the 2nd episode of the 2nd block). The 2nd group are considered as un-surprising transitions, while the 1st and the 3rd groups contain the surprising transitions - mild surprise for 1st group and huge surprise for the 3rd one. Therefore, we can group whole trials to two sets of surprising and un-surprising. Therefore, we can group all the trials in the first episode of both blocks into two conditions: SURPRISING and UN-SURPRISING conditions. By comparing the ERPs averaged in each group, we found that the time interval from 150 to 300ms after the state onset was a potential interval that reflects the magnitude of surprise.

We then extracted the mean amplitudes in the interval of interest and regressed the amplitudes with estimated surprise (computed by our computational model) on a trial-by-trial and participant-by-participant basis for all trials in the 1st episodes of both blocks. The regression between mean amplitudes in the time interval 150-300ms and estimated surprise was not significant ($p = 0.86$, $t(18) = -0.17$, $sd = 2.61$). However, to see whether the bio-marker of surprise could be hidden inside this interval, we used a sliding window of 50ms (10ms per step) to test the regression between the mean amplitudes in the sliding window and the surprise computed by our model. The ensemble window was determined using the earliest time point of the first sliding window, whose mean amplitudes correlated significantly with surprise, and the latest time point of the last sliding window, whose mean amplitudes correlated significantly with surprise. We found that the regression is significant between surprise and the mean amplitude in the interval of 150-210ms after the state onset ($p = 0.003$, $t(10) = -3.83$, $sd = 0.21$, mean slope $= -0.25$). The results indicate that the interval from 150 to 210ms after the state onset is a potential bio-marker for surprise.

However, to see whether the bio-marker of surprise could be hidden inside this interval, we used the sliding window to search for the potential time course of Surprise on a trial-by-trial and participant-by-participant basis (Details in Method). We found that the regression is significant between surprise and the mean amplitude in the interval of 150-210ms after the state onset ($p = 0.003$, $t(10) = -3.83$, $sd = 0.21$, mean slope $= -0.25$). The results indicate that the interval from 150 to 210ms after the state onset is a potential bio-marker for surprise.

2.4 Reward presented in late P3 component

To investigate where in the EEG an indicator of the external reward is presented, we compared the averaged ERPs of the goal states and a non-goal state (state 1) in both blocks. We selected state 1 as the control state because the transitions to and from state 1 were not disrupted in either

blocks. The ERP comparison (Fig. 5) shows that in the interval between 400 and 450ms after the state onset the curves differed significantly ($p = 0.01$, $t(10) = -2.9$, $sd = 2.14$).

3 Discussion

When the external reward is sparse and delayed, surprise and novelty can be considered as internal feedback for learning. In this study we built a learning model (SurNoR) to solve the learning situation where surprise, novelty, reward are all involved in learning . The SurNoR model outperformed the other models in explaining participants behaviours. One of the important factor that makes SurNoR different from other models is that SurNoR considers novelty as an intrinsic reward. Previous studies [42, 43, 44, 45] have shown that the dopamine level affects both novelty and reward processing. When a state of high reward appears, the dopamine level increases. Similarly, when a state of high novelty appears, the dopamine level increases. These studies provide physiological evidence for the SurNoR model.

Furthermore, we found that the novelty signal is reflected in EEG recording around 80-130ms after the state onset, and that the surprise signal is reflected around 150-210ms after the state onset. In the SurNoR model, the novelty of a state is defined as the global probability of seeing that state, and the surprise of a state-action transition is defined as the changes in the local transition probability (details see Appendix). In corresponding to the EEG result, the global signal (novelty) occurs earlier than the local signal (surprise). This finding is in line with previous findings. In [29], although the study used MEG recording in an oddball task, the authors showed that the brain activity around 60-130ms is sensitive to global changes in the stimuli. In [28], the authors built a Bayesian inference model to explain physiological data from previous researchers. Meyniel's results showed that the P300 component in EEG is a bio-marker that reflects local transition probability changes. Here in our study, we confirmed the previous findings in Maheu's and Meyniel's work, and generalise the conclusions to a truly sequential decision making task.

4 Methods

4.1 Computational Model and Behavioral Data Analysis

See supplementary materials for the details of the computational model and corresponding pseudo codes as well fitting procedures.

4.2 EEG Analysis

4.2.1 Experiment set up

Experiments were conducted on a Phillips 201B4 monitor, running at a screen resolution of 1980×1080 pixels and a refresh rate of 100 Hz, using a 2.8 GHz Intel Pentium 4 processor workstation running Windows 7. Experiments were scripted in Matlab R 7.11 using custom software and extensions from the Psychophysics Toolbox for Windows XP ([46]).

4.2.2 Participants

14 paid participants joined the experiment. Two participants quit the experiment, hence, we analysed data for 12 participants (5 females, aged 20-26 years, mean = 22.8, sd = 1.7). All participants were right-handed and naïve to the purpose of the experiment. All participants had normal or corrected-to-normal visual acuity. All participants provided written consent. The experiment was approved by the local ethics committee.

4.2.3 Stimuli and general procedure

Before starting the experiment, we showed the participants the goal image that they were required to find. Next, participants were presented, in random order, the other images that they may encounter during the experiment. After seeing the images presented on the screen, participants clicked the 'start' button to start the experiment proper. At each trial, participants were presented an image (state) and four grey disks below the image (Fig. 4). Clicking on one of the disks (action) led participants to a subsequent image. Participants clicked through the environment until they found the goal state. An episode was finished when participants found the goal state and thereafter the next episode started.

4.2.4 EEG recording and processing

EEG signals were recorded using BioSemi equipment with 128 electrodes at a 2048Hz sampling rate. Recorded data were band pass filtered from 0.1Hz to 40Hz and down sampled to 256Hz. Common average referencing was applied for re-referencing. "Bad" channels were visually inspected and interpolated using the EEGLAB toolbox ([47]). Eye movements and electromyography (EMG) artefacts were removed by using independent component analysis (ICA). Trials in which the change in voltage at any channel exceeded 35 μV per sampling point were discarded. For each trial, a time window was extracted from 200ms before to 700ms after the image onset. The baseline activity was remove by subtracting the mean calculated over the interval from 200ms to 0ms before the image onset. Prefrontal Event-Related Potentials (ERPs) were computed by averaging the EEG data of selected prefrontal electrodes (Fz, F1, F2, AFz, FCz) for Event-Related Potential (ERP) analysis. Data was analyzed during the time window from 0 to 700ms after image onset (green interval in Figure 1A).

128

4.2.5 EEG time course search for novelty and surprise

- Group analysis
  To find the time course where the novelty signal is reflected in EEG, we compared the averaged ERPs of least frequently visited states (states with high novelty) and most frequently visited states (states with low novelty) in the first episode of the first block. The least frequently visited states in the first episode of the first block were states 5,6,7 (HIGH-NOVELTY CONDITION). The most frequently visited states in the first episode of the first block were states 8,9,10 (LOW-NOVELTY CONDITION). We averaged the ERPs of the selected states in each condition for each participant. Fig. 5 showed the averaged the ERPs in each condition over all participants. To search for the time course where the two conditions were significantly different, we applied t-test on each time point of the ERPs from 0 to 700ms after the state onset. Each time point on the ERP presents $1/256 = 3.9ms$ and there were in total 179 points compared.
  The same procedure were applied for the surprise vs. non-surprise conditioned ERPs, and for the reward vs. non-reward conditioned ERPs.

- Participant-based analysis
  In the group analysis, ERPs of all participants were averaged together to find the novelty-related time course. The average across participants removed physiological and instrumental noise and improved the signal-to-noise ratio. However, it also removed the participant-specific information. We wanted to test if the ERP amplitudes in this time course can reflect novelty in a participant-by-participant basis. To do this, we used a model-based EEG analysis to identify the bio-markers of novelty and surprise signals. The term 'model-based' here is different from 'model-based models' in reinforcement learning. The model-based EEG analysis uses the signals predicted by a known model (novelty and surprise signals from the SUNOR model in our study), and searches for the EEG components that can reflect those signals.

  To find the EEG time course that can reflect novelty signal in a participant-by-participant basis, we analysed the EEG amplitudes in an interval from 50ms to 150ms after the state onset. We chose this interval because it covered the time course where the high-novelty condition ERP differed significantly from the low-novelty condition ERP. Inside this interval, we used a sliding time window with the width 50ms from the leftmost time point of the interval, and moved 10ms at each step until the window reached the rightmost time point of the interval (detailed window configuration, see Supplementary Materials). There were in total 6 windows of 50ms tested. At each 50ms window, we averaged the mean amplitude of the ERP at each trial for each participant. Then we correlated the mean amplitudes with the estimated novelty signals over all trials in the first episode of the first block for each participant.

  After testing all the sliding windows within the selected interval, we combined the small windows into a big window if the correlations were significant in two consecutive windows. The combined big window started from the leftmost time point of the earlier window and ended at the rightmost point of the later window. Then we used the mean amplitude in the combined big window to correlate with the estimated novelty signals in the first episode of the first block for each participant.

The same procedure was applied for the surprise signal analysis. The selected time interval was from 150ms to 300ms after the state onset, and 10 sliding windows with 50ms width and 10ms step were applied for the analysis.

130

5 Supplementary Materials

5.1 Surprise-Novelty-Reward (SurNoR) algorithm

The SurNoR algorithm combines surprise signals with novelty and reward so as to explore and learn the environment, and exploit rewards. A simple block diagram of the algorithm is shown in **??**, and its pseudocode is shown in Algorithm 1. As it is shown in **??**, SurNoR algorithm has two branches of model-based and model-free which are interacting with each other. Given an agent's perception of novelty as internal reward and its estimation of external reward in the environment, the output of each branch is a value corresponding to a pair of state and action. Then, actions are made by a policy using a convex combination of these two values, so called hybrid policy - see [4, 5] for similar approaches. In this section, we describe our algorithm SurNoR with details, explain how each branch computes the value of pairs of states and actions, and how the policy is shaped as a result.

Formalization of the environment. The state and the action at time t are random variables S_t and A_t which take values in the finite sets $\mathcal{S}$ and $\mathcal{A}$, respectively. In the particular case of our experiment, we have $\mathcal{S} = \{1, ..., 11\}$ and $\mathcal{A} = \{1, ..., 4\}$. From a Bayesian perspective, we consider the transition probability matrix as another random variable Θ, i.e.

$$\mathbf{P}(S_{t+1} = s' | S_t = s, A_t = a, \Theta = \theta) = \theta_{s,a}(s'). \tag{1}$$

Since our environment is deterministic, except for the switch of two states before the start of the second block, the transition probabilities are

$$\theta_{s,a}(s') = \delta(s', T(s, a)), \tag{2}$$

where $T(s, a)$ denotes the target state of the transition from state s given action a, and the Kronecker δ is defined as $\delta(x, x') = 1$ if $x = x'$ and zero otherwise. The target state depends on the block number. Note that $T(s, a)$ is unknown to the participants and to SurNoR as well.

Definition of novelty. While the participant moves in the environment, the count $C_s^{(t)} = |\{t' : 1 \leq t' \leq t \text{ and } s_{t'} = s\}|$ indicates how often state s has been encountered up to time t. We assume that at each time t participants are able to estimate the empirical frequency $p_N^{(t)}(s)$ of encountering state $s \in \mathcal{S}$, formally defined as

$$p_N^{(t)}(s) = \frac{C_s^{(t)} + 1}{t + |\mathcal{S}|}, \tag{3}$$

where $|\mathcal{S}|$ is number of states (i.e. 11 for our experiment). The empirical frequency in Eq. 3 is equal to the expected probability of observing state s given $s_{1:t}$ under the assumption of a uniform prior over states.

The novelty of the state s at time t is defined as the negative logarithm of the empirical frequency

$$N^{(t)}(s) = -\log(p_N^{(t)}(s)). \tag{4}$$

In our algorithm, novelty acts as an exploration bonus (see subsection 'Formalizing model-based Q-values'). The main difference between previously proposed measures of exploration bonus [22, 23, 24, 48, 49] and our approach is that we define our bonus on to states rather than pairs of states and actions which is more consistent with the beahvior of subjects in our experiment.

131

5.1.1 SurNoR model-based branch

World Model. Participants know that there are 11 states and four possible actions in each state, but are not aware of the actual transition probability matrix. In particular, they do not know whether the environment is deterministic or stochastic. A subject's model of the world is therefore summarized as an approximation q of the posterior distribution of the transition probability matrix,

$$q^{(t)}(\theta) \approx \mathbf{P}(\Theta = \theta | S_{1:t} = s_{1:t}, A_{1:t-1} = a_{1:t-1}). \tag{5}$$

In the following, we call q the belief of the subject. We assume that a participant estimates the transition probabilities by a weighted average

$$\hat{\theta}^{(t)} = \mathbb{E}_{q^{(t)}}[\Theta], \tag{6}$$

where the weighting factor is given by the belief $q^{(t)}$.

Since participants do not know the generative model of the environment, exact Bayesian inference is not possible. Rather than making explicit assumptions about the generative model as a starting point for exact Bayesian inference, we work with a general distribution $q^{(t)}$ which is updated by an appropriate learning algorithm after each observation.

Beliefs as Dirichlet distributions. We assume that the transition probabilities from different pairs of states and actions are independent of each other, i.e.

$$q^{(t)}(\theta) = \prod_{s \in \mathcal{S}, a \in \mathcal{A}} q^{(t)}(\theta_{s,a}), \tag{7}$$

where $\theta_{s,a}$ is defined as in Eq. 1. As a natural[1] choice for a probability distribution over transition probabilities, we consider the belief $q^{(t)}(\theta_{s,a})$ to be a Dirichlet distribution with parameter $\alpha^{(t)}_{s,a}$ as

$$q^{(t)}(\theta_{s,a}) = \mathrm{Dir}(\theta_{s,a}; \alpha^{(t)}_{s,a}). \tag{8}$$

As a result, at each time t, the belief of subjects about their environment can be summarized in the set $\alpha^{(t)} = \{\alpha^{(t)}_{s,a}, \forall (s,a) \in \mathcal{S} \times \mathcal{A}\}$. We consider the parameter of the prior belief $q^{(1)}$ (i.e. $\alpha^{(1)}$) to be the same for all transitions as

$$\alpha^{(1)} = \{\alpha^{(1)}_{s,a}(s') = \epsilon, \quad \forall (s, s', a) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}\} \tag{9}$$

where $\epsilon > 0$ is a free-parameter of the model. With this choice of prior, $\hat{\theta}^{(1)}_{s,a}$ (i.e. estimation of the transition probabilities from the pair of state s and action a) is a uniform distribution over states. Furthermore, the free parameter ϵ expresses how deterministic the transitions are from the point of view of a participant, i.e. smaller values of ϵ indicate a more deterministic interpretation of the environment.

Using Dirichlet distribution for the belief $q^{(t)}$ and Eq. 6, a subject's estimation of the transition probabilities is found by

$$\hat{\theta}^{(t)}_{s,a}(s') = \frac{\alpha^{(t)}_{s,a}(s')}{\sum_{\tilde{s}' \in \mathcal{S}} \alpha^{(t)}_{s,a}(\tilde{s}')}. \tag{10}$$

[1] If transition probabilities are stationary and have a uniform prior, exact Bayesian inference yields a Dirichlet distribution.

132

Definition of surprise. We work with the "Generative Model" surprise $\mathbf{S}_{GM}$ [35]. Consider the transition $(S_t = s, A_t = a) \to (S_{t+1} = s')$. The Generative Model surprise corresponding to this transition is [35]

$$\mathbf{S}_{GM}^{(t+1)} = \frac{\hat{\theta}_{s,a}^{(1)}(s')}{\hat{\theta}_{s,a}^{(t)}(s')}. \tag{11}$$

Due to the particular form of the prior $q^{(1)}$ that we chose, $\hat{\theta}_{s,a}^{(1)}(s')$ is constant. As a result, $\mathbf{S}_{GM}^{(t+1)}$ is proportional to the inverse of the probability of the mentioned transition $\hat{\theta}_{s,a}^{(t)}(s')$. Note that in the particular case of our work, the Shannon surprise [50] is just the shifted logarithm of the "Generative Model" surprise, and hence the surprise modulation in SurNoR can be re-written solely in terms of Shannon surprise.

Surprise modulated update of the belief. Learning the world-model is equivalent to updating the parameters of the Dirichlet distribution after each transition. Consider the transition $(S_t = s, A_t = a) \to (S_{t+1} = s')$ with a surprise equal to $\mathbf{S}_{GM}^{(t+1)}$. The surprise modulated learning rate [35] is defined as

$$\gamma(\mathbf{S}_{GM}^{(t+1)}, m) = \frac{m\mathbf{S}_{GM}^{(t+1)}}{1 + m\mathbf{S}_{GM}^{(t+1)}}, \tag{12}$$

where $m > 0$ is a positive free parameter. Note that γ is a sigmoidal function of surprise with values in the range $0 \leq \gamma \leq 1$. The parameter m controls the sharpness of the transition.

With this modulated learning rate, the change in a participant's belief is given by an update of the Dirichlet parameters $\alpha_{\tilde{s},\tilde{a}}^{(t+1)}(\tilde{s}')$ for all $(\tilde{s}, \tilde{s}', \tilde{a}) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$ [35]

$$\alpha_{\tilde{s},\tilde{a}}^{(t+1)}(\tilde{s}') = \begin{cases} (1 - \gamma_{t+1})\alpha_{\tilde{s},\tilde{a}}^{(t)}(\tilde{s}') + \gamma_{t+1}\alpha^{(1)}(\tilde{s}') + \delta(s', \tilde{s}') & \text{if} \quad \tilde{s} = s, \tilde{a} = a \\ \alpha_{\tilde{s},\tilde{a}}^{(t)}(\tilde{s}') & \text{otherwise} \end{cases}, \tag{13}$$

where $\gamma_{t+1} = \gamma(\mathbf{S}_{CC}^{(t+1)}, m)$. The update rule expresses the new belief as a mix between two possibilities, represented by the current parameters $\alpha_{\tilde{s},\tilde{a}}^{(t)}(\tilde{s}')$ and the prior $\alpha^{(1)}(\tilde{s}')$, weighted with $1 - \gamma_{t+1}$ and γ_{t+1}, respectively. In the case of a large surprise, the value of γ_{t+1} is close to one and the current parameters are forgotten. The update makes a step based on the currently observed transition, expressed by the Kronecker-δ in the first line. The parameters of transitions from the pairs of the states and actions different form the current one (i.e. s and a) are not changed (second line). The update rule Eq. 13 has been called Variational Surprise Minimizing Learning (VarSMiLe) in [35].

Formalizing model-based Q-values. The world model of the participants is summarized by their beliefs $q^{(t)}(\theta)$ about the transition matrix of the environment. For the model-based branch:

Generative-Model-Surprise $\mathbf{S}_{GM}$ is used to modulate the learning rate for the update of the world model. Since the world model is summarized by parameters of Dirichlet distributions, the surprise enters into the update equation Eq. 13 of the Dirichlet parameters $\alpha_{\tilde{s},\tilde{a}}^{(t+1)}(\tilde{s}')$. With these Dirichlet parameters, participants estimate the transition probabilities $\hat{\theta}_{s,a}^{(t)}(s')$ at time t; cf. Eq. 6.

Novelty $N^{(t)}(s)$ of state s at time t (cf. Eq. 4) is used to guide exploration. Analogous to TD-learning where information of a reward at state s' is propagated by the Bellman equation to states $s \neq s'$, we use a Bellman equation to propagate the novelty of state s' to other states $s \neq s'$.

133

More specifically, for the model-based branch, we assign to each state-action pair a novelty-based value $Q^{(t)}_{MB,N}(s,a)$ which is an estimation of the accumulated future discounted novelty that can be gained by taking action a in state s. The Bellman equation is

$$Q^{(t)}_{MB,N}(s,a) = \sum_{s' \in \mathcal{S}} \hat{\theta}^{(t)}_{s,a}(s') \Big(N^{(t)}(s') + \lambda_N \max_{a' \in \mathcal{A}} Q^{(t)}_{MB,N}(s',a') \Big), \tag{14}$$

where $\hat{\theta}^{(t)}_{s,a}(s')$ are the estimated transition probabilities and $\lambda_N \in [0,1]$ is a discount factor for novelty. The Bellman equation assigns a value to the action a in state s as long as a novel state is likely to be reached within the next few steps - even if the immediately neighboring states are not novel. The discount rate λ_N controls the time horizon of 'future novelty'. For $\lambda_N \to 0$ only the novelty of the immediately following state matters; for $\lambda_N \to 1$ the time horizon becomes infinitely long.

Rewards $R(s)$ of states $s \in \mathcal{S}$ guide behavior during exploitation. In the theory of reinforcement learning, reward information is summarized in values $Q^{(t)}_{MB,R}(s,a)$ that are estimations of the accumulated future discounted reward that can be collected when starting at state s with action a. The Q-values are given by by the Bellman equation

$$Q^{(t)}_{MB,R}(s,a) = \sum_{s' \in \mathcal{S}} \hat{\theta}^{(t)}_{s,a}(s') \Big(R(s') + \lambda_R \max_{a' \in \mathcal{A}} Q^{(t)}_{MB,R}(s',a') \Big), \tag{15}$$

where $\lambda_R \in [0,1]$ is the discount factor for reward, which is not necessarily equal to the discount factor for novelty λ_N. Note that in our environment $R(s) = 0$ at all states except at the goal. Since the scale of the reward is arbitrary we set $R(s) = \delta(s, s_{Goal})$.

Total model-based Q-value is a linear combination of the Q-values for novelty $Q^{(t)}_{MB,N}(s,a)$ and reward $Q^{(t)}_{MB,N}(s,a)$ as

$$Q^{(t)}_{MB}(s,a) = \beta_R Q^{(t)}_{MB,R}(s,a) + \beta_N Q^{(t)}_{MB,N}(s,a), \tag{16}$$

where $\beta_R \geq 0$ and $\beta_N \geq 0$ are inverse temperature controlling exploitation and exploration, respectively - see subsection 'Hybrid Policy' for details.

In our model, β_R is fixed for all episodes, but β_N depends on whether subjects are in the exploration phase or the exploitation phase. This dependency was simplified as follows: Since novelty is the main drive in the 1st episode of the 1st block, we keep β_N fixed at a value β_{N1} throughout this episode. However, at the end of the 1st episode of the 1st block, subjcets find the goal, and hence there is no need for exploration, so we set $\beta_N = 0$ for remaining episodes of the 1st block. Since surprise increases rapidly after the first action of the second episode (which starts in state 3, now located one step before the goal), subjects find out the goal is lost; therefore we set $\beta_N = \beta_{N2}$ for the 1st episode of the 2nd block, and with the same arguments as for the 1st block, zero for the remaining episodes. Our statistical analysis shows that β_{N1} and β_{N2} are very close to each other.

Computing model-based Q-value. Since solving the non-linear sets of equations 15 and 14 for computing two separate sets of Q-values (i.e. $Q^{(t)}_{MB,N}(s,a)$ and $Q^{(t)}_{MB,R}(s,a)$ for all $(\tilde{s},\tilde{a}) \in \mathcal{S} \times \mathcal{A}$) is extremely computationally costly, we use a variant of the Prioritized Sweeping algorithm [51, 1] for computing model-based Q-values. The pseudocode of the new algorithm is shown in Algorithm 2. There is a free parameter $T_{PS} \in \mathbb{N}$ for this algorithm.

134

The idea of the algorithm, for example for updating $Q^{(t)}_{MB,R}(s,a)$, is to define a set of $|\mathcal{S}|$ new variables $U^{(t)}_R(s)$, and rewrite Eq. 15 as

$$\begin{aligned} Q^{(t)}_{MB,R}(s,a) &= \sum_{s' \in \mathcal{S}} \hat{\theta}^{(t)}_{s,a}(s') \Big(R(s') + \lambda_R U^{(t)}_R(s') \Big) \\ U^{(t)}_R(s') &= \max_{a' \in \mathcal{A}} Q^{(t)}_{MB,R}(s',a'). \end{aligned} \tag{17}$$

Then at each time-step, the intuitive explanation of the algorithm is to update $Q^{(t)}_{MB,R}(s,a)$ using the old values of $U^{(t)}_R(s)$ by the 1st equation, and update the values $U^{(t)}_R(s)$ for a finite number (T_{PS}) of the most "effective" states using new values of $Q^{(t)}_{MB,R}(s,a)$ by the 2nd equation. For details, see Algorithm 2.

5.1.2 SurNoR model-free branch

Formalizing model-free Q-values. Similar to what we did for the model-based branch, we define $Q^{(t)}_{MF,R}(s,a)$ and $Q^{(t)}_{MF,N}(s,a)$ as values of pairs of states and actions corresponding to external reward and novelty (internal reward), respectively. The main variation of the model-free Q-values from the model-based Q-values is in using TD-learning for their computation, in which the model of the world is not directly used - see the part "Computing model-free Q-values" for details.

Similar to total model-based Q-values, we define total model-free Q-values as

$$Q^{(t)}_{MF}(s,a) = \beta_R Q^{(t)}_{MF,R}(s,a) + \beta_N Q^{(t)}_{MF,N}(s,a), \tag{18}$$

where $\beta_R \geq 0$ and $\beta_N \geq 0$ are has the same value as the ones used in Eq. 16.

Reward and novelty prediction error. A crucial signal in model-free reinforcement learning is the reward prediction error, defined as the difference between the expected "reward" of a pair of state and action and its real "reward" [1]. Since we defined two separate sets of Q-values, one for the external reward and one for novelty (internal reward), we hence define two separate corresponding prediction errors.

Consider the transition $(S_t = s, A_t = a) \to (S_{t+1} = s')$, the reward prediction error at time $t+1$ is defined as

$$RPE_{t+1} = R(s') + \lambda_R \max_{a' \in \mathcal{A}} Q^{(t)}_{MF,R}(s',a') - Q^{(t)}_{MF,R}(s,a), \tag{19}$$

and similarly, the novelty prediction error at time $t+1$ is defined as

$$NPE_{t+1} = N(s') + \lambda_N \max_{a' \in \mathcal{A}} Q^{(t)}_{MF,N}(s',a') - Q^{(t)}_{MF,N}(s,a), \tag{20}$$

where λ_R and λ_N are the same discount factors as the ones used in the model-based branch.

Eligibility Trace. To keep track of the previously chosen pairs of states and actions, and to include them in the update rule, we use eligibility trace [1, **?**]. To have the most general setting, we define two separate eligibility traces, one for the external reward $e^{(t)}_R(s,a)$ and one for novelty (internal reward) $e^{(t)}_N(s,a)$ for all pairs of states and actions (s,a). The eligibility traces are initialized by zero, i.e. $e^{(1)}_R(s,a) = e^{(1)}_N(s,a) = 0 \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. Now, consider the transition

$(S_t = s, A_t = a) \rightarrow (S_{t+1} = s')$, eligibility traces are update as

$$e_R^{(t+1)}(s'', a'') = \begin{cases} 1 & \text{if} \quad s'' = s, a'' = a \\ \lambda_R \mu_R e_R^{(t)}(s'', a'') & \text{if} \quad \text{o.w.} \end{cases}$$
$$e_N^{(t+1)}(s'', a'') = \begin{cases} 1 & \text{if} \quad s'' = s, a'' = a \\ \lambda_N \mu_N e_N^{(t)}(s'', a'') & \text{if} \quad \text{o.w.}, \end{cases} \tag{21}$$

where λ_R and λ_N are the discount factors defined above, and $\mu_N \in [0, 1]$ and $\mu_R \in [0, 1]$ are free parameters expressing how fast eligibility traces decay in time.

Surprise modulation of model-free learning rate. Usual TD learning algorithms use a constant learning rate for updating Q-values [1]. However, the model-free branch of our SurNoR algorithm modulates its learning rate with surprise computed by the model-based branch. This novel interaction between model-based and model-free modules have not been explored by previous hybrid models in neuroscience, e.g. [4, 5].

We define the surprise modulated model-free learning rate ρ_t as

$$\rho_t = \rho_b + \gamma(\mathbf{S}_{GM}^{(t)}, m)\delta\rho, \tag{22}$$

where $\gamma(\mathbf{S}_{GM}^{(t)}, m)$ is the surprise modulated learning rate of the model-based branch defined in Eq. 12, $\rho_b \in [0, 1]$ is the baseline learning rate (when there is no surprise, i.e. $\mathbf{S}_{GM}^{(t)} = 0$), and $\delta\rho \in [0, 1 - \rho_b]$ is the maximum possible variation of the learning rate due to the surprise modulation. As a result, the learning ρ_t is between ρ_b (when $\mathbf{S}_{GM}^{(t)} = 0$) and $\rho_b + \delta\rho$ (when $\mathbf{S}_{GM}^{(t)} \rightarrow \infty$).

Computing model-free Q-value. The model-free Q-values for external reward are initialized by zero as $Q_{MF,R}^{(1)}(s, a) = 0$. The reason is to only have novelty as the exploration drive during the 1st episode of the 1st block. We also analyzed the alternative algorithm which uses the optimistic initialization for $Q_{MF,R}^{(1)}(s, a)$ (instead of novelty) for exploration [1] - see section "Alternative Algorithms" for details. However, to consider the most general case, we initialize the model-free Q-values for novelty with a free parameter $Q_{N0} \geq 0$ as $Q_{MF,N}^{(1)}(s, a) = Q_{N0}$.

Then, at each time step $t + 1$, the model-free Q-values are updated with a simple TD-learning algorithm as

$$Q_{MF,R}^{(t+1)}(s, a) = Q_{MF,R}^{(t)}(s, a) + \rho_{t+1} e_R^{(t+1)}(s, a) RPE_{t+1}$$
$$Q_{MF,N}^{(t+1)}(s, a) = Q_{MF,N}^{(t)}(s, a) + \rho_{t+1} e_N^{(t+1)}(s, a) NPE_{t+1}. \tag{23}$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

5.1.3 Hybrid policy

The policy for action selection is based on a convex combination of Q-values, similar to what is proposed by [4, 5]. Let us define the softmax function $\sigma(x(a)) = \exp(x(a))/\sum_{a'=1}^{4} \exp(x(a'))$. The action a is chosen in state s with probability

$$\pi(A_t = a | S_t = s) = \sigma\Big(\omega Q_{MF}^{(t)}(s, a) + (1 - \omega) Q_{MB}^{(t)}(s, a)\Big), \tag{24}$$

where $\omega \in [0, 1]$ is a free parameter balancing between the effect of the model-based and the model-free branches. When $\omega = 1$, the policy is purely model-free (except for the effect of surprise

modulation on TD-learning learning rate), and when $\omega = 0$, the policy is purely model-based. The reverse temperatures β_R and β_N, used in equations 16 and 18, control the sharpness of policy (the greater the inverse temperature the sharper the policy) and balance the exploration against exploitation.

As it was shown by [4], ω does not need to be fixed over time. Therefore, specific to our experiment, we consider ω to be piecewise constant in time: 1. $\omega = \omega_{11}$ for the 1st episode of the 1st block, when subjects are in the pure exploration phase, 2. $\omega = \omega_{12}$ for the 1st episode of the 2nd block, when the goal is lost, and 3. $\omega = \omega_0$ for the rest of the experiments (i.e. episodes 2 to 5 for both blocks), when subjects are in the exploitation phase.

5.1.4 Summary of free parameters

SurNoR has 16 free parameters, summarized as

$$\eta = \{\epsilon, m, \lambda_R, \lambda_N, \beta_R, \beta_{N1}, \beta_{N2}, T_{PS}, \mu_R, \mu_N, Q_{N0}, \rho_b, \delta\rho, \omega_0, \omega_{11}, \omega_{12}\}. \tag{25}$$

ϵ is used for initialization of the belief in Eq. 9. m is used for modulation of learning rate in Eq. 12. λ_R and λ_N are discount factors using in definitions and updates of Q-values. β_R, β_{N1}, and β_{N2} are used for balancing novelty against external reward in equations 16 and 18 and controlling the sharpness of the hybrid policy in Eq. 24. T_{PS} is used for Prioritized Sweeping in Algorithm 2. μ_R and μ_N are used for controling the decay of eligibility traces in Eq. 21. Q_{N0} is used for initialization of $Q_{MF,N}$. ρ_b and $\delta\rho$ are used for baseline TD-learning learning rate and its surprise modulation in Eq. 22. ω_0, ω_{11}, and ω_{12} are used for balancing model-free against model-based in the hybrid policy of Eq. 24.

5.2 Alternative algorithms

To statistically test the effect of surprise and novelty, we implemented 8 alternative algorithms explained as follows. Their key features are summarized in Table 1.

Model-based alternatives. Three out of 8 algorithms are purely model-based. In the same fashion as SurNoR, they all use novelty as an intrinsic motivation for exploration, but their approaches for learning the world model are different, and not necessarily surprise-modulated. These three algorithms are as follows.

(i) "Perf+N": 'Perf.' is an abbreviation for 'Perfect Integration', and 'N' is supposed to express that this algorithm uses novelty. Instead of a surprise-modulated learning rule, Perf+N uses perfect integration for learning the world model. This means that the main difference to the model-based branch of SurNoR algorithm is the update equation for the Dirichlet parameters (in Eq. 13) which is now

$$\alpha^{(t+1)}_{\tilde{s},\tilde{a}}(\tilde{s}') = \begin{cases} \alpha^{(t)}_{\tilde{s},\tilde{a}}(\tilde{s}') + \delta(s', \tilde{s}') & \text{if} \quad \tilde{s} = s, \tilde{a} = a \\ \alpha^{(t)}_{\tilde{s},\tilde{a}}(\tilde{s}') & \text{otherwise} \end{cases}, \tag{26}$$

which basically is identical to considering $m = 0$ in SurNoR, i.e. independent of surprise value we have $\gamma_t = 0$. Perf+N is equivalent to doing the exact Bayesian inference with the assumption that the underlying transition probabilities are fixed in time. This algorithm has 7 free parameters $\{\epsilon, \lambda_R, \lambda_N, T_{PS}, \beta_R, \beta_{N1}, \beta_{N2}\}$, and can be considered as a model nested in SurNoR by assuming $m = \mu_R = \mu_N = Q_{N0} = \rho_b = \delta\rho = \omega_{11} = \omega_{12} = \omega_0 = 0$.

137

Algorithm 1 Pseudocode for SurNoR

Specify $\mathcal{S}$ and $\mathcal{A}$
Specify Episode and Block
Specify free parameters $\eta = \{\epsilon, m, \lambda_R, \lambda_N, \beta_R, \beta_{N1}, \beta_{N2}, T_{PS}, \mu_R, \mu_N, Q_{N0}, \rho_b, \delta\rho, \omega_0, \omega_{11}, \omega_{12}\}$
if Episode $= 1$ and Block $= 1$ **then**
    $\omega = \omega_{11}$ and $\beta_N = \beta_{N1}$
if Episode $= 1$ and Block $= 2$ **then**
    $\omega = \omega_{12}$ and $\beta_N = \beta_{N2}$
if Episode $\neq 1$ **then**
    $\omega = \omega_0$ and $\beta_N = 0$
Initialize $e_R^{(1)}(s,a) = e_N^{(1)}(s,a) = 0$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
if Episode $= 1$ and Block $= 1$ **then**
    Initialize $C_s^{(1)} = 0$, $U_R^{(1)}(s) = 0$, $U_N^{(1)}(s) = \frac{\log(|\mathcal{A}|)}{1-\lambda}$, $\forall s \in \mathcal{S}$.
    Initialize $Q_{MB,R}^{(1)}(s,a) = 0$, $Q_{MB,N}^{(1)}(s,a) = U_N^{(1)}(s)$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
    Initialize $Q_{MF,R}^{(1)}(s,a) = 0$, $Q_{MF,N}^{(1)}(s,a) = Q_{N0}$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
    Initialize $\alpha_{s,a}^{(1)}(s') = \epsilon$, $\forall (s,s',a) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$.
else
    Initialize $C_s^{(1)}$, $U_R^{(1)}(s)$, $U_N^{(1)}(s)$, $Q_{MB,R}^{(1)}(s,a)$, $Q_{MB,N}^{(1)}(s,a)$, $Q_{MF,R}^{(1)}(s,a)$, $Q_{MF,N}^{(1)}(s,a)$ and $\alpha_{s,a}^{(1)}(s')$ with their latest values in the previous Episode.
Initialize state $S_1 = s_1$, update counts $C_s^{(1)} \leftarrow C_s^{(1)} + \delta(s, s_1)$, and put $t \leftarrow 1$
while $s_t \neq s_{Goal}$ **do**
    Compute $Q_{MF}^{(t)}(s,a) = \beta_R Q_{MF,R}^{(t)}(s,a) + \beta_N Q_{MF,N}^{(t)}(s,a)$.
    Compute $Q_{MB}^{(t)}(s,a) = \beta_R Q_{MB,R}^{(t)}(s,a) + \beta_N Q_{MB,N}^{(t)}(s,a)$.
    Sample a_t from $\pi(A_t = a | S_t = s) = \sigma\Big(\omega Q_{MF}^{(t)}(s,a) + (1-\omega) Q_{MB}^{(t)}(s,a)\Big)$
    Observe $S_{t+1} = s_{t+1}$.
    Compute $RPE_{t+1} = R(s_{t+1}) + \lambda_R \max_{a' \in \mathcal{A}} Q_{MF,R}^{(t)}(s_{t+1}, a') - Q_{MF,R}^{(t)}(s_t, a_t)$
    Compute $NPE_{t+1} = N^{(t)}(s_t + 1) + \lambda_N \max_{a' \in \mathcal{A}} Q_{MF,R}^{(t)}(s_{t+1}, a') - Q_{MF,R}^{(t)}(s_t, a_t)$
    Update counts $C_s^{(t+1)} = C_s^{(t)} + \delta(s, s_{t+1})$ and novelty $N^{(t+1)}(s) = \log \frac{t + |\mathcal{S}|}{C_s^{(t+1)} + 1}$.
    Compute $\mathbf{S}^{(t+1)} = \frac{\hat{\theta}_{s_t,a_t}^{(1)}(s_{t+1})}{\hat{\theta}_{s_t,a_t}^{(t)}(s_{t+1})}$, $\gamma_{t+1} = \frac{m\mathbf{S}^{(t+1)}}{1 + m\mathbf{S}^{(t+1)}}$, and $\rho_{t+1} = \rho_b + \gamma_{t+1}\delta\rho$.
    Update $e_N^{(t+1)}(s_t, a_t) = 1$, and $e_N^{(t+1)}(s,a) = \lambda_N \mu_N e_N^{(t)}(s,a)$ $\forall s \neq s_t, a \neq a_t$.
    Update $e_R^{(t+1)}(s_t, a_t) = 1$, and $e_R^{(t+1)}(s,a) = \lambda_R \mu_R e_R^{(t)}(s,a)$ $\forall s \neq s_t, a \neq a_t$.
    Update $Q_{MF,R}^{(t+1)}(s,a) = Q_{MF,R}^{(t)}(s,a) + \rho_{t+1} e_R^{(t+1)}(s,a) RPE_{t+1}$
    Update $Q_{MF,N}^{(t+1)}(s,a) = Q_{MF,N}^{(t)}(s,a) + \rho_{t+1} e_N^{(t+1)}(s,a) NPE_{t+1}$.
    Update $\alpha_{s_t,a_t}^{(t+1)}(s) = (1-\gamma_{t+1})\alpha_{s_t,a_t}^{(t)}(s) + \gamma_{t+1}\epsilon + \delta(s_{t+1}, s)$, and $\alpha_{s,a}^{(t+1)} = \alpha_{s,a}^{(t+1)}$ $\forall s \neq s_t, a \neq a_t$.
    Update $\hat{\theta}^{(t+1)}$ as $\hat{\theta}_{s,a}^{(t+1)}(s') = \frac{\alpha_{s,a}^{(t+1)}(s')}{\sum_{\tilde{s}' \in \mathcal{S}} \alpha_{s,a}^{(t+1)}(\tilde{s}')}$.
    Update $Q_{MB,N}^{(t+1)}(s,a)$ and $U_N^{(t+1)}(s)$ using Alg. 2 and $N^{(t+1)}(s)$ as rewards.
    if Episode $= 1$ and Block $= 1$ and $s_t \neq s_{Goal}$ **then**
        Update $Q_{MB,R}^{(t+1)}(s,a) = U_R^{(t+1)}(s) = 0$.
    else
        Update $Q_{MB,R}^{(t+1)}(s,a)$ and $U_R^{(t+1)}(s)$ using Alg. 2 and $R(s) = \delta(s, s_{Goal})$ as rewards.
    $t \leftarrow t + 1$.

138

Algorithm 2 Pseudocode for the modified version of Prioritized Sweeping Algorithm of [51, 1] for one time-step at time $t+1$

Free parameters: λ (i.e. λ_R for reward and λ_N for novelty) and T_{PS}.
Input: $\mathcal{S}$, $\mathcal{A}$, $\hat{\theta}^{(t+1)}$, $Q^{(t)}$ (i.e. $Q^{(t)}_{MB,R}$ for reward and $Q^{(t)}_{MB,N}$ for novelty), $U^{(t)}$ (i.e. $U^{(t)}_R$ for reward and $U^{(t)}_N$ for novelty), and $Reward$ (i.e. R for reward and $N^{(t+1)}$ for novelty)
for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
    $Q^{(t+1)}(s,a) = \sum_{s' \in \mathcal{S}} \hat{\theta}^{(t+1)}_{s,a}(s') \Big(Reward(s') + \lambda U^{(t)}(s') \Big)$
for $s \in \mathcal{S}$ **do**
    $U^{(t+1)}(s) = U^{(t)}(s)$
    $Prior(s) = |U^{(t+1)}(s) - \max_{a \in \mathcal{A}} Q^{(t+1)}(s,a)|$
for T_{PS} iterations **do**
    $s' = \arg\max_{s \in \mathcal{S}} Prior(s)$
    $\Delta V = \max_{a \in \mathcal{A}} Q^{(t+1)}(s',a) - U^{(t+1)}(s')$
    $U^{(t+1)}(s') = \max_{a \in \mathcal{A}} Q^{(t+1)}(s',a)$
    for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
        $Q^{(t+1)}(s,a) \leftarrow Q^{(t+1)}(s,a) + \lambda \hat{\theta}^{(t+1)}_{s,a}(s') \Delta V$
    for $s \in \mathcal{S}$ **do**
        $Prior(s) = |U^{(t+1)}(s) - \max_{a \in \mathcal{A}} Q^{(t+1)}(s,a)|$

(ii) "Leak+N": 'Leak.' is an abbreviation for 'Leaky Integration', and 'N' is supposed to express that this algorithm uses novelty. Leak+N is similar to Perf+N, but it uses leaky integration for the update equation for the Dirichlet parameters (in Eq. 13) which is now

$$\alpha^{(t+1)}_{\tilde{s},\tilde{a}}(\tilde{s}') = \begin{cases} \kappa_{\text{Leak}} \alpha^{(t)}_{\tilde{s},\tilde{a}}(\tilde{s}') + \delta(s',\tilde{s}') & \text{if} \quad \tilde{s}=s, \tilde{a}=a \\ \alpha^{(t)}_{\tilde{s},\tilde{a}}(\tilde{s}') & \text{otherwise} \end{cases}, \tag{27}$$

where $\kappa_{\text{Leak}} \in [0,1]$ is a constant free parameter. Such a learning rule has been used previously to model human behavior [28, 30, 29, 52]. Overall, Leak+N has 8 free parameters as $\{\epsilon, \kappa_{\text{Leak}}, \lambda_R, \lambda_N, T_{PS}, \beta_R, \beta_{N1}, \beta_{N2}\}$. It cannot be considered fully as a model nested in SurNoR, but it is equivalent to SurNoR by using Eq. 27 instead of Eq. 13 for updating the belief, and by assuming $m = \mu_R = \mu_N = Q_{N0} = \rho_b = \delta\rho = \omega_{11} = \omega_{12} = \omega_0 = 0$.

(iii) "SMB+N": 'SMB' is an abbreviation for 'Surprise-modulated Model Based', and 'N' is supposed to express that this algorithm uses novelty. This algorithm is a reduced version of SurNoR with $\mu_R = \mu_N = Q_{N0} = \rho_b = \delta\rho = \omega_{11} = \omega_{12} = \omega_0 = 0$, which is equivalent to the model-based branch of SurNoR. It has 8 free parameters as $\{\epsilon, m, \lambda_R, \lambda_N, T_{PS}, \beta_R, \beta_{N1}, \beta_{N2}\}$.

Model-free alternatives. Three out of 8 algorithms are purely model-free, explained as follows.

(iv) "MF+Q0": 'MF' is an abbreviation for 'Model-Free', and 'Q0' is supposed to express that this algorithm uses optimistic initialization for exploration (instead of novelty). MF+Q0 is equivalent to what is usually called $Q(\lambda)$ [1], with $\lambda = \mu_R$ in our notation. It can be seen a modified version of SurNoR with initializing $Q^{(0)}_{MF,R} = Q_{R0}$ (where Q_{R0} is a free parameter) and assuming $m = \lambda_N = \beta_{N1} = \beta_{N2} = T_{PS} = \mu_N = Q_{N0} = \delta\rho = 0$ and $\omega_{11} = \omega_{12} = \omega_0 = \epsilon = 1$. It has overall 5 free parameters as $\{\lambda_R, Q_{R0}, \rho_b, \beta_R, \mu_R\}$.

139

	Algorithm	Model-based	Model-free	Novelty	Surprise	Free-Param.
i	Perf+N	Y	N	Y	N	7
ii	Leak+N	Y	N	Y	N	8
iii	MBS+N	Y	N	Y	Y	8
iv	MF+Q0	N	Y	N	N	5
v	MF+N	N	Y	Y	N	9
vi	MF+NS	N	Y	Y	Y	12
vii	Hyb+N	Y	Y	Y	N	15
iix	Random Choice	N	N	N	N	0

Table 1: Summary of the key features of alternative models.Y: Contains; N: Does not contain.

(v) "MF+N": 'MF' is an abbreviation for 'Model-Free', and 'N' is supposed to express that this algorithm uses novelty. MF+N is a reduced version of SurNoR with assuming $m = T_{PS} = \delta\rho = 0$ and $\omega_{11} = \omega_{12} = \omega_0 = \epsilon = 1$, which is equivalent to the model-free branch of SurNoR without any surprise modulation. It has overall 9 free parameters as $\{\lambda_R, \lambda_N, \beta_R, \beta_{N1}, \beta_{N2}, \mu_R, \mu_N, Q_{N0}, \rho_b\}$.

(vi) "MF+NS": 'MF' is an abbreviation for 'Model-Free', and 'NS' is supposed to express that this algorithm uses both novelty and surprise. In fact, the model-based branch of SurNoR is used in MF+NS, but only for computing surprise and modulating the learning rate of TD-learner and not for the hybrid policy. MF+NS can be seen as a reduced version of SurNoR with assuming $T_{PS} = 0$ and $\omega_{11} = \omega_{12} = \omega_0 = 1$, which is equivalent to the model-free branch of SurNoR but with surprise modulation. It has 12 free parameters as $\{\epsilon, m, \lambda_R, \lambda_N, \beta_R, \beta_{N1}, \beta_{N2}, \mu_R, \mu_N, Q_{N0}, \rho_b, \delta\rho\}$.

Hybrid alternative.

(vii) "Hyb+N": 'Hyb.' is an abbreviation for 'Hybrid' meaning both model-based and model-free branches of SurNoR are used for the policy, and 'N' is supposed to express that this algorithm uses novelty. Therefore, the main difference between this algorithm and SurNoR is in using Eq. 27 instead of Eq. 13 for updating the belief (introducing the new free parameter $\kappa_{\text{Leak}} \in [0, 1]$), and by assuming $\delta\rho = m = 0$, i.e. there is no modulation of the learning rate for TD-learner. Similar to Leak+N, Hyb+N is not fully nested in SurNoR, because of its particular shape of the update rule. It has overall 15 free parameters as $\{\epsilon, \kappa_{\text{Leak}}, \lambda_R, \lambda_N, \beta_R, \beta_{N1}, \beta_{N2}, T_{PS}, \mu_R, \mu_N, Q_{N0}, \rho_b, \omega_{11}, \omega_{12}, \omega_0\}$.

Null alternative.

(iix) "Random Choice": According to this algorithm, subjects choose actions with uniform distribution, i.e. each action is selected with a probability equal to $\frac{1}{|\mathcal{A}|} = 0.25$. We used this model to analyze particularly the effect of our novelty-seeking exploration in the 1st episode of the 1st block. This algorithm does not have any free parameter.

140

5.3 Fitting to human behavior

Setup for statistical inference Let us show the behavioral data of subject i with $\mathcal{D}_i$, and the whole set of behavioral data with $\mathcal{D} = \{\mathcal{D}_i, 1 \leq i \leq 12\}$. Note that $\mathcal{D}_i$ consists of the sequences of the states and actions for all episodes of both blocks corresponding to subject i. As discussed in the previous subsection, we compare different computational models, indexed with j where $1 \leq j \leq 9$, i.e. SurNoR plus 8 alternative algorithms. Let us also denote the computational model j with $\mathcal{M}_j$ and its corresponding parameter set with η_j. Our whole analysis is based on a Bayesian model selection approach [41] and is based on computing each model log-evidence given behavioral data as

$$\log \mathbf{P}(\mathcal{D}_i|\mathcal{M}_j) = \log \int \mathbf{P}(\mathcal{D}_i|\mathcal{M}_j, \eta_j)\mathbf{P}(\eta_j|\mathcal{M}_j)d\eta_j, \tag{28}$$

where $\mathbf{P}(\mathcal{D}_i|\mathcal{M}_j, \eta_j)$ is the likelihood function, and $\mathbf{P}(\eta_j|\mathcal{M}_j)$ is the prior distribution over parameter set. To estimate model log-evidence, we need to have the prior $\mathbf{P}(\eta_j|\mathcal{M}_j)$. By fitting models to behavioral data, we mean finding $\mathbf{P}(\eta_j|\mathcal{M}_j)$ with a Cross-validated empirical Bayesian approach explained in the next part.

Cross-Validated Empirical Bayes: We considered the prior distribution as a delta distribution $\mathbf{P}(\eta_j|\mathcal{M}_j) = \delta(\eta_j - \eta_j^*)$, which is identical to assuming that the parameter of the model is fixed. An empirical Bayesian approach [53] to estimate η_j^* is to maximize $\log \mathbf{P}(\mathcal{D}|\mathcal{M}_j, \eta_j^*) = \sum_{i=1}^{12} \log \mathbf{P}(\mathcal{D}_i|\mathcal{M}_j, \eta_j^*)$ over η_j^* - which is basically equivalent to finding the maximum likelihood estimation over $\mathcal{D}$. The result of such an approach is to have the total log-evidence equal to maximum log-likelihood. However, to avoid over-fitting (i.e. over-estimating log-evidence), we combined the idea of empirical Bayes with 3-fold cross-validation, similar to the approach of [**?**]: (i) We divided data to 3 folds, each of which consists of four subjects, (ii) To compute the log-evidence for subject i, we estimated the parameter η_j^* by maximizing the likelihood function of the folds which do not include subject i. The maximization process were done using coordinate ascent (using grid search for each coordinate). For each model and fold, we ran the maximization algorithm from 50 different random initial points until full convergence.

Bayesian model selection (BMS) and accuracy rate: Using the computed log-evidences for each subject, we used the Bayesian Model Selection (BMS) approach proposed by [41] to compare different models. The idea of BMS is to: (i) Assume that the model j is selected for each subject with a probability $P_{\mathcal{M}_j}$, (ii) Estimate the expected posterior probability $\hat{P}_{\mathcal{M}_j} = \mathbb{E}[P_{\mathcal{M}_j}|\mathcal{D}]$ as well as the model exceedance probability $\phi_{\mathcal{M}_j} = \mathbf{P}(\{P_{\mathcal{M}_j} > P_{\mathcal{M}_i}, \forall i \neq j\}|\mathcal{D})$ using log-evidences. [41] uses a variational approach for estimating these statistics of interests.

We used the functions developed by [41] for BMS in MATLAB toolbox SPM12, and computed model posterior and exceedance probabilities for each of our 9 models. Results are reported.

Having our models fitted to the behavioral data, we can compute each subject's policy at a given time t (given the sequences of states and actions for that subject until time t). Given the policy, we can predict the subject's next action, and then compute the accuracy rate of our predictions. Results for such an analysis for SurNoR (i.e. the best model) is shown in Fig. 2 in the main text.

5.4 EEG Sliding Window Analysis

Sliding window for novelty analysis: The comparison between LOW-NOVELTY and HIGH-NOVELTY conditioned ERPs showed that in the time window from 80 to 110ms after the state

141

onset, the ERP amplitudes of the two conditions differed significantly. We chose the time interval from 50 to 150ms after the state onset as the search area for participant-based analysis. We used a sliding window of 50ms width and 10ms step inside the chosen interval. There were in total 6 sliding windows, which were 50-100ms, 60-110ms, 70-120ms, 80-130ms, 90-140ms and 100-150ms after the state onset. We used the mean amplitudes of each sliding window to correlate with the estimated novelty for each participant in the first episode of the first block. The 80-130ms time window showed a significant result. Thus we considered the time course from 80 to 130ms after the state onset as a bio-marker for the novelty signal.

Sliding window for surprise analysis: Similar to the analysis done with novelty bio-marker analysis. We choose the time interval from 150 to 300ms as the interval of interest for participant-based surprise analysis. There were in total 10 sliding windows, which were 150-200ms, 160-210ms, 170-220ms, 180-230ms, 190-240ms, 200-250ms, 210-260ms, 220-270ms, 230-280ms and 290-300ms after the state onset. The mean amplitudes in only two windows, 150-200ms and 160-210ms, showed significant correlations on participant-by-participant basis. Thus we combined the two windows as 150-210ms, and used it for participant-based analysis. We considered this time course as a bio-marker for the surprise signal.

References

[1] Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).

[2] Hurley, M. M., Dennett, D. C., Adams Jr, R. B. & Adams, R. B. *Inside jokes: Using humor to reverse-engineer the mind* (MIT press, 2011).

[3] Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience* **8**, 1704 (2005).

[4] Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).

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

[6] Niv, Y., Edlund, J. A., Dayan, P. & O'Doherty, J. P. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience* **32**, 551–562 (2012).

[7] O'Doherty, J. P., Cockburn, J. & Pauli, W. M. Learning, reward, and decision making. *Annual review of psychology* **68**, 73–100 (2017).

[8] O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H. & Dolan, R. J. Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329–337 (2003).

[9] Niv, Y. *et al.* Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience* **35**, 8145–8157 (2015).

[10] Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J. & Frith, C. D. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* **442**, 1042 (2006).

[11] Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual review of psychology* **68**, 101–128 (2017).

[12] Lehmann, M. *et al.* Evidence for eligibility traces in human learning. *arXiv preprint arXiv:1707.04192* (2017).

[13] Jaegle, A., Mehrpour, V. & Rust, N. Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. *arXiv preprint arXiv:1901.02478* (2019).

[14] Gottlieb, J. & Oudeyer, P.-Y. Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience* 1 (2018).

[15] Murayama, K., FitzGibbon, L. & Sakaki, M. Process account of curiosity and interest: A reward-learning perspective. *Educational Psychology Review* 1–21 (2019).

[16] Ryan, R. M. & Deci, E. L. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology* **25**, 54–67 (2000).

[17] Holm, L., Wadenholt, G. & Schrater, P. Episodic curiosity for avoiding asteroids: Per-trial information gain for choice outcomes drive information seeking. *Scientific reports* **9** (2019).

143

[18] Blanchard, T. C., Hayden, B. Y. & Bromberg-Martin, E. S. Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. *Neuron* **85**, 602–614 (2015).

[19] Oudeyer, P.-Y., Gottlieb, J. & Lopes, M. Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies. In *Progress in brain research*, vol. 229, 257–284 (Elsevier, 2016).

[20] Juechems, K. & Summerfield, C. Where does value come from? *Trends in cognitive sciences* (2019).

[21] Oudeyer, P.-Y. & Kaplan, F. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics* **1**, 6 (2009).

[22] Kolter, J. Z. & Ng, A. Y. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 513–520 (ACM, 2009).

[23] Bellemare, M. *et al.* Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, 1471–1479 (2016).

[24] Martin, J., Narayanan, S. S., Everitt, T. & Hutter, M. Count-based exploration in feature space for reinforcement learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2471–2478 (AAAI Press, 2017).

[25] Preuschoff, K., t Hart, B. M. & Einhauser, W. Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in neuroscience* **5**, 115 (2011).

[26] Ostwald, D. *et al.* Evidence for neural encoding of bayesian surprise in human somatosensation. *NeuroImage* **62**, 177–188 (2012).

[27] Mars, R. B. *et al.* Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *Journal of Neuroscience* **28**, 12539–12545 (2008).

[28] Meyniel, F., Maheu, M. & Dehaene, S. Human inferences about sequences: A minimal transition probability model. *PLoS computational biology* **12**, e1005260 (2016).

[29] Maheu, M., Dehaene, S. & Meyniel, F. Brain signatures of a multiscale process of sequence learning in humans. *Elife* **8**, e41541 (2019).

[30] Modirshanechi, A., Kiani, M. M. & Aghajan, H. Trial-by-trial surprise-decoding model for visual and auditory binary oddball tasks. *NeuroImage* **196**, 302–317 (2019).

[31] Nassar, M. R., Wilson, R. C., Heasly, B. & Gold, J. I. An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience* **30**, 12366–12378 (2010).

[32] Nassar, M. R. *et al.* Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience* **15**, 1040 (2012).

[33] Behrens, T. E., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. Learning the value of information in an uncertain world. *Nature neuroscience* **10**, 1214 (2007).

[34] Faraji, M., Preuschoff, K. & Gerstner, W. Balancing new against old information: the role of puzzlement surprise in learning. *Neural computation* **30**, 34–83 (2018).

144

[35] Liakoni, V., Modirshanechi, A., Gerstner, W. & Brea, J. An approximate bayesian approach to surprise-based learning. *arXiv preprint arXiv:1907.02936* (2019).

[36] Yu, A. J. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681–692 (2005).

[37] Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D. & Brea, J. Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules. *Frontiers in neural circuits* **12** (2018).

[38] Frémaux, N. & Gerstner, W. Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Frontiers in neural circuits* **9**, 85 (2016).

[39] Barto, A., Mirolli, M. & Baldassarre, G. Novelty or surprise? *Frontiers in psychology* **4**, 907 (2013).

[40] Tartaglia, E. M., Clarke, A. M. & Herzog, M. H. What to choose next? a paradigm for testing human sequential decision making. *Frontiers in psychology* **8**, 312 (2017).

[41] Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *Neuroimage* **46**, 1004–1017 (2009).

[42] Beaufour, C. C., Le Bihan, C., Hamon, M. & Thiébot, M.-H. Extracellular dopamine in the rat prefrontal cortex during reward-, punishment-and novelty-associated behaviour. effects of diazepam. *Pharmacology Biochemistry and Behavior* **69**, 133–142 (2001).

[43] Bunzeck, N., Dayan, P., Dolan, R. J. & Duzel, E. A common mechanism for adaptive scaling of reward and novelty. *Human brain mapping* **31**, 1380–1394 (2010).

[44] Bódi, N. *et al.* Reward-learning and the novelty-seeking personality: a between-and within-subjects study of the effects of dopamine agonists on young parkinson's patients. *Brain* **132**, 2385–2395 (2009).

[45] Bevins, R. A. *et al.* Novel-object place conditioning: behavioral and dopaminergic processes in expression of novelty reward. *Behavioural brain research* **129**, 41–50 (2002).

[46] Pelli, D. G. & Vision, S. The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision* **10**, 437–442 (1997).

[47] Delorme, A. & Makeig, S. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods* **134**, 9–21 (2004).

[48] Little, D. Y.-J. & Sommer, F. T. Learning and exploration in action-perception loops. *Frontiers in neural circuits* **7**, 37 (2013).

[49] Mobin, S. A., Arnemann, J. A. & Sommer, F. Information-based learning by agents in unbounded state spaces. In *Advances in Neural Information Processing Systems*, 3023–3031 (2014).

[50] Shannon, C. A mathematical theory of communication. *Bell System Technical Journal 27: 379-423 and 623–656* **20** (1948).

[51] Van Seijen, H. & Sutton, R. S. Efficient planning in mdps by small backups. In *Proc. 30th Int. Conf. Mach. Learn.*, 1–3 (2013).

145

[52] Yu, A. J. & Cohen, J. D. Sequential effects: superstition or rational behavior? In *Advances in neural information processing systems*, 1873–1880 (2009).

[53] Efron, B. & Hastie, T. *Computer age statistical inference*, vol. 5 (Cambridge University Press, 2016).

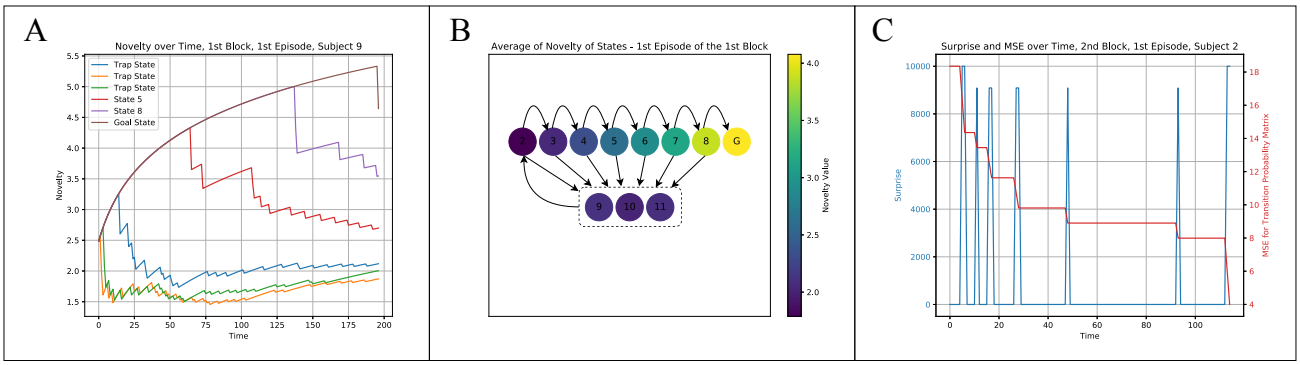


Figure 1: **A.** Novelty time-series during the 1st episode of the 1st block: Data is for a single subject. **B.** Novelty heatmap at the end of the 1st episode of the 1st block: The values are averaged over all subjects. **C.** Time-series for surprise and mean square error of model-estimation: Data is for a single subject.

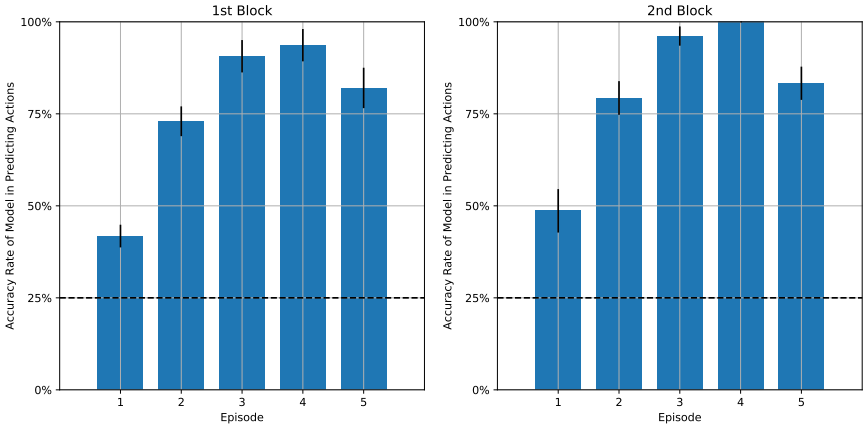


Figure 2: Average (over subject) accuracy rate of predicting actions for each episode of each block. The accuracy is corresponding to the model with highest posterior probability, see Supplementary Materials. The error bars stands for the mean errors. The dash-line is corresponding to the accuracy rate of the random choice (null) model, i.e. 25%.

147

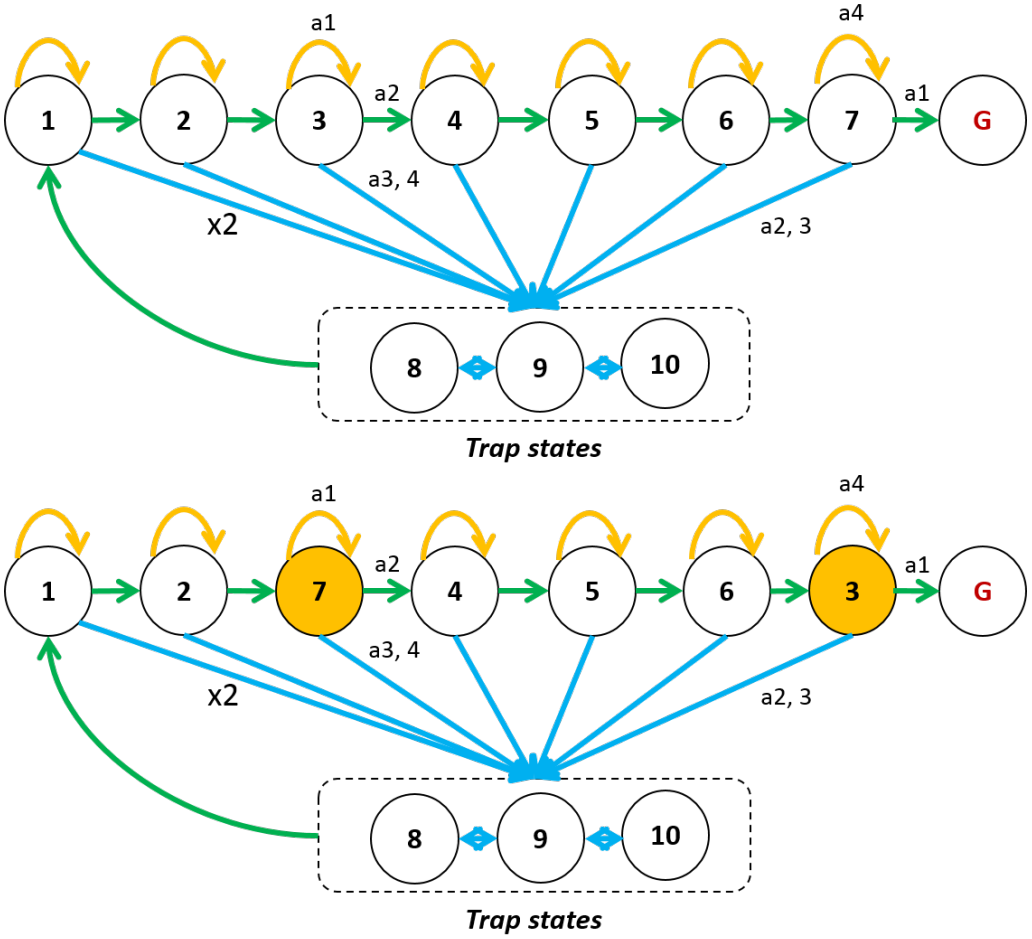


Figure 3: Structure of the learning environment used in the experiment. *Upper Panel*: Environment used in block 1. There are 11 states and 4 actions for each state. States 1-7 are the *progressing states*, states 8-10 are the *trap states*, goal state is presented by the red G. At each progressing state, there are one action that leads participants to the next progressing state (green arrow), two actions lead participants to one of the trap states (blue arrow), and one action let participant stay at current state (yellow arrow). At each trap state, there are three actions that lead participant to one of the trap states (blue arrow), and one action (green arrow) leads participants to the beginning of the progressing state (state 1). Action arrows are not fully drawn for the trap state because of limited space. *Lower Panel*: Environment used in block 2. The image presenting state 3 and state 7 are swapped in block 2. Other transitions stays the same.

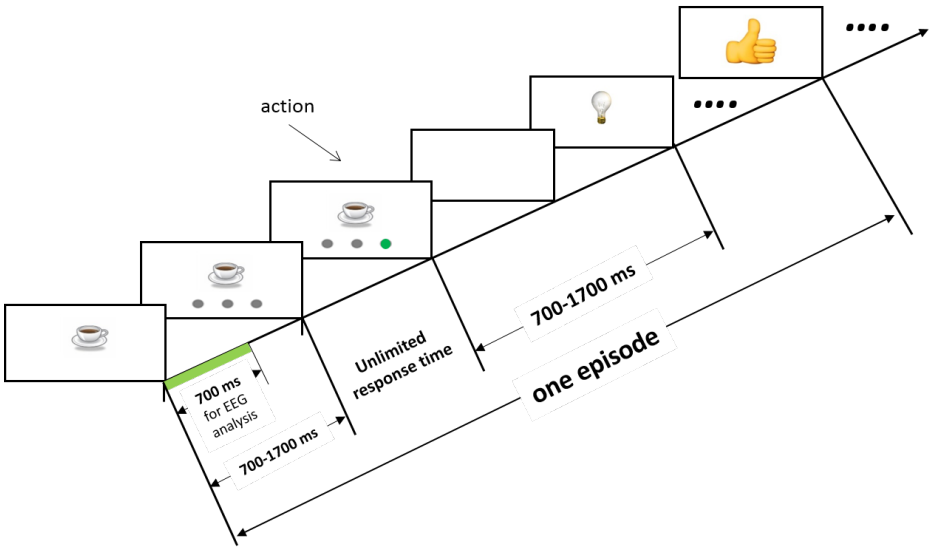


Figure 4: After an image (state) was presented, participants needed to wait for 700-1700ms, randomly chosen, until grey disks were presented at the bottom of the image. After clicking on one disk (action), a blank screen was shown for 700 to 1700ms, randomly chosen, and then the next image appeared. The environment was deterministic, e.g., clicking on the left disk in the house image always brought the participant to the coffee cup. The goal image is a 'thumb-up' image in this example. Different observers saw different images. Green intervals indicate the time window for which EEG was analysed.

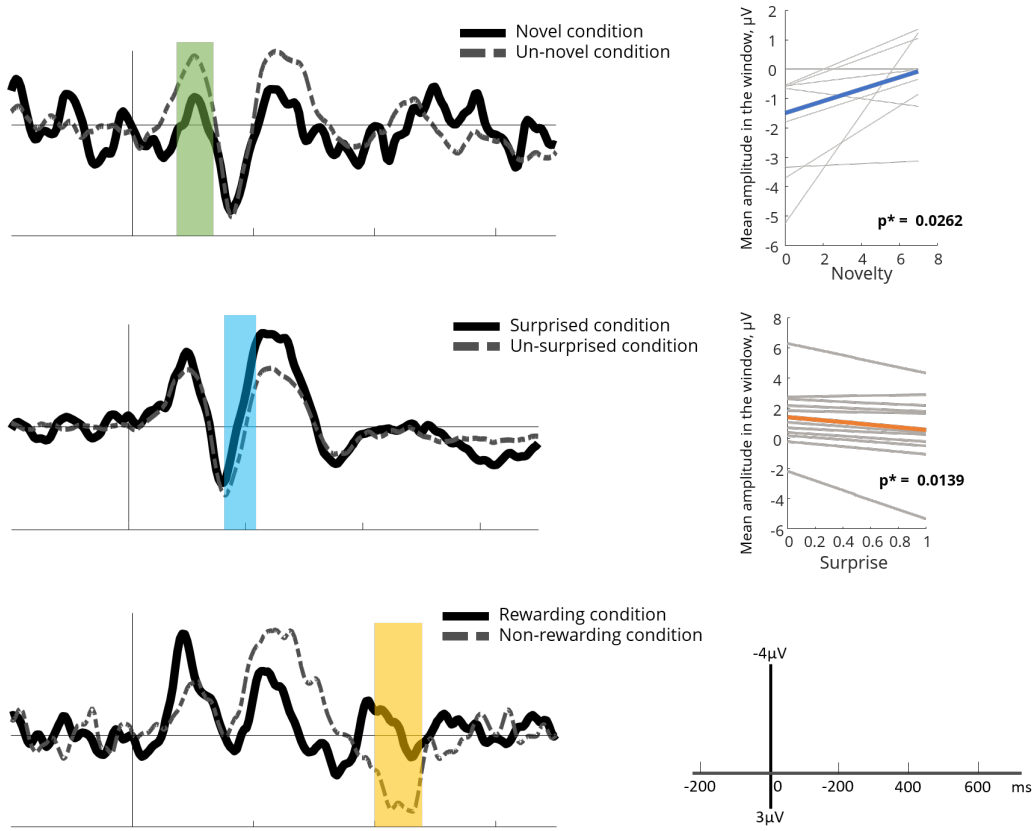


Figure 5: *Left Panel*: ERP comparisons to search for the time courses of Novelty, Surprise and Reward response. Colored region presents the time course where the mean amplitudes in the interval reflects the corresponding signals. *Right Panel*: Participant-based linear regressions (grey lines) and averaged linear regression (colored lines) between mean amplitude in the interval of interest and novelty (upper plot) and surprise (lower plot).